

T.C.  
İSTANBUL 29 MAYIS ÜNİVERSİTESİ  
SOSYAL BİLİMLER ENSTİTÜSÜ  
MÜTERCİM VE TERCÜMANLIK ANABİLİM DALI

**EVALUATING STATE-OF-THE-ART NMT ON FOOTBALL  
TWEETS WITH HIGH CONTEXT: CORRELATING  
HUMAN (MQM) AND AUTOMATED (COMETKIWI-XL)  
METRICS**

(YÜKSEK LİSANS TEZİ)

**Aziz Berkan KARADELİ**

**Danışman:**

**Prof. Dr. Ayşe Banu KARADAĞ**

**İSTANBUL**

**2025**



T. C.  
İSTANBUL 29 MAYIS ÜNİVERSİTESİ  
SOSYAL BİLİMLER ENSTİTÜSÜ  
MÜTERCİM VE TERCÜMANLIK ANABİLİM DALI

**EVALUATING STATE-OF-THE-ART NMT ON  
FOOTBALL TWEETS WITH HIGH CONTEXT:  
CORRELATING HUMAN (MQM) AND AUTOMATED  
(COMETKIWI-XL) METRICS**

(YÜKSEK LİSANS TEZİ)

**Aziz Berkan KARADELİ**

**Danışman:**

**Prof. Dr. Ayşe Banu KARADAĞ**

**İSTANBUL**

**2025**

## TEZ ONAY SAYFASI

T. C.

İSTANBUL 29 MAYIS ÜNİVERSİTESİ

SOSYAL BİLİMLER ENSTİTÜSÜ MÜDÜRLÜĞÜNE

Mütercim ve Tercümanlık Anabilim Dalı, Çeviribilim (İngilizce) Tezli Yüksek Lisans Programı'nda 200522YL01 numaralı Aziz Berkan KARADELİ'nin hazırladığı “*Evaluating State-of-the-Art NMT on Football Tweets with High Context: Correlating Human (MQM) and Automated (COMET-Kiwi-XL) Metrics*” konulu yüksek lisans tezi ile ilgili tez savunma sınavı, 09/09/2025 günü 14:00 – 14.45 saatleri arasında yapılmış, sorulan sorulara alınan cevaplar sonunda adayın tezinin başarılı olduğuna oy birliği ile karar verilmiştir.

Prof. Dr. Ayşe Banu KARADAĞ  
Yıldız Teknik Üniversitesi  
(Tez Danışmanı ve Sınav Komisyonu Başkanı)

Prof. Dr. Işın ÖNER  
İstanbul 29 Mayıs Üniversitesi

Dr. Öğr. Üyesi Muhammed BAYDERE  
Karadeniz Teknik Üniversitesi

## **BEYAN**

Bu tezin yazılmasında bilimsel ahlak kurallarına uyulduğunu, başkalarının eserlerinden yararlanılması durumunda bilimsel normlara uygun olarak atıfta bulunulduğunu, kullanılan verilerde herhangi bir tahrifat yapılmadığını, tezin herhangi bir kısmının bu üniversite veya başka bir üniversitedeki başka bir tez çalışması olarak sunulmadığını beyan ederim.

Aziz Berkan KARADELİ

09/09/2025

## ABSTRACT

Adopting a correlation-centered evaluation framework, this thesis investigates how well the reference-free metric COMETKiwi-XL aligns with human MQM 2.0 judgements. Within this framework, the study compares segment- and system-level relationships between normalized MQM points (RQS) and COMETKiwi-XL scores. The sample of the study is English↔Turkish football UGC—tweets rich in slang, abbreviations, emojis, and culture-specific references. A small, domain-focused dataset of 20 tweets (10 per direction) was translated by three machine translation systems (DeepL, ChatGPT-o3, and X Translate), yielding 60 outputs. Three expert annotators applied MQM; category×severity counts were converted to APT → PWPT → RQS (0–1) and then averaged per segment. We compared these human scores with COMETKiwi-XL using Pearson’s  $r$  and Kendall’s  $\tau$ , at both the segment and system levels. Findings suggest moderate yet inconsistent segment-level correlations, partly driven by floor effects (multiple items with RQS = 0.000) that compress variability. At the system level, MQM and COMETKiwi-XL converge to a larger extent in EN-TR but diverge substantially in TR-EN, yielding a reversed ordering of systems. Regarding MQM error tags, the distribution was balanced across translation directions. Although ChatGPT-o3 generally achieves higher RQS than the other systems, the central takeaway of this study is methodological rather than system-specific. MQM, by design, can reduce poor translations to very low scores, sometimes reaching 0.000, which indicates complete failure. COMETKiwi-XL, in contrast, distributes outputs on a smoother scale, placing scores closer together and reflecting quality differences in a more gradual manner. Given the limited size and single domain, the results should be interpreted cautiously. Even so, they provide some indications of how human-in-the-loop MQM and a reference-free automatic metric might be used together, while also pointing to the need for larger, multi-domain datasets to better examine MQM–COMET(Kiwi-XL) alignment.

**Keywords:**

machine translation, MQM, COMETKiwi-XL, user-generated content, football tweets,  
translation quality assessment



## ÖZ

Bu tez, korelasyon merkezli bir değerlendirme çerçevesi benimseyerek referanssız metrik COMETKiwi-XL'nin insan temelli MQM 2.0 değerlendirmeleriyle ne ölçüde örtüştüğünü incelemektedir. Bu çerçevede, normalleştirilmiş MQM puanları (RQS) ile COMETKiwi-XL skorları arasındaki ilişkiler hem segment hem de sistem düzeyinde karşılaştırılmıştır. Bu çalışmada, İngilizce↔Türkçe futbol alanına ilişkin, kullanıcılar tarafından üretilen içeriklerden (UGC) oluşan 20 tweetlik (iki dil yönünde 10'ar tane) bir veri seti kullanılmıştır. Bu tweetler, argo ifadeler, kısaltmalar, emojiler ve kültüre özgü göndermeler içermektedir. İlgili tweet'ler DeepL, ChatGPT-o3 ve X Translate (yerleşik Google Translate) tarafından çevrilmiş ve toplam 60 çıktı elde edilmiştir. Üç uzman değerlendirici MQM 2.0 üzerinden değerlendirmelerde bulunmuş; ilgili değerlendirmeler kategori×önem puanları  $APT \rightarrow PWPT \rightarrow RQS$  (0–1) hesaplamalarına dönüştürülmüş ve her segment için ortalama alınmıştır. Elde edilen MQM ham skorları Pearson's  $r$  ve Kendall's  $\tau$  ölçümleri kullanılarak COMETKiwi-XL sonuçlarıyla karşılaştırılmıştır. Bulgular, segment düzeyinde orta derecede ancak tutarsız bir korelasyon ortaya koymaktadır. Bu durum kısmen, varyansı daraltan ve birden çok çıktının  $RQS = 0.000$  almasıyla ortaya çıkan taban etkisinden kaynaklanmaktadır. Sistem düzeyinde ise, MQM ve COMETKiwi-XL EN-TR yönünde yüksek ölçüde örtüşürken, TR-EN yönünde belirgin biçimde ayrışmakta ve bahsi geçen sistemleri tamamen zıt biçimde sıralamaktadır. Ek olarak, insan değerlendiricilerin MQM hata seçim sayılarına bakıldığında ise dağılım çeviri yönleri arasında dengeli olmuştur. ChatGPT-o3 genellikle diğer sistemlerden daha yüksek RQS elde etse de bu çalışmanın temel çıkarımı sistem üzerine olmaktan ziyade yöntemsel bir gözlem içermektedir. Doğası gereği MQM, düşük kaliteli çevirileri 0.000 gibi minimum puanlara indirip tamamen başarısız çeviri olarak puanlayabiliyorken, COMETKiwi-XL, çevirileri kademeli bir şekilde değerlendirerek puanları birbirine daha yakın konumlandırmakta ve kalite farklılıklarını daha aşamalı bir biçimde yansıtmaktadır. Bu çalışmada kullanılan veri setinin sınırlı büyüklüğü ve tek alanlı yapısı nedeniyle elde edilen sonuçlar dikkatle değerlendirilmelidir. Buna rağmen, sınırlı veri

seti ile bu çalışma, insan temelli MQM ve referanssız otomatik metriklerin birlikte nasıl kullanılabileceğine dair çıkarımlar sağlarken, MQM–COMET (Kiwi-XL) uyumunu daha iyi incelemek adına daha büyük ve çok alanlı veri setlerine ihtiyaç olduğuna da işaret etmektedir.

**Anahtar Kelimeler:**

makine çevirisi, MQM, COMETKiwi-XL, kullanıcılar tarafından üretilen içerikler, futbol tweetleri, çeviri kalite değerlendirmesi



## ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my thesis advisor, Prof. Dr. Ayşe Banu KARADAĞ, for being a constant source of wisdom, support, and inspiration. Her insightful feedback and unwavering encouragement have profoundly shaped not only this thesis but also my approach to lifelong learning. Being her student for three years has been a privilege, and her mentorship has left an enduring mark on my academic and personal development.

Words cannot fully express my gratitude to my life mentor, Asst. Prof. Dr. Muhammed BAYDERE, with whom I've had the privilege of working on various projects for nearly seven years. Throughout my MA thesis, he supported me as a true co-supervisor, offering invaluable expertise in Machine Translation and Artificial Intelligence. His guidance, both academic and personal, has been a constant source of motivation. Completing this thesis with his unwavering support has been one of my greatest fortunes.

My heartfelt thanks go to Prof. Dr. Işın ÖNER, whose intellectual presence has always challenged us to think deeper and aim higher. Her insightful remarks—even a single sentence during a session—have often been transformative, encouraging us to exceed our limits.

I would also like to express my sincere appreciation to Assoc. Prof. Dr. Nilüfer ALİMEN, whose guidance in research methodologies and unwavering support throughout this journey have been truly invaluable.

I am forever grateful to my beloved family, Gülizar KARADELİ, Turgay KARADELİ, and my brother, Burak KARADELİ for their unconditional love, sacrifices, and faith in me. Their constant support has been the foundation of everything I have achieved.

Special thanks to İrem SİRKECİOĞLU, who has stood by me through every challenge. Her belief in me and continuous encouragement gave me strength when I needed it most.

Lastly, I am deeply grateful to the annotators for their support throughout the thesis.

## TABLE OF CONTENTS

|                                                                        |           |
|------------------------------------------------------------------------|-----------|
| TEZ ONAY SAYFASI.....                                                  | ii        |
| BEYAN.....                                                             | iii       |
| ABSTRACT .....                                                         | iv        |
| ÖZ .....                                                               | vi        |
| ACKNOWLEDGEMENTS.....                                                  | viii      |
| TABLE OF CONTENTS .....                                                | ix        |
| LIST OF ABBREVIATIONS.....                                             | xii       |
| LIST OF TABLES.....                                                    | xv        |
| LIST OF FIGURES .....                                                  | xvii      |
| <br><b>CHAPTER 1: INTRODUCTION .....</b>                               | <b>1</b>  |
| 1.1 General Remarks.....                                               | 1         |
| 1.2 Statement of the Problem.....                                      | 2         |
| 1.3 Purpose of the Study .....                                         | 3         |
| 1.4 Research Questions.....                                            | 4         |
| 1.5 Methodology .....                                                  | 5         |
| 1.6 Significance of the Study .....                                    | 6         |
| 1.7 Organization of the Thesis .....                                   | 8         |
| <br><b>CHAPTER 2: LITERATURE REVIEW .....</b>                          | <b>10</b> |
| 2.1 Machine Translation Systems: From Rule-Based to LLMs .....         | 10        |
| 2.1.1 Rule-Based Systems.....                                          | 11        |
| 2.1.2 Corpus-Based Systems.....                                        | 12        |
| 2.1.3 Statistical Machine Translation .....                            | 13        |
| 2.1.4 Example-Based Machine Translation .....                          | 13        |
| 2.1.5 Neural-Based Systems.....                                        | 14        |
| 2.1.6 Hybrid Machine Translation .....                                 | 15        |
| 2.1.7 Large Language Models in Translation.....                        | 16        |
| 2.1.8 Recent Advancements in Machine Translation Technologies .....    | 17        |
| 2.2 Power of Social Media.....                                         | 18        |
| 2.2.1 Social Media and Football.....                                   | 18        |
| 2.3 Machine Translation Quality Evaluation.....                        | 20        |
| 2.3.1 Early Approaches of MT Evaluation.....                           | 21        |
| 2.3.2 Neural-Based Evaluation Metrics - COMET Family.....              | 22        |
| 2.4 Human Evaluation - Multidimensional Quality Metrics (MQM).....     | 23        |
| 2.4.1 Details on Reference-Free Metrics and the COMETKIWI Family ..... | 24        |
| 2.4.2 COMETKIWI.....                                                   | 25        |
| 2.5 Related Studies on MT Evaluation .....                             | 27        |
| <br><b>CHAPTER 3: METHODOLOGY .....</b>                                | <b>34</b> |

|                                                                       |           |
|-----------------------------------------------------------------------|-----------|
| 3.1 Research Design .....                                             | 34        |
| 3.2 Rationale for Dual Approach.....                                  | 35        |
| 3.2.1 A Novel Evaluation Design: COMETKiwi-XL Combined with MQM ..... | 36        |
| 3.2.2 Dataset and Translation Output Collection .....                 | 38        |
| 3.2.3 Translation Systems Used .....                                  | 39        |
| 3.3 Human Annotation Through MQM.....                                 | 39        |
| 3.3.1 MQM Error Typologies Explanation .....                          | 40        |
| 3.3.2 MQM 2.0 .....                                                   | 43        |
| 3.3.3 COMETKiwi-XL Evaluation .....                                   | 45        |
| 3.4 Statistical Tools .....                                           | 46        |
| 3.4.1 Inter-Rater Agreement.....                                      | 46        |
| 3.4.2 Comparative Analysis .....                                      | 47        |
| 3.5 Translation Systems Used.....                                     | 50        |
| 3.5.1 X (formerly Twitter).....                                       | 50        |
| 3.5.2 DeepL Translation.....                                          | 52        |
| 3.5.3 ChatGPT-o3 .....                                                | 53        |
| <b>CHAPTER 4: RESULTS .....</b>                                       | <b>54</b> |
| 4.1 Inter-Rater Agreement (IRA).....                                  | 54        |
| 4.2. MQM Error Analysis.....                                          | 56        |
| 4.3 Overall MQM Error Tags .....                                      | 62        |
| 4.3.1 Number of Error Tags by System .....                            | 64        |
| 4.3.2 Number of Error Tags by Category.....                           | 65        |
| 4.3.3 Number of Severity Tags .....                                   | 67        |
| 4.3.4 Number of Error Tags by Category and System .....               | 69        |
| 4.4 Representative Segment-Level Examples from MQM Annotations .....  | 70        |
| 4.5 COMETKiwi-XL Results.....                                         | 74        |
| 4.6 Correlation Analysis .....                                        | 78        |
| <b>CHAPTER 5: DISCUSSION .....</b>                                    | <b>81</b> |
| 5.1 Inter-Rater Agreement .....                                       | 81        |
| 5.2 MQM Error Analysis.....                                           | 82        |
| 5.2.1 Normalized MQM RQS Results .....                                | 82        |
| 5.2.2 Number of Error Tags in MQM Annotation .....                    | 84        |
| 5.2.3 General Error Performance of the Systems.....                   | 84        |
| 5.2.4 Direction-Based Error Distribution .....                        | 85        |
| 5.2.5 Per-System MQM Performance in Context .....                     | 85        |
| 5.3 Segment Level Correlation .....                                   | 86        |
| 5.4 System Level Correlation .....                                    | 88        |
| <b>CHAPTER 6: CONCLUSION .....</b>                                    | <b>90</b> |
| REFERENCES .....                                                      | 94        |

|                                                          |     |
|----------------------------------------------------------|-----|
| APPENDICES .....                                         | 108 |
| APPENDIX A – ENGLISH TWEETS .....                        | 108 |
| APPENDIX B – TURKISH TWEETS .....                        | 113 |
| APPENDIX C – ANNOTATION FRAMEWORK .....                  | 118 |
| APPENDIX D – IRA CALCULATION .....                       | 121 |
| APPENDIX E – COMETKiwiXL RESULTS FRAMEWORK .....         | 122 |
| APPENDIX F – SEGMENT LEVEL CORRELATION CALCULATION ..... | 123 |
| APPENDIX G – SYSTEM LEVEL AVERAGES .....                 | 124 |
| APPENDIX H – SYSTEM LEVEL CORRELATION CALCULATION .....  | 124 |
| APPENDIX I – MQM RAW QUALITY SCORE FRAMEWORK .....       | 125 |
| APPENDIX J – MQM RAW QUALITY CALCULATION FORMULAS .....  | 127 |
| APPENDIX K – PYTHON PIPELINE FOR RATER AGGREGATION ..... | 128 |
| ÖZGEÇMİŞ .....                                           | 132 |

## LIST OF ABBREVIATIONS

| Abbreviations | Bibliographic Information                                               |
|---------------|-------------------------------------------------------------------------|
| AI            | Artificial Intelligence                                                 |
| ALPAC         | Automatic Language Processing Advisory Committee (Report)               |
| API           | Application Programming Interface                                       |
| AWS           | Amazon Web Services                                                     |
| BERT          | Bidirectional Encoder Representations from Transformers                 |
| BLEU          | Bilingual Evaluation Understudy                                         |
| BLEURT        | Bilingual Evaluation Understudy with Representations from Transformers  |
| BPE           | Byte-Pair Encoding                                                      |
| CBMT          | Corpus-Based Machine Translation                                        |
| CNN           | Convolutional Neural Network                                            |
| COMET         | Cross-lingual Optimized Metric for Evaluation of Translation            |
| COMETKiwi     | COMET with the KIWI predictor–estimator quality-estimation architecture |
| COMETKiwi-XL  | COMETKiwi model scaled to XLM-R, Extra Large parameters                 |
| EBMT          | Example-Based Machine Translation                                       |
| EMA           | European Medicines Agency                                               |
| FDA           | Food and Drug Administration                                            |
| FIFA          | Fédération Internationale de Football Association                       |
| GNMT          | Google Neural Machine Translation                                       |
| GPT           | Generative Pre-trained Transformer                                      |
| HOPE          | Human-Oriented and Professional Evaluation                              |
| ISO           | International Organization for Standardization                          |
| IRA           | Inter-Rater Agreement                                                   |

|        |                                                                                                             |
|--------|-------------------------------------------------------------------------------------------------------------|
| IQR    | Interquartile range                                                                                         |
| LGP    | Language for General Purposes                                                                               |
| LEPOR  | Length Penalty, Precision, Recall and n-gram Order metric                                                   |
| LLM    | Large Language Model                                                                                        |
| LSTM   | Long Short-Term Memory                                                                                      |
| METEOR | Metric for Evaluation of Translation with Explicit ORdering                                                 |
| ML     | Machine Learning                                                                                            |
| MQM    | Multidimensional Quality Metrics                                                                            |
| MT     | Machine Translation                                                                                         |
| NLP    | Natural Language Processing                                                                                 |
| PFSMB  | Parallel French Social Media Bank                                                                           |
| PIL    | Patient Information Leaflet                                                                                 |
| PMUMT  | Parallel <i>Multi-domain</i> User-Generated-Content Machine-Translation corpus (Rosales Núñez et al., 2021) |
| Q&A    | Question and Answer                                                                                         |
| QE     | Quality Estimation                                                                                          |
| RBP    | Role-Based Prompt                                                                                           |
| RBMT   | Rule-Based Machine Translation                                                                              |
| RNN    | Recurrent Neural Network                                                                                    |
| SMT    | Statistical Machine Translation                                                                             |
| TER    | Translation Edit Rate                                                                                       |
| TM     | Translation Memory                                                                                          |
| TQA    | Translation Quality Assessment                                                                              |
| UGC    | User-Generated Content                                                                                      |
| WMT    | Workshop on Machine Translation                                                                             |
| XLm-R  | Cross-lingual Language Model – RoBERTa                                                                      |
| XL     | Extra Large (model size)                                                                                    |

|                             |                                                                                   |
|-----------------------------|-----------------------------------------------------------------------------------|
| $\tau$ -b (Kendall's Tau-b) | Non-parametric correlation coefficient, used for ordinal data in annotation tasks |
| $r$ (Pearson's $r$ )        | Pearson correlation coefficient measures linear correlation between MQM scores    |
| $\tau$ (Kendall's Tau)      | Kendall's Tau correlation coefficient, used for rank correlation                  |
| ChatGPT-o3                  | OpenAI's GPT-o3 large language model, used for translation evaluation             |
| X Translate                 | Twitter's (X) in-app translation system, powered by Google Translate              |

## LIST OF TABLES

|                                                                                   |    |
|-----------------------------------------------------------------------------------|----|
| <b>Table 1:</b> Accuracy Errors .....                                             | 40 |
| <b>Table 2:</b> Linguistic Conventions Errors .....                               | 40 |
| <b>Table 3:</b> Style Errors .....                                                | 41 |
| <b>Table 4:</b> Audience Appropriateness Errors .....                             | 41 |
| <b>Table 5:</b> Terminology Errors .....                                          | 41 |
| <b>Table 6:</b> Inter-Rater Agreement (Kendall's Tau-b) .....                     | 55 |
| <b>Table 7:</b> Overall Agreement of Three Annotators (Kendall's Tau-b) .....     | 55 |
| <b>Table 8:</b> Mean MQM RQS Scores (by system and direction) .....               | 59 |
| <b>Table 9:</b> Segment-level MQM RQS Scores (TR-EN) .....                        | 61 |
| <b>Table 10:</b> Segment-level MQM RQS Scores (EN-TR) .....                       | 62 |
| <b>Table 11:</b> Total Number of Error Tags by Translation Direction .....        | 63 |
| <b>Table 12:</b> Total Number of Error Tags by System .....                       | 63 |
| <b>Table 13:</b> Error Tags by System in EN-TR Translations .....                 | 64 |
| <b>Table 14:</b> Error Tags by System in TR-EN Translations .....                 | 65 |
| <b>Table 15:</b> Error Tags by Category .....                                     | 65 |
| <b>Table 16:</b> Error Tags by Category in EN-TR Translations .....               | 66 |
| <b>Table 17:</b> Error Tags by Category in TR-EN Translations .....               | 67 |
| <b>Table 18:</b> Error Tags by Severity (All Systems, Both Directions) .....      | 67 |
| <b>Table 19:</b> Error Tags by Severity in EN-TR Translations .....               | 68 |
| <b>Table 20:</b> Error Tags by Severity in TR-EN Translations .....               | 69 |
| <b>Table 21:</b> Error Tags by Category and System .....                          | 69 |
| <b>Table 22:</b> Representative Segment (TR-EN) with Inline MQM Annotations ..... | 70 |

|                                                                                      |    |
|--------------------------------------------------------------------------------------|----|
| <b>Table 23:</b> Representative Segment (TR-EN) with Inline MQM Annotations .....    | 71 |
| <b>Table 24:</b> Representative Segment (EN-TR) with Inline MQM Annotations .....    | 72 |
| <b>Table 25:</b> Representative Segment (EN-TR) with Inline MQM Annotations .....    | 73 |
| <b>Table 26:</b> System-Level COMETKiwi-XL Scores.....                               | 75 |
| <b>Table 27:</b> Segment-Level COMETKiwi-XL Scores (EN-TR) .....                     | 76 |
| <b>Table 28:</b> Segment-Level COMETKiwi-XL Scores (TR-EN) .....                     | 77 |
| <b>Table 29:</b> Segment-Level Correlation between MQM and COMETKiwi-XL.....         | 78 |
| <b>Table 30:</b> System-Level MQM vs. COMETKiwi-XL Averages.....                     | 79 |
| <b>Table 31:</b> System-Level Correlation Results between MQM and COMETKiwi-XL ..... | 79 |

## LIST OF FIGURES

|                                                                                             |    |
|---------------------------------------------------------------------------------------------|----|
| <b>Figure 1:</b> Screenshot of Turkish Tweet with Slang and Informal Expressions .....      | 38 |
| <b>Figure 2:</b> Screenshot of English Tweet with Slang and Informal Expression.....        | 38 |
| <b>Figure 3:</b> MQM Selected Error Types .....                                             | 39 |
| <b>Figure 4:</b> Screenshot of the Annotation Framework.....                                | 42 |
| <b>Figure 5:</b> MQM 2.0. Evaluation Scorecard (Lommel et al., 2024) .....                  | 44 |
| <b>Figure 6:</b> Raw Quality Score Calculation (Lommel et al., 2024).....                   | 49 |
| <b>Figure 7:</b> Distribution of MQM Raw Quality Scores (RQS) by System and Direction ..... | 60 |

# CHAPTER 1: INTRODUCTION

## 1.1 General Remarks

People's lives have been profoundly influenced by social media in this digital age. In the moments of both joy and sorrow, individuals share their experiences, ideas, passions, hobbies, favorite music, and feelings online, which means exposing almost every part of their life. Today Cambridge Dictionary (2025) defines social media “websites and computer programs that allow people to communicate and share information, opinions, pictures, videos, etc. on the internet, especially social networking websites”.

Social media is in heavy use and has been a versatile tool for modern people. Numbers also prove that social media is used by billions. According to the data provided by DataReportal’s chief analyst and CEO of Kepios, Simon Kemp (2024), the number of global internet users has reached 5.44 billion in 2024 with the addition of 178 million new users, resulting in a global penetration rate of 67.1 percent and indicating that more than two-thirds of the world’s population now accesses the internet. In terms of sectoral contexts, social media is mostly used by journalists, broadcasters, public relations professionals, advertisers, marketers and people in business and non-profit sectors (cf. Lipschultz, 2018: 13), for the reasons of advertising, brand popularity, visibility and sales (Özevin, 2021). Directly or indirectly, these factors mentioned above are also relevant within the sports industry. For instance, global events such as the Super Bowl, FIFA World Cup, and the Olympics, along with professional teams like Manchester United and Real Madrid, internationally recognized brands like Converse, and influential figures such as Lionel Messi, and Cristiano Ronaldo, all invest significant resources to incorporate social media into their major marketing strategies (cf. Filo et al., 2015:166). Considering that the use of social media is growing extremely day by day, and users globally produce a great amount of content on it, accurate translation of this content is quite essential to prevent misinterpretation and misconceptions across various fields. Among these fields, the necessity of accurate translations is particularly

critical in football, a highly globalized sport where mistranslations are avoided, which might cause significant issues (cf. Baines, 2019:24), which constitute the core of this study.

Therefore, employing the state-of-the-art, reference-free metric COMET-KIWI-XL (Rei et al., 2023) together with human annotation under Multidimensional Quality Metrics (MQM) (Lommel et al., 2024), this study conducts a comparative translation-quality evaluation of DeepL, ChatGPT-o3, and X's built-in translator (rebranded as X) on Turkish ↔ English football-related user-generated content (UGC, hereafter), tweets posted on X.

## 1.2 Statement of the Problem

Even though the amount of UGC is climbing up day by day, translation of this content in social media remains a significant challenge. The emergence and proliferation of UGC on social media have significantly impacted cross-linguistic communication. UGC is inherently informal and often includes non-standard spelling, emoticons, slang, and other noise features, all of which pose substantial challenges for machine translation (MT) systems. Even the latest MT technology, Neural Machine Translation (NMT) (Poibeau, 2017), faces significant problems in translating noisy UGC (Rosales-Núñez et al., 2021a) since "they contain linguistic challenges that are endemic to the medium, including non-standard punctuation, capitalization, spelling, vocabulary, and syntax." (Eisenstein, 2013:359) and UGC features "graphical stretching (yesssss → yes), replication of punctuation marks (????? → ?), emoticons, sometimes used as a verb (Je t'<3 → Je t'aime)" (Sanguinetti et al., 2020:496), which results in a non-standard language challenging MT systems.

In the football domain, the stakes are even higher since social media is used to express emotionally charged content in real time. The study of Alsubhi (2024) demonstrated that translating football-related tweets is still challenging. Similarly, Saadany et al. (2021a) have underlined that MT systems frequently mistranslate the emotions conveyed in Arabic football-related UGC, and the outputs lack grammatical and semantic coherence, which might change the emotional polarity due to specific linguistic or contextual factors. Despite

the engagement of the football industry with diverse professionals and fans globally, the phenomenon of cross-linguistic communication remains underexplored in academic research (cf. Baines, 2019:27). This gap highlights the need for further studies elaborating on football-related UGC.

From an evaluation perspective, existing MT quality metrics might overlook domain-specific failure cases. Candel-Mora (cf., 2022:133-134) stated that a uniform, universal grading system for evaluating translation quality does not exist, and the evaluation of translations mostly depends on the nature of the content and its intended purpose. This concern is also evidenced by the utilization of more sophisticated automatic evaluation instruments such as a reference-free COMETKIWI-XL metric with over 3.5b parameters (Rei et al., 2022), designed to be used in combination with human evaluation with MQM (Lommel et al., 2014, 2024), and together deeply analyze and categorize translation problems in noisy, domain-specific datasets. In this study, the said combination is expected to facilitate the comprehension of translation problems encountered in football-related UGC.

### **1.3 Purpose of the Study**

Tweets have unique linguistic characteristics shaped by a localized, culture-specific language (Imran et al., 2016). Furthermore, studies show that there is an extensive use of slang on X, causing the language to be context-specific, which reflects the informal and playful nature of the platform (Izazi and Sepora, 2020). Parallely, fan-generated football tweets often contain emotionally charged and linguistically irregular expressions. For that reason, football-related UGC content is quite hard to translate not only for translation systems, but also for human translators. The increasing complexity of UGC, particularly on platforms such as X, causes further challenges in the evaluation of MT. Nagarajan and Gamon (2011) suggest that “a large portion of language found in UGC is in the Informal English domain — a blend of abbreviations, slang and context-specific terms; lacking in sufficient context and regularities; and delivered with an indifferent approach to grammar and spelling” (p. iii). This phenomenon demonstrates the unique qualities of social media content compared with other

text genres, affecting the accuracy of MT systems. Similarly, Ding et al. (2021) underline that non-professional content creators on social media, with their limited proficiency in any language pair, provide translated content or subtitles and break down linguistic borders. Yet, it is another factor damaging translation accuracy.

These critical points highlight the need for more comprehensive and domain-sensitive evaluation frameworks. As noted by Freitag et al. (cf., 2022:56), traditional metrics such as BLEU and chrF frequently fail to align with human judgments in particular domains such as news, social media, and e-commerce, thereby underscoring the importance of adopting more sophisticated evaluation strategies. Therefore, combining MQM 2.0 with COMETKIWI-XL is expected to offer a more robust evaluation alternative.

Building on this phenomenon, the study aims at using a two-part novel evaluation method that combines the reference-free neural metric COMETKIWI-XL (Rei et al., 2023) with expert human scoring via MQM 2.0 (Lommel et al., 2024). We apply this approach to three major systems: OpenAI’s new ChatGPT-o3 (2025), the widely used DeepL engine (2024), and X’s in-app translator, whose effectiveness is still debated. Since football tweets are loaded with slang, casual phrasing, and fast-changing context, they are quite challenging for any machine-translation system. With this methodology, the study targets to demonstrate how well each engine deals with real-world, highly informal language of football-related UGC and, at the same time, explores how closely scores from COMETKIWI-XL correlate with human annotation of MQM scores.

## **1.4 Research Questions**

1. Which machine translation system demonstrates the highest overall performance in translating football-related tweets between English and Turkish, as measured by MQM and COMETKIWI-XL?
2. Which machine translation system receives the highest MQM score from expert human annotators for football tweet translations in both English–Turkish and Turkish–English directions?

3. Which machine translation system receives the highest COMETKIWI-XL score for football tweet translations in both English–Turkish and Turkish–English directions?
4. How do DeepL, ChatGPTo3 and X Translate differ in terms of MQM error categories, particularly with respect to slang, idiomatic expressions, and informal language, in English–Turkish and Turkish–English football tweet translations?
5. To what extent do COMETKIWI-XL scores correlate with human-assigned MQM scores in evaluating football tweet translations between English and Turkish?

## 1.5 Methodology

This case study employs a mixed-methods evaluation technique to evaluate the effectiveness of the translation of football-related UGC on social media. Evaluating UGC translation in a specific domain through traditional metrics might be misleading. For instance, Núñez et al. (2021b) suggested that standard metrics are unable to demonstrate translation quality clearly. Therefore, they created specific parallel data, PMUMT, which is made of real-world UGC, for evaluating MT of UGC content. With this parallel data, they evaluated NMT models by mainly using BLEU and chrF2. Their study proves that researchers need to initiate solutions for the evaluation of UGC to compensate the informal nature of it.

In light of this, the current study also initiates a novel method by combining automatic scoring techniques with manual human assessment to provide a comprehensive evaluation of translation performance. Specifically, it integrates human annotation utilizing the latest version of the Multidimensional Quality Metrics (MQM 2.0) (Lommel et al., 2024) with the scores of COMETKIWI-XL, a newly introduced fine-grained reference-free neural metric (Rei et al., 2023).

For the evaluation, 20 football-related tweets were selected on X, focusing on characteristics that often pose difficulties for MT systems, including slang, informal phrases, acronyms, and football-specific terminology. Three prominent MT systems—DeepL, ChatGPT-o3, and the in-built translation of X—were employed to translate each tweet into

the target language. This resulted in a total of 60 translations, which were evaluated by human annotators through MQM and the automatic metric, COMETKiwi-XL.

Among the error typologies of MQM (Multidimensional Quality Metrics, n.d.), the current study opts for Accuracy, Linguistic conventions, Style, Terminology and Audience appropriateness. In order to apply these error typologies consistently and smoothly, the annotators were trained using the latest version of MQM and provided with particular specifications, which are explained in Chapter 3 in detail. Then, the three expert annotators who have over 10 years of experience in Translation Studies evaluated the three sets of outputs. With the aim of securing inter-rater reliability, we employed Kendall's Tau-b to evaluate the agreement among the annotators.

Simultaneously, all translations were evaluated using COMETKIWI-XL, which assigns a quality score between 0 and 1. The combination of human judgments and neural quality estimation, supported by correlation analysis between the two, enables a comprehensive assessment of translation performance for informal and domain-specific UGC. Further methodological details, including metric configurations, rater procedures, and statistical analysis techniques, are provided in Chapter 3.

## **1.6 Significance of the Study**

Being a global game, Football has over 5 billion football fans, as the number of total fan engagements in FIFA World Cup Qatar 2022 was reportedly over 5 billion (FIFA, n.d.). Football is beyond a simple game or sport; it is a social and cultural object (Moreau et al., 2021). Although football unites many different professional communities interacting with worldwide audiences, studies on the issue of cross-linguistic communication have received almost no attention (Baines, 2019). However, since football is a global sport that depends heavily on cross-cultural communication, it is essential that the translated content be clear and accurate, without critical errors that could lead to misunderstandings.

This global nature of football has led to a significant increase in multilingual discourse, especially on digital platforms like X, where football fans generate huge amounts

of content daily. For instance, with the signings of international stars, Saudi Arabian football clubs have been trying hard to promote their global outreach, and this leads to a significant growth in the multilingual content often combining Arabic and English (Alsubhi, 2024). Similarly, clubs in Türkiye have been signing footballers and managers who have global influence in football, resulting in a similar increase in multilingual content production. This increased football-related UGC content is usually in a unique and informal language. For example, Lohar (2020:38) demonstrates the challenges in translating football tweets, “a tweet from the football domain... here is full of both grammar and spelling errors. The user wrote BT instead of But, the small letter i instead of the capital I, and dnt instead of don’t.” These facts mentioned above prove the need for actions to be taken for improving translation quality in the social media context.

Correspondingly, the study addresses the need for analyzing and eventually improving the translation quality of UGC in the football domain with a very unique methodology and by specifically focusing on tweets on X posted by football fans.

Therefore, the current study is significant in three striking ways:

- (1) Although prior studies have investigated translation quality across diverse domains—such as literary texts on social media (Fakih et al., 2024), Chinese classical poetry (Gao et al., 2024), user-generated microblogs (McNamee and Duh, 2022), Classical Arabic verse (Farghal and Haider, 2024), ChatGPT vs. NMT output alignment (Jiang et al., 2024), literalness in GPT outputs (Raunak et al., 2023), and general translation performance of GPT models across language pairs (Hendy et al., 2023)—to the best of our knowledge, none have adopted the distinctive methodology proposed in this study—namely, the combination of the novel reference-free metric COMETKiwi-XL, which has been shown to outperform traditional lexical overlap metrics such as BLEU and chrF (Freitag et al., 2024; Agrawal et al., 2024), with MQM-based human annotation.

- (2) The study provides insights into the performance of widely used translation platforms when translating the football-related UGC posted on X, which is quite informal and context-specific.
- (3) Lastly, the study not only signifies the efficiency of combining human annotation through MQM and a state-of-the-art, fine-tuned, neural, reference free metric, COMETKIWI-XL in terms of how they correlate in this specific football domain, but also reveals the strengths and weaknesses of ChatGPT-o3, DeepL and X’s in-built translation system, which might help developers of these systems to improve them for better use.

## 1.7 Organization of the Thesis

This thesis is structured into seven chapters, each systematically exploring the translation quality of football-related UGC through a dual assessment methodology combining human annotation and reference-free evaluation.

**Chapter 1** introduces the research context. It outlines the statement of the problem, research objectives, key questions, significance of the study, and the rationale behind the chosen methodology.

**Chapter 2** presents a comprehensive literature review. It traces the evolution of machine translation systems—from rule-based models to large language models (LLMs)—and explores the linguistic nature of social media texts, particularly in the football domain. The chapter also evaluates the tools used for assessing machine translation quality, with a focus on Multidimensional Quality Metrics (MQM) and the COMETKIWI-XL model. Lastly, this chapter provides a concise overview of previous studies in the field, highlighting gaps and informing the research design of this thesis.

**Chapter 3** details the methodological framework of the study. It discusses the research design, data collection process, dual evaluation approach, translation systems used, the MQM 2.0 annotation process, the COMETKIWI-XL scoring procedure, and the statistical tools employed for inter-rater agreement and comparative analysis.

**Chapter 4** presents the results. It includes system-level and segment-level findings for each translation system in both English→Turkish and Turkish→English directions, based on MQM annotation and COMETKIWI-XL scores. It also compares the performance of DeepL, ChatGPT-o3, and X Translate.

**Chapter 5** discusses these findings in depth. The discussion includes error typology patterns, direction-based observations, and comparative insights, along with a reflection on the implications of using combined human and automatic evaluation methods.

**Chapter 6** concludes the thesis by summarizing the key findings, reflecting on the contributions and limitations of the study, and offering suggestions for future research in the field of machine translation evaluation, particularly in the context of social media and UGC.

## CHAPTER 2: LITERATURE REVIEW

### 2.1 Machine Translation Systems: From Rule-Based to LLMs

Dating back to the 17th century, the concept of MT has a long intellectual history, with early ideas rooted in the universal language movements of thinkers like Descartes and Leibniz (Hutchins and Somers, 1992). Even though the idea of an automated translation goes back to the letters between Descartes and Mersenne (Cottingham et al., 1991) and to the discussions on building a “universal language” (Schwartz, 2018), the first concrete steps were not taken until the 19<sup>th</sup> century. After the WW2, the very first research on MT was initiated by prominent scholars of the era, Andrew Booth in the UK and Warren Weaver in the US (cf. Poibeau, 2017: 30). Only known experiments before the memorandum were the two initiatives, one in France by French-Armenian George Artsrouni and the other one in Russia by the Russian researcher Petr Smirnov-Troyanskii (Hutchins and Somers, 1992). Nevertheless, these attempts received skeptical views, since there wasn’t any electronical computer back then (cf. Nirenburg et al., 2002: 61).

Warren Weaver, “The father of machine translation” (Poibeau, 2017: 32), published a memorandum in 1949, suggesting the possibility of using computers for MT (Schwartz, 2018). Even though the memorandum had drawbacks, the idea pushed researchers to carry out experiments. For example, an Israeli researcher, Joshua Bar-Hillel, acted as a quite important figure in the development of MT in the US during the 1950s (cf. Poibeau, 2017: 36). In these years, the UK, US and USSR governments started to support MT research significantly (Poibeau, 2017). However, things would not go as presumed in the coming years.

With the infamous ALPAC Report in 1966, the significant financial support given by the US was cut off, and the message was given, MT was hopeless (Hutchins, 1996). Following this, there was a long pause over 20 years in the field (Poibeau, 2017). Yet, after the late 1980s, new studies started to be published in the MT field (Hutchins and Somers, 1992). While the initial visions were philosophical and the field had some setbacks in the

past, today MT is a fast, accurate, and cost-efficient technology. As defined by the Cambridge Dictionary (2025), MT is “the process of changing text from one language into another language using a computer”. Modern MT systems have undergone several significant phases since their initial introduction in the 1950s (Rikters, 2019). Even though neural networks constitute the majority of today's best systems, they are founded on earlier rule-based and statistical versions.

### **2.1.1 Rule-Based Systems**

The original systems were rule-based machine translation (RBMT), employing manually built linguistic rules (Jolley and Maimone, 2022). In the next step, statistical (SMT) and phrase-based machine translation (PBMT) emerged, followed by NMT, which uses deep learning technology (Burchardt et al., 2020). Among these earlier versions, RBMT originated in the field of MT by manually constructing language rules. Initial MT studies favored these systems for their consistent outcomes and ability to provide comprehensible outputs (Angelone et al., 2020). Until the 1990s, RBMT was the dominant approach, relying heavily on bilingual dictionaries and hand-coded linguistic rules; however, the process of creating and maintaining these rules was quite labor-intensive, and their limited transferability across domains and languages limited the scalability of RBMT for open-domain and multilingual translation tasks (Wang et al., 2022).

RBMT approaches are typically grouped into direct, transfer, and interlingua models (Hutchins, 2003; Way, 2020).

1. The direct method translates word-by-word with minimal linguistic analysis (Schwartz, 2018).
2. The transfer approach generates intermediate syntactic and semantic structures in both source and target languages (Hutchins, 2003).
3. The Interlingua paradigm, dating back to Weaver’s famous memorandum in 1949, is the method of removing transfer stage and using interlingual intermediate representations instead (cf. Schwartz, 2018:26-27). Poibeau (2017) describes this

model as “very ambitious since they need both a complete understanding of the sentence to be translated and accurate generation components” (p. 30).

RBMT had significant advantages: It was transparent, straightforward to debug, and did not rely on extensive multilingual corpora (Rikters, 2019). Kenny (2020) indicates that developing a completely functional RBMT system was costly and challenging due to the substantial requirement for linguistic data and manual rule formulation. By the 1980s, parallel corpora became increasingly prevalent, and statistical and corpus-based methodologies began to outperform RBMT. This process resulted in the creation of data-driven systems and the subsequent emergence of NMT (Kenny, 2020).

### **2.1.2 Corpus-Based Systems**

By the end of the 1980s, major developments began to change the field of MT research. These changes may have marked the beginning of a fresh stage in its development. The introduction of "corpus-based" methods, alongside traditional "rule-based" systems, was one of the most significant advances (Hutchins, 1993:12). These methods rely heavily on big sets of multilingual data, called "corpora" (Rikters, 2019), instead of hand-coded linguistic rules. Contrary to RBMT that basically employs syntactic rules, CBMT matches linguistic units between a source and a target text within a singular language pair (Quah, 2006).

Poibeau (2017) states that since the late 1990s, aligned bilingual corpora have served as the foundation for research aimed at identifying word- or phrase-level translation equivalents. He adds that they also have contributed to the development of completely automated translation systems working with statistical analysis, which remained the predominant approach and pushed significant research advancements in the domain at that time. Way (2020) supports this view, stating that CBMT became central by the 1990s. CBMT is often considered in two sub-categories: Statistical Machine Translation (SMT, hereafter) and Example-Based Machine Translation (EBMT, hereafter) (Carl and Way, 2003). As their

name suggests, in the most basic terms, SMT works with statistical analysis and EBMT with examples in the reference translation. Next part elaborates on the two approaches.

### **2.1.3 Statistical Machine Translation**

SMT perceives the source text as a corrupted representation of an unknown phrase in an unfamiliar target language and attempts to reconstruct the original sentence by reverse decoding; thus, it attempts to detect the target-language text that is most probable to have originated from the source-language input (Samuelsson, 2012). Brown et al. (1993), in their *The Mathematics of Statistical Machine Translation: Parameter Estimation*, established and explained the foundation for SMT by employing expectation-maximization techniques to determine the probability of word alignment and translation in bilingual datasets. Wolk and Marasek (2019) also assert that SMT relies on statistical patterns seen in large parallel datasets. Therefore, these systems not only depend on models that match words between languages but also on language models by ensuring translations are grammatically accurate and sound natural in the target language.

### **2.1.4 Example-Based Machine Translation**

A conference presentation in 1981, later published by Nagao in 1984, is widely regarded as the formal beginning of the EBMT paradigm (cf. Hutchins, 2005:197). Thus, initiated and called “machine translation by example-guided inference” by Nagao (1984:346), EBMT relies on using pre-translated cases rather than explicit rules to identify matches. In other words, it uses a method similar to a dictionary to determine the similarity of the input to the examples (Somers, 2003a). EBMT is mostly found similar to Translation Memory (Quah, 2006), as both EBMT and Translation Memory (TM) involve matching input with stored authentic instances, which is the reason why these two are frequently perceived as similar to and occasionally confused with each other (Somers, 2003b). One significant disadvantage of this method is that if there are no close matches found in the bilingual

example set or if there is a wordplay in the input sentences, the method inevitably fails (Quah, 2006).

Overall, despite the differences between these two translation approaches, both use extensive bilingual datasets as the foundation for translation rather than relying on formal grammar-based processing. Furthermore, SMT employs both linguistic and translation models to generate output, whereas EBMT often relies solely on one linguistic model (Quah, 2006). Although both belong to the CBMT, they require distinct resources and employ different methodologies; hence, they ought to be considered as distinct yet interconnected translation methods.

### **2.1.5 Neural-Based Systems**

NMT emerged in the mid-2010s as a groundbreaking approach to translation, distinct from the segmented, pipeline-based framework of SMT. It provided a comprehensive, integrated neural framework, increasing the level of translation quality. Bahdanau et al. (2015) introduced the first influential attention-based NMT system, stating that it “aims at building a single neural network that can be jointly tuned to maximize the translation performance” (p. 1). They also proposed an attention mechanism that enables the model to “automatically (soft-)search for parts of a source sentence that are relevant to predicting a target word” (p. 1), which significantly helped overcome the limitations of fixed-length encoding vectors.

Prior to this, Sutskever et al (2014) developed an RNN-based encoder-decoder architecture, in which a multilayered Long Short-Term Memory (LSTM) first encodes the input sequence into a fixed vector, followed by another LSTM that decodes it into a target sequence.

The Transformer model by Vaswani et al. (2017) was another significant advancement, as the authors “propose[d] a new simple network architecture ... based solely on attention mechanisms, dispensing with recurrence and convolutions entirely” (p. 1). They highlighted its benefits, claiming it was “more parallelizable and [required] significantly less time to train” (p. 1), and it outperformed the previously named best models by reaching a

new the state-of-the-art performance with BLEU scores of 28.4 for English–German and 41.8 for English–French in WMT 2014 (p. 6).

In the meantime, the MT industry was also shifting to NMT models, as Google launched its GNMT system using RNN with attention mechanism in 2016 (Turovsky, 2016). Wu et al. (2016) developed the Google Neural Machine Translation (GNMT) system, which employs two recurrent neural networks: one for encoding the source text and the other for generating the target translation. They reported that the system requires substantial computational system requirements for both training and inference; yet GNMT significantly enhanced translation quality, reducing errors by over 60% in comparison to PBMT. Today, widely used MT systems such as Google Translate, DeepL and Microsoft Translator work with neural architecture.

### **2.1.6 Hybrid Machine Translation**

Hybrid machine translation, working with the combination of multiple approaches, today is used by several MT systems e.g., Yandex. Translate (Sukhoverkhov et al., 2019) and SYSTRAN (Poibeau, 2017). Basically, it combines MT approaches, including rule-based, statistical, example-based, and neural approaches, in order to use their advantages and cover up drawbacks. Way (2008) demonstrates that using a statistical language model in combination with an example-based approach may enhance translation quality. This report indicates that including statistical components into EBMT enhances the overall quality of the result. Artetxe et al. (2015) proposed a hybrid model in which an example-based preprocessor generates high-quality partial translations. The main SMT engine uses these segments as input to finish the output. The researchers claim that the preprocessor works like a translation memory, which improves the quality of translations by reusing segments and filling up gaps with statistics. Several hybrid MT frameworks employ a more layered approach. In 2017, Dhariya et al. proposed a three-tier architecture in which example-based and statistical components produce first draft translations, further refined by rule-based modules. This rule-

based post-processing is believed to enhance translations, making them more fluent, accurate, and grammatically correct compared to systems employing a singular translation approach.

Zhou et al. (2017) also developed a method to integrate neural systems that combines the outputs of both SMT and NMT models into a single structure. Their approach outperformed the leading single system by 5.3 BLEU points and exceeded the traditional method combination by 3.4 BLEU points, indicating a significant enhancement in translation quality. Numerous evaluations confirm the effectiveness of hybrid MT. Alves et al. (2006) assert that the integration of several paradigms enables systems to use their strengths in manners unachievable by a singular approach. They emphasize that sticking too tightly to a single approach might hinder progress, whereas integrating the most effective elements from other models produces superior translation outputs. Overall, one of the best effective characteristics of hybrid systems is that they are used when needed for a specific language pair. As Poibeau (2017) also suggests, for Indo-European languages like French and Spanish, which differ significantly from English, hybrid systems are likely to be favored as they account for both statistical patterns and language-specific features. These examples together demonstrate how hybrid systems continue to play a key role in translation quality development across MT research.

### **2.1.7 Large Language Models in Translation**

Recently, Large Language Models (LLMs) such as GPT-3, GPT-4, PaLM, and ChatGPT have been shown to be highly effective for MT, outperforming traditional NMT systems in certain respects. Vilar et al. (2023) assert that despite being predominantly trained on next-token prediction using extremely English and non-parallel multilingual data, LLMs showed quite significant proficiency in MT tasks. Similarly, Xu et al. (2023) highlighted that fine-tuning LLMs are quite good at translating, as they improve by over 12 BLEU and 12 COMET points across 10 language pairs and become comparable to the leading MT systems. Among LLMs, PaLM, heavily trained on multilingual non-parallel data, outperformed previous LLMs in language translation (Vilar et al., 2023) marking a significant advancement in zero-

shot translation. Another study comparing GPT-4 and human translators to depict GPT-4's translation performance demonstrated that it produced almost as identical number of errors as junior human translators, yet it still could not catch the output quality level of more experienced, mid- and senior-level professionals (Yan et al., 2024).

From a user-centered approach, Son and Kim (2023) carried out a comparative analysis on LLMs and NMT models, finding that despite having higher BLEU chrF and TER scores than ChatGPT overall, Google Translate and Microsoft Translator were outperformed by GPT-4 in specific language pairs such as German-English. Another example study focusing on specific language pairs (Chandra et al., 2025) reported that in translating the Bhagavad Gita (a sacred Hindu scripture) from Sanskrit into English, GPT-4o and GPT-3.5 performed better than Google Translate.

Therefore, it might be agreed that LLMs may perform better than NMT systems in certain language pairs. Naturally, LLMs also have some drawbacks, such as poor-quality training examples might reduce the output quality (Vilar et al., 2023) and translating in low-resource languages might be challenging (Yan et al., 2024). With the latest model, ChatGPT-o3, launched on April 16, 2025, OpenAI sets a new standard by integrating advanced reasoning into the model.

### **2.1.8 Recent Advancements in Machine Translation Technologies**

Over the past few years there have been tremendous improvements in MT technologies with the improvements in DeepL and the advent of ChatGPT (especially ChatGPT-4o and ChatGPT-o3) that take up the leading positions with the use of complex algorithms/architectures coupled with big data sets to enable translation across many languages. Thus, there is a constant increase in the user number of these platforms for the purpose of translation. As the data highlight, ChatGPT reportedly has over 200 million weekly active users (Backlinko, 2024). DeepL, however, states in their press release in Cologne (2024) that it is the most used MT provider among language provider companies

with a huge difference, reportedly having 82% of users, while Google has 46%, Microsoft 32%, and Amazon AWS 17%.

## **2.2 Power of Social Media**

The meaning people often attributed to social media is that it is an internet-based platform encouraging interaction among users, including forums, blogs, wikis, podcasts, social networks and content-sharing sites (Page et al., 2014). Naturally, it is not an inaccurate characterization; however, in today's context, social media might mean more than it is commonly perceived. For example, social media can be used for agenda-shifting or -setting to influence people. This is particularly evident in Taras (2022), where he suggests that elections are intensely fought through the media, with each candidate attempting to control the campaign agenda by drawing public attention to matters that benefit them while diverting attention away from ones that may benefit their opponents. Another example is the use of social media in the sports industry. Sports organizations invest large amounts of money to promote digital interaction and build strong relationships with their consumers due to the popularity of social media (Filo et al., 2015). Social media offers football teams several avenues to interact with their supporters, resulting in a significant growth in UGC in this domain. Moreover, with this increasing UGC, it offers club management insightful analysis of the motivations, loyalties, and relationships inside their supporter base (Kaya et al., 2024). Given the extensive use of social media within the football industry, including clubs, players, and fans, it is increasingly essential to understand how they interact in digital spaces (Guzman et al., 2021).

### **2.2.1 Social Media and Football**

The dynamic interaction between football teams and their fan base has been significantly strengthened by social media, with platforms such as Facebook, X, Instagram, and YouTube serving as the primary channels for the new generation to express themselves and engage

with the sport (cf., Kaya et al., 2024:70). Among these platforms, X is predominantly used by fans for textual communication, as it enables them to express their emotions and opinions through the platform's unique language. Also highlighted by Romero-Jara et al. (2024), X is recognized as the major social media platform within the football industry, showcasing the largest number of posts among the eight professional football leagues analyzed, surpassing both Facebook and Instagram in total activity. Given that it is a platform containing most football-related content and that the sentimental and emotional nuances of tweets may change during translation between languages, accurate translations of the posts are critical (Alhadlaq and Alnuaim, 2023). For instance, due to the potential for poor output from X's automatic translation tool, which may result in confusion and negatively impact brand image, professionals in the football industry usually prefer to hire human translators to ensure accuracy and therefore protect their reputation (Baines, 2019). In addition to the already complex and informal style of X, football-related tweets are filled with sport-specific vocabulary, which creates an even greater challenge for translation systems to produce accurate and contextually relevant translations. However, since football tweets reach a global audience, including fans, clubs, players, etc. (Price et al., 2013), it is extremely important to provide accurate and fluent translations, which is a prominent strategy addressing the commercial goals of the football industry (Baines, 2019).

Since the majority of tweets originate from fans expressing criticism of players, coaches, referees, and other clubs the analysis towards the quality of football-related tweets' translation is particularly useful for coaches, match analysts, and broadcasters, and also sheds light on the role of randomness in football outcomes (Wunderlich and Memmert, 2022). Nevertheless, this analysis is not easy, since Tweets are difficult to translate because they frequently contain emoticons, phrasal abbreviations such as "lol," "smh," or "ikr," expressive use of words such as "cooolll," shorter forms, typos, or words written in non-standard formats that follow irregular vocabulary and informal syntax (Neagu et al., 2022:2). Yet, a detailed analysis of UGC in the football domain could offer valuable insights into the dynamics of the game, helping to better understand - how in-play events shape the progression of a match (Wunderlich and Memmert, 2022).

### **2.2.2 Translating UGC**

X has become among the most extensively used platforms for individuals to share ideas on various topics (Pla and Hurtado, 2017). People use X for various purposes in numerous fields, including the sports industry where it has a huge impact (Kaya et al., 2024). Recent research in MT has been centered in analyzing UGC, particularly tweets, due to the unique challenges caused by informal language, shortness and emotional complexity in the posts. Saadany et al. (2021a) assert that the linguistic characteristics that are unique to X significantly hinder the accurate expression of emotions in other languages. These challenges arise from factors such as unusual spelling, colloquial language, the use of emojis and hashtags, and language that is sarcastic or filled with emotions, which usually doesn't exist in formal writing. In the PhD study of Lohar (2020), it is evident that football-related tweet translation is quite challenging due to the informal nature of the content. In addition, Saadany et al. (2023) emphasize that translating UGC, such as tweets, has many challenges, especially in effectively representing slang, emotional tone, and figurative language, including sarcasm and euphemisms. Considering the aforementioned facts, domain specific UGC translations are subject to inaccuracies and need to be analyzed deeply.

## **2.3 Machine Translation Quality Evaluation**

The criteria for evaluating MT technology have evolved as its performance has improved day by day. The evaluation of MT has always been accepted as quite necessary since their first emergence (Poibeau, 2017). The evaluation methods have significantly evolved from early traditional approaches to contemporary neural-based and quality rating standards. Apart from them, human evaluation started to be used to obtain information in the 1990s, even though it was quite costly and inconsistent (Poibeau, 2017). Today, the advanced versions of automatic measures and human evaluation methods are in use not only separately but also in combination.

This following section analyzes the evolution of MT evaluation methods, emphasizing significant developments during the early 2000s. It emphasizes the outcomes of the yearly WMT (Workshop on Machine Translation) shared tasks, which serve as a primary means to compare and evaluate assessment methodologies.

### **2.3.1 Early Approaches of MT Evaluation**

BLEU (Bilingual Evaluation Understudy), introduced by Papineni et al. in 2002, emerged as the predominant automated assessment in the early 2000s. The researchers noted that the primary component of BLEU involves determining the quantity of n-gram overlaps between the candidate translation and the reference translations, irrespective of the sequencing of the words. Eventually, a translation is of superior quality if it has a larger number of these matches. Despite its widespread usage, BLEU has several drawbacks. For instance, Liu and Zhu (2023) assert the primary issue is its emphasis on the precision of individual phrases rather than the overall coherence of the translated material. Moreover, Callison-Burch et al. (cf. 2006:252) also highlight that BLEU is incapable of matching between synonyms or lexical variations of words, resulting in zero contribution to the BLEU score.

To address the limitations of BLEU, several metrics were brought about. For example, in 2004, METEOR was designed to evaluate MT quality at the sentence level, with a particular focus on aligning closely with human judgment (Lavie and Denkowski, 2009). In addition, Translation Edit Rate (TER) was launched in 2006, evaluating the translation quality by calculating the extent of editing required for a system output and its alignment with a reference translation (cf. Snover et al., 2006:223). Another notable advancement has arisen with LEPOR in 2012, a robust metric using rich and augmented evaluation factors and producing scores similar to human evaluation (Han et al., 2012), which was proven in 2013 (Han et al., 2013) when the metric received the highest average score of 0.86.

### 2.3.2 Neural-Based Evaluation Metrics - COMET Family

The introduction of neural-based evaluation metrics has made a huge change in MT evaluation literature. COMET (Cross-lingual Optimized Metric for Evaluation of Translation) was created by a research group in 2020, namely Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. It was first presented as a neural-based framework for evaluating MT, mainly to increase the precision and dependability of automated MT evaluations (Rei et al., 2020). They explain that it is a PyTorch-based platform that works on evaluating MT output by using multilingual models to compare MT output with the source and reference translation, achieving a similar level with human quality assessments. Guerreiro et al. (2023) highlighted that COMET performed effectively in terms of sentence and word-based evaluation, it fails to detect and explain specific translation errors, causing a major limitation that could be filled by combining the automatic metric with human annotation.

Since its introduction, COMET has established itself as the new benchmark for assessing translations. In the 2020 article which Rei et al. presented COMET, stating it is “a neural framework ... which obtains new the state-of-the-art levels of correlation with human judgements” (p. 1). It was a quite significant advancement at that time. Notably, this earlier version of COMET was submitted to the WMT 2019 Metrics Shared Task, where it outperformed traditional metrics such as BLEU and chrF ranked among the top systems (Rei et al., 2020). The growing trend in the COMET family continued in 2022, on which Freitag et al. published their findings under a striking title, “Stop using BLEU – neural metrics are better and more robust.” Their metric comparison study in 2022 proved that neural-based metrics significantly outperform BLEU, spBLEU and chrF in correlation with human scores. This was a significant turning point for the MT evaluation field, illustrating the necessity of moving beyond traditional metrics.

In the same WMT 2022 task study, COMETKiwi performed the best reference-free metrics, outperforming other reference-free metrics such as UNITE-SRC, COMET-QE etc., and also performed better than various neural and traditional reference-based metrics (Freitag

et al., 2022). At WMT 2023, metrics such as COMET-22 and the more recent COMETKIWI shown superior performance in addressing contrastive test and MQM challenge sets (Freitag et al., 2023).

COMETKIWI, in particular, was built upon previous COMET models and incorporated improved multilingual generalization (Freitag et al., 2023). Moreover, with their study titled *Scaling up COMETKIWI: Unbabel-IST 2023 Submission for the Quality Estimation Shared Task*, Rei et al. (2023) launched the novel COMETKIWI model improved with XLM-R XL and XXL models, which outperformed the latest model, COMETKIWI-22, up to 10 Spearman points. In the WMT24 results, which is the last metrics shared task, the domination of fine-tuned neural based the metric continued even while evaluating LLM-based translation systems (Freitag et al., 2024).

## **2.4 Human Evaluation - Multidimensional Quality Metrics (MQM)**

Started “In the mid 1990s” with only “comprehension, adequacy, and fluency” (Poibeau, 2017:124), human evaluation of MT quality assessment is widely accepted as the gold standard (cf. Licht et al., 2022:309), and now with systematic frameworks the error typologies moved over 100 (see Lommel et al., 2014; Lommel et al., 2024). The first WMT campaign in 2006 (Koehn and Monz) included a 5-point rating scale to evaluate participants’ assessments of “fluency and adequacy” (p.102), however, using a 1-5 scale to evaluate translation was found “very hard to perform” (p. 110) at the end. Addressing the limitations in previous methods, In 2014, Lommel et al. published an EU-funded QTLaunchPad project introducing MQM framework, presenting a comprehensive set of over 100 error types to evaluate translation across various quality assessment metrics. This framework has been used for years. Freitag et al. (2021a) carried out “the largest MQM research study to date” (p. 1460) where they evaluated common systems in 2020 WMT shared task and revealed a “substantially different ranking of evaluated systems from the one established by the WMT crowd workers, exhibiting a clear preference for human over machine output” (p. 1460). Following the success, MQM was used in the coming years WMT tasks to evaluate metric

correlation (Freitag et al., 2021b; Freitag et al., 2022; Freitag et al., 2023; Freitag et al., 2024). Lommel et al. launched MQM 2.0 in 2024, which is an improved version of the framework that incorporates more complex evaluation methodologies, including a Linear Scoring Model with Calibration and a Non-Linear Scoring Model. Therefore, as in automatic metrics, human evaluation methods are also in active improvement.

Overall, while human evaluation is considered the most accurate method for assessing MT quality, it remains "labor-intensive, time-consuming, and expensive" (Moghe et al., 2022:3). Therefore, a hybrid approach combining human and automatic evaluation could be an effective practice in MT evaluation – balancing pros and cons, as in the case of this current study.

#### **2.4.1 Details on Reference-Free Metrics and the COMETKIWI Family**

Quality Estimation (QE), or reference-free metrics, emerged in the early 2000s to assess the quality of MT without relying on human-generated translations as a reference. Basically, QE refers to the task of automatically evaluating the quality of machine-translated output without relying on reference translations, making it particularly useful in post-editing and translation quality control for MT (Huang et al., 2023). Blatz et al. (2004) originated the concept of QE, titling his study "Confidence Estimation for Machine Translation." They employed posterior probabilities and word-level predictors to assess the reliability of MT outputs without reference at both the sentence and word levels. In 2009, Raybaud et al. significantly advanced the Confidence Estimation field by creating confidence metrics that employed mutual information, n-gram language modeling, and lexical markers to precisely assess the reliability of MT outputs with only partial reliance in reference, as they stated that their measure “do not directly take into account word deletion nor word order, and neither do our reference corpus (in which words are labelled as correct or incorrect, but missing words are not indicated)” (p. 67-68). Following studies introduced more thorough Quality Estimation frameworks, like QuEst++ by Specia et al. (2013), which employed linguistic and statistical indicators to construct a feature-rich architecture for predicting translation quality. Another

one is a modern, PyTorch-based framework developed by Kepler et al. (2019), OPENKIWI, a multilingual neural-based system that integrated sentence-level regression and word-level tagging, further enhancing QE performance across tasks.

Throughout WMT shared tasks studies, the use and performance of reference-free metrics have been significantly increased. For example, it was concluded in the WMT 2019 findings (Fonseca et al., 2019) that reference-free metrics showed promise but were not able to completely replace reference-based ones. With the WMT 2020 QE shared task dataset, Sun et al. (2020) found that multilingual QE models offer superior performance compared to single-language systems, particularly in contexts with limited resources and uneven label distributions. In the coming year, the WMT21 results showed that reference-free metrics such as COMET-QE and OPENKIWI had very good performance when human translators were involved, yet they were still not reliable enough to replace reference-based ones (Freitag et al., 2021b). In the same year, Tuan et al. (2021) suggested a self-supervised quality estimation approach using synthetic data to reduce the dependency on costly human-labeled data. They produced the data by translating mined Wikipedia sentence pairs with NMT models and reformulating reference sentences using masked language models (MLMs). In the end, the synthetic mistakes in the data facilitated the construction of QE models that showed similar performance to those trained on human-annotated data. The improvements in reference-free metric performance continued in the coming years of WMT tasks.

#### **2.4.2 COMETKIWI**

Building on the success of neural reference-based metrics like COMET, Rei et al. (2022) introduced COMETKIWI, a hybrid architecture that merges COMET’s multilingual representations with the predictor-estimator design of OPENKIWI. By integrating these strengths, the model was able to offer both interpretable outputs and multilingual evaluation capabilities. Rei et al. (2022) described this approach as “combine the strengths of COMET and OPENKIWI” (p. 634) enabling the system to deliver state-of-the-art performance across sentence-level, word-level, and explainable QE tracks. As a result, COMETKIWI

outperformed competing systems and dominated the WMT 2022 Quality Estimation tasks in Rei et al., achieving high alignment with human judgments. Moreover, COMETKIWI was scaled up in 2023 with the introduction of XLM-R XL and XXL models, increasing performance up to 10 spearman points (Rei et al., 2023), and performed better than various metrics in the study of Agrawal et al. (2024).

Researchers are utilizing COMETKIWI and its more sophisticated variant, COMETKIWI-XL, alongside WMT shared task studies. They are gaining significant popularity in both scholarly research and practical MT operations. Anugraha et al. (2024) discussed the dual applications of COMETKIWI in their MetaMetrics-MT study. They indicated the existence of two distinct versions: wmt22-cometkiwi-da and wmt23-cometkiwi-da-xl, each associated with its respective optimized checkpoint. This indicates that the metric is becoming modular and complex, and it is widely recognized as a robust benchmarking method across various translation contexts. In addition, Kocmi et al. (2024) elaborated on the integration of COMETKIWI-XL into the General MT pipeline at WMT 2024. It was used to assist with model-based reranking and to function as a benchmark for evaluating various quality estimation (QE) methodologies. This further illustrates its significance.

Overall, Quality Evaluation has emerged as a crucial component of MT evaluation, from initial Confidence Estimation to transformer-scale models. The emergence of COMETKIWI and COMETKiwi-XL – XXL signifies a transformation in methodologies: reference-free metrics perform similarly to, if not better than, the reference-based metrics in some cases, particularly when combined with human evaluation. They are important components of modern MT systems thanks to their comprehensibility, real-time applicability, and multilingual availability. Future QE improvements may include model uncertainty, an error taxonomy grounded in measurement quality metrics, and active deep learning to enhance the system’s accuracy, fairness, and applicability across many domains.

## 2.5 Related Studies on MT Evaluation

This thesis addresses a critical gap in Translation Quality Assessment (TQA) for UGC on social media—specifically football-related tweets—by integrating MQM-style human annotation with advanced reference-free metrics like COMETKiwi-XL. While no existing study combines all these elements, five key works offer valuable insights into emotion preservation, domain-specific errors, human versus automatic evaluation, and modern QE approaches.

Saadany et al. (2023) evaluate the performance of online MT systems in conveying emotion in tweets written in English, French, and Arabic. The researchers established a multilingual compilation of 30,000 tweets in the said languages. Employing this source annotated data, they created two datasets with the use of the Google Translate API, and they used these datasets for extracting cases to use in the analysis. In addition, a Roberta XLM model was used to classify and predict the emotions of MT outputs. They used BLEU and METEOR scores to quantify general MT performance. Their findings reveal a notable mismatch between high BLEU/METEOR scores and accurate emotion representation, demonstrating that standard metrics fail to penalize emotionally distorted translations. The study gives multiple examples where fluent and lexically faithful translations receive high automatic metric scores despite significantly altering or neutralizing the intended emotion of the original tweet. The authors identify certain language features that frequently result in misconceptions in MT by meticulous manual error analysis. These include misplacing negatives, employing unconventional punctuation, utilizing idioms, and referencing culturally specific references. The elements often seen in tweets significantly hinder the efficacy of NMT systems. Saadany et al. (2023) emphasize the ethical problems associated with such errors, cautioning that users on online platforms may receive outputs that seem accurate yet misleading, such as changing or omitting the original meaning. They finally assert that traditional metrics frequently fail to adequately penalize translations of emotionally charged UGC, such as tweets, which poses a challenge since it might influence users' emotions.

Another similar study is from Rosales Núñez et al. (2021a), who conducted one of the most comprehensive evaluations of character-based NMT systems applied to real-world UGC, focusing on their robustness under zero-shot conditions. Their study is particularly significant for MT research involving noisy and informal inputs, as it systematically questions the prevailing assumption that character-level models are inherently robust due to their open-vocabulary nature. To do this, they developed and annotated a dataset of 400 UGC sentences drawn from two test sets: the MTNT corpus adapted from Michel and Neubig (2018), which was collected on Reddit, and the Parallel French Social Media Bank (PFSMB) adapted from Rosales Núñez et al. (2019), which includes data from platforms such as Facebook, X, and health-related forums. The annotation followed a fine-grained 12-category typology adapted from Sanguinetti et al. (2020), capturing common UGC phenomena like phonetic spellings, emoji use, letter repetitions, inconsistent casing, and named entities. The authors trained four different NMT models: two character-level models—charCNN and char2char—and two BPE-based baselines—a seq2seq model and a Transformer. Training data consisted of canonical corpora (WMT and OpenSubtitles), while all evaluation was done under strict zero-shot conditions, without fine-tuning on noisy data.

Their experimental results revealed that character-based models were consistently outperformed by BPE-based models, even on noisy UGC inputs. This counterintuitive finding was further explored through a “copy task” experiment, which involved training a char2char model on artificial identity mappings (where source and target sequences are identical) and then testing it with varying character vocabulary sizes. The results demonstrated that character-level systems are highly sensitive to vocabulary size—rare and unseen characters substantially degraded translation quality. By reducing the character vocabulary size from the conventional 300 down to around 70, the char2char model exhibited significant improvements in BLEU scores on noisy UGC test sets while maintaining comparable performance on canonical data. Moreover, error analysis showed that certain UGC-specific elements, such as hashtags, inconsistent casing, and named entities, posed the greatest challenges, particularly for char2char. This clearly shows the effects of UGC on translation quality.

The Transformer model benefited from BPE tokenization, often copying unknown tokens more effectively than its character-level counterparts. The study also implemented an unknown-token replacement strategy in post-processing, which further boosted performance, especially in scenarios where OOV entities appeared identically in both source and reference translations. The authors concluded that character-level models are not intrinsically resilient to noisy UGC, and that careful tuning of the character vocabulary hyperparameter can play a critical role in enhancing robustness. Their findings underscore a methodological gap in prior research, where vocabulary size has often been overlooked in the evaluation of character-based architectures. This work not only provides a benchmark dataset for UGC translation evaluation but also highlights the limitations of open-vocabulary models in zero-shot translation scenarios. The publicly available annotated corpus and model outputs contribute substantially to the field by facilitating reproducibility and further research on MT robustness in the face of naturally occurring UGC noise.

In their comprehensive report for the WMT24 Metrics Shared Task, Freitag et al. (2024) critically examine the effectiveness of contemporary automatic MT evaluation metrics in assessing increasingly fluent and near-human outputs produced by large language models (LLMs). The study addresses a core concern in MT evaluation—whether the state-of-the-art metrics, including COMET, BLEURT, and newly proposed LLM-based or hybrid models, can continue to correlate reliably with human judgments in a landscape where the performance gap between systems is narrowing. This concern is particularly relevant for studies like the present one, which employ both expert-based annotation frameworks (e.g., MQM) and reference-free neural metrics (e.g., COMETKiwi-XL). The shared task focused on three language pairs—English→German, English→Spanish (Latin America), and Japanese→Chinese—and utilized paragraph-level test data drawn from four domains: news, social, speech, and literary. Human evaluations were conducted using the Multidimensional Quality Metrics (MQM) framework only, with annotations performed by professional translators or trained raters. Over 25 metrics were submitted, encompassing traditional lexical similarity metrics (e.g., BLEU, chrF), neural models (e.g., COMET-22, BLEURT-20, COMETKIWI), and LLM-based metrics such as GEMBA-ESA. Additionally, a challenge

set subtask targeted metric robustness against difficult translation phenomena, such as hallucinations, negation mismatches, and named entity errors—phenomena that are also prominent in UGC.

The findings underscore a growing mismatch between the output quality of advanced MT systems and the diagnostic sensitivity of many existing metrics. While several fine-tuned neural and LLM-based metrics correlated well with average system-level performance, they exhibited reduced alignment with MQM scores, particularly when tasked with distinguishing among high-performing systems. Notably, MQM remained more effective than scalar scoring methods at capturing fine-grained translation errors. Challenge set results revealed persistent metric vulnerabilities, especially in semantically critical contexts. Ensemble metrics such as MetaMetrics-MT and MetricX-24-Hybrid demonstrated increased robustness, yet no single metric was optimal across all domains and tasks. Their conclusion highlighted the domination of neural metrics (e.g. COMET) over overlap-based metrics (e.g. BLEU and chrF), which strongly supports the present study’s hybrid evaluation approach, which combines MQM annotations with a reference-free neural metric (COMETKiwi-XL) to assess translation quality in user-generated, domain-specific texts.

In another metrics study, Freitag et al. (2023) – “*Results of the WMT23 Metrics Shared Task: Metrics Might Be Guilty but References Are Not Innocent*” – investigate how closely today’s automatic machine-translation evaluation metrics correlate with professional human judgements through MQM at both system and segment levels, while foregrounding an overlooked phenomenon: the quality of the reference translations themselves. They benchmark a broad collection of reference-based and reference-free metrics on three language pairs—English-German (paragraph-level, 557 segments), Hebrew-English (sentence-level, 1910 segments), and Chinese → English (sentence-level, 1976 segments)—all annotated by professional translators whose weighted MQM scores constitute the gold standard. Because the Chinese-English references received unexpectedly low MQM ratings, the authors mitigated reference noise by constructing synthetic references, drawing on the pool of MT outputs and their MQM scores to assemble higher-quality surrogates for every segment. To keep evaluation tractable, the original 201 subtasks were distilled into ten core

tasks—system-level Pearson correlation, system-level pairwise accuracy (triple-weighted for its practical relevance), segment-level Pearson correlation, and tie-aware segment-level pairwise accuracy ( $\text{acc}^{\text{eq}}$ )—augmented by an additional contrastive challenge set that probes specific metric weaknesses. In the Hebrew-English track two human references were supplied, yet only *Reference B* affected scoring; *Reference A* was provided solely as optional input to participating metrics.

Results significantly favour fine-tuned neural metrics: XCOMET-Ensemble posts the top overall meta-evaluation score (0.825), significantly outclassing all competitors and underscoring the enduring gap between neural embedding-based approaches and surface-overlap measures such as BLEU or chrF. Crucially, the study shows that strong segment-level correlations do not guarantee reliable system-level rankings, highlighting the need for multi-granular assessment. The poor quality of the Chinese→English references dramatically depressed the performance of reference-based metrics, whereas reference-free metrics remained comparatively robust—a finding corroborated by the challenge-set analysis, which further demonstrates the resilience of multilingual embedding-driven metrics.

By introducing a transparent, balanced scoring protocol, rigorously quantifying the impact of reference quality, and releasing openly available tools and datasets, Freitag et al. provide a benchmark that is poised to guide future research in MT evaluation.

Lastly, Baines (2019) conducts one of the initial studies by an industry insider examining the impact of translation on the X presence of professional football. It illustrates how clubs, players, and people at the management in the football industry deal with worldwide brand expansion, local authenticity, and the (in)visibility of translators within a very commercial environment. The article addresses the relationship between sport communication and the problems in social media translation. It examines the impact of brand demands, glocalization strategies, and the conflict between fluency and authenticity on the translation of tweets.

The study employs a qualitative, multi-faceted approach. The primary data consists of five semi-structured interviews with professional translators employed in premier leagues, including the English Premier League, Serie A, La Liga, and Ligue 1. These translators

operated in diverse linguistic environments. One was translated from English to French, while the others were translated into English from French, Spanish, Italian, and Portuguese. They assisted athletes, trainers, and organizations. Three interviews were conducted in person, while two were conducted over Skype. The researcher used semi-structured interviews due to their flexibility. The primary objective was to gather experiences, perspectives, and factual data on the translation of tweets in football. Five open-ended questions were sent via email to four initial respondents and one extra participant associated with a governing organization. This provided us with further information. The questionnaires addressed topics raised in the interviews, including the construction of identity on X, the visibility of translators, the translation of hashtags and emojis, the management of taboo and technical language (particularly concerning injuries), adherence to client instructions, evaluation of translation quality, and the handling of time constraints during live events. He conducted preliminary interviews and questionnaires with the first participant from each group to ensure the clarity of the questions. The responses were documented and analyzed thematically. The study confirms its findings by including tweets from migrant players and clubs, which were gathered over time from the public domain to illustrate certain trends identified during the analysis. The findings demonstrate that brand protection significantly influences translation decisions. To protect their reputation, most clubs need their tweets to be composed in flawless English. Nonetheless, several prominent clients prefer a more informal and "genuine" tone, even at the expense of minor grammatical errors. A wide variety of glocalization strategies exists for translation. Some firms employ multilingual accounts, whilst others maintain distinct monolingual feeds. Numerous brands retain hashtags such as #UCL or #ForzaJuve in their original form to maximize global visibility. Interviewees emphasized that even a single poorly translated or machine-translated tweet may jeopardize a career, particularly in high-stakes contexts as live matches or press conferences. Individuals consistently translate this type of content even under significant time pressure, which reduces the visibility of their effort and increases the possibility of errors. Overall, Baines argues that translation practices in football tweeting prioritize managing commercial risk, shaping public identity, and engaging global audiences more than adhering to strict linguistic accuracy. The

study calls for more research into how multilingual communication operates in sports and other fast-paced digital environments.



## CHAPTER 3: METHODOLOGY

This chapter outlines the methodological framework employed in the present study. It first explains the overall research design, followed by a detailed description of the evaluation techniques and tools applied. By combining human annotation with automatic metrics, the study aims to provide a comprehensive and reliable analysis of translation quality across different systems. Each subsection introduces the rationale for methodological choices and highlights how they contribute to addressing the research questions.

### 3.1 Research Design

The current study opts for a mixed-methods evaluation technique by combining human annotation through the latest version of the Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2024) with the fine-grained neural reference-free metric COMETKiwi-XL (Rei et al., 2023), with the purpose of evaluating the translation performance of DeepL, ChatGPT-o3, and Twitter’s in-built translation system (X Translate) on tweets posted by football fans.

A reference-free quality evaluation (QE) metric is crucial in practical scenarios such as social media, where human references are often unavailable. Specia et al. (2018:1) note "the primary motivation is to enhance the utility of applications in real-world contexts, where information regarding the quality of each output is essential and reference outputs are unavailable." They also emphasize that QE can improve natural language processing systems by providing accurate predictions in situations when reference translations are unreliable, particularly in dynamic, real-world environments. This is particularly applicable to tweets on football, which are casual, rapidly posted, and domain specific. Therefore, creating reference translations for these tweets would be both unfeasible and labor-intensive. Huang et al. (2023) support this by stating that reference-free evaluation models are less dependent on manual annotations, making them more flexible across many domains, including social media. Rei et al. (2022) do not explicitly mention UGC; instead, their COMETKIWI models

are trained on source-target pairs and human evaluations without the necessity of reference translations. This illustrates the potential use of QE in circumstances when references are unavailable. The various perspectives underscore the significance of reference-free standards for evaluating the quality of translations in UGC. This is because platforms such as X, which have significant immediacy and linguistic diversity in nature, challenge the use of traditional evaluation methodologies.

Since there is no single ultimate metric that can be applied to all cases, MT evaluation is quite a complex task by nature (Anugraha et al., 2024). There are two types of MT quality evaluation: human annotation vs. automated evaluation. Each has its pros and cons. Although inferior in quality, automatic evaluation offers a cost-effective and user-friendly way for comparing multiple systems; human evaluation, however, serves as a reliable measure of translation quality but is costly, time-intensive, and constrained by the necessity for reliable bilingual annotators and consistent inter-rater reliability (Lee et al., 2023). Given that they both have their strengths and weaknesses, researchers ought to opt for the method(s) that is best for their study.

### **3.2 Rationale for Dual Approach**

The performance and effectiveness of a specific metric might significantly vary based on the task and language pairs, highlighting the fact that depending on a single specific metric might be insufficient (Anugraha et al., 2024). This phenomenon was also supported by the results of Jiang et al. (2024), suggesting that it is quite challenging to depend completely on automatic metrics since their grasp of norms, cultural nuances and clarity is questionable. They further suggest using human annotation for capturing contextual and culture-specific nuances. Considering the factors above and knowing that using automatic metrics in combination with human annotation could significantly speed up the process and overcome drawbacks related to time (Anugraha et al., 2024), the current study opts for a dual approach to evaluate MT outputs: (1) human annotation grounded in the MQM framework (Lommel et al., 2024) and (2) automated evaluation through COMETKiwi-XL (Rei et al., 2023), which

has shown superior correlation with human assessment compared to traditional lexical overlap metrics such as BLEU and chrF (Freitag et al., 2024; Agrawal et al., 2024; Rei et al., 2023) and LLM-based metrics such as GEMBA-GPT4 (Rei et al., 2023).

### **3.2.1 A Novel Evaluation Design: COMETKiwi-XL Combined with MQM**

The primary methodological enhancement of this study is the integration of COMETKIWI-XL, the latest and most robust iteration of the COMETKIWI metric family, with the latest version of Multidimensional Quality Metrics (MQM), a prevalent framework for assessing human-based translation through error annotations (Lommel et al., 2024). COMETKIWI was first introduced as a quality assessment system that operates without the necessity for references. It integrates the optimal features of COMET and OPENKIWI by employing COMET’s training methodologies and OPENKIWI’s predictor-estimator framework, facilitating tagging at both the word and sentence levels inside a singular model. Rei et al. (2022) assert that COMETKIWI is a hybrid model that integrates the multilingual generalization capabilities of COMET with the predictor-estimator framework of OPENKIWI. This proves that it is a practical paradigm that integrates the most effective elements of both techniques.

This research opts for COMETKiwi-XL, an enhanced and more scalable version of COMETKIWI. It improves upon its predecessor by using bigger pre-trained encoders (XLM-R XL and XXL) and attaining much better alignment with human annotations (Rei et al., 2023). COMETKiwi-XL significantly enhances its alignment with human assessments, achieving an increase of up to 10 points in Spearman correlation relative to the previous version (Rei et al., 2023). This strong correlation is not confined to a certain granularity level, as stated “our multilingual approaches are ranked first for all tasks, achieving the state-of-the-art performance for quality estimation at word, span, and sentence levels” (Rei et al., 2023).

COMETKiwi-XL is significantly more advantageous since it has strong generalization capabilities to novel domains and low-resource languages. Rei et al. (2023)

assert that the model may enhance performance on novel language pairings with a minimum of 500 training samples, without adversely affecting its correlation scores on existing pairs. This significantly benefits the evaluation of MT in certain areas or languages, such as Turkish in social media.

COMETKIWI is both flexible and aligns effectively with human evaluations, particularly those grounded in MQM frameworks. Rei et al. (2023) assert that COMETKIWI consistently outperforms other quality estimation models across all evaluation metrics, demonstrating its reliability in alignment with expert MQM-based assessments. The WMT22 Shared Task findings further validate this performance. They demonstrate that neural fine-tuned metrics, such as COMETKIWI, not only have superior accuracy but also maintain consistent performance across several domains (Freitag et al., 2022).

Recent evaluations indicate that these metrics remain effective when used on translations generated by large language models (LLMs). The findings of WMT24 indicate that “fine-tuned neural metrics consistently demonstrate strong performance and effectively assess quality, even for LLM-based translations” (Freitag et al., 2024:47).

Since COMET KIWI is a quite novel version of the COMET family, there have been only two studies combining COMET KIWI family and MQM to the best of our knowledge, which were conducted in different text domains (e.g., WMT 24 Metrics Shared Task - the domains cover news, literary, speech, and social [Freitag et al., 2024] and WMT 22 – 23 Metrics and Chat datasets – the domains cover news, TED talks, social, e-commerce, and chat [Agrawal et al., 2024]). Also, while there are some studies combining human evaluation and automatic metrics in Turkish (e.g., Doğru, 2022) or employing only MQM with human evaluation (e.g., Baydere and Karadeli, 2023, and Çetin and Duran, 2024), to the best of our knowledge, there is no study combining COMETKiwi-XL and MQM in the context of Türkiye.

### 3.2.2 Dataset and Translation Output Collection

Total 20 tweets (10 in English and 10 in Turkish) were manually selected on X by the researchers based on their inclusion of slang, abbreviations, informal expressions, and football-specific terminology, which are thought to cause significant challenges for MT. Then, these tweets were translated into the target language on DeepL, ChatGPT-o3 and X Translate, producing a total of 60 translation outputs.

Below you may see example tweets in English and Turkish (see the rest in Appendix A & B).



**Figure 1:** Screenshot of Turkish Tweet with Slang and Informal Expressions



**Figure 2:** Screenshot of English Tweet with Slang and Informal Expression

As highlighted in the latest version of MQM (Lommel et al., 2024), the 60 total output sample size falls within the medium-range category. This makes the current study suitable for linear raw scoring rather than small-sample statistical control methods.

### 3.2.3 Translation Systems Used

As mentioned in the previous parts, DeepL and ChatGPT are among the most popular systems used for translation nowadays. Each of the three translation services operates on a distinct basic mechanism, making this research significant by offering a comparative analysis of these popular systems built on distinct concepts. DeepL is an NMT, using artificial neural networks and trained by millions of translated texts (DeepL Team, 2021). ChatGPT-o3, which is originally an AI model powered by Large Language Models (LLMs, hereafter), is the latest update by OpenAI with the integration of advanced reasoning (2025). Lastly, the X translation feature is powered by Google Translate (Udin, 2023), which has also used NMT structure since their announcement in 2016 (Turovsky, 2016). Consequently, evaluating these three separate translation systems is quite valuable, as it highlights particular areas where each system needs improvement.

### 3.3 Human Annotation Through MQM

The current paper employs the latest version of MQM (Lommel et al., 2024), which is a dynamic and flexible framework (Lommel et al., 2014) allowing researchers to apply the errors based on their needs. Among the error typologies of MQM (Lommel et al., 2015), the current study opts for Accuracy, Linguistic conventions, Style, Terminology and Audience appropriateness. The figure below shows the selected subtypes for each error.

| Accuracy         | Linguistic conventions | Style              | Terminology                     | Audience appropriateness.  |
|------------------|------------------------|--------------------|---------------------------------|----------------------------|
| Mistranslation   | Grammar                | Language Register  | Inconsistent Use of Terminology | Culture-specific Reference |
| Overtranslation  | Punctuation            | Awkward Style      | Wrong Term                      | Offensive                  |
| Undertranslation | Spelling               | Inconsistent Style |                                 |                            |
| Addition         | Unintelligible         | Unidiomatic Style  |                                 |                            |
| Omission         | Textual Conventions    |                    |                                 |                            |
| Untranslated     |                        |                    |                                 |                            |

**Figure 3:** MQM Selected Error Types

### 3.3.1 MQM Error Typologies Explanation

Below is the explanation of each selected main/sub error type gathered from the website of MQM (n.d.).

**Table 1:** Accuracy Errors

| <b>Subcategory</b> | <b>Definition</b>                                                           |
|--------------------|-----------------------------------------------------------------------------|
| Mistranslation     | The target text does not accurately reflect the meaning of the source text. |
| Overtranslation    | The target text adds content not present in the source.                     |
| Undertranslation   | The target text omits content that is present in the source.                |
| Addition           | Unnecessary or unjustified information is added in the target text.         |
| Omission           | Important source information is left out in the translation.                |
| Untranslated       | The source text remains untranslated in the target text.                    |

**Table 2:** Linguistic Conventions Errors

| <b>Subcategory</b>  | <b>Definition</b>                                                                                 |
|---------------------|---------------------------------------------------------------------------------------------------|
| Grammar             | Issues with syntax, agreement, tense, or other grammatical norms of the target language.          |
| Punctuation         | Incorrect, missing, or excessive punctuation marks.                                               |
| Spelling            | Misspelled words according to target language orthography.                                        |
| Unintelligible      | The translation is incomprehensible or nonsensical.                                               |
| Textual Conventions | Violation of formatting, discourse, or stylistic norms like quotation marks or numbering systems. |

**Table 3: Style Errors**

| <b>Subcategory</b> | <b>Definition</b>                                                                                     |
|--------------------|-------------------------------------------------------------------------------------------------------|
| Language Register  | Use of an inappropriate level of formality or tone for the context, genre, or audience.               |
| Awkward Style      | The translation sounds unnatural or clumsy in the target language.                                    |
| Inconsistent Style | Inconsistent tone, formality, or style throughout the text, breaking coherence or genre expectations. |
| Unidiomatic Style  | Style that is grammatical, but unnatural.                                                             |

**Table 4: Audience Appropriateness Errors**

| <b>Subcategory</b>         | <b>Definition</b>                                                                               |
|----------------------------|-------------------------------------------------------------------------------------------------|
| Culture-specific Reference | Failure to adapt cultural references or idioms appropriately for the target audience.           |
| Offensive                  | Content that may be disrespectful, culturally insensitive, or offensive to the target audience. |

**Table 5: Terminology Errors**

| <b>Subcategory</b>              | <b>Definition</b>                                                                                              |
|---------------------------------|----------------------------------------------------------------------------------------------------------------|
| Inconsistent Use of Terminology | Use of different terms for the same concept where consistency is expected, especially in technical contexts.   |
| Wrong Term                      | Use of an incorrect or inappropriate term that may distort the meaning or function of the original expression. |

Annotators were provided with a rigorously structured MQM-based annotation framework, implemented in Excel format via Google Sheets, to ensure both efficiency and methodological consistency. This framework incorporated standardized drop-down menus through which annotators could systematically assign error categories and severity levels to each highlighted segment. To complement this, an extra column was allocated for free-text

input, allowing annotators to provide clarifications, justifications, or to record error types not pre-highlighted by the researchers.

Furthermore, to promote inter-annotator consistency, all additional comments and error identifications submitted by one annotator were reported to the others. This process fostered a transparent environment in which annotators could review potential errors collaboratively, while retaining autonomy to either confirm, reject, or override such suggestions by selecting the “no error” option where appropriate. Overall, this approach ensured a balance between systematic categorization and annotators’ independent judgment, thereby enhancing both the reliability and validity of the annotation process. See Figure 4 below showing the annotation framework used.

|   |                                                                                                                                                                                                |                                                                                                                                                             |                                       |                 |   |             |   |
|---|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------|-----------------|---|-------------|---|
| 6 | İstifa et yeter düğ ulan yakamızdan Allah'ın belas! Allah'ın laneti yeter bırak gıt Allah'ın galisi lanet olsun senin gibi yüzüsize davalık olucam yele r lan yeter bıktık ilallah etik senden | Resign, enough shower, God's curse, God's curse, enough let go, God's gs person, damn it, I'll sue the brazen like you, enough, we're sick and tired of you | enough shower                         | Metatranslation | - | Critical    | - |
|   |                                                                                                                                                                                                |                                                                                                                                                             | God's curse, God's curse              | -               | - | -           | - |
|   |                                                                                                                                                                                                |                                                                                                                                                             | God's gs person                       | Metatranslation | - | Major       | - |
|   |                                                                                                                                                                                                |                                                                                                                                                             | damn it, I'll sue the brazen like you | Metatranslation | - | Critical    | - |
|   |                                                                                                                                                                                                |                                                                                                                                                             | I'll sue                              | Metatranslation | - | Critical    | - |
|   |                                                                                                                                                                                                |                                                                                                                                                             | enough, enough                        | -               | - | -           | - |
|   |                                                                                                                                                                                                |                                                                                                                                                             | of you                                | -               | - | Punctuation | - |
|   |                                                                                                                                                                                                |                                                                                                                                                             | galisi - gs                           | -               | - | -           | - |
|   |                                                                                                                                                                                                |                                                                                                                                                             | galisi - gs                           | -               | - | -           | - |
|   |                                                                                                                                                                                                |                                                                                                                                                             | damn it                               | -               | - | -           | - |
|   |                                                                                                                                                                                                |                                                                                                                                                             |                                       | -               | - | -           | - |
|   |                                                                                                                                                                                                |                                                                                                                                                             |                                       | -               | - | -           | - |

**Figure 4:** Screenshot of the Annotation Framework

To ensure consistent application of the framework, annotators received a brief MQM guideline sheet and two instructional videos—one demonstrating the usage of the Google Sheets interface and the other illustrating the annotation process. This study follows standard MQM procedures by employing three professional annotators per segment, consistent with practices in WMT-style MQM evaluations. For example, Freitag et al. (2021) explicitly state that they used ‘3 professional’ raters per segment in their MQM setup. This design supports robust inter-rater agreement and is widely adopted in high-quality MT evaluation research. Therefore, the three expert annotators who have over 10 years of experience in Translation

Studies evaluated the three sets of outputs. In accordance with MQM standards, annotators identified and categorized issues in each translation by kind and severity (Minor = 1, Major = 5 and Critical = 25). This ensures that the annotation process is systematic, reproducible, and comparable to COMETKiwi-XL scores with the use of the Linear Raw Scoring Model (Lommel et al., 2024).

### 3.3.2 MQM 2.0

In alignment with the XL model, the recently proposed MQM 2.0 version (Lommel et al., 2024) is used in this study. The latest MQM model significantly facilitates consistent quality evaluation across varying sample sizes using calibrated scoring models and enhanced error typologies for MQM. Contrary to the 1.0 version of MQM (Lommel et al., 2014), MQM 2.0 allows researchers to have better accuracy with an advanced scoring model, which provides three scoring models based on purpose and the sample size (Lommel et al., 2024). Among the scoring models, this study employs the Linear Raw Scoring Model (without calibration) in accordance with the updated MQM principles, as it uses a medium-sized dataset. All in all, even though XL models perform better than the traditional automatic metrics, they still have weaknesses in that even though the errors can be categorized by severity (minor, major, critical), they cannot capture the broad range of error types covered by MQM (Guerreiro et al., 2023). Therefore, the use of human annotation through MQM is expected to provide better understanding of MT performances.

**Converting Human Annotation Scores into Quality Scores** In order to convert human annotation scores into quantitatively interpretable and comparable scores with COMETKiwi-XL scores, the current study opts for the Linear Raw Scoring Model, presented by Lommel et al. (2024) with MQM 2.0. For each output taken from the translation systems, the formula below is used to calculate the scores:  $\text{MQM Score} = 1 - (\text{Penalty Points} / \text{Word Count})$ . In this formula, Penalty Points means the total weighted error scores for a translation (e.g. Minor x 1 + Major x 5), and Word Count means the total number of words in one output in question. With this method, the human annotation scores are normalized onto a 0-1 scale, where 1

represents a perfect output. In this way, the human annotation scores can be compared with COMETKiwi-XL scores, where the score range is 0-1.

Lommel et al. (2024) present three scoring models: (1) Linear Raw Scoring, (2) Linear Scoring with Calibration, and (3) Non-Linear Scoring. Each scoring model, with and without calibration, serves unique purposes (Lommel et al., 2024). The reason why the current study does not use the other two is that the objective is not to determine the acceptability of elements in a production environment but to gather overall scores through a scoring system that is transparent, comprehensible, and can be statistically analyzed, converted and thus compared to the scores of COMETKiwi-XL.

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

MQM 2.0 Evaluation Scorecard

| Client                                                                | Project Name              |                          | Translator                      |                               | Evaluator                       |                                |                          |                   |
|-----------------------------------------------------------------------|---------------------------|--------------------------|---------------------------------|-------------------------------|---------------------------------|--------------------------------|--------------------------|-------------------|
| XXX                                                                   | XXX                       |                          | XXX                             |                               | XXX                             |                                |                          |                   |
|                                                                       | Source Language           | Target Language          | Total Word Count                |                               | Evaluation Word Count (EWC)     | Content Type                   |                          |                   |
|                                                                       | en-US                     | ja-JP                    | 3000                            |                               | 1500                            | XXX                            |                          |                   |
| Score Calculation Parameters                                          | Maximum Score Value (MSV) | Passing Threshold Raw QS | Passing Threshold Calibrated QS |                               | Reference Word Count (RWC)      | Acceptable Penalty Points/RWC  |                          |                   |
|                                                                       | 100                       | 95                       | 90                              |                               | 1000                            | 10                             |                          |                   |
| Evaluation Results                                                    | Quality Rating (QR)       | Raw Quality Score (RQS)  | Absolute Penalty Total (APT)    | Per-Word Penalty Total (PWPT) | Normed Penalty Total (NPT)      | Calibrated Quality Score (CQS) |                          |                   |
|                                                                       | PASS                      | 99.2                     | 12                              | 0.008                         | 8.00                            | 92.00                          |                          |                   |
| General Comment                                                       |                           |                          |                                 |                               |                                 |                                |                          |                   |
| Please provide a detailed comment on quality of the evaluated sample. |                           |                          |                                 |                               |                                 |                                |                          |                   |
| Severity Penalty Multiplier                                           | 0                         | 1                        | 5                               | 25                            |                                 |                                |                          |                   |
| Error Type                                                            | Neutral Errors Count      | Minor Errors Count       | Major Errors Count              | Critical Errors Count         | Error Type Penalty Total (ETPT) | Weighted ET Penalty Points     | Normed ET Penalty Points | Error Type Weight |
| Terminology                                                           | 0                         | 1                        | 1                               | 0                             | 6                               | 6                              | 4.00                     | 1                 |
| Accuracy                                                              | 0                         | 0                        | 1                               | 0                             | 5                               | 5                              | 3.33                     | 1                 |
| Linguistic conventions                                                | 0                         | 0                        | 0                               | 0                             | 0                               | 0                              | 0.00                     | 1                 |
| Style                                                                 | 0                         | 1                        | 0                               | 0                             | 1                               | 1                              | 0.67                     | 1                 |
| Locale conventions                                                    | 0                         | 0                        | 0                               | 0                             | 0                               | 0                              | 0.00                     | 1                 |
| Audience appropriateness                                              | 0                         | 0                        | 0                               | 0                             | 0                               | 0                              | 0.00                     | 1                 |
| Design and markup                                                     | 0                         | 0                        | 0                               | 0                             | 0                               | 0                              | 0.00                     | 1                 |
| Grand Total                                                           | 0                         | 2                        | 2                               | 0                             | 12                              | 12                             | 8.00                     |                   |

**Figure 5:** MQM 2.0. Evaluation Scorecard (Lommel et al., 2024)

Among the scoring models, we opted for The Linear Raw Scoring Model for some reasons. First of all, the model works with basic values and parameters, which are the word count of the annotated translation, and the total of severity-weighted error points assigned during annotation (Lommel et al., 2024). Secondly, this scoring model “calculates the portion of the evaluated text with errors and subtracts this value from 100 to get the value” (Lommel

et al., 2024:6), which offers insights with the overall score of the translation. These two critical factors allow this study to have a straightforward numerical representation of translation quality, which is what is needed to compare human annotation scores of MQM with the quality scores generated by COMETKiwi-XL. Lastly, these raw results gathered by this model are easier to calculate (Lommel et al., 2024).

The outputs taken from DeepL and ChatGPT-o3 were evaluated by COMETKiwi-XL and by translation scholars with over 10 years of experience in the field through the use of error typologies listed in MQM. X language is not only multilingual, it is also shaped by cultural nuances and context-specific items (Saha and Menezes, 2016). Also, the language used on X consists of informal characteristics, which makes it hard to understand and translate tweets. Thus, the outputs were analyzed in terms of domain-specific and culture-specific linguistic items, abbreviations, and slang.

### **3.3.3 COMETKiwi-XL Evaluation**

The COMETKiwi-XL-based ensemble achieved a Spearman correlation of 0.594 in sentence-level QE, outperforming the second-best system (Yan et al., 2023) by approximately 4 points and significantly surpassing GPT-4 (0.319). While performance in error span detection was comparable to GPT-4, the COMETKIWI models offered greater control and computational efficiency (Rei et al., 2023:845–847).

Among the latest released models of COMET (Cross-lingual Optimized Metric for Evaluation of Translation) (Rei et al., 2020), COMETKiwi-XL has 3.5B parameters, while COMETKiwi-XXL has 10.7B parameters (Rei et al., 2023). XXL model performs better than XL model as expected (Agrawal et al., 2024). In the initial testing and decision phase of this study, it was found that the COMETKiwi-XXL model, comprising approximately 10.7 billion parameters, required a minimum of 44 GB of GPU VRAM, in addition to substantial system memory and storage capacity, making it challenging to operate on standard academic or personal research hardware (Hugging Face, 2024a; Stack Overflow, 2023). The COMETKiwi-XL model has around 3.5 billion parameters and performs optimally on

computers equipped with a minimum of 32 GB of RAM and 16–24 GB of GPU VRAM, particularly when SSD storage and optional swap memory are utilized (Hugging Face, 2024b; GitHub Discussions, 2024). The XL version was selected for this study since it has more manageable hardware requirements and superior disk management efficiency.

Overall, COMETKiwi-XL, an NMT evaluation tool trained on MQM-labeled data, is used to examine all 60 translation outputs alongside human annotation with MQM. This metric consistently yields a quality score ranging from 0 to 1, contingent upon the source input and the MT output. The WMT Task studies (2022, 2023 and 2024) proved that COMETKiwi-XL exhibits a closer correlation to human judgment compared to conventional metrics such as BLEU or chrF. Moreover, Jiang et al. (2024) highlight that even though BLEU and Chrf metrics are widely employed for translation quality evaluation, they have limitations with LLMs; however, neural metrics such as COMET and BERTScore produce more nuanced evaluations focusing on semantic aspects and fluency.

### **3.4 Statistical Tools**

To ensure the robustness of the findings, the study employs statistical tools that account for both the consistency of human annotations and the relationship between human and automatic evaluation scores. These tools provide a systematic means of testing reliability, identifying patterns, and supporting the validity of the results. The following subsections outline the specific statistical measures adopted in this research.

#### **3.4.1 Inter-Rater Agreement**

In order to ensure the reliability among the raters, inter-rater agreement was measured using Kendall’s Tau-b. All annotators in the current study used the same ordinal scale (Minor, Major, Critical), and it is quite possible that they may have tied scores. Therefore, Tau-b has been a model that has been widely used for assessing the strength of the link between two ordinal variables, particularly when their ranks are analyzed in a contingency table (Berry et

al., 2009). In this current paper, the annotators score the outputs using an identical ordinal scale (Minor, Major, Critical), and the results are subsequently converted into normalized MQM quality scores.

Therefore, Kendall’s Tau-b is a statistically optimal measure for inter-rater agreement since it accommodates both ordinal data and tied ranks. As Berry et al. (2009) highlights, “Under many circumstances,  $\tau_c$  reduces to  $\tau_b$  when  $r = c$ , and always reduces to  $\tau_a$  when  $r = c$  and no values are tied.” (p. 1144), meaning that in cases where annotators have the same rating categories and the data includes tied values, Tau-b offers the most direct and interpretable estimate of ordinal association. Furthermore, they note that “When  $r \neq c$ ,  $\tau_c$  may not properly norm between  $-1$  and  $+1$ .” (Berry et al., 2009:1144), making Tau-b more stable and appropriate also for system-level comparison of translation quality scores.

### 3.4.2 Comparative Analysis

This study used two distinct statistical methods, Pearson’s correlation coefficient ( $r$ ) and Kendall’s Tau ( $\tau$ ), to evaluate the correlation between human annotation and automatic MT scores. These specific methods are employed to evaluate correlations at both the system and segment levels between human MQM ratings and automated COMETKiwi-XL scores. With the combination of two said methods, the current study examines the reliability, consistency, and alignment of expert human evaluations and metric predictions about the translation of football tweets.

**Figure 6:** Kendall Tau Formula

$$\tau = \frac{\text{Concordant} - \text{Discordant}}{\text{Concordant} + \text{Discordant}}$$

(Freitag et al., 2021a:11)

**Figure 7: Pearson's r Formula**

$$r = \frac{SP}{\sqrt{SS_X SS_Y}}$$

(Gravetter and Wallnau, 2015:492)

Pearson's  $r$  calculates the extent of linear correlation between human-annotated MQM ratings and COMETKiwi-XL scores. It is used at both the system and segment levels to evaluate the alignment of COMETKiwi-XL with human numerical annotations. At the system level, Pearson correlation assesses whether the average scores provided by COMETKiwi-XL for each MT system (ChatGPT-o3, DeepL, X Translate) align with the average human MQM assessments. At the segment level, however, it provides insight into COMETKiwi-XL's sensitivity to variations in translation quality on a tweet-by-tweet basis. Therefore, the study sheds light on the trustworthiness and reliability of COMETKiwi-XL at both segment and system levels in comparison with human annotation scores. As stated in both WMT22 and WMT 21 Metrics Shared Task: "Pearson correlation tests linear fit with MQM scores, a stringent but reasonable criterion since we expect these scores to conform to a linear scale" (Freitag et al., 2022:55). This approach is parallel to WMT21, reporting: "We measure correlation using the Pearson-r statistic at the system level and the Kendall-tau statistic at the segment level" (Freitag et al., 2021b:14). In addition, Guerreiro et al. (2023) employed system- and segment-level Pearson correlation for evaluation. Therefore, the current study also employs Kendall's Tau to enhance the evaluation by examining the consistency of ranks across human and automated assessments. This is particularly advantageous when the absolute scores may vary; however, the relative rankings of the systems remain consistent. Therefore, it is aimed at showing if COMETKiwi-XL identifies the best system the same way (or in the same order) humans do.

As emphasized by Freitag et al. (2021a), in the case of rating a few systems, Kendall’s Tau serves as a superior statistic since it directly pertains to the primary objective of comparing and assessing system performance. Also, Kendall’s Tau functions by conducting pairwise comparisons between systems in use, measuring the consistency of their rankings (Freitag et al., 2022) by comparing all possible pairings of systems. This serves as an excellent statistical method to verify and rationalize the consistency of orderings, a prevalent approach for comparing systems. Therefore, this study benefits from Kendall’s Tau in evaluating whether COMETKiwi-XL can show not just score similarity but also human-like system rankings.

In addition to correlation, a comprehensive analysis of MQM error categories (Lommel et al., 2024) across translation systems is conducted. With this analysis, the study reveals the highest frequency of errors among the MT systems in question. The study uses a Linear Scoring Model (Lommel et al., 2024) to normalize error frequencies by the number of segments per system. Even though each system translates the same set of tweets, normalization is still important to express and compare error frequency as a rate instead of a raw count, since it helps compare across systems and categories in a clearer way. For the normalization process, the study uses the framework below provided by Lommel et al. (2024).

| Step | Raw Quality Score Calculation | Formulas                                                                                                                                                                                  |
|------|-------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1    | Absolute Penalty Total (APT)  | $\sum_{ij} Error\ Count_{ij} \times Severity\ Multiplier_j \times Error\ Type\ Weight_i$ <p>Where: <math>i</math> = index for Error Types, <math>j</math> = index for Severity Level.</p> |
| 2    | Per-Word Penalty Total (PWPT) | $\frac{Absolute\ Penalty\ Total}{Evaluation\ Word\ Count}$                                                                                                                                |
| 3    | Raw Quality Score (RQS)       | $1 - PWPT$ <p>(for a scale with a maximum of 1), or</p> $100 - (PWPT \times 100)$ <p>(for a scale with a maximum of 100)</p>                                                              |

**Figure 8:** Raw Quality Score Calculation (Lommel et al., 2024)

Overall, with the use of Pearson correlation for absolute score alignment, Kendall's Tau to check rank consistency between human annotation and COMETKiwi-XL scores, and error-type-based analysis, this comparative study provides a multi-dimensional, statistically grounded evaluation of DeepL, ChatGPT-o3 and X Translate in a domain-sensitive and annotation-rich context.

### **3.5 Translation Systems Used**

The current study employed a novel methodology by comparing MQM and COMETKiwi-XL scores in MT translations of Tweets. Therefore, the in-built system of X, DeepL and ChatGPT-o3 translation outputs were evaluated, as the first is the default translation feature on X and the latter ones are the most used ones.

#### **3.5.1 X (formerly Twitter)**

There are various social media platforms such as Facebook, Instagram, YouTube, and X. According to the *Digital 2025: Turkey* report by DataReportal (Kemp, 2025), Instagram is the most used social media platform, followed by WhatsApp, Facebook and X. The results of the report indicate a shift in the rankings under the category of favorite social media platforms, with Instagram leading, followed by WhatsApp and X (formerly Twitter). X occupies third place, yet it is seen as a niche platform and is not for everyone (Woodward, 2025). Among other social media platforms, X is a unique one where Tweets are brief, casual and often characterized by peculiar language use such as slang (Izazi and Sepora, 2020), abbreviations (Imran et al., 2016) and hashtags, which make it difficult for MT systems to perform correctly (Gotti et al., 2014). It is therefore important that as social media continues to grow in global reach, accurate translations of these posts should be made available, especially for cross-cultural interactions, international business and worldwide news sharing. This perfectly addresses the nature of X, a worldwide platform where people from many language backgrounds may participate, interact, and exchange ideas in real time (Alhadlaq

and Alnuaim, 2023). Thus, the translation performance of social media platforms is still under study.

X's in-app translation feature has developed through three major phases. It first emerged in 2013, when the platform began experimenting with Microsoft's Bing Translator, enabling users to translate tweets directly within the interface (Soper, 2013). That same year, X highlighted that users could choose from 54 available languages in the app's settings, with translations powered by Bing (Twitter, 2013). However, its reach remained somewhat limited; when Bing was reintegrated in 2015, X stated it now supported tweet translations in more than 40 languages (Schwartz, 2015). Due to ongoing user dissatisfaction regarding translation quality, X made a significant change in 2019 by shifting to Google Translate for its Progressive Web App (Carrasqueira, 2019). Scholars have since observed that users can activate a translation feature on tweets, with X indicating that this function is supported by Google Translate (Saadany et al., 2021b). Google's system significantly broadened linguistic coverage—reportedly supporting over 130 languages (Caswell, 2022)—and demonstrated performance advantages over other engines; for instance, it was found to handle informal or non-standard language better than Naver Papago (Andriani and Setyaningsih, 2024).

However, not all updates were well received. A test run in 2020, which automatically translated entire timelines, was heavily criticized by users who disapproved of removing manual control, labeling the update as misguided (Savić, 2020). As a result, the feature remained optional. Following Twitter's rebranding as X in 2023, financial considerations reportedly influenced another shift: by mid-2025, the platform had transitioned from Google Translate to its proprietary Grok Translation system for its web interface, offering real-time translation into users' preferred languages (Team Latestly, 2025).

Overall, X's translation journey—from Bing's initial support for 54 languages, through Google's expanded capabilities, to the latest use of Grok—reflects the platform's continual attempts to optimize translation tools to meet the multilingual demands of its global user base.

### 3.5.2 DeepL Translation

Launched in 2017 by Linguee's founders, DeepL entered the NMT market with a narrow language roster (seven EU languages) but quickly attracted attention and was mentioned as more accurate and nuanced compared to Google, Microsoft and Facebook translation systems (Coldewey and Lardinois, 2017). Also, in an August 2017 blind test, professional translators were asked to rank translations of 100 sentences (English ↔ German/French/Spanish) from DeepL versus Google, Microsoft, and Facebook without knowing which engine produced them. DeepL's outputs were chosen three times more often than Google's because they "sounded more natural" to the human evaluators (Smolentceva, 2018). DeepL has since become widely used in both consumer and professional settings; by early 2025 it had been used by over one billion people worldwide (Phrase, 2025) and was adopted by 82% of language service companies (far ahead of Google's 46%) in a 2024 industry survey (DeepL Team, 2024). Today, DeepL supports over 30 languages, including Turkish (DeepL, n.d.). DeepL's neural architecture has been a key to its superior translations. Unlike earlier statistical or simple recurrent NMT systems, DeepL embraced advanced deep learning techniques. Notably, the team incorporated convolutional neural networks (CNNs) in its model – an approach that, with proper adjustments, proved especially effective for translating long, complex sentences with better global context, and this CNN was combined with attention mechanisms (Coldewey and Lardinois, 2017). In summary, high-quality training data plus a specialized neural network design (and substantial computing power) enabled DeepL to produce remarkably human-like translations.

In 2024, DeepL introduced a next-generation LLM-powered model, designed to enhance fluency and contextual precision, and reduce the need for post-editing. This model outperformed ChatGPT-4, Google and Microsoft in translation quality (DeepL Team, 2024). Thus, DeepL has become one of the most widely known translation systems in recent years.

### 3.5.3 ChatGPT-o3

OpenAI's development of GPT-3 in 2020 marked a significant milestone in MT, demonstrating that a large-scale transformer model—comprising 175 billion parameters and trained on a diverse, multilingual web corpus—could perform competitively using only in-context learning rather than traditional fine-tuning approaches (Brown et al., 2020). Subsequent empirical studies have validated this capability. For instance, Hendy et al. (2023) showed that GPT models perform quite competitively in translation, especially in high-resource languages. Similarly, the study of Neves et al. (2023) demonstrated that GPT-3.5 performed quite well in comparison with 18 systems in biomedical text translation. These results suggest that GPT models might be used particularly for handling user-generated, informal text through few-shot prompting, which allows for dynamic adaptation to varying registers and styles without requiring parameter updates.

OpenAI has continued to refine its models with frequent iterations. GPT-3.5-Turbo, released in March 2023, introduced improved steerability, reliable output formatting and custom tone aiming to improve model performance in various use areas (OpenAI, 2023). The latest advancement, ChatGPT-o3 (launched on 16 April 2025), retains GPT-3's multilingual foundation while incorporating extended reinforcement learning stages, tool integration capabilities, and broader context windows. According to OpenAI's internal evaluations, the o3 model reduces major error rate by approximately 20% compared to OpenAI o1, on challenging real-world tasks (OpenAI, 2025). As reported by the same publishing of OpenAI, with the state-of-the-art advanced reasoning integrated into ChatGPT-o3, translation performance is also expected to increase.

Overall, the current study considered some key factors while choosing the translation systems to compare. In-built system of X was added to the comparison as it is the default system – and powered by Google. DeepL, however, was opted since it has been widely used and studies have confirmed its quality in recent years as mentioned above. Lastly, the latest GPT model of OpenAI, ChatGPT-o3 was selected for the comparison.

## CHAPTER 4: RESULTS

This chapter presents the results of the study by reporting the MQM 2.0 annotations, the COMETKiwi-XL scores, and the correlation analyses between them. The purpose of this section is to provide a clear and systematic account of the empirical outcomes without engaging in interpretation, which will be addressed in the following chapter.

The analysis focuses on football-related tweets as a form of UGC, which are characterized by their informal style, domain-specific jargon, abbreviations, and culture-bound references. Given these features, evaluating translations of such data provides unique insights into the capabilities and limitations of the three MT systems under investigation: DeepL, ChatGPT-o3, and X Translate.

The findings are presented in four stages. First, the inter-rater agreement (IRA) among the three expert annotators is reported to establish the reliability of the MQM evaluations. Second, the distribution of error categories and severity levels is analyzed to highlight the main challenges encountered by each system. Third, segment-level correlations between MQM scores and COMETKiwi-XL predictions are provided to assess consistency at the individual tweet level. Finally, system-level correlations are examined to compare the overall rankings of the systems across the two language directions (EN-TR and TR-EN).

### 4.1 Inter-Rater Agreement (IRA)

To assess the reliability of the MQM 2.0 annotations, inter-rater agreement (IRA) among the three expert evaluators was calculated using Kendall’s Tau-b, a non-parametric correlation coefficient suitable for ordinal data and tied rankings, which are common in error annotation tasks (cf. Berry et al., 2009:1144). The resulting coefficients will be interpreted with reference to the widely cited thresholds proposed by Landis and Koch (1977), whereby values between 0.61–0.80 indicate substantial agreement and values between 0.81–1.00 denote almost perfect agreement; a more detailed discussion of these interpretations is provided in Chapter 5.

**Table 6:** Inter-Rater Agreement (Kendall’s Tau-b)

| Direction      | Annotator Pair | Kendall’s $\tau$ -b |
|----------------|----------------|---------------------|
| <b>Overall</b> | A–B            | 0.864               |
|                | A–C            | 0.685               |
|                | B–C            | 0.691               |
| <b>EN→TR</b>   | A–B            | 0.759               |
|                | A–C            | 0.619               |
|                | B–C            | 0.603               |
| <b>TR→EN</b>   | A–B            | 0.976               |
|                | A–C            | 0.710               |
|                | B–C            | 0.720               |

The pairwise results in Table 6 demonstrate substantial to very high levels of agreement. Overall, Annotators A and B achieved the strongest correlation ( $\tau$ -b = 0.864), while the A–C and B–C pairs reached moderately strong agreement ( $\tau$ -b = 0.685 and 0.691, respectively). In terms of translation direction, the TR-EN direction revealed particularly strong consistency: A–B approached almost perfect agreement ( $\tau$ -b = 0.976), and both A–C and B–C also produced high correlations ( $\tau$ -b = 0.710 and 0.720). In contrast, the EN-TR direction yielded somewhat lower but still substantial agreement, ranging from  $\tau$ -b = 0.603 to  $\tau$ -b = 0.759.

**Table 7:** Overall Agreement of Three Annotators (Kendall’s Tau-b)

| Direction      | Kendall’s $\tau$ -b |
|----------------|---------------------|
| <b>Overall</b> | 0.747               |
| <b>EN→TR</b>   | 0.660               |
| <b>TR→EN</b>   | 0.802               |

In addition to pairwise scores, an aggregated measure was calculated to capture the overall consistency among all three annotators. As shown in Table 7, the overall Kendall’s Tau-b value reached 0.747 across the dataset, indicating substantial agreement. When broken down by translation direction, the TR-EN direction again outperformed EN-TR, achieving  $\tau\text{-b} = 0.802$  compared to  $\tau\text{-b} = 0.660$ .

## 4.2. MQM Error Analysis

To gain a more fine-grained understanding of the translation performance of the three systems, all 60 outputs (30 EN-TR and 30 TR-EN) were annotated according to the MQM 2.0 framework. Unlike aggregate quality scores, error analysis provides insights into the specific linguistic dimensions where systems succeed or fail. In this study, errors were classified into five main categories—Accuracy, Terminology, Linguistic Conventions, Style, and Audience Appropriateness—and weighted according to their severity (Minor = 1, Major = 5, Critical = 25). This design allows both the frequency and the impact of errors to be considered, thus offering a more nuanced measure of translation quality than raw counts alone. Therefore, as stated by Lommel et al. (2024:5) “The Raw Linear score calculates the portion of the evaluated text with errors and subtracts this value from 100 to get the value which directly represents the error-free portion of the evaluated sample”, the use of Linear Raw Scoring Model calculation helps to convert MQM results into normalized 0-100 scores to compare with COMETKiwi-XL scores.

The analysis in this section is organized as follows. First, the Raw Quality Score (RQS) calculations of MQM results are presented. Then, the overall number of error tags for each system are reported across both language directions. Lastly, the distribution of errors by main category is examined, highlighting which linguistic aspects posed the greatest challenges. Finally, the breakdown of error severities with examples are presented to show the proportion of Minor, Major, and Critical issues in each system’s output.

#### 4.2.1 Normalized MQM Raw Quality Scores Calculation

This study computes human quality scores following MQM 2.0 as described by Lommel et al. (2024). Each translation segment was annotated for five error categories—Accuracy, Terminology, Linguistic Conventions, Style, and Audience Appropriateness—and three severities (minor, major, critical). To avoid imposing category preferences, all categories received an equal type weight of 1, whereas severities were scaled with the MQM standard multipliers (minor = 1, major = 5, critical = 25). Thus, category contributes uniformly, while severity governs the magnitude of penalty. Below is the explanation of normalization process through formulas.

1. Let  $i$  index error categories and  $j \in \{\text{minor, major, critical}\}$  index severity. For a given segment, the Absolute Penalty Total (APT) is:

$$APT = \sum_i \sum_j (count_{\{i,j\}} \times severity_j \times type\_weight_i)$$

2. Because  $type\_weight_i = 1$  for all categories in our setup, APT reduces to a severity-weighted sum of counts. APT is then normalized by the length of the target text to obtain the Per-Word Penalty Total (PWPT):

$$PWPT = \frac{APT}{\text{word\_count}}$$

3. Finally, we map penalties to a unit-interval RQS:

$$RQS = \max(0, 1 - PWPT), \quad RQS \in [0,1].$$

RQS equation implements the standard floor effect (Lommel et al., 2024): Segments whose penalty per word equals or exceeds 1 receive  $RQS = 0$ . For rater aggregation, each segment–system–direction triple was annotated independently by three raters. For comparisons with automatic metrics, we collapsed rater variation by taking the mean RQS per item:

$$\overline{RQS}_s = \frac{1}{3} \sum_{r=1}^3 RQS_{r,s},$$

where  $s$  indexes a unique (segment, system, direction) and  $r$  a rater. The per-item standard deviation was also computed descriptively where relevant.

#### 4.2.1.1 Implementation of the Analysis

The scoring pipeline was divided between Google Sheets and Python. All primary calculations (APT, PWPT, RQS) were executed in Google Sheets to ensure transparency and auditability. With the annotation scores of each annotator, gathered through the annotation framework, we created six distinct workbooks with three pages in it representing each MT system. In these workbooks, each row represented one segment, with columns storing error counts per category×severity, the target text, and computed fields. The exact Google Sheet spreadsheet is provided in Appendix C. All spreadsheet formulas were implemented in Google Sheets under the Turkish locale, which uses semicolons as function separators and commas as decimal markers. For subsequent processing in Python, numeric fields were changed into standard dot-decimal format (e.g., “0,9056” → “0.9056”). Word counts were computed using a whitespace-based tokenization procedure implemented in Google Sheets. Each space character was treated as a delimiter, meaning that hashtags, numbers, and emojis were counted as individual tokens whenever separated by spaces.

After per-rater computation, the results were exported from Google Sheets as CSV files (Annotator A/B/C). These files were merged and averaged in Google Colab using Python to produce two consolidated datasets—MQM\_Averages – EN2TR and MQM\_Averages – TR2EN—containing (segment\_id, system, RQS). These normalized

human scores serve as the gold standard in subsequent segment- and system-level correlations with COMETKiwi-XL.

We (i) verified that all error counts were non-negative integers (comparable with COMETKiwi-XL); (ii) flagged rows with empty target text ( $\text{word\_count} = 0$ ) and assigned  $PWPT = 0$  and  $RQS = 1$  only when  $APT = 0$ ; (iii) confirmed that  $0 \leq RQS \leq 1$  for all rows; and (iv) reproduced mean RQS values in Python to validate exports (see Appendix P).

$$0 \leq RQS \leq 1$$

#### 4.2.2 MQM Raw Quality Score Results

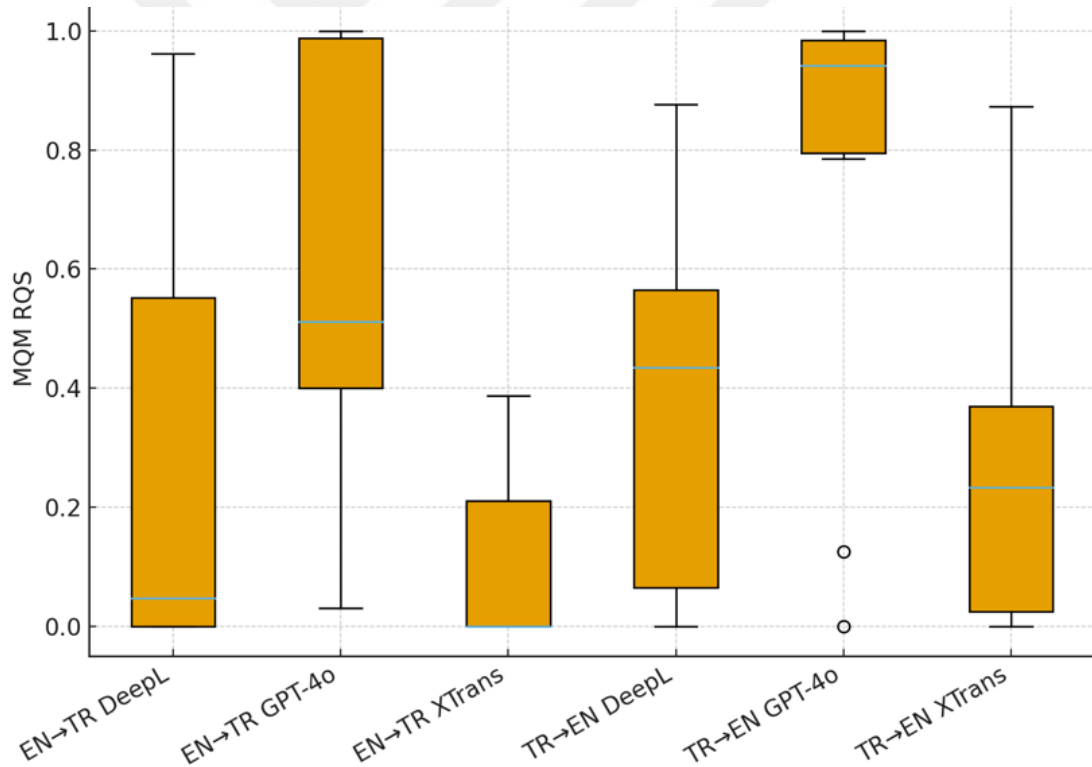
The normalized RQS derived from the MQM framework are presented in this section. Scores represent the mean of three independent annotators per segment–system–direction triple (see section 5.2.1 and Appendix P). Standard deviations are reported to provide a sense of variability across items. Table 8 below summarizes the average RQS values by system for both translation directions.

**Table 8:** Mean MQM RQS Scores (by system and direction)

| Direction | System      | Mean RQS | SD    |
|-----------|-------------|----------|-------|
| EN-TR     | DeepL       | 0.277    | 0.372 |
| EN-TR     | GPT-4o      | 0.607    | 0.357 |
| EN-TR     | X Translate | 0.104    | 0.149 |
| TR-EN     | DeepL       | 0.387    | 0.319 |
| TR-EN     | GPT-4o      | 0.758    | 0.375 |
| TR-EN     | X Translate | 0.274    | 0.285 |

As shown in the table, average RQS values varied considerably across systems. In the EN-TR direction, GPT-4o obtained the highest average RQS (0.607), followed by DeepL (0.277), while X Translate scored markedly lower (0.104). In the TR-EN direction, GPT-4o again outperformed the other systems (0.758), whereas DeepL and X Translate obtained lower averages (0.387 and 0.274, respectively). Standard deviations ranged from 0.149 to 0.375, suggesting some degree of variance across segments but generally stable patterns across annotators.

**Figure 9:** Distribution of MQM Raw Quality Scores (RQS) by System and Direction



Each box shows the interquartile range (IQR), with the horizontal line marking the median. Whiskers extend to  $1.5 \times \text{IQR}$ , and circles indicate outliers. The figure illustrates

how system performance varies not only in central tendency but also in score dispersion, with clearer separation between systems in TR-EN than in EN-TR.

While system-level averages and overall distributions provide a useful summary, they also conceal important segment-level variation. In particular, several items received a RQS of 0.000, reflecting total translation failure. These extreme cases are not fully visible in the aggregated view but have substantial implications for downstream correlation analyses. To capture this variation more transparently, Tables 9 and 10 present the full segment-level RQS results for TR-EN and EN-TR directions, respectively.

**Table 9:** Segment-level MQM RQS Scores (TR-EN)

| <b>Segment</b> | <b>DeepL</b> | <b>ChatGPT-o3</b> | <b>X Translate</b> |
|----------------|--------------|-------------------|--------------------|
| <b>1</b>       | 0.780        | 1.000             | 0.277              |
| <b>2</b>       | 0.877        | 0.986             | 0.348              |
| <b>3</b>       | 0.503        | 0.983             | 0.576              |
| <b>4</b>       | 0.000        | 0.000             | 0.000              |
| <b>5</b>       | 0.439        | 0.942             | 0.873              |
| <b>6</b>       | 0.000        | 0.125             | 0.000              |
| <b>7</b>       | 0.000        | 0.825             | 0.000              |
| <b>8</b>       | 0.259        | 0.943             | 0.190              |
| <b>9</b>       | 0.586        | 0.992             | 0.098              |
| <b>10</b>      | 0.430        | 0.785             | 0.376              |

Table 9 suggests that ChatGPT-o3 performs best with consistently high scores, while DeepL is uneven—strong in some segments but failing in others. X Translate generally scores lowest, with several complete breakdowns.

**Table 10:** Segment-level MQM RQS Scores (EN-TR)

| Segment | DeepL | ChatGPT-o3 | X Translate |
|---------|-------|------------|-------------|
| 1       | 0.030 | 0.030      | 0.121       |
| 2       | 0.061 | 1.000      | 0.000       |
| 3       | 0.000 | 0.444      | 0.000       |
| 4       | 0.000 | 0.952      | 0.000       |
| 5       | 0.764 | 0.250      | 0.000       |
| 6       | 0.673 | 0.000      | 0.388       |
| 7       | 0.000 | 0.578      | 0.000       |
| 8       | 0.962 | 0.000      | 0.295       |
| 9       | 0.326 | 0.426      | 0.241       |
| 10      | 0.000 | 0.000      | 0.000       |

Table 10 suggests that the performance is more uneven across systems compared to TR-EN. ChatGPT-o3 shows strong results in some segments (e.g., 2, 4) but drops to zero in three segments. DeepL is inconsistent—ranging from very low scores to one near-perfect. X Translate mostly scores zero, with only a few modest results.

### 4.3 Overall MQM Error Tags

The MQM annotation process produced a total of 899 error tags across both translation directions. It is important to emphasize that these figures do not represent the number of unique errors or a direct measure of quality, but rather the total count of tags assigned during annotation. Therefore, these findings of error tag counts play a supporting role to see how they align with the raw scores of MQM. The findings of error tag counts are shown below in tables.

**Table 11:** Total Number of Error Tags by Translation Direction

| Direction      | Total Errors |
|----------------|--------------|
| EN-TR          | 447          |
| TR-EN          | 452          |
| <b>Overall</b> | <b>899</b>   |

Table 11 above shows that the number of error tags is quite close to each other in each direction.

**Table 12:** Total Number of Error Tags by System

| System         | Total Errors |
|----------------|--------------|
| DeepL          | 373          |
| ChatGPT-o3     | 109          |
| X Translate    | 417          |
| <b>Overall</b> | <b>899</b>   |

Table 12 above demonstrates that X Translate received the most error tags, followed by DeepL. ChatGPT-o3, however, received significantly lower number of error tags.

As shown in Tables 11–12, the distribution of error tags was relatively balanced across directions (447 in EN→TR and 452 in TR→EN), while variation among systems was more pronounced. For example, X Translate received 417 tags, DeepL 373, and ChatGPT-o3 109. The overall results thus highlight clear disparities among the systems, with direction-specific totals remaining stable but system-level outcomes showing sharp contrasts. The following subsections will provide a more detailed breakdown of these errors by category and severity.

These counts should not be interpreted as conclusive evidence of one system's superiority over another. Instead, they serve as supporting evidence for the MQM RQS, which provide the primary basis for comparing system performance. In other words, error tag totals illustrate annotation tendencies and the relative density of flagged issues but are not in themselves sufficient to determine overall translation quality.

By combining RQS with error-tag distributions, the analysis ensures that both the weighted impact of errors (via RQS) and their frequency across categories (via error tags) are taken into account, giving a more complete picture of the systems' strengths and weaknesses.

#### 4.3.1 Number of Error Tags by System

In this part, the total number of error tags in EN-TR and TR-EN translations across systems is presented.

**Table 13:** Error Tags by System in EN-TR Translations

| <b>System</b> | <b>Errors</b> |
|---------------|---------------|
| DeepL         | 179           |
| ChatGPT-o3    | 51            |
| X Translate   | 217           |
| <b>Total</b>  | <b>447</b>    |

Table 13 shows the number of error tags in EN-TR translations. ChatGPT-o3 produced the fewest errors with the number of 51, while X Translate had the most with 217, followed by DeepL with 179.

**Table 14:** Error Tags by System in TR-EN Translations

| <b>System</b> | <b>Errors</b> |
|---------------|---------------|
| DeepL         | 194           |
| ChatGPT-o3    | 58            |
| X Translate   | 200           |
| <b>Total</b>  | <b>452</b>    |

Table 14 presents the error tag counts for TR-EN translations. ChatGPT-o3 again shows the fewest errors, while X Translate and DeepL produced substantially higher counts.

#### 4.3.2 Number of Error Tags by Category

In this part, the overall number error tags by category is displayed and explained in detail.

**Table 15:** Error Tags by Category

| <b>Error Category</b>    | <b>Total Errors</b> |
|--------------------------|---------------------|
| Accuracy                 | 395                 |
| Linguistic Conventions   | 263                 |
| Style                    | 113                 |
| Audience Appropriateness | 61                  |
| Terminology              | 67                  |
| <b>Total</b>             | <b>899</b>          |

Table 15 shows that most error tags fell under Accuracy and Linguistic Conventions, confirming that meaning transfer and grammatical correctness were the main challenges in this dataset. Other categories such as Style, Terminology, and Audience Appropriateness

appeared less frequently. These distributions are useful for illustrating annotation tendencies but should be read as supportive to MQM RQS, which remains the primary quality indicator.

#### 4.3.2.1 Number of Error Tags by Category – EN-TR

In this part, the overall number of error tags is given for EN-TR direction.

**Table 16:** Error Tags by Category in EN-TR Translations

| Error Category           | Errors     |
|--------------------------|------------|
| Accuracy                 | 186        |
| Linguistic Conventions   | 114        |
| Style                    | 80         |
| Audience Appropriateness | 36         |
| Terminology              | 31         |
| <b>Total</b>             | <b>447</b> |

Table 16 above shows that 447 tags were assigned in this direction, mostly in Accuracy and Linguistic Conventions. This indicates that English fan discourse, with its idiomatic and compressed style, posed difficulties in meaning preservation and grammaticality. Other categories were less frequent but still present, reflecting occasional stylistic and domain-specific issues.

#### 4.3.2.2 Number of Error Tags by Category – TR-EN

In this part, the overall number of error tags is given for TR-EN direction in tables below.

**Table 17:** Error Tags by Category in TR-EN Translations

| Error Category           | Errors     |
|--------------------------|------------|
| Accuracy                 | 209        |
| Linguistic Conventions   | 149        |
| Style                    | 33         |
| Audience Appropriateness | 25         |
| Terminology              | 36         |
| <b>Total</b>             | <b>452</b> |

A similar pattern emerged with 452 tags in total. Accuracy and Linguistic Conventions again dominated, highlighting challenges in rendering Turkish football slang and colloquial language into English while maintaining correctness. Style, Terminology, and Audience Appropriateness tags were fewer but underline specific weaknesses in capturing register and jargon.

#### 4.3.3 Number of Severity Tags

In this part, the overall number of severity tags is given.

**Table 18:** Error Tags by Severity (All Systems, Both Directions)

| Severity Level | Total Errors |
|----------------|--------------|
| Critical       | 213          |
| Major          | 227          |
| Minor          | 459          |
| <b>Total</b>   | <b>899</b>   |

Table 18 shows the severity distribution of error tags across both directions: Minor (459), Major (227), and Critical (213). Most tags were Minor, reflecting surface-level issues

such as punctuation or stylistic deviations. However, nearly half of the remaining tags were Major or Critical, indicating high-impact meaning-related problems. These figures are best read as supportive evidence that complements the weighted MQM RQS scores, which already incorporate severity into the overall calculation.

#### 4.3.3.1 Number of Severity Tags in EN-TR

In this part, the overall number of severity tags in EN-TR direction is given.

**Table 19:** Error Tags by Severity in EN-TR Translations

| Severity Level | Errors     |
|----------------|------------|
| Critical       | 108        |
| Major          | 123        |
| Minor          | 216        |
| <b>Total</b>   | <b>447</b> |

Out of 447 tags, nearly half were Minor (216), while Major (123) and Critical (108) tags together made up 52%. This suggests that although many issues were relatively low-level, a substantial portion directly impaired meaning transfer and readability, underscoring the difficulty of rendering idiomatic English fan discourse in Turkish.

#### 4.3.3.2 Number of Severity Tags in TR-EN

In this part, the overall number of severity tags in TR-EN direction is demonstrated. In both language directions, the presence of Major and Critical tags confirms that simple counts are not enough for quality assessment; instead, RQS provides the more reliable measure by weighting severity appropriately.

**Table 20:** Error Tags by Severity in TR-EN Translations

| Severity Level | Errors     |
|----------------|------------|
| Critical       | 105        |
| Major          | 104        |
| Minor          | 243        |
| <b>Total</b>   | <b>452</b> |

Among 452 tags, Minor errors (243) were more frequent, while Major (104) and Critical (105) were balanced at around 23% each. This indicates that Turkish-to-English translations contained a high volume of surface issues but were also prone to severe adequacy failures, especially when dealing with colloquial football expressions.

#### 4.3.4 Number of Error Tags by Category and System

In this part, the overall number of error tags by category and system is given.

**Table 21:** Error Tags by Category and System

| Error Category           | DeepL      | ChatGPT-o3 | X Translate | Total      |
|--------------------------|------------|------------|-------------|------------|
| Accuracy                 | 154        | 73         | 168         | 395        |
| Linguistic Conventions   | 110        | 7          | 146         | 263        |
| Style                    | 48         | 12         | 53          | 113        |
| Audience Appropriateness | 31         | 10         | 20          | 61         |
| Terminology              | 30         | 7          | 30          | 67         |
| <b>Total</b>             | <b>373</b> | <b>109</b> | <b>417</b>  | <b>899</b> |

Table 21 shows the distribution of error tags across systems and categories. The majority of tags fell under Accuracy (395) and Linguistic Conventions (263). Both DeepL and X Translate accumulated high counts in these categories, while ChatGPT-o3 received far

fewer tags overall. Across all categories, ChatGPT-o3 consistently attracted fewer tags, suggesting relatively stronger performance in meaning preservation, grammar, and style. By contrast, DeepL and X Translate showed weaknesses particularly in semantic fidelity and formal correctness. These distributions illustrate annotation tendencies and align with the broader MQM RQS findings, but they should not be read as independent proof of quality differences. Instead, they provide supportive evidence by showing where systems were more prone to flagged issues.

#### 4.4 Representative Segment-Level Examples from MQM Annotations

In order to complement the numerical results presented in the previous sections, this subsection provides a few representative examples from the MQM annotations. These examples are included to support the quantitative data by illustrating typical error types and to show how errors occurred in specific cases.

**Table 22:** Representative Segment (TR-EN) with Inline MQM Annotations

| Item        | Text                                                                                                                           |
|-------------|--------------------------------------------------------------------------------------------------------------------------------|
| Source      | “Sen utanmaz bir herifsin lan şeref yoksunu <b>yahı çocuğu!...</b> ”                                                           |
| DeepL       | “You’re a shameless bastard, you dishonorable <b>son of a bitch!...</b> ”<br>(Mistranslation — Critical; cultural nuance lost) |
| ChatGPT-o3  | “You’re a shameless bastard, you honor-less <b>mansion brat!...</b> ” (No Error — preserves intended insult)                   |
| X Translate | “You are a shameless man, you dishonorable <b>son of a bitch!...</b> ”<br>(Mistranslation — Critical; cultural nuance lost)    |

This TR-EN example shows how the culturally loaded insult *yahı çocuğu* (“spoiled rich kid from a mansion”) was mistranslated by DeepL and X Translate as “son of a bitch,” leading to a critical accuracy error and a loss of cultural nuance. Such renderings replace a

socio-cultural insult tied to wealth and privilege with a generic vulgarity, thereby altering the pragmatic effect of the original. ChatGPT-o3, although slightly awkward, produced “mansion brat,” which better preserved the intended meaning and conveyed the social implication behind the expression more effectively. This highlights the particular difficulty MT systems face when dealing with culture-specific insults, where literal or generic substitutions often fail to capture the underlying nuance.

**Table 23:** Representative Segment (TR-EN) with Inline MQM Annotations

| Item               | Text                                                                                                                                                                                                                                                                                                                |
|--------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Source</b>      | “İstifa et yeter düş ulan yakamızdan Allah’ın belası Allah’ın laneti yeter bırak git”                                                                                                                                                                                                                               |
| <b>DeepL</b>       | “Resign, enough <b>shower</b> , God’s curse, God’s curse, enough let go, God’s <b>gs person</b> , damn it...” (“enough shower” – Mistranslation <b>Critical</b> ; “God’s curse, God’s curse” – Style <b>Minor</b> ; “God’s gs person” – Mistranslation <b>Major</b> )                                               |
| <b>ChatGPT-o3</b>  | “Just resign already and get off our backs, <b>you curse of God!</b> Enough—leave and go. Damn you, you shameless thing...” (“you curse of God” – Style <b>Minor</b> )                                                                                                                                              |
| <b>X Translate</b> | “Resign, enough is enough, get off our backs, God’s curse, God’s curse is enough, <b>leave me, God’s curse be upon me...</b> ” (Punctuation issues – <b>Minor</b> ; “leave me” – Mistranslation <b>Major</b> ; “God’s curse be upon me” – Mistranslation <b>Major</b> ; “awkward repetition” – Style <b>Minor</b> ) |

This example illustrates how noisy, emotionally charged football tweets expose multiple error types simultaneously and challenge MT systems in different ways. DeepL’s output contained a critical mistranslation of *düş* as “shower,” along with awkward repetition and a misrendering of *yakadan düşmek* (“get off our backs”), leading to a severe distortion of meaning. X Translate showed similar weaknesses, with added punctuation noise and two major mistranslations that disrupted both adequacy and fluency. By contrast, ChatGPT-o3 produced a more fluent and idiomatic rendering, with only a minor stylistic issue in the phrase “curse of God.” Although not perfect, it managed to preserve the intended tone and pragmatic force of the original insult more effectively. This case underlines the difficulty of translating emotionally loaded and culturally specific expressions, where accuracy, fluency, and style must all be preserved simultaneously.

**Table 24:** Representative Segment (EN-TR) with Inline MQM Annotations

| Item   | Text                                                                                                                                                                                                         |
|--------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Source | “Na man i think MSN was an atg front 3 all balling out”                                                                                                                                                      |
| DeepL  | “Hayır dostum, bence MSN, hepsi top oynayan bir <b>atg ön 3’lüsüydü...</b> ”<br>(MSN – Untranslated <b>Minor</b> ; “top oynayan” – Undertranslation <b>Minor</b> ;<br>“atg” – Untranslated <b>Critical</b> ) |

**Table 24 - continue**

|                    |                                                                                                                                                                                      |
|--------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>X Translate</b> | “Hayır dostum, bence MSN bir <b>atg ön cephesiydi</b> ve hepsi <b>topla oynuyordu...</b> ” (“atg” – Untranslated <b>Critical</b> ; “topla oynuyordu” – Mistranslation <b>Major</b> ) |
| <b>ChatGPT-o3</b>  | “Yok kanka, bence MSN gelmiş geçmiş en iyi hücum üçlüsü—hepsi de sahada coşuyordu.” (MSN – Culture-specific reference <b>Major</b> )                                                 |

This EN-TR example highlights how football-specific slang “atg = all-time great” and abbreviations create major challenges for MT systems. Both DeepL and X Translate failed to render *atg*, leaving it untranslated, which was annotated as a critical error. X Translate compounded this with a major mistranslation of “balling out” as “topla oynuyordu.” ChatGPT-o3, on the other hand, opted for a culturally adapted translation—“gelmiş geçmiş en iyi hücum üçlüsü”—which, while annotated as a major culture-specific reference error due to interpretive freedom, conveyed the intended meaning more effectively and fluently in Turkish.

**Table 25:** Representative Segment (EN-TR) with Inline MQM Annotations

| <b>Item</b>       | <b>Text</b>                                                                                                                                                          |
|-------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Source</b>     | “No energy, no passing sequence.. Except Yıldız this team has nothing affront...”                                                                                    |
| <b>DeepL</b>      | “Enerji yok, <b>pas sırası</b> yok. Yıldız dışında bu takımda hiçbir şey yok.” (“pas sırası” – Accuracy <b>Major</b> ; under-informative rendering of football term) |
| <b>ChatGPT-o3</b> | “Ne enerji var, ne <b>pas trafiği</b> . Yıldız dışında hücumda hiçbir şey yok...” (No Error – fluent and contextually appropriate)                                   |

**Table 25 - continue**

|                        |                                                                                                                                                                                     |
|------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>X<br/>Translate</b> | “Enerji yok, <b>pas dizisi</b> yok.. Yıldız dışında bu takımda hiçbir <b>hakaret</b> yok...” (“pas dizisi” – Wrong Term <b>Minor</b> ; “hakaret” – Mistranslation <b>Critical</b> ) |
|------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

This EN-TR example illustrates how football terminology challenges system performance. DeepL mistranslated “passing sequence” as “pas sırası”, which is not idiomatic in Turkish football discourse, and was annotated as a major accuracy error. X Translate produced “pas dizisi” (minor wrong term) and, more severely, mistranslated “affront” as “hakaret”, marked as a critical error due to the distortion of meaning. In contrast, ChatGPT-o3 correctly rendered “passing sequence” as “pas trafiği” and conveyed the intended football-specific nuance without errors.

In sum, the MQM analysis not only quantified error distributions across systems, categories, and severity levels but also revealed concrete patterns through segment-level examples. The representative cases demonstrated how football-related UGC challenged MT systems, particularly with cultural insults, idiomatic slang, and domain-specific terminology. While ChatGPT-o3 generally preserved meaning and style more effectively, DeepL and X Translate frequently produced major and critical mistranslations that altered the intended message. These illustrative examples reinforce the aggregate findings and highlight the importance of qualitative analysis alongside quantitative results. Building on this human-evaluation baseline, the next section presents the automatic evaluation results obtained with COMETKiwi-XL.

#### **4.5 COMETKiwi-XL Results**

To complement the human evaluation with MQM, this section presents the results of the automatic evaluation using COMETKiwi-XL, a state-of-the-art reference-free quality estimation model (Rei et al., 2023). COMETKiwi-XL predicts translation quality scores on

a continuous scale between 0 and 1, with higher values indicating better adequacy and fluency. Results are reported at both the segment level (individual tweets) and the system level (average scores across all segments) for EN-TR and TR-EN directions. While segment-level scores reveal variability in performance across different tweets, the system-level averages provide a robust basis for cross-system comparison.

**Table 26:** System-Level COMETKiwi-XL Scores

| <b>System</b>      | <b>EN-TR Avg. Score</b> | <b>TR-EN Avg. Score</b> |
|--------------------|-------------------------|-------------------------|
| <b>DeepL</b>       | 0.627                   | 0.506                   |
| <b>ChatGPT-o3</b>  | 0.638                   | 0.496                   |
| <b>X Translate</b> | 0.602                   | 0.527                   |

As shown in Table 26, in the EN-TR direction, ChatGPT-o3 achieved the highest average COMET score (0.638), followed closely by DeepL (0.627). X Translate performed less strongly (0.602), though the margin separating the three systems remained relatively narrow. By contrast, in the TR-EN direction, the ranking shifted: X Translate obtained the highest score (0.527), while DeepL (0.506) and ChatGPT-o3 (0.496) trailed slightly behind. These results suggest that while ChatGPT-o3 and DeepL were stronger in EN-TR translations, X Translate appeared relatively more competitive in TR-EN, even surpassing the other two systems in terms of COMET’s automatic predictions.

Tables 27 and 28 display the segment-level COMETKiwi-XL scores for EN-TR and TR-EN translations of the ten football-related tweets. These results reveal several noteworthy trends.

**Table 27:** Segment-Level COMETKiwi-XL Scores (EN-TR)

| <b>Tweet</b> | <b>DeepL</b> | <b>ChatGPT-o3</b> | <b>X Translate</b> |
|--------------|--------------|-------------------|--------------------|
| <b>1</b>     | 0.548        | 0.652             | 0.695              |
| <b>2</b>     | 0.644        | 0.636             | 0.635              |
| <b>3</b>     | 0.512        | 0.534             | 0.509              |
| <b>4</b>     | 0.417        | 0.493             | 0.403              |
| <b>5</b>     | 0.699        | 0.739             | 0.665              |
| <b>6</b>     | 0.747        | 0.711             | 0.703              |
| <b>7</b>     | 0.660        | 0.649             | 0.629              |
| <b>8</b>     | 0.690        | 0.705             | 0.621              |
| <b>9</b>     | 0.688        | 0.656             | 0.639              |
| <b>10</b>    | 0.663        | 0.610             | 0.522              |

In the EN-TR direction (Table 27), scores fluctuated across tweets but followed a relatively consistent ranking pattern. ChatGPT-o3 and DeepL generally achieved higher values than X Translate, particularly in Tweets 5–8, where both systems surpassed 0.70, while X Translate scored slightly lower. The differences, however, were not uniform. For instance, in Tweet 1 X Translate (0.695) outperformed both DeepL (0.548) and ChatGPT-o3 (0.652), showing that system performance could vary depending on lexical or contextual challenges within individual tweets. Overall, ChatGPT-o3 achieved the strongest performance in most cases, while X Translate tended to lag behind, though occasional tweet-level outliers reversed the ranking.

**Table 28:** Segment-Level COMETKiwi-XL Scores (TR-EN)

| <b>Tweet</b> | <b>DeepL</b> | <b>ChatGPT-o3</b> | <b>X Translate</b> |
|--------------|--------------|-------------------|--------------------|
| <b>1</b>     | 0.588        | 0.624             | 0.599              |
| <b>2</b>     | 0.598        | 0.578             | 0.577              |
| <b>3</b>     | 0.564        | 0.503             | 0.563              |
| <b>4</b>     | 0.360        | 0.316             | 0.455              |
| <b>5</b>     | 0.578        | 0.602             | 0.602              |
| <b>6</b>     | 0.261        | 0.312             | 0.316              |
| <b>7</b>     | 0.524        | 0.478             | 0.528              |
| <b>8</b>     | 0.472        | 0.407             | 0.540              |
| <b>9</b>     | 0.596        | 0.595             | 0.579              |
| <b>10</b>    | 0.513        | 0.541             | 0.506              |

In the TR-EN direction (Table 28), the picture was less uniform. Here, X Translate often performed comparably to or better than the other two systems, as seen in Tweets 5, 7, and 8, where its scores exceeded both DeepL and ChatGPT-o3. DeepL, by contrast, displayed greater inconsistency, ranging from relatively strong scores (0.596 in Tweet 9) to much lower ones (0.261 in Tweet 6). ChatGPT-o3 showed similar instability, with outputs that occasionally matched DeepL but also dropped below in certain tweets (e.g., Tweet 3 and Tweet 4). This variability suggests that TR-EN translation posed greater challenges for all three systems, producing more volatile segment-level outcomes than EN-TR.

Taken together, the segment-level results demonstrate that while system rankings were broadly stable in EN-TR, performance was far more variable in TR-EN. These observations align with the human annotation results reported earlier, where TR-EN also produced greater disagreement and more frequent critical errors. Nevertheless, as will be

shown in the next subsection, averaging across tweets provides a clearer view of the systems’ relative strengths and weaknesses at the system level.

#### 4.6 Correlation Analysis

Having presented the MQM and COMETKiwi-XL results separately, this section examines the extent to which the two evaluation approaches align. Correlation analysis serves as a critical step in validating whether automatic quality estimation scores mirror human judgments of translation quality. Specifically, correlations were calculated at both the segment level (tweet-by-tweet MQM RQS vs. COMET scores) and the system level (average MQM scores vs. average COMET scores for each MT system). Pearson’s  $r$  was used to capture linear associations, while Kendall’s Tau-b was applied to assess ordinal consistency, particularly in cases where rankings and ties were relevant. By combining these complementary measures, the analysis provides a robust account of the degree of overlap—and divergence—between human and automatic evaluations across both language directions.

**Table 29:** Segment-Level Correlation between MQM and COMETKiwi-XL

| Direction | Pearson’s $r$ | Kendall’s Tau-b ( $\tau$ ) |
|-----------|---------------|----------------------------|
| EN-TR     | 0.3732        | 0.3514                     |
| TR-EN     | 0.5626        | 0.4336                     |

As shown in Table 29, segment-level correlations between MQM and COMETKiwi-XL scores were moderate but varied across directions. In the EN-TR direction, correlations were relatively modest (Pearson’s  $r = 0.373$ ; Kendall’s  $\tau = 0.351$ ), suggesting only partial alignment between human judgments and COMET predictions at the tweet level. By contrast, in the TR-EN direction, stronger correlations emerged (Pearson’s  $r = 0.563$ ; Kendall’s  $\tau = 0.434$ ), indicating a clearer association between the two evaluation methods. These results imply that COMETKiwi-XL was more consistent with human annotators when evaluating

Turkish-to-English translations, while its alignment was weaker in the English-to-Turkish direction.

**Table 30:** System-Level MQM vs. COMETKiwi-XL Averages

| Direction | System      | MQM Avg. Score | COMET Avg. Score |
|-----------|-------------|----------------|------------------|
| EN-TR     | DeepL       | 0.277          | 0.627            |
|           | ChatGPT-o3  | 0.607          | 0.638            |
|           | X Translate | 0.104          | 0.602            |
| TR-EN     | DeepL       | 0.387          | 0.506            |
|           | ChatGPT-o3  | 0.760          | 0.496            |
|           | X Translate | 0.274          | 0.527            |

Table 30 presents the average MQM and COMETKiwi-XL scores across systems. In the EN-TR condition, the rankings produced by the two methods were identical: ChatGPT-o3 achieved the highest scores, followed by DeepL and then X Translate. This convergence suggests strong alignment between human and automatic evaluation for English-to-Turkish football tweets.

In contrast, in the TR-EN condition, the two evaluation methods diverged completely. MQM placed ChatGPT-o3 substantially higher than the other systems, while COMET scored it lowest, instead ranking X Translate highest. This discrepancy indicates that correlations at the system level are likely to be positive for EN-TR but negative for TR-EN. To examine this relationship more directly, Table 31 reports the correlation coefficients between MQM and COMET averages.

**Table 31:** System-Level Correlation Results between MQM and COMETKiwi-XL

| Translation Direction | Pearson's $r$ | Kendall's Tau-b ( $\tau$ ) |
|-----------------------|---------------|----------------------------|
| EN-TR                 | 0.928         | 1.000                      |
| TR-EN                 | -0.877        | -1.000                     |

At the system level, the correlation analysis revealed a striking directional asymmetry. In the EN-TR direction, Pearson’s  $r$  was 0.928 and Kendall’s Tau reached a perfect 1.000, indicating near-complete agreement between MQM and COMETKiwi-XL in ranking the three systems. This result confirms that, for English-to-Turkish football tweets, COMET closely mirrored human annotations not only in relative rankings but also in score distributions.

By contrast, in the TR-EN direction, correlations were strongly negative (Pearson’s  $r = -0.877$ ; Kendall’s  $\tau = -1.000$ ). Here, the two methods produced completely inverted system rankings: MQM rated ChatGPT-o3 highest and X Translate lowest, whereas COMETKiwi-XL favored X Translate and placed ChatGPT-o3 last. This inverse relationship demonstrates a fundamental misalignment between human and automatic evaluations for Turkish-to-English translations.

In summary, the correlation analysis revealed a clear directional asymmetry. At the segment level, COMETKiwi-XL showed moderate alignment with MQM, with stronger associations in TR-EN than in EN-TR. At the system level, the contrast was more pronounced: in EN-TR, MQM and COMET were in near-perfect agreement, whereas in TR-EN, the two methods produced completely inverted rankings, yielding strong negative correlations. These findings suggest that while COMETKiwi-XL can serve as a reliable proxy for human judgments in English-to-Turkish translations, its performance in Turkish-to-English evaluations is far less dependable. This asymmetry underscores the importance of combining human annotation with automatic metrics, particularly in the evaluation of UGC content where cultural and linguistic nuances play a decisive role.

## CHAPTER 5: DISCUSSION

This chapter provides a critical interpretation of the empirical findings by triangulating human evaluation through MQM 2.0 with automatic quality estimation using COMETKiwi-XL on football-related user-generated tweets. The discussion addresses the observed error patterns—particularly the predominance of Accuracy and Linguistic Conventions—and explores the direction-specific differences between EN-TR and TR-EN translations. In addition, the relationships between MQM and COMETKiwi-XL scores at both the segment and system levels are examined to evaluate the extent to which automatic metrics approximate human judgments. The chapter concludes by reflecting on the implications of these findings for translation evaluation in UGC domains, outlining limitations of the present study, and suggesting directions for future research.

### 5.1 Inter-Rater Agreement

The inter-rater agreement (IRA) observed in this study indicates that the annotators applied the MQM framework with a high degree of internal consistency. Kendall’s Tau-b coefficients ranged from 0.603 to 0.976, with overall agreement scores between annotator pairs reaching  $\tau = 0.864$  (A–B), 0.685 (A–C), and 0.691 (B–C). Directional analysis showed that in the EN-TR condition, agreement was substantial ( $\tau = 0.603$ –0.759), whereas in the TR-EN condition, agreement reached good to excellent levels ( $\tau = 0.710$ –0.976).

According to the well-known interpretative scheme proposed by Landis and Koch (1977), which classifies categorical agreement levels based on the “Agreement Measures for Categorical Data” table, values between 0.61–0.80 are considered *substantial* and those between 0.81–1.00 are considered *almost perfect* (p. 165). Similarly, Koo and Li (2016) provide reliability guidelines for correlation-based measures, noting that “values between 0.75 and 0.90 indicate good reliability, and values greater than 0.90 indicate excellent reliability” (p. 162). Applying these thresholds, the results here place EN-TR in the *substantial* category and TR-EN in the *good to excellent* range.

This outcome stands out when compared to earlier MQM evaluations. Freitag et al. (2021) report that “MQM annotations ... between annotators themselves, show relatively low levels of agreement” (p. 3), typically in the 0.20–0.60 range. Against this background, the comparatively high IRA values in this study indicate that annotators applied the error typology and severity weights with consistency, thereby producing a dataset more reliable than typically documented in MQM-based research. This provides a strong empirical foundation for the subsequent error analysis and for the comparison between MQM and COMETKiwi-XL.

Also, Lee et al. (2023) suggest that despite its strengths, human evaluation faces several persistent challenges, such as inter-annotator agreement, subjectivity, and its time-consuming nature. In this study, however, the agreement values were considerably higher than commonly reported, which can be attributed to several methodological factors. Annotators were provided with detailed training materials tailored both to MQM and to the aims of this study. In addition, the Google Sheets annotation interface offered an extensive framework with drop-down menus, highlighted source and target segments for potential errors, and even an additional column to capture unforeseen errors. These measures minimized ambiguity and fostered consistency among annotators, ultimately leading to stronger agreement and a more robust dataset.

## **5.2 MQM Error Analysis**

In this section, the MQM error findings and analysis are discussed in detail referring to current literature. Firstly, normalized MQM RQS results will be discussed, followed by number of error tags, and correlation.

### **5.2.1 Normalized MQM RQS Results**

The RQS results obtained in this study reveal a complex interplay between system-level consistency and segment-level variability. At the system level, GPT-o3 outperformed the

other systems across both EN-TR and TR-EN, whereas X Translate lagged considerably behind. However, segment-level analysis exposed substantial variance, including multiple instances of  $RQS = 0.000$ , reflecting complete translation breakdowns. These extremes exemplify the floor effect (Michalos, 2014:631,632) whereby a high concentration of minimum scores compresses variance and undermines correlation at the segment level.

However, segment-level distributions revealed a more complex picture. Several translations received  $RQS = 0.000$ , indicating total breakdowns (Tables 10–11). These outcomes exemplify the floor effect, whereby the accumulation of minimum values compresses variance and undermines statistical correlation.  $RQS$  is computed according to the MQM raw linear scoring model:  $RQS \dots 1 - PWPT \dots$  (or  $100 - PWPT \times 100$ ) (Lommel et al., 2024:11). From this formula it follows that when  $PWPT \geq 1$ ,  $RQS = 0$ . This explains why multiple segments in our dataset received  $RQS = 0.000$ , indicating total translation breakdowns and compressing score variance — a classic floor effect.

The results also showed partial rater divergence, particularly in football-related UGC translations containing hashtags, jargon, and noisy markers. Such disagreements are a well-documented feature of linguistic annotation, which is quite natural in human annotation. Furthermore, the differences extended across translation directions. In TR-EN, the spread of scores was wider and the systems were more clearly distinguished, whereas in EN-TR the outputs clustered more closely together (see Figure 7). Similar tendencies were observed in the WMT22 Metrics Task, which reported that surface-overlap measures such as BLEU, spBLEU, or chrF aligned only weakly with human assessments, while neural metrics provided more robust correlations (Freitag et al. 2022:48). Likewise, the WMT23 report emphasized that strong segment-level correlations do not necessarily translate into equally strong correlations at the system level (Freitag et al. 2023:602).

Finally, the segment-level outcomes in this study displayed notable variation among annotators. This observation is consistent with Lommel et al. (2024:3), who remark that “on a segment level two linguists are bound to have three opinions on the same issue.” Such variability reflects the inherent subjectivity of annotation rather than a flaw in the evaluation

process, underscoring the challenges of achieving stable inter-rater agreement in domains dominated by UGC.

### **5.2.2 Number of Error Tags in MQM Annotation**

The MQM analysis produced a nearly balanced number of error tags (EN-TR: 447; TR-EN: 452). It is important to clarify that these values represent the count of annotation tags, not the total number of unique errors. Since a single translation may receive multiple tags for different dimensions, these figures cannot be treated as absolute error counts or direct indicators of system quality. Error-tag statistics illustrate annotation patterns—for example, Accuracy tags were more frequent in EN-TR, while Linguistic Conventions were relatively higher in TR-EN—but they cannot establish which system performed better or worse. Previous MQM studies show similar tendencies. Freitag et al. (2021) report that accuracy/mistranslation tags dominate both human and machine outputs, while Lu et al. (2023:22) find that accuracy/mistranslation (39%), style/awkward (12%), and fluency/grammar (9%) account for most tags across languages.

In terms of severity, Minor tags were most common ( $n = 459$ ), though Major (227) and Critical (213) tags had greater impact on adequacy and fluency. This pattern is consistent with prior work, where surface issues such as punctuation are less decisive (Freitag et al., 2021:1470), while high-severity mistranslation or omission tags drive more substantial quality differences (Freitag et al., 2022:52).

In sum, error-tag distributions highlight where annotators flagged issues most often, but they remain complementary to RQS, which weight severity and normalize the scores across text length, provide more reliable basis for interpreting translation quality.

### **5.2.3 General Error Performance of the Systems**

The aggregate MQM results indicated variation across systems, but these outcomes should be read only as supportive to RQS, not as direct proof of superiority. ChatGPT-o3 received

fewer error tags overall—particularly in Accuracy and Style—while DeepL and X Translate showed relatively more tags in Accuracy and Linguistic Conventions.

Comparable tendencies are reported in prior research. For instance, ChatGPT was found to handle meaning and nuance better than DeepL and Google Translate in poetry translation (Gao et al., 2024), while studies on social media note recurring adequacy and lexical issues in Google Translate outputs (Sabtan et al., 2021). Other work observes that GPT systems often produce less literal, more idiomatic renderings (Raunak et al., 2023). MQM-based evaluations similarly highlight that accuracy, fluency, and style remain persistent challenges for MT systems, especially in UGC (Fakih et al., 2024).

Overall, these findings suggest broad patterns consistent with the present dataset. Still, since error-tag counts are not equivalent to actual error numbers, they should not be used in isolation to rank systems. The more reliable assessment comes from the weighted RQS scores, with error-tag distributions serving only as complementary context.

#### **5.2.4 Direction-Based Error Distribution**

The directional breakdown showed similar totals (EN-TR: 447; TR-EN: 452), but category profiles differed. EN-TR contained more Accuracy (186) and Style (80) tags, while TR-EN showed slightly higher Accuracy (209) and notably more Linguistic Conventions (149), with fewer Style issues (33).

This pattern aligns with prior MQM research, where Accuracy and Fluency (Linguistic Conventions) dominate across languages (Freitag et al., 2021). Directional variation is also consistent with findings that error distributions shift depending on system type and translation context (Ye & Toral, 2020; Klubička et al., 2018).

#### **5.2.5 Per-System MQM Performance in Context**

At the system level, Accuracy and Linguistic Conventions tags were the most frequent across all outputs. DeepL (154 / 110) and X Translate (168 / 146) accumulated higher counts in

these categories, while ChatGPT-o3 had considerably fewer tags (73 / 7). Other categories such as Style, Terminology, and Audience Appropriateness were less common overall but again showed higher annotation density for DeepL and X Translate.

Similar patterns have been observed in comparative studies. GPT systems are reported to achieve greater fluency (Hendy et al., 2023), though sometimes at the cost of over-inserting punctuation. Research also notes that while DeepL is generally accurate, it may miss nuance, and Google Translate (on which X Translate is based) often struggles with complex structures (Çetin & Duran, 2024). MQM shared-task reports likewise confirm that accuracy-related issues dominate system evaluations (Freitag et al., 2021, 2022).

In this context, the present tag distributions align with earlier findings, showing where systems tend to attract more annotation. Still, they should not be read as conclusive rankings of system quality. Instead, they provide descriptive context, helping to illustrate the types of challenges each system faced when handling football-related UGC.

### 5.3 Segment Level Correlation

At the segment level, COMETKiwi-XL's scores showed moderate alignment with MQM 2.0: EN-TR yielded Pearson  $\approx 0.37$ ,  $\tau_b \approx 0.35$ , while TR-EN reached Pearson  $\approx 0.56$ ,  $\tau_b \approx 0.43$ . The reason of the difference between the correlations of TR-EN and EN-TR direction might be because of the floor effect in MQM RQS data, which occurs “When data hits the bottom end of the distribution range, the effect is called the “floor effect,” ... they represent a measurement problem when attempting to identify changes.” (Michalos, 2014:631,632). In EN-TR MQM RQS results, there are 13 0.000 points while it is 7 in TR-EN. Therefore, it is possible that the floor effect damaged the correlation especially between EN-TR.

On the other hand, COMET family did not produce any 0.000 score. Furthermore, COMETKiwi-XL did not even produce scores near 0.000. In Zouhar et al. (2024:4), even empty translations did not receive 0.000, noted as “this is not the proper behavior because an empty hypothesis should receive a score of 0, to match no meaning preserved. This however

is unsurprising with COMET, which has likely not seen empty hypotheses during training that would have received a score of 0 from a human annotator.”

Similarly, the magnitude and directionality of these correlations are consistent with findings from recent shared tasks and QE studies. First, large-scale MQM meta-evaluations have long noted that segment-level agreement is noisier than system-level. For WMT23, the organizers explicitly caution that “High correlations between automatic metrics and human judgments at the segment level do not necessarily guarantee high correlations at the system level” (Freitag et al. 2023: 80). Building on this, WMT23 replaced Kendall’s  $\tau$  at the segment level with an accuracy formulation that “gives metrics credit for correctly predicting ties in human scores, while automatically calibrating for each metric’s natural scale,” to better reflect segment-level decision properties (Freitag et al. 2023:580).

Historical benchmarks also contextualize the present results. In WMT21, average metric–human correlations at the segment level were reported in the range of 0.17–0.30, indicating the inherent difficulty of capturing fine-grained judgments (Freitag et al., 2021, 1470). Similarly, the WMT22 report noted that differences among metrics were especially pronounced at the segment level, reflecting higher variability in fine-grained evaluation (Freitag et al., 2022).

Recent analyses have further shown why segment-level correlations often plateau. Agrawal et al. (2024) observe that current metrics remain insensitive to nuanced quality differences and struggle to distinguish between outputs when both are of high quality. At the same time, reference-free QE models like COMETKiwi have consistently demonstrated strong segment-level alignment with human judgments, with the IST–Unbabel submission noting “strong results” in this regard (Rei et al., 2022: 7). The WMT23 leaderboard confirmed this performance, ranking COMETKiwi among the top metrics with an average correlation of 0.78 (Freitag et al. 2023). Furthermore, Rei et al. (2021) showed that segment-level metric–human correlations for English–German approached the same range as inter-annotator correlations, suggesting a practical ceiling effect.

Against this backdrop, the observed segment-level correlations suggest that COMETKiwi-XL is capturing a meaningful portion of the MQM signal on football-related

UGC—especially for TR-EN—while also reflecting well-documented limitations of segment-level meta-evaluation (tie prevalence, subtle preference judgments, and quality ceilings). The move toward tie-aware accuracy in WMT23 strengthens the case that moderate  $\tau_b$  values, coupled with consistent system-level behavior (analyzed in the next section), are an expected and informative outcome for contemporary, high-performing QE metrics in challenging, slang-heavy domains.

#### 5.4 System Level Correlation

At the system level, the results revealed a striking directional asymmetry: while EN-TR showed near-perfect alignment between MQM and COMETKiwi-XL (Pearson  $\approx 0.93$ ;  $\tau = 1.00$ ), TR-EN produced an inverse correlation (Pearson  $\approx -0.88$ ;  $\tau = -1.00$ ). This significant inverse correlation when the language direction changes could be explained by the fact that the Turkish literature lacks studies utilizing COMET family as automatic metric. Supported by Freitag et al. (2022:47), “Correlations with human scores have different ranges in different settings”. This similar case was evident in Freitag et al. (2021b:1466) who put “When comparing the system ranks based on WMT for both language pairs with the ones generated by MQM, we can see low correlation for English→German ... and even negative correlation for Chinese→English ... We also see very low segment-level correlation for both language pairs”. Also, again connecting the reason to the language origin “Another possible explanation is the nature of the source sentences. Most sentences come from Chinese government news pages that have a formal style that may be difficult to render in English” (Freitag et al. (2021b:1469)).

Nevertheless, in the broader literature the system-level metric–human agreement is typically reported as substantially stronger and more stable than segment-level. The WMT21 Metrics Shared Task noted that system-level correlations with MQM usually fall in the 0.60–0.80 range across language pairs, significantly higher than segment-level associations (Freitag et al., 2021). Similarly, the WMT22 report emphasized that correlations at the system level were overall higher than in the previous year, with COMET and related neural metrics

consistently emerging as the best-performing approaches (Freitag et al., 2022). WMT23 further reinforced this view by observing that neural metrics such as COMET and CometKiwi tend to achieve the strongest alignment with expert MQM scores across conditions (Freitag et al., 2023).

Both domain and evaluation setup are known to influence system-level results. WMT21 documented that metric performance varies considerably depending on the test domain and that rankings can shift substantially when moving from one dataset (e.g., news) to another (e.g., TED) (Freitag et al., 2021). Since the present dataset comprises football-related UGC tweets—arguably more informal and culturally specific than standard benchmarks—the inverted TR-EN correlation plausibly reflects such domain effects. Another factor concerns the nature of reference-free QE metrics: WMT21 found that while COMET-QE and OpenKiwi performed competitively when human translations were included, they tended to underperform relative to reference-based metrics when only MT outputs were evaluated (Freitag et al., 2021). As the current study compares only MT systems, this limitation may have amplified the divergence in TR-EN.

Meanwhile, COMETKiwi itself has consistently ranked among the top QE metrics in shared tasks. The 2022 Unbabel-IST submission reported that COMETKiwi achieved state-of-the-art results in three out of four language pairs, while the 2023 submission highlighted the scalability of COMETKiwi-XL and its strong performance across multiple tracks (Rei et al., 2022; Rei et al., 2023). Earlier, the 2021 system showed notable gains from task-specific tuning, yielding over 0.1 improvement in Pearson correlation compared to baseline models (Rei et al., 2021). Most recently, the WMT24 Metrics Shared Task stressed that large language models are reshaping MT evaluation but cautioned that persistent limitations remain unaddressed, even with neural and LLM-based metrics (Freitag et al. 2024, 18). Overall, these insights suggest that while COMETKiwi-XL can replicate human-derived rankings with high fidelity (EN-TR), directionality, domain-specific challenges, and the constraints of QE models explain the observed inversion in TR-EN.

## CHAPTER 6: CONCLUSION

This thesis set out to examine the relationship between human-based MQM 2.0 annotation and the reference-free COMETKiwi-XL metric in the evaluation of football-related UGC. While the study also assessed the outputs of three widely used MT systems, namely ChatGPT-o3, DeepL, and X Translate, the central focus was not on ranking these systems but on analyzing the extent to which COMETKiwi-XL aligns with and complements structured human judgments through MQM. Since this metric had not previously been systematically evaluated for the English–Turkish language pair, the research makes an original contribution by offering one of the first accounts of MQM–COMETKiwi-XL correlation in this context, using football discourse as a domain marked by slang, cultural references, and informal expressions.

The findings suggest that this dual framework can provide a more comprehensive account of translation quality than either dimension alone. Human annotation offered fine-grained insights into the types and severities of errors, especially in areas such as accuracy, linguistic conventions, and style. COMETKiwi-XL, in contrast, produced smoother score distributions that pointed to relative quality differences while avoiding complete collapses. Together, these perspectives indicate that triangulation between human and automatic evaluation can strengthen the reliability of translation assessment in noisy and domain-specific contexts.

At the same time, the results point to clear directional differences. In English-to-Turkish, COMETKiwi-XL scores and human judgments through MQM tended to converge at the system level, suggesting that neural metrics can capture a substantial portion of human evaluation in this direction. In Turkish-to-English, however, the two diverged markedly, even reversing the relative system ranking. This asymmetry highlights that directionality remains a critical factor in machine translation evaluation and that automatic metrics may not be equally reliable across language pairs.

As with any research, this study is bounded by its design choices. The dataset was deliberately small—only 20 tweets—and restricted to one domain. While this ensured

detailed coverage of football-specific slang, idiomaticity, and cultural references, it also prevents broader generalization. The results should therefore be seen as exploratory rather than definitive.

A second limitation lies in annotation itself. Even with relatively strong inter-rater agreement, error tagging remains partly subjective, especially for culturally loaded or offensive expressions typical of football discourse. Broader annotator pools or cross-institutional collaborations could help increase consistency and validity.

Metric-specific challenges also emerged. As Zouhar et al. (2024:10) remind us, COMET is a statistical learning model that inherits the biases of its training data and risks of misconfiguration. The lack of targeted training for Turkish in UGC contexts may partly explain the unstable system–segment correlation observed here. In addition, the cluster of “0.000” RQS values (13 in EN-TR, 7 in TR-EN) suggests a floor effect, where extreme low scores reduce variability and obscure meaningful differences. Both training bias and scale effects must be considered when interpreting COMET correlations.

Building on these limitations, several avenues for future work emerge. Expanding the dataset to include a wider range of user-generated genres and larger volumes of data would allow for more robust statistical validation. Diversifying the pool of annotators across institutions could mitigate subjectivity and strengthen reliability. Incorporating hybrid or fine-tuned metrics tailored to Turkish social media discourse may reduce directional asymmetries observed in this study. Most importantly, extending evaluation frameworks to integrate sentiment and stance preservation would address one of the most pressing challenges in translating informal, emotionally charged content.

Despite these constraints, the study offers several implications. These implications of extend beyond the immediate findings. For theoretical and methodological debates in translation studies, the study demonstrates the feasibility and value of combining structured human evaluation with neural metrics. This dual approach bridges the gap between interpretability and scalability, suggesting a productive way forward for future evaluation research. It also draws attention to directionality as a central variable in MT evaluation,

reinforcing the idea that system performance cannot be assumed to be symmetrical across language pairs.

For educational contexts, the thesis provides a model that can be directly incorporated into translation training programs. By integrating MQM annotation and metric comparison into classroom practice, educators can help students critically engage with authentic, informal texts, better preparing them to handle the complexities of digital communication. Moreover, the inter-rater agreement process illustrated here demonstrates how evaluator training itself can become an instructive pedagogical tool, fostering methodological awareness and critical literacy during the evaluation.

For industry and institutional practice, the findings underline the potential of embedding reference-free metrics into translation workflows for social media and other informal domains. Automatic scores, when complemented by human oversight, can support quality monitoring, risk detection, and system development in real time. Particularly in sectors where reputation and immediacy matter—such as sports media, broadcasting, and customer engagement—such frameworks can provide actionable insight into translation quality while ensuring that context-sensitive errors do not go unnoticed. Therefore, by considering the fact that a minor translation problem can cause big issues in football industry, it is quite important embed reference-free metrics into social media translation platforms, where real-time quality scores could inform users and increase transparency. Furthermore, this practice could be benefited not only by sports or football industry, but also by other fields.

Finally, the study opens a pathway toward sentiment-aware translation evaluation. Football tweets are emotionally charged, and the preservation of tone and polarity is often as crucial as lexical accuracy. By exposing how current systems struggle with the affective dimensions of language, this research lays the groundwork for integrating sentiment analysis into translation quality assessment, ensuring that future evaluations address not only linguistic form but also pragmatic and emotional fidelity.

In conclusion, this thesis contributes to translation studies and MT evaluation by demonstrating the strengths and weaknesses of combining MQM with COMETKiwi-XL on

football-related user-generated text. It shows how structured human annotation and reference-free metrics can complement one another. By foregrounding the challenges of noisy, affective discourse, it highlights not only where translation systems fall short but also how evaluation itself might be considered in the future. Ultimately, with its limited data, the study suggests the use of hybrid evaluation practices that bring together human expertise and automatic metrics to advance translation quality in the rapidly evolving landscape of digital communication.



## REFERENCES

- Agrawal, Swapnil, André Farinhas, Rico Rei, and António F. T. Martins. "Can Automatic Metrics Assess High-Quality Translations?" *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* 1 (2024): 10798-10817. <https://doi.org/10.18653/v1/2023.acl-long.615>.
- Alhadlaq, Abdullah, and Ali Alnuaim. "A Twitter-Based Comparative Analysis of Emotions and Sentiments of Arab and Hispanic Football Fans." *Applied Sciences* 13, no. 11 (2023): 6729. <https://doi.org/10.3390/app13116729>.
- Alsubhi, Waleed O. "Translation of Social Media Posts: Saudi Football Clubs' Social Media Posts as Case Study." *International Journal of Religion* 5, no. 11 (2024): 5441-5457. <https://doi.org/10.61707/5anv6v28>.
- Alves, Fábio, João Luís Gonçalves, and Raquel de Sousa. "Assessing Translation Quality: Ideas for a General Framework." In *Interfaces in Translation*, edited by Fábio Alves, 103-25. Amsterdam: John Benjamins, 2006. <https://doi.org/10.1075/btl.63.08alv>.
- Andriani, Dwi, and Dian Setyaningsih. "Comparative Performance of Google Translate vs Naver Papago on Non-Standard Word Translation." ResearchGate preprint, 2024. <https://www.researchgate.net>.
- Angelone, Erik, Maureen Ehrensberger-Dow, and Gary Massey, eds. *The Bloomsbury Companion to Language-Industry Studies*. London: Bloomsbury, 2020.
- Anugraha, Daniel, Genta Kuwanto, Leonardi Susanto, Dominikus T. Wijaya, and Genta Indra Winata. "METAMETRICS-MT: Tuning Meta-Metrics for Machine Translation via Human Preference Calibration." *arXiv* preprint, 2024. <https://arxiv.org/abs/2411.00390>.
- Artetxe, Mikel, Gorka Labaka, and Kepa Sarasola. "Building Hybrid Machine Translation Systems by Using an EBMT Pre-Processor to Create Partial Translations." In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, 11-18. 2015.
- Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate." *arXiv* preprint, 2015. <https://arxiv.org/abs/1409.0473>.

- Baines, Rebecca. "Translating Tweets in the Soccer Industry: Identity Management and Visibility in a Global Game." *Papers in Language and Communication Studies* 2, no. 2 (2019): 21-45.
- Backlinko Team. "ChatGPT / OpenAI Statistics: How Many People Use ChatGPT?" Backlinko (blog), 22 October 2024. <https://backlinko.com/chatgpt-stats>.
- Blatz, John, Erin Fitzgerald, George Foster, Sylvain Gandrabur, Cyril Goutte, Amanda Kulesza, Adrià Sanchis, and Nicola Ueffing. "Confidence Estimation for Machine Translation." In *Proceedings of the 20th International Conference on Computational Linguistics*, 315-21. 2004. <https://aclanthology.org/C04-1045>.
- Brown, Peter E., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. "The Mathematics of Statistical Machine Translation: Parameter Estimation." *Computational Linguistics* 19, no. 2 (1993): 263-311. <https://aclanthology.org/J93-2003/>.
- Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, et al. "Language Models Are Few-Shot Learners." *arXiv preprint*, 2020. <https://arxiv.org/abs/2005.14165>.
- Burchardt, Aljoscha, Arle Lommel, and Verena Macketanz. "A New Deal for Translation Quality." *Universal Access in the Information Society* 20 (2020): 701-715. <https://doi.org/10.1007/s10209-020-00736-5>.
- Callison-Burch, Chris, Miles Osborne, and Philipp Koehn. "Re-Evaluating the Role of BLEU in Machine Translation Research." In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 249-56. 2006. <https://aclanthology.org/E06-1032/>.
- Cambridge University Press. "Machine Translation." *Cambridge Dictionary*. Accessed 6 July 2025. <https://dictionary.cambridge.org/dictionary/english/machine-translation>.
- Cambridge University Press. "Social Media." *Cambridge Dictionary*, 2025. <https://dictionary.cambridge.org/dictionary/english/social-media>.
- Candel-Mora, Miguel Ángel. "Fine-Tuning Machine Translation Quality-Rating Scales for New Digital Genres: The Case of User-Generated Content." *ELUA – Estudios de Lingüística Aplicada* 38 (2022): 117-136. <https://doi.org/10.14198/ELUA.21900>.
- Carl, Michael, and Andy Way. *Recent Advances in Example-Based Machine Translation*. Dordrecht: Springer, 2003.
- Carrasqueira, Diogo. "Twitter PWA Replaces Bing with Google Translate." *Neowin*, 21 February 2019. <https://www.neowin.net>.

- Caswell, Isaac. "Google Translate Learns 24 New Languages." Google Blog, 11 May 2022. <https://blog.google/products/translate/24-new-languages/>.
- Chandra, Rahul, Abhijit Kumar, and Dipanjan Mukherjee. "Evaluating LLMs on Classical Texts: A Comparative Study of GPT-4o, GPT-3.5, and Google Translate on Bhagavad Gita Translations." In *Proceedings of the 2025 Conference on Computational Linguistics and Ancient Texts*, 88-102. 2025.
- Coldewey, Devin, and Frederic Lardinois. "DeepL Schools Other Online Translators with Clever Machine Learning." *TechCrunch*, 29 August 2017. <https://techcrunch.com/2017/08/29/deepl-schools-other-online-translators-with-clever-machine-learning/>.
- Çetin, H., and Duran, H. 2024. *A Comparative Analysis of the Performances of ChatGPT, DeepL, Google Translate and a Human Translator in Community-Based Settings*. *Journal of Language and Linguistic Studies* 20 (1): 127–141.
- Descartes. *Discourse on Method: Part One*. Translated by John Cottingham, Robert Stoothoff, Dugald Murdoch, and Anthony Kenny. New York: Cambridge University Press, 1991 (orig. 1637).
- DeepL. "Discover DeepL Translator." Accessed 6 July 2025. <https://www.deepl.com/en/products/translator>.
- DeepL Team. "How Does DeepL Work?" DeepL Blog, 1 November 2021. <https://www.deepl.com/en/blog/how-does-deepl-work>.
- DeepL Team. "The Numbers Are In: Translation Experts Weigh In on Translation Expertise." DeepL Blog, 7 May 2024. <https://www.deepl.com/en/blog/alc-survey-results>.
- DeepL Team. "DeepL's Next-Gen LLM Outperforms ChatGPT-4, Google, and Microsoft for Translation Quality." DeepL Blog, 19 November 2024. <https://www.deepl.com/en/blog/next-gen-llm-translation-quality>.
- Dhariya, Jatin, Shreya Malviya, and U. S. Tiwary. "A Novel Hybrid Machine Translation System Guided by Rule-Based Translation." *Procedia Computer Science* 122 (2017): 442-449. <https://doi.org/10.1016/j.procs.2017.11.392>.
- Ding, Ning, Zhilin Yang, Shuozhi Li, and Aixin Zhang. "Where Translation Impacts: The Non-Professional Community on Chinese Online Social Media—A Descriptive Case Study on the User-Generated Translation Activity of Bilibili Content Creators." *Global*

- Media and China* 6, no. 2 (2021): 171-190. <https://doi.org/10.1177/20594364211000645>.
- Eisenstein, Jacob. "What to Do about Bad Language on the Internet." In *Proceedings of the First Workshop on Noisy User-Generated Text*, 1-9. 2013. <https://aclanthology.org/N13-1037/>.
- Fakih, Ahmad, Mahdi Ghassemiazghandi, Abdul-Hadi Fakih, and Mohinder K. M. Singh. "Evaluation of Instagram's Neural Machine Translation for Literary Texts: An MQM-Based Analysis." *GEMA Online Journal of Language Studies* 24, no. 1 (2024): 214-233. <https://doi.org/10.17576/gema-2024-2401-13>.
- Farghal, Muhammad, and Ahmed S. Haider. "Translating Classical Arabic Verse: Human Translation vs AI Large Language Models (Gemini and ChatGPT)." *Cogent Social Sciences* 10, no. 1 (2024): 2410998. <https://doi.org/10.1080/23311886.2024.2410998>.
- FIFA. *FIFA World Cup Qatar 2022™: Global Engagement and Audience Report (Executive Summary)*. Accessed 16 May 2025. <https://inside.fifa.com/tournament-organisation/audience-reports/qatar-2022>.
- Filo, Kevin, Daniel Lock, and Andrew Karg. "Sport and Social Media Research: A Review." *Sport Management Review* 18, no. 2 (2015): 166-181. <https://doi.org/10.1016/j.smr.2014.11.001>.
- Fonseca, Erick, Liudmila Yankovskaya, António F. T. Martins, Maarit Fishel, and Chris Federmann. "Findings of the WMT 2019 Shared Tasks on Quality Estimation." In *Proceedings of the Fourth Conference on Machine Translation* 3: 1-10. 2019. <https://aclanthology.org/W19-5401>.
- Freitag, Markus, Jonathan Albrecht, Rachel Bawden, Marta R. Costa-jussà, Ondřej Bojar, et al. 2021a. "Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation." *Transactions of the Association for Computational Linguistics* 9 (2021): 1460-1474. <https://aclanthology.org/2021.tacl-1.87>.
- Freitag, Markus, Rico Rei, Nitika Mathur, Chun-Kai Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. "Results of the WMT 21 Metrics Shared Task: Evaluating Metrics with Expert-Based Human Evaluations on TED and News Domain." In *Proceedings of the Sixth Conference on Machine Translation*, 733-747. 2021. <https://aclanthology.org/2021.wmt-1.85>.
- Freitag, Markus, Yonatan Al-Onaizan, and Chinnadhurai Sankar. "Stop Using BLEU: Towards Better Automatic Evaluation Metrics for Machine Translation." *Proceedings of the 2022 Conference of the North American Chapter of the ACL: Human Language Technologies* (2022): 4823-4835. <https://doi.org/10.18653/v1/2022.naacl-main.354>.

- Freitag, Markus, Felipe Alva-Manchego, Enrique Alfonseca, Isaac Caswell, António C. Farinha, Maria Fomicheva, et al. "Findings of the WMT 2023 Shared Task on Machine Translation Metrics." In *Proceedings of the Eighth Conference on Machine Translation*, 911-957. 2023. <https://www.statmt.org/wmt23/pdf/2023.wmt-1.51.pdf>.
- Freitag, Markus, Nitika Mathur, Daniel Deutsch, Chun-Kai Lo, Eleftherios Avramidis, Rico Rei, Alon Lavie, et al. "Are LLMs Breaking MT Metrics? Results of the WMT 24 Metrics Shared Task." In *Proceedings of the Ninth Conference on Machine Translation*, 47-81. 2024. <https://doi.org/10.18653/v1/2024.wmt-1.2>.
- Gao, Ruoyu, Yitao Lin, Ning Zhao, Xiaoya Zhang, et al. "Machine Translation of Chinese Classical Poetry: A Comparison among ChatGPT, Google Translate, and DeepL Translator." *Humanities and Social Sciences Communications* 11 (2024): 835. <https://doi.org/10.1057/s41599-024-03363-0>.
- Gotti, Fabien, Philippe Langlais, and Atefeh Farzindar. "Hashtag Occurrences, Layout and Translation: A Corpus-Driven Analysis of Tweets Published by the Canadian Government." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 1000-1007. 2014.
- Guerreiro, Nuno M., Rico Rei, Dirk van Stigt, Luísa Coheur, Paul Colombo, and António F. T. Martins. "xCOMET: Transparent Machine Translation Evaluation through Fine-Grained Error Detection." *Transactions of the Association for Computational Linguistics* 11 (2023): 110-130.
- Han, An-Liang F., Daniel F. Wong, and Lidia S. Chao. "LEPOR: A Robust Evaluation Metric for Machine Translation with Augmented Factors." In *Proceedings of the 24th International Conference on Computational Linguistics*, 441-450. 2012. <https://aclanthology.org/C12-2044/>.
- Han, An-Liang F., Daniel F. Wong, Lidia S. Chao, Yi Lu, Lihua He, Yue Wang, and Jingyi Zhou. "A Description of Tunable Machine Translation Evaluation Systems in WMT 13 Metrics Task." In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, 414-421. 2013. <https://aclanthology.org/W13-2246/>.
- Hendy, Ahmed, Mahmoud Abdelrehim, Ahmed Sharaf, Vishrav Raunak, Mohamed Gabr, Hiroaki Matsushita, et al. "How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation." *arXiv preprint*, 2023. <https://arxiv.org/abs/2302.09210>.
- Huang, Han, Sen Wu, Kai Chen, Xiaohua Liang, Hui Di, Min Yang, and Tong Zhao. "Multi-View Fusion for Universal Translation Quality Estimation." *Neurocomputing* 531 (2023): 126292. <https://doi.org/10.1016/j.neucom.2023.126292>.

- Hutchins, John. "ALPAC: The (In)Famous Report." *MT News International*, no. 14 (June 1996): 9-12.
- Hutchins, John. "Example-Based Machine Translation: A Review and Commentary." *Machine Translation* 19, nos. 3-4 (2005): 197-211. <https://doi.org/10.1007/s10590-006-9003-9>.
- Hutchins, John. "Latest Developments in Machine Translation Technology: Beginning a New Era in MT Research." In *Proceedings of the MT Summit IV*, 11-34. 1993.
- Hutchins, W. John, and Harold L. Somers. *An Introduction to Machine Translation*. London: Academic Press, 1992.
- Hutchins, W. John. "Machine Translation: General Overview." In *The Oxford Handbook of Computational Linguistics*, edited by Ruslan Mitkov, 883-900. Oxford: Oxford University Press, 2003.
- Imran, Muhammad, Prasenjit Mitra, and Carlos Castillo. "Twitter as a Lifeline: Human-Annotated Twitter Corpora for NLP of Crisis-Related Messages." *arXiv* preprint, 2016. <https://arxiv.org/abs/1605.05894>.
- Izazi, Zulkifli Z., and Tengku Mahadi Tengku-Sepora. "Reasons to Code-Switch: A Case Study of Malaysian Twitter Users." *Rangsit Journal of Social Sciences and Humanities* 7, no. 2 (2020): 161-176.
- Jiang, Ziye, Qiang Lv, Zhenguo Zhang, and Liang Lei. "Convergences and Divergences between Automatic Assessment and Human Evaluation: Insights from Comparing ChatGPT-Generated Translation and Neural Machine Translation." *arXiv* preprint, 2024.
- Johnson, Melvin, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, et al. "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation." *Transactions of the Association for Computational Linguistics* 5 (2017): 339-351. <https://aclanthology.org/Q17-1024/>.
- Jolley, Joshua R., and Leigh M. Maimone. "Machine Translation: A Concise Overview." In *The Routledge Handbook of Translation and Technology*, 27-42. London: Routledge, 2022. <https://doi.org/10.4324/9780429286803-3>.
- Kaya, Sefa, Mehmet Argan, and Mehmet T. Argan. "Like, Share or Comment 'Till We Die: Analyzing Football Fans' Engagements, Identity and Loyalty on Social Media." *Sport i Turystyka. Środkowoeuropejskie Czasopismo Naukowe* 7, no. 1 (2024): 69-91.
- Kepler, Fabio, Josselin Trénous, Marco Treviso, Marta Vera, and António F. T. Martins. "OpenKiwi: An Open-Source Framework for Quality Estimation." In *Proceedings of*

- the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 117-122. 2019. <https://aclanthology.org/P19-3020/>.
- Kemp, Simon. *Digital 2024 April Global Statshot Report*. DataReportal, 2024. <https://datareportal.com/reports/digital-2024-april-global-statshot>.
- Kemp, Simon. *Digital 2025: Turkey*. DataReportal, 2025. <https://datareportal.com/reports/digital-2025-turkey>.
- Kenny, Dorothy. "Machine Translation." In *Routledge Encyclopaedia of Translation Studies*, 3rd ed., 305-310. London: Routledge, 2020.
- Klubička, Filip, Antonio Toral, and Víctor M. Sánchez-Cartagena. 2018. "Quantitative Fine-Grained Human Evaluation of Machine Translation Systems: A Case Study on English to Croatian." *Machine Translation* 32 (5): 507–532. <https://doi.org/10.1007/s10590-018-9220-1>.
- Kocmi, Tomáš, Brian Thompson, Antonio Raganato, Rui Wang, Kai Sun, Loïc Barrault, et al. "Findings of the WMT 24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet." *arXiv preprint*, 2024. <https://arxiv.org/abs/2406.11559>.
- Koehn, Philipp, and Christof Monz. "Manual and Automatic Evaluation of Machine Translation between European Languages." In *Proceedings of the Workshop on Statistical Machine Translation*, 102-121. 2006. <https://aclanthology.org/W06-3119/>.
- Koo, Terry K., and Mae Y. Li. 2016. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine* 15 (2): 155–63.
- Landis, J. Richard, and Gary G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33 (1): 159–74.
- Lavie, Alon, and Michael J. Denkowski. "The METEOR Metric for Automatic Evaluation of Machine Translation." *Machine Translation* 23 (2009): 105-115. <https://doi.org/10.1007/s10590-009-9059-4>.
- Lee, Seungjun, Jungseob Lee, Hyeonseok Moon, Chanjun Park, Jaehyung Seo, Sugyeong Eo, Seonmin Koo, and Heuseok Lim. 2023. "A Survey on Evaluation Metrics for Machine Translation." *Mathematics* 11 (2): 1006. <https://doi.org/10.3390/math11041006>
- Licht, Daniel, Chen Gao, Justin Lam, Francisco Guzmán, Mona Diab, and Philipp Koehn. "Consistent Human Evaluation of Machine Translation across Language Pairs." In

- Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas*, 309-316. 2022. <https://aclanthology.org/2022.amta-research.24/>.
- Lipschultz, Jeremy Harris. *Social Media Communication: Concepts, Practices, Data, Law and Ethics*. 2nd ed. New York: Routledge, 2018.
- Liu, Siqi, and Wenbo Zhu. “An Analysis of the Evaluation of the Translation Quality of Neural Machine Translation Application Systems.” *Applied Artificial Intelligence* 37, no. 1 (2023): 1-18. <https://doi.org/10.1080/08839514.2023.2214460>.
- Lohar, Praful. *Machine Translation of User-Generated Content*. Master’s thesis, Dublin City University, 2020. <https://doras.dcu.ie/24988/>.
- Lommel, Arle R., Serge Gladkoff, Alan Melby, Sharon Wright, Ivana Strandvik, Katarzyna Gasova, et al. “The Multi-Range Theory of Translation Quality Measurement: MQM Scoring Models and Statistical Quality Control.” *arXiv preprint*, 2024. <https://doi.org/10.48550/arXiv.2405.16969>.
- Lommel, Arle R., Serge Gladkoff, Alan Melby, Sharon Wright, Ivana Strandvik, Katarzyna Gasova, et al. *Multidimensional Quality Metrics 2.0 Scoring Guide: Linear and Non-Linear Models*. *arXiv preprint*, 2024. <https://arxiv.org/pdf/2405.16969.pdf>.
- Lommel, Arle R., Aljoscha Burchardt, and Maja Popović. “Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics.” In *Proceedings of LREC 2014*. 2014. [https://www.dfki.de/fileadmin/user\\_upload/import/7445\\_LREC-Lommel-Burchardt-Popovic.pdf](https://www.dfki.de/fileadmin/user_upload/import/7445_LREC-Lommel-Burchardt-Popovic.pdf).
- Lommel, Arle. “The MQM Core Typology.” *The MQM Framework*. Accessed 11 July 2025. <https://themqm.org/the-mqm-typology/>.
- Méndez Guzman, Eduardo, Zhewei Zhang, and Waqar Ahmed. *Towards Understanding a Football Club’s Social Media Network: An Exploratory Case Study of Manchester United*. Report, n.d.
- McNamee, Paul, and Kevin Duh. “The Multilingual Microblog Translation Corpus: Improving and Evaluating Translation of User-Generated Text.” In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 910-918. 2022. <https://aclanthology.org/2022.lrec-1.96/>.
- Michel, Paul, and Graham Neubig. “MTNT: A Testbed for Machine Translation of Noisy Text.” In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 543-553. 2018. <https://doi.org/10.18653/v1/D18-1050>.

- Moghe, Naman, Arash Fazla, Christopher Amrhein, Tomáš Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Lucia Guillou. “Machine Translation Meta-Evaluation through Translation Accuracy Challenge Sets.” In *Proceedings of the Seventh Conference on Machine Translation*, 3-18. 2022. <https://aclanthology.org/2022.wmt-1.2>.
- Moreau, Nicolas, Maxime Roy, Aimee Wilson, and Laurence Atlani-Duault. ““Life Is More Important Than Football”: Comparative Analysis of Tweets and Facebook Comments Regarding the Cancellation of the 2015 African Cup of Nations in Morocco.” *International Review for the Sociology of Sport* 56, no. 2 (2021): 252-275. <https://doi.org/10.1177/1012690219899618>.
- Michalos, Alex C., ed. 2014. *Encyclopedia of Quality of Life and Well-Being Research*. Dordrecht: Springer.
- Nagao, Makoto. “A Framework of a Mechanical Translation between Japanese and English by Analogy Principle.” In *Artificial and Human Intelligence*, edited by Amit Elithorn and R. Banerji, 173-180. Amsterdam: North-Holland, 1984.
- Nagarajan, Mythili, and Michael Gamon. “Introduction.” In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, ii-iii. 2011.
- Neagu, Delia-Camélia, Ana-Bogdana Rus, Marius Grec, Mircea-Adrian Boroianu, Nicolae Bogdan, and Adrian Gal. “Towards Sentiment Analysis for Romanian Twitter Content.” *Algorithms* 15, no. 10 (2022): 357. <https://doi.org/10.3390/a15100357>.
- Neves, Mariana, Antonio Jimeno Yepes, and Aurélie Névéol. “Findings of the WMT 2023 Biomedical Translation Shared Task: Evaluation of ChatGPT 3.5 as a Comparison System.” In *Proceedings of WMT 2023*, 1-12. 2023. <https://aclanthology.org/2023.wmt-1.2.pdf>.
- Nirenburg, Sergei. “Historical Introduction.” In *Readings in Machine Translation*, edited by Sergei Nirenburg, Harold L. Somers, and Yorick Wilks, 3-11. Cambridge, MA: MIT Press, 2003.
- OpenAI. “GPT-3.5-Turbo and the System-Message API.” OpenAI Blog, 1 March 2023. <https://openai.com/index/gpt-3-5-turbo-fine-tuning-and-api-updates/>.
- OpenAI. “Introducing OpenAI o3 and o4-mini.” OpenAI Blog, 16 April 2025. <https://openai.com/index/introducing-o3-and-o4-mini/>.

- Özevin, Ozan. “Investigation of Social Media Usage of Companies by Network, Industry and Financial Performance: BIST Application.” *FESA Dergisi* 6, no. 3 (2021): 408-420. <https://doi.org/10.29106/fesa.916015>.
- Page, Ruth, David Barton, Johann W. Unger, and Michele Zappavigna. *Researching Language and Social Media: A Student Guide*. London: Routledge, 2014. <https://doi.org/10.4324/9781315771786>.
- Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. “BLEU: A Method for Automatic Evaluation of Machine Translation.” In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311-318. 2002. <https://doi.org/10.3115/1073083.1073135>.
- Perez, Sarah. “Google Translate Adds Support for 24 New Languages, Including Its First Indigenous American Languages.” *TechCrunch*, 11 May 2022. <https://techcrunch.com>.
- Phrase. “Exploring DeepL for Machine Translation: How It Works, and How Accurate It Is.” Phrase Blog, 2025. <https://phrase.com/blog/posts/deepl/>.
- Pla, Ferran, and Luis-Fernando Hurtado. “Language Identification of Multilingual Posts from Twitter: A Case Study.” *Knowledge and Information Systems* 51, no. 3 (2017): 965-989. <https://doi.org/10.1007/s10115-016-0997-x>.
- Poibeau, Thierry. *Machine Translation*. Cambridge, MA: MIT Press, 2017. <https://doi.org/10.7551/mitpress/9780262037551.001.0001>.
- Price, Jon, Neil Farrington, and Lee Hall. “Changing the Game? The Impact of Twitter on Relationships between Football Clubs, Supporters and the Sports Media.” *Soccer & Society* 14, no. 4 (2013): 446-461. <https://doi.org/10.1080/14660970.2013.810431>.
- Quah, Christine K. *Translation and Technology*. New York: Palgrave Macmillan, 2006.
- Raunak, Vishrav, Arul Menezes, Matt Post, and Hany Hassan Awadalla. “Do GPTs Produce Less Literal Translations?” Microsoft Azure AI Technical Report, 2023. <https://arxiv.org/pdf/2305.16806>.
- Raybaud, Sylvain, Chloé Lavecchia, Daniel Langlois, and Kamel Smaïli. “New Confidence Measures for Statistical Machine Translation.” *arXiv* preprint, 2009. <https://arxiv.org/pdf/0902.1033>.
- Rei, Rico, Nuno M. Guerreiro, João Pombal, Dirk van Stigt, Marco Treviso, Luísa Coheur, João Graça de Souza, and António F. T. Martins. “Scaling Up COMETKiwi: Unbabel IST 2023 Submission for the Quality Estimation Shared Task.” In *Proceedings of the*

- Eighth Conference on Machine Translation*, 841-848. 2023. <https://doi.org/10.18653/v1/2023.wmt-1.73>.
- Rei, Rico, Craig Stewart, António C. Farinha, and Alon Lavie. “COMET: A Neural Framework for MT Evaluation.” In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2685-2702. 2020. <https://doi.org/10.18653/v1/2020.emnlp-ain.213>.
- Rikters, Markus. “Hybrid Machine Translation by Combining Output from Multiple Machine Translation Systems.” *Baltic Journal of Modern Computing* 7, no. 3 (2019): 301-341. <https://doi.org/10.22364/bjmc.2019.7.3.01>.
- Romero-Jara, Eduardo, Ferran Solanellas, Jaume Muñoz, and Susana López-Carril. “Connecting with Fans in the Digital Age: An Exploratory and Comparative Analysis of Social-Media Management in Top Football Clubs.” *Managing Sport and Leisure* 28, nos. 1-2 (2023): 1-21. <https://doi.org/10.1080/23750472.2022.2050807>.
- Romero-Jara, Eduardo, Ferran Solanellas, Susana López-Carril, Dimitrios Kolyperas, and Christos Anagnostopoulos. “The More We Post, the Better? A Comparative Analysis of Fan Engagement on Social Media Profiles of Football Leagues.” *International Journal of Sports Marketing and Sponsorship* 23, no. 4 (2024): 695-714. <https://doi.org/10.1108/IJSMS-06-2021-0134>.
- Rosales Núñez, José Carlos, Djame Seddah, and Guillaume Wisniewski. “Comparison between NMT and PBSMT Performance for Translating Noisy User-Generated Content.” In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, 2-14. 2019. <https://aclanthology.org/W19-6101/>.
- Rosales Núñez, José Carlos, Djame Seddah, and Guillaume Wisniewski. “Noisy UGC Translation at the Character Level: Revisiting Open-Vocabulary Capabilities and Robustness of Char-Based Models.” In *Proceedings of the Seventh Workshop on Noisy User-Generated Text*, 199-211. 2021. <https://doi.org/10.18653/v1/2021.wnut-1.23>.
- Rosales Núñez, José Carlos, Djame Seddah, and Guillaume Wisniewski. “Understanding the Impact of UGC Specificities on Translation Quality.” *arXiv preprint*, 2021. <https://doi.org/10.48550/arXiv.2110.12551>.
- Saadany, Hoda, Constantin Orăsan, Rocío Caro Quintana, Fernanda do Carmo, and Luciana Zilio. 2021a. “Challenges in Translation of Emotions in Multilingual User Generated Content: Twitter as a Case Study.” *arXiv preprint*. <https://doi.org/10.48550/arXiv.2106.10719>.

- Saadany, Ahmed, Jong-Ho Kim, and Youn Park. 2021b. "User Enabled Click 'Translate Tweet' Feature Powered by Google Translate: A Case Study." *arXiv preprint*. <https://arxiv.org>.
- Saadany, Hoda, Constantin Orăsan, Rocío Caro Quintana, Fernanda do Carmo, and Luciana Zilio. "Analysing Mistranslation of Emotions in Multilingual Tweets by Online MT Tools." In *Proceedings of the Eighth Conference on Machine Translation (WMT 2023)*, 1045-1058. 2023. <https://doi.org/10.18653/v1/2023.wmt-1.93>.
- Schwartz, Barry. "Twitter Partners with Bing to Offer Tweet Translation in over 40 Languages." *SERoundTable*, 2015. <https://www.seroundtable.com>.
- Schwartz, Lynne. "The History and Promise of Machine Translation." In *Innovation and Expansion in Translation Process Research*, edited by Isabel Lacruz and Riitta Jääskeläinen, 161-190. Amsterdam: John Benjamins, 2018. <https://doi.org/10.1075/ata.18.08sch>.
- Taras, David. "How Digital Media Has Changed Elections: An Introduction." In *Electoral Campaigns, Media, and the New World of Digital Politics*, edited by David Taras and Richard Davis, 1-17. Ann Arbor: University of Michigan Press, 2022. <https://doi.org/10.3998/mpub.12439646>.
- Team Latestly. "X Replaces Google Translate with Grok Translation on Web Version." *Latestly*, 12 June 2025. <https://www.latestly.com>.
- Tuan, Yi-Ling, Ahmed El-Kishky, Akshay Renduchintala, Vaibhav Chaudhary, Francisco Guzmán, and Lucia Specia. "Quality Estimation without Human-Labeled Data." In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10154-10164. 2021. <https://doi.org/10.18653/v1/2021.emnlp-main.793>.
- Turovsky, Barak. "Ten Years of Google Translate." Google Blog, 11 May 2016. <https://www.blog.google/products/translate/ten-years-of-google-translate/>.
- Twitter. "New Features on Twitter for Windows Phone 3.0." Twitter Blog, 2 July 2013. [https://blog.x.com/official/en\\_us/a/2013/new-features-on-twitter-for-windows-phone-30.html](https://blog.x.com/official/en_us/a/2013/new-features-on-twitter-for-windows-phone-30.html).
- Udin, Ebube. "What Is Happening to Twitter's Translate Feature?" *Gizchina*, 2 June 2023. <https://www.gizchina.com/2023/06/02/what-is-happening-to-twitters-translate-feature/>.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need." *Advances*

- in *Neural Information Processing Systems* 30 (2017): 5998-6008. <https://arxiv.org/abs/1706.03762>.
- Vilar, David, Melvin Johnson, Arvind Parmar, Aakanksha Chowdhery, Sharan Narang, and Marc-Aurelio Ranzato. "The Utility of Translationese in Large Language Models for Machine Translation." In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 15001-15018. 2023. <https://aclanthology.org/2023.emnlp-main.933>.
- Wang, Hao, Jianguo Zhang, and Qun Liu. "Progress in Machine Translation." *Engineering* 8, no. 11 (2022): 1477-1490. <https://doi.org/10.1016/j.eng.2022.07.017>.
- Way, Andy. "A Hybrid Approach to Machine Translation." In *Computational Linguistics and Intelligent Text Processing*, vol. 4919, 527-538. Berlin: Springer, 2008. [https://doi.org/10.1007/978-3-540-78135-6\\_49](https://doi.org/10.1007/978-3-540-78135-6_49).
- Way, Andy. "Machine Translation: Where Are We at Today?" In *The Bloomsbury Companion to Language-Industry Studies*, 285-309. London: Bloomsbury, 2020.
- Woodward, Michael. "Social Media User Statistics: How Many People Use Social Media?" *Search Logistics*, 2025. <https://www.searchlogistics.com/learn/statistics/social-media-user-statistics/>.
- Wolk, Krzysztof, and Krzysztof Marasek. "Survey on Neural Machine Translation into Polish." In *Multimedia and Network Information Systems*, edited by Krzysztof Choroś, Maciej Kopel, Elżbieta Kukla, and Andrzej Siemiński, 260-272. Cham: Springer, 2019. [https://doi.org/10.1007/978-3-319-98678-4\\_27](https://doi.org/10.1007/978-3-319-98678-4_27).
- Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, et al. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." *arXiv preprint*, 2016. <https://arxiv.org/abs/1609.08144>.
- Wunderlich, Florian, and Daniel Memmert. "A Big-Data Analysis of Twitter Data during Premier League Matches: Do Tweets Contain Information Valuable for In-Play Forecasting of Goals in Football?" *Social Network Analysis and Mining* 12 (2022): 23. <https://doi.org/10.1007/s13278-021-00842-z>.
- Xu, Haoran, Yoon-Jeong Kim, Ahmed Sharaf, and Hany Hassan Awadalla. "A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models." *arXiv preprint*, 2023. <https://arxiv.org/abs/2309.11674>.
- Yan, Jiayi, Peng Yan, Yujia Chen, Jiacheng Li, Xiaotian Zhu, and Yichao Zhang. "GPT-4 vs. Human Translators: A Comprehensive Evaluation of Translation Quality across

Languages, Domains, and Expertise Levels.” *arXiv* preprint, 2024.  
<https://doi.org/10.48550/arXiv.2407.03658>.

Zhou, Long, Shaohan Huang, Zhaopeng He, and Hua Wu. “A Neural Combination Model for Machine Translation.” In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 548-557. 2017.  
<https://doi.org/10.18653/v1/P17-1051>.

Zouhar, Vilém, Pinzhen Chen, Tsz Kin Lam, Ninad Moghe, and Barry Haddow. 2024. “Pitfalls and Outlooks in Using COMET.” *arXiv* (version 3), posted September 30, 2024. <https://doi.org/10.48550/arXiv.2408.15366>.

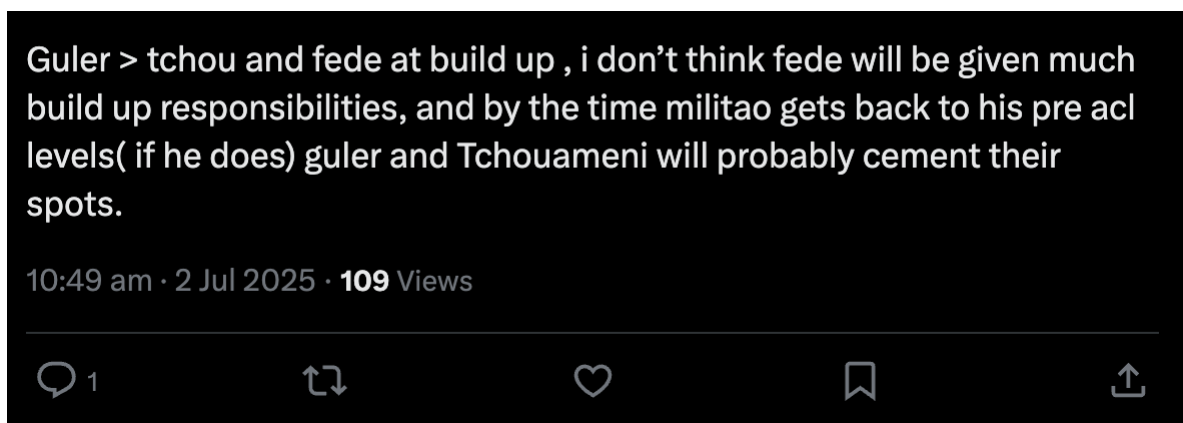
## APPENDICES

### APPENDIX A – ENGLISH TWEETS

#### Tweet 1



#### Tweet 2



### Tweet 3

The idiot just denied me 50odds game stake with 1K  
Last game to be precise  
[REDACTED] is a stup!d coach..  
Once he's winning  
He will start playing defensive game..  
Playing against a team that just got promoted to the league..  
he's just overrated I swear

12:07 am · 18 Aug 2024 · **111** Views



### Tweet 4

Na man i think MSN was an atg front 3 all balling out  
To win a league title against that guardiola side with centurions and 46  
league goals is mad  
He didnt win euros but if we compare season for season i acc think that  
the major tournaments shd heavily weigh in  
Add 10% bias too

9:53 pm · 2 Jul 2025 · **51** Views



### Tweet 5

well, they called him delusional when he said that... It all coming from [redacted] and their stupidity and insecurities. I don't want to hear [redacted] and the [redacted] in the same sentence ever! 🙄

12:25 pm · 1 Jul 2025 · 65 Views



### Tweet 6

Fantastic goalkeeping but equally Savio should be converting this. Football IQ sometimes is what makes players like Messi exceptional. At this point the GK had higher intelligence and anticipation cos naturally every player will play it like Savio. But again he should score this

2:42 pm · 1 Jul 2025 · 17.8K Views



#### Tweet 7

Have you been living under a rock, Wdym 'Unless'. It's wide known that we're targeting a ST and it's between Ekitike, Osimhen and Isak. Not only Isak. We are 100% selling Nunez and buying a ST. If we sell Diaz or Jota we will buy another attacker. Wirtz is playing the no.10. Stop

4:05 pm · 2 Jul 2025 · **284** Views



#### Tweet 8

Ball hits defender's body, from there it hits his hand.. referee checks VAR and gives penalty decision, gifting YB the win... BRAVO! Then don't frigg'n use VAR at all

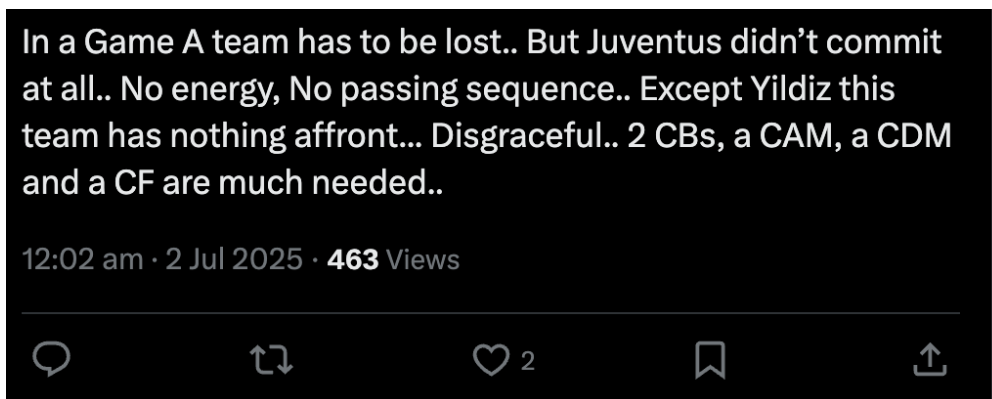
12:21 am · 22 Aug 2024 · **79** Views



### Tweet 9



### Tweet 10



## APPENDIX B – TURKISH TWEETS

### Tweet 1

[Translate post](#)

Cuesta'yı zararına İspanya ya da yine Belçika falan gibi Top7 dışı bi lige kitlesek, maaş yükünü düşününce, katkı da almıyor ve alamayacakken en doğrusunu yapmış oluruz. Her teklif değerlendirilmeli.

Metehan Baltacı, Cuesta kadar performans veriyor zaten.

3:39 pm · 23 May 2025 · **298** Views



### Tweet 2

Aston Villa gelip senin gözünün önündeki Kasımpaşa'dan 19'luk Yasin Özcan'ı alıp elindeki ıskartaya çıkmış [REDACTED] Diego Carlos'u 30m€ maliyetle sana çakıyor. Sırf şu olay bile istifayı basıp gitmeniz için yeterli.

[Translate post](#)

4:50 AM · May 20, 2025 · **137** Views



### Tweet 3

[Translate post](#)

[REDACTED] denilen  
ağlak bunağın ne seviyesi kalmış da hala onu savunuyorsun?Adam en  
son italya da rezilliklerinden dolayı adı kırmızı kart kralıydı.Bu ligde  
yaptığı rezillikleri bu güne dek hiçbir teknik direktör de görmedik.

2:46 am · 8 May 2025 · **22** Views



### Tweet 4

carlos diye topçu aldın 11 m devre arası büyük para sezon bitti 5 maça  
çıkmadı çağları aldın 2 haftada bi adelesini tutuyor ismail yüksek geçen  
senenin fred ismail pandemi müthişti mourinho sağolsun o pandemiye  
bitirdi amrabat diye topçu aldılar ben amrabat gibi 10 topç sayrm+

[Translate post](#)

12:45 AM · May 8, 2025 · **64** Views



### Tweet 5

denen zat,şampiyonluk hevesini berbat oyun akli ve futbolcu tercihleri ile bitirmiştir tek suçlu İsteddiği her topçu fazlasıyla alındı ama bir ilk 11 bile yapamadı. Tüm derbileri kaybetti kayseriyi samsunu evinde yenemedi gs yi 3 maçta da yenemedi

[Translate post](#)

1:17 PM · May 24, 2025 · **730** Views



### Tweet 6

İstifa et yeter duş ulan yakamizdan Allah'ın belası Allah'ın laneti yeter bırak git Allah'ın gslisi lanet olsun senin gibi yüz süze davalık olucam yeter lan yeter bıktık illallah ettik senden

[Translate post](#)

12:57 PM · May 24, 2025 · **315** Views



### Tweet 7

[Translate post](#)

43 oranlık kuponum yattı [REDACTED] yüzünden dünyanın en küçük takımı 1 gol atsa kg den alıyordum. Monchengladbach bile 3 0 dan döndürdü bu ezik [REDACTED] gol atamadı

11:28 pm · 17 Sep 2023 · **218** Views



### Tweet 8

Ayrıca twitin saçma olmuş Ş.G geldiğinde siz 11 maçın 7 dep yapmış fikstür avantajını eline almıştınız bizim içerde oynayacağımız tek derbi gs kalmıştı ve G alamamıştık üstüne ist. Hatay ve Alanya'ya puanlar bıraktık yani avantajı biz Şenol hoca gelmeden kaybetmiştik

[Translate post](#)

1:34 PM · May 9, 2023 · **70** Views



### Tweet 9

Lemina gibi adamı listeye almazsan [redacted] gibi çocukla [redacted] gibi ruhsuz antrenman topçusu ile [redacted] n gibi kazma ile [redacted] gibi hantal vurdumduymaz biriyle böyle rezil olursun [#Okanburukistifa](#)

[Translate post](#)

2:25 AM · Feb 14, 2025 from Elbistan, Kahramanmaraş · **139** Views



### Tweet 10

[redacted] r acil şark görevine gönderilmeli. Bi maç var bi maç yok. Morata diye aldığın adam top kontrol edemiyor. Kaan Ayhanın oynadığı pozisyon yanlış. Ayrıca ilk yarıdan sarı kart gören defans oyuncusu mu olur ? Kafanı yastığa nasıl rahat koyacaksın Okan hoca ?

[Translate post](#)

2:26 AM · Feb 14, 2025 · **191** Views



## APPENDIX C – ANNOTATION FRAMEWORK

| No       | Source                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | Output                                                                                                                                                                                                                                                                                                                                   | Errors                       | Annotators  |          | Annotators             |                   | Annotators |          | Annotators               |             | Annotators  |          |  |
|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------|-------------|----------|------------------------|-------------------|------------|----------|--------------------------|-------------|-------------|----------|--|
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          |                              | Accuracy    | Severity | Linguistic Conventions | Severity          | Style      | Severity | Audience Appropriateness | Severity    | Terminology | Severity |  |
| 1        | Cuestia'y zaranna lapanya ya da yine Beljika talan gbi Top7 d'ga bi lige <b>lissawa</b> , masay yikunni diguninco, kaka da almyor ve alanyayakalen en digunuru yayin duruz. Her tallif dajanyindinwal. Maishan Batazo, Cuesta katar performs veriyor zalen.                                                                                                                                                                                                                                                                                                                  | If we <b>liss</b> Cuesta in a non Top7 league like Spain or Belgium again at a loss, it would be the right thing to do considering his heavy burden and the fact that he doesn't and can't contribute. Every offer should be considered. Maishan Batazo is already performing as much as Cuesta.                                         | look                         | Nonexistent | None     |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          |                              |             |          |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          |                              |             |          |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          |                              |             |          |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          |                              |             |          |                        |                   |            |          |                          |             |             |          |  |
| 2        | Aston Villa gelip senin giccinin onindaki Kasimpaga'dan 19luk Yasin Ozcan'i atip elindaki <b>akurayya</b> <b>plam</b> <b>gip</b> <b>qdu</b> <b>qip</b> 32lik Diego Carlos'u 30m€ maliyyete sana <b>qakayor</b> . Sif gu dayi bile alitayy haang ghwanz qin yelari.                                                                                                                                                                                                                                                                                                           | Aston Villa come and take Yasin Ozcan, a 19-year-old from Kasimpaga, right in front of your eyes, and then they bring you a 32-year-old Diego Carlos, a <b>discarded piece of garbage</b> , at a cost of €30m. This incident alone is enough for you to resign and leave.                                                                | bring                        | Nonexistent | None     |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          | a discarded piece of garbage | Nonexistent | None     |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          |                              |             |          |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          |                              |             |          |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          |                              |             |          |                        |                   |            |          |                          |             |             |          |  |
| 3        | Sen ulamatz bir harfilin lan garel yoksunu yeli <b>qoqig</b> mourinho <b>qamun</b> aglar bawagen <b>n</b> <b>aweyas</b> <b>lurmu</b> da haka enu <b>kanunayunusu</b> ?Aden en son italya da naziliklerinden dolay ad kumuz kart kralaydi.Bu ligitte yapigir reziliken bu game dek hycir tehnik dinstatir da girmelik.                                                                                                                                                                                                                                                        | You're a shameless bastard, you dishonorable <b>son of a bitch</b> ! <b>What level</b> is left of the whiny old man called Mourinho that you're still defending him? The last time he was in Italy his name was the red card king because of his disgrace. We've never seen any manager in this league do the disgrace he's done.        | son of a bitch               | Nonexistent | None     |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          | What level is left           | Nonexistent | None     |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          | devil - called               | Nonexistent | None     |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          |                              |             |          |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          |                              |             |          |                        |                   |            |          |                          |             |             |          |  |
| 4        | <b>carloz</b> <b>diye</b> <b>toppi</b> <b>aldin</b> 11 m <b>deve</b> <b>arasa</b> <b>buyuk</b> <b>para</b> <b>sezon</b> <b>bili</b> 5 <b>maye</b> <b>qikmadi</b> <b>qajari</b> <b>aldin</b> 2 <b>haffada</b> <b>bi</b> <b>adelesini</b> <b>kutayor</b> <b>senal</b> <b>qakayor</b> <b>gwan</b> <b>senante</b> <b>fred</b> <b>senal</b> <b>pandemic</b> <b>midigip</b> <b>mourinho</b> <b>qajayor</b> <b>q</b> <b>pandemiyayla</b> <b>biradi</b> <b>arimabul</b> <b>diye</b> <b>toppi</b> <b>aldilar</b> <b>ben</b> <b>arimabul</b> <b>gbi</b> 10 <b>toppi</b> <b>sayim</b> + | you brought a ball player called carloz. 11 m half time big money season ended he didn't play 5 games you brought qajari he holds his muscle in 2 weeks small high last year's fred senal pandemic was great thanks to mourinho he ended that pandemic they brought a ball player called arimabul i count 10 ball players like arimabul+ | ball player                  | Nonexistent | None     |                        |                   |            |          |                          | calling him | None        |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          | 11 m                         |             |          |                        | Local Conventions | None       |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          | toppi                        |             |          |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          | senal                        |             |          |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          | half time                    |             |          |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          | deve arasa                   | Nonexistent | None     |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          | buyuk para                   | Nonexistent | None     |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          | sezon bili                   | Nonexistent | None     |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          | maye                         | Nonexistent | None     |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          | qikmadi                      | Nonexistent | None     |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          | qajari                       | Nonexistent | None     |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          | aldin                        | Nonexistent | None     |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          | haffada                      | Nonexistent | None     |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          | bi                           | Nonexistent | None     |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          | adelesini                    | Nonexistent | None     |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          | kutayor                      | Nonexistent | None     |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          | senal                        | Nonexistent | None     |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          | qajayor                      | Nonexistent | None     |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          | q                            | Nonexistent | None     |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          | pandemiyayla                 | Nonexistent | None     |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          | biradi                       | Nonexistent | None     |                        |                   |            |          |                          |             |             |          |  |
|          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                          | arimabul                     | Nonexistent | None     |                        |                   |            |          |                          |             |             |          |  |
| diye     | Nonexistent                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | None                                                                                                                                                                                                                                                                                                                                     |                              |             |          |                        |                   |            |          |                          |             |             |          |  |
| toppi    | Nonexistent                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | None                                                                                                                                                                                                                                                                                                                                     |                              |             |          |                        |                   |            |          |                          |             |             |          |  |
| aldilar  | Nonexistent                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | None                                                                                                                                                                                                                                                                                                                                     |                              |             |          |                        |                   |            |          |                          |             |             |          |  |
| ben      | Nonexistent                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | None                                                                                                                                                                                                                                                                                                                                     |                              |             |          |                        |                   |            |          |                          |             |             |          |  |
| arimabul | Nonexistent                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | None                                                                                                                                                                                                                                                                                                                                     |                              |             |          |                        |                   |            |          |                          |             |             |          |  |
| gbi      | Nonexistent                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | None                                                                                                                                                                                                                                                                                                                                     |                              |             |          |                        |                   |            |          |                          |             |             |          |  |
| 10       | Nonexistent                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | None                                                                                                                                                                                                                                                                                                                                     |                              |             |          |                        |                   |            |          |                          |             |             |          |  |
| toppi    | Nonexistent                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | None                                                                                                                                                                                                                                                                                                                                     |                              |             |          |                        |                   |            |          |                          |             |             |          |  |
| sayim    | Nonexistent                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | None                                                                                                                                                                                                                                                                                                                                     |                              |             |          |                        |                   |            |          |                          |             |             |          |  |
| +        | Nonexistent                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | None                                                                                                                                                                                                                                                                                                                                     |                              |             |          |                        |                   |            |          |                          |             |             |          |  |

## Appendix C Continued

[illegible]

## Appendix C Continued

[illegible]

## APPENDIX D – IRA CALCULATION

```
from scipy.stats import kendalltau

tau_AB, _ = kendalltau(df_wide["A"], df_wide["B"])
tau_AC, _ = kendalltau(df_wide["A"], df_wide["C"])
tau_BC, _ = kendalltau(df_wide["B"], df_wide["C"])

print("Annotator A-B tau:", tau_AB)
print("Annotator A-C tau:", tau_AC)
print("Annotator B-C tau:", tau_BC)
```

```
Annotator A-B tau: 0.10373207751977687
Annotator A-C tau: 0.6848833002372047
Annotator B-C tau: 0.09369064885646072
```

```
[ ] for d in ["EN2TR", "TR2EN"]:
    sub = df_wide.reset_index()
    sub = sub[sub["direction"]==d]
    if len(sub) > 2:
        tau_AB, _ = kendalltau(sub["A"], sub["B"])
        tau_AC, _ = kendalltau(sub["A"], sub["C"])
        tau_BC, _ = kendalltau(sub["B"], sub["C"])
        print(f"\n=== {d} ===")
        print("A-B:", tau_AB)
        print("A-C:", tau_AC)
        print("B-C:", tau_BC)
```

```
=== EN2TR ===
A-B: 0.16421004995023789
A-C: 0.6188327312808924
B-C: 0.25190385342279265

=== TR2EN ===
A-B: 0.11590203349898415
A-C: 0.7095979144542457
B-C: 0.07817734021220872
```

## APPENDIX E – COMETKiwiXL RESULTS FRAMEWORK

### Segment Level COMETKiwi-XL EN-TR Scores

| Tweet Number | DeepL Raw Score      | GPTo3 Raw Score      | X Translate Raw Score |
|--------------|----------------------|----------------------|-----------------------|
| 1            | 0.5473774075508118 . | 0.6515794396400452 . | 0.694537341594696 .   |
| 2            | 0.6436865329742432.  | 0.6361086368560791.  | 0.6354755759239197.   |
| 3            | 0.5115372538566589.  | 0.5343095064163208.  | 0.5090376734733582.   |
| 4            | 0.41732144355773926. | 0.49291926622390747. | 0.40250760316848755.  |
| 5            | 0.6993419528007507.  | 0.7391961216926575.  | 0.6653890609741211.   |
| 6            | 0.7474977374076843.  | 0.7111332416534424.  | 0.7037166357040405.   |
| 7            | 0.6598144769668579.  | 0.6490976810455322.  | 0.6287176609039307.   |
| 8            | 0.6904576420783997.  | 0.7045637369155884.  | 0.6207448840141296.   |
| 9            | 0.6879998445510864.  | 0.6555473208427429.  | 0.6388864517211914.   |
| 10           | 0.6633468866348267.  | 0.6104163527488708.  | 0.5222848653793335.   |

### Segment Level COMETKiwi-XL TR-EN Scores

| Tweet Number | DeepL Raw Score     | GPTo3 Raw Score      | X Translate Raw Score |
|--------------|---------------------|----------------------|-----------------------|
| 1            | 0.5884928107261658. | 0.623889148235321.   | 0.5990288257598877.   |
| 2            | 0.5982646942138672. | 0.5776396989822388.  | 0.5774787664413452.   |
| 3            | 0.5642539262771606. | 0.5029257535934448.  | 0.5631089210510254.   |
| 4            | 0.3599887788295746. | 0.3158932626247406.  | 0.4552450180053711.   |
| 5            | 0.5788437724113464. | 0.6019689440727234.  | 0.6023834943771362.   |
| 6            | 0.2606627643108368. | 0.31190744042396545. | 0.31526559591293335.  |
| 7            | 0.5236371159553528. | 0.47840502858161926. | 0.5277572870254517.   |
| 8            | 0.4722481966018677. | 0.40683045983314514. | 0.5400225520133972.   |
| 9            | 0.5958633422851562. | 0.5953834056854248.  | 0.5793527960777283.   |
| 10           | 0.5130795836448669. | 0.5410372614860535.  | 0.5057713985443115.   |

## System Level COMETKiwi-XL EN-TR Scores

| DeepL Raw Score             | GPTo3 Raw Score                       | X Translate Raw Score                       |  |
|-----------------------------|---------------------------------------|---------------------------------------------|--|
| DeepL system average: 0.627 | GPTo3 system average: $\approx 0.638$ | X Translate system average: $\approx 0.602$ |  |

## System Level COMETKiwi-XL TR-EN Scores

| DeepL Raw Score             | GPTo3 Raw Score             | X Translate Raw Score             |
|-----------------------------|-----------------------------|-----------------------------------|
| DeepL system average: 0.506 | GPTo3 system average: 0.496 | X Translate system average: 0.527 |

## APPENDIX F – SEGMENT LEVEL CORRELATION CALCULATION

```
from scipy.stats import pearsonr, kendalltau, spearmanr

def corr_report(df, label):
    x = df["MQM_score_avg"]
    y = df["COMET_score"]

    p, _ = pearsonr(x, y)
    k, _ = kendalltau(x, y)
    s, _ = spearmanr(x, y)

    print(f"{label} segment-level:")
    print(f"    Pearson:    {p:.4f}")
    print(f"    KendallTau: {k:.4f}")

corr_report(en2tr, "EN2TR")
corr_report(tr2en, "TR2EN")
```

```
EN2TR segment-level:
    Pearson:    0.3732
    KendallTau: 0.3514
TR2EN segment-level:
    Pearson:    0.5626
    KendallTau: 0.4336
```

## APPENDIX G – SYSTEM LEVEL AVERAGES

```
0s ✓ # Sistem bazında ortalamaları hesapla
en2tr_sys = en2tr.groupby("system")[["MQM_score_avg", "COMET_score"]].mean()
tr2en_sys = tr2en.groupby("system")[["MQM_score_avg", "COMET_score"]].mean()

print("EN→TR system-level averages:\n", en2tr_sys)
print("\nTR→EN system-level averages:\n", tr2en_sys)
```

↗ EN→TR system-level averages:

|            | MQM_score_avg | COMET_score |
|------------|---------------|-------------|
| system     |               |             |
| deepl      | 0.277294      | 0.626838    |
| gpto3      | 0.607210      | 0.638487    |
| xtranslate | 0.104482      | 0.602130    |

↗ TR→EN system-level averages:

|            | MQM_score_avg | COMET_score |
|------------|---------------|-------------|
| system     |               |             |
| deepl      | 0.387276      | 0.505533    |
| gpto3      | 0.757970      | 0.495588    |
| xtranslate | 0.273740      | 0.526541    |

## APPENDIX H – SYSTEM LEVEL CORRELATION CALCULATION

```
0s ✓ from scipy.stats import pearsonr, kendalltau

def system_level_corr(df, label):
    pearson, _ = pearsonr(df["MQM_score_avg"], df["COMET_score"])
    kendall, _ = kendalltau(df["MQM_score_avg"], df["COMET_score"])
    print(f"{label} system-level → Pearson: {pearson:.4f}, KendallTau: {kendall:.4f}")

system_level_corr(en2tr_sys, "EN→TR")
system_level_corr(tr2en_sys, "TR→EN")
```

↗ EN→TR system-level → Pearson: 0.9276, KendallTau: 1.0000  
TR→EN system-level → Pearson: -0.8771, KendallTau: -1.0000

## APPENDIX I – MQM RAW QUALITY SCORE FRAMEWORK

Since the framework is not readable when put full, it is shown in parts below:

### Part 1

| A          | B         | C          | D         | E         | F            | G          | H          |
|------------|-----------|------------|-----------|-----------|--------------|------------|------------|
| Segment_id | direction | system     | acc_minor | acc_major | acc_critical | ling_minor | ling_major |
| 1          | TR2EN     | deepl      | 0         | 1         | 0            | 0          | 0          |
| 1          | TR2EN     | gpto3      | 0         | 0         | 0            | 0          | 0          |
| 1          | TR2EN     | xtranslate | 0         | 1         | 1            | 0          | 0          |
| 2          | TR2EN     | deepl      | 1         | 1         | 0            | 0          | 0          |
| 2          | TR2EN     | gpto3      | 1         | 0         | 0            | 0          | 0          |
| 2          | TR2EN     | xtranslate | 0         | 1         | 1            | 0          | 0          |

## Part 2

| I             | J           | K           | L              | M              | N              | O                 | P          | Q          | R             | S             |
|---------------|-------------|-------------|----------------|----------------|----------------|-------------------|------------|------------|---------------|---------------|
| ling_critical | style_minor | style_major | style_critical | audience_minor | audience_major | audience_critical | term_minor | term_major | term_critical | total_penalty |
| 0             | 0           | 0           | 0              | 0              | 0              | 0                 | 0          | 0          | 0             | 5             |
| 0             | 0           | 0           | 0              | 0              | 0              | 0                 | 0          | 0          | 0             | 0             |
| 0             | 1           | 0           | 0              | 0              | 0              | 0                 | 0          | 0          | 0             | 31            |
| 0             | 0           | 0           | 0              | 0              | 0              | 0                 | 0          | 0          | 0             | 6             |
| 0             | 0           | 0           | 0              | 0              | 0              | 0                 | 0          | 0          | 0             | 1             |
| 0             | 0           | 0           | 0              | 0              | 0              | 0                 | 0          | 0          | 0             | 30            |

## Part 3

| T                                                                                                                                                                                                                                                                                                     | U     | V             | W            |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|---------------|--------------|
| target_text                                                                                                                                                                                                                                                                                           | words | PWPT          | RQS_1        |
| If we lock Cuesta in a non-Top7 league like Spain or Belgium at a loss, considering the salary burden, and since he is not and will not receive any contributions, it would be the right thing to do. Every offer should be evaluated. Metehan Baltacı already gives a performance as good as Cuesta. | 53    | 0,09433962264 | 0,9056603774 |
| If we lock Cuesta in a non-Top7 league like Spain or Belgium at a loss, considering the salary burden, and since he is not and will not receive any contributions, it would be the right thing to do. Every offer should be evaluated. Metehan Baltacı already gives a performance as good as Cuesta. | 53    | 0             | 1            |
| If we lock Cuesta in a non-Top7 league like Spain or Belgium at a loss, considering the salary burden, and since he is not and will not receive any contributions, it would be the right thing to do. Every offer should be evaluated. Metehan Baltacı already gives a performance as good as Cuesta. | 53    | 0,5849056604  | 0,4150943396 |
| Aston Villa comes and takes Yasin Özcan, aged 19, from Kasımpaşa, right before your eyes, and gives you their trash son of a bitch, Diego Carlos, aged 32, for a cost of 30m € . This incident alone is enough for you to resign and leave.                                                           | 46    | 0,1304347826  | 0,8695652174 |
| Aston Villa comes and takes Yasin Özcan, aged 19, from Kasımpaşa, right before your eyes, and gives you their trash son of a bitch, Diego Carlos, aged 32, for a cost of 30m € . This incident alone is enough for you to resign and leave.                                                           | 46    | 0,02173913043 | 0,9782608696 |
| Aston Villa comes and takes Yasin Özcan, aged 19, from Kasımpaşa, right before your eyes, and gives you their trash son of a bitch, Diego Carlos, aged 32, for a cost of 30m € . This incident alone is enough for you to resign and leave.                                                           | 46    | 0,652173913   | 0,347826087  |
| You are a shameless man. you dishonorable son of a bitch! What level                                                                                                                                                                                                                                  |       |               |              |

## APPENDIX J – MQM RAW QUALITY CALCULATION FORMULAS

Above the parts from the RQS calculation framework were given. Below Google Sheet formulas which were used during calculations are given.

### Listing J.1 — Absolute Penalty Total (APT)

```
=acc_minor*1 + acc_major*5 + acc_critical*25  
+ ling_minor*1 + ling_major*5 + ling_critical*25  
+ style_minor*1 + style_major*5 + style_critical*25  
+ audience_minor*1 + audience_major*5 + audience_critical*25  
+ term_minor*1 + term_major*5 + term_critical*25
```

Note: In practice, these formulas were implemented with cell references (e.g., C2, D2, E2) in Google Sheets; here they are presented with variable-style labels for readability.

### Listing J.2 — Word Count

```
=IF(LEN(TRIM(T2))=0; 0; COUNTA(SPLIT(TRIM(T2); " ")))
```

### Listing J.3 — PWPT

```
=IF(U2=0; 0; S2/U2)
```

### Listing J.4 — RQS

```
=MAX(0; 1 - V2)
```

## APPENDIX K – PYTHON PIPELINE FOR RATER AGGREGATION

### Listing K.1 — Reading Annotator CSVs with Python

```
▶ import pandas as pd

# Dosya isimlerini kendi yüklediğin isimlere göre yaz
df_A = pd.read_csv("Annotator A - Sayfa1.csv")
df_B = pd.read_csv("Annotator B - Sayfa1.csv")
df_C = pd.read_csv("Annotator C - Sayfa1.csv")

# Kontrol edelim
print(df_A.head())
print(df_B.head())
print(df_C.head())
```

```
↔
```

|   | Segment_id | direction | system     | RQS_1         |
|---|------------|-----------|------------|---------------|
| 0 | 1          | TR2EN     | deepl      | 0,9056603774  |
| 1 | 1          | TR2EN     | gpto3      | 1             |
| 2 | 1          | TR2EN     | xtranslate | 0,4150943396  |
| 3 | 2          | TR2EN     | deepl      | 0,8695652174  |
| 4 | 2          | TR2EN     | gpto3      | 0,9782608696  |
|   | Segment_id | direction | system     | RQS_1         |
| 0 | 1          | TR2EN     | deepl      | 0,09090909091 |
| 1 | 1          | TR2EN     | gpto3      | 0,09090909091 |
| 2 | 1          | TR2EN     | xtranslate | 0,3939393939  |
| 3 | 2          | TR2EN     | deepl      | 0             |
| 4 | 2          | TR2EN     | gpto3      | 0,5416666667  |
|   | Segment_id | direction | system     | RQS_1         |
| 0 | 1          | TR2EN     | deepl      | 0,5283018868  |
| 1 | 1          | TR2EN     | gpto3      | 1             |
| 2 | 1          | TR2EN     | xtranslate | 0             |
| 3 | 2          | TR2EN     | deepl      | 0,8913043478  |
| 4 | 2          | TR2EN     | gpto3      | 1             |

## Listing K.2 — Numeric conversion and annotator tagging

```
import pandas as pd
from scipy.stats import kendalltau

# 1) Dosyaları oku (isimleri seninkilerle aynı kalsın)
df_A = pd.read_csv("Annotator A - Sayfa1.csv")
df_B = pd.read_csv("Annotator B - Sayfa1.csv")
df_C = pd.read_csv("Annotator C - Sayfa1.csv")

# 2) Kolon adlarını standardize et (büyük/küçük, boşluk vs.)
def fix_cols(df):
    df = df.rename(columns={
        "Segment_id": "segment_id",
        "Segment_Id": "segment_id",
        "SegmentID": "segment_id",
        "Direction": "direction",
        "System": "system",
        "RQS_1": "RQS_1"
    })
    # direction ve system'i temizleyelim
    df["direction"] = df["direction"].astype(str).str.upper().str.replace("-", "2").str.replace("-", "").str.strip()
    df["system"] = df["system"].astype(str).str.lower().str.strip()
    return df

df_A = fix_cols(df_A)
df_B = fix_cols(df_B)
df_C = fix_cols(df_C)

# 3) Sayısal dönüştürme (virgöl -> nokta; string -> float)
def coerce_score(df, col="RQS_1"):
    # '0,9056' -> '0.9056'
    df[col] = (df[col]
               .astype(str)
               .str.replace(",", ".", regex=False)
               .str.strip())
    df[col] = pd.to_numeric(df[col], errors="coerce")
    return df

df_A = coerce_score(df_A, "RQS_1")
df_B = coerce_score(df_B, "RQS_1")
df_C = coerce_score(df_C, "RQS_1")
```

### Listing K.3 — Pivot Table of Mean RQS Scores per Annotator

```
df_wide = df_all.pivot_table(
    index=["segment_id", "system", "direction"],
    columns="annotator",
    values="RQS_1",
    aggfunc="mean" # varsa tekrarları ortalıyor
).dropna()

print(df_wide.head())
```

| annotator  |            |           | A       | B        | C        |
|------------|------------|-----------|---------|----------|----------|
| segment_id | system     | direction |         |          |          |
| 1          | deepl      | EN2TR     | 0.00000 | 0.090909 | 0.000000 |
|            |            | TR2EN     | 0.90566 | 0.905660 | 0.528302 |
|            | gpto3      | EN2TR     | 0.00000 | 0.090909 | 0.000000 |
|            |            | TR2EN     | 1.00000 | 1.000000 | 1.000000 |
|            | xtranslate | EN2TR     | 0.00000 | 0.363636 | 0.000000 |

### Listing K.4 — Exporting averaged datasets

```
from google.colab import files
import pandas as pd

uploaded = files.upload()
print("Uploaded:", list(uploaded.keys()))

def pick(name_part):
    for fn in uploaded.keys():
        if name_part.lower() in fn.lower():
            return fn
    return None

en2tr_path = pick("EN-TR") or pick("EN2TR") or pick("EN_TR")
tr2en_path = pick("TR-EN") or pick("TR2EN") or pick("TR_EN")
print("EN2TR file:", en2tr_path, " | TR2EN file:", tr2en_path)

def load_csv_auto(path):
    df = pd.read_csv(path)
    if df.shape[1] == 1: # TR yerel CSV ise
        df = pd.read_csv(path, sep=";", decimal=",")
    return df

en2tr_raw = load_csv_auto(en2tr_path)
tr2en_raw = load_csv_auto(tr2en_path)
en2tr_raw.head(), tr2en_raw.head()
```

Choose Files No file chosen Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

Saving MQM\_Averages - EN2TR - Averages.csv to MQM\_Averages - EN2TR - Averages (1).csv  
 Saving MQM\_Averages - TR2EN - Averages.csv to MQM\_Averages - TR2EN - Averages (1).csv  
 Uploaded: ['MQM\_Averages - EN2TR - Averages (1).csv', 'MQM\_Averages - TR2EN - Averages (1).csv']  
 EN2TR file: MQM\_Averages - EN2TR - Averages (1).csv | TR2EN file: MQM\_Averages - TR2EN - Averages (1).csv

| Segment_id | direction | system | MQM_score_avg           | COMET_score   |
|------------|-----------|--------|-------------------------|---------------|
| 0          | 1         | EN2TR  | 0,0303030303            | 0,5473774076  |
| 1          | 1         | EN2TR  | 0,0303030303            | 0,6515794396  |
| 2          | 1         | EN2TR  | xtranslate 0,1212121212 | 0,6945373416  |
| 3          | 2         | EN2TR  | deepl 0                 | 0,643686533   |
| 4          | 2         | EN2TR  | gpto3 0,4444444444      | 0,6361086369, |
| Segment_id | direction | system | MQM_score_avg           | COMET_score   |
| 0          | 1         | TR2EN  | deepl 0,7798742138      | 0,5884928107  |
| 1          | 1         | TR2EN  | gpto3 1                 | 0,6238891482  |
| 2          | 1         | TR2EN  | xtranslate 0,2767295597 | 0,5990288258  |
| 3          | 2         | TR2EN  | deepl 0,8768115942      | 0,5982646942  |
| 4          | 2         | TR2EN  | gpto3 0,9855072464      | 0,577639699)  |

## Listing K.5 — Inter-rater agreement

### Overall IRA Agreement

```
from scipy.stats import kendalltau

tau_AB, _ = kendalltau(df_wide["A"], df_wide["B"])
tau_AC, _ = kendalltau(df_wide["A"], df_wide["C"])
tau_BC, _ = kendalltau(df_wide["B"], df_wide["C"])

print("Annotator A-B tau:", tau_AB)
print("Annotator A-C tau:", tau_AC)
print("Annotator B-C tau:", tau_BC)
```

```
⇒ Annotator A-B tau: 0.8640011139917744
   Annotator A-C tau: 0.6848833002372047
   Annotator B-C tau: 0.6900256823985352
```

### IRA Agreement by Language Direction

```
▶ for d in ["EN2TR", "TR2EN"]:
    sub = df_wide.reset_index()
    sub = sub[sub["direction"]==d]
    if len(sub) > 2:
        tau_AB, _ = kendalltau(sub["A"], sub["B"])
        tau_AC, _ = kendalltau(sub["A"], sub["C"])
        tau_BC, _ = kendalltau(sub["B"], sub["C"])
        print(f"\n=== {d} ===")
        print("A-B:", tau_AB)
        print("A-C:", tau_AC)
        print("B-C:", tau_BC)
```

```
⇒
=== EN2TR ===
A-B: 0.758369717360841
A-C: 0.6188327312808924
B-C: 0.602663125548284

=== TR2EN ===
A-B: 0.9757869249394674
A-C: 0.7095979144542457
B-C: 0.7197350275178778
```

## ÖZGEÇMİŞ

|                                                  |                                                                                                                                                                                                                                                                                                                                                                                                                                |                                                  |                                                 |
|--------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------|-------------------------------------------------|
| <b>Adı, Soyadı</b>                               | Aziz Berkan                                                                                                                                                                                                                                                                                                                                                                                                                    |                                                  | KARADELİ                                        |
| <b>Doğum Yeri ve Yılı</b>                        |                                                                                                                                                                                                                                                                                                                                                                                                                                |                                                  |                                                 |
| <b>Bildiği Yabancı Diller</b>                    | İngilizce                                                                                                                                                                                                                                                                                                                                                                                                                      |                                                  | Rusça                                           |
| <b>ve Düzeyi</b>                                 | İleri                                                                                                                                                                                                                                                                                                                                                                                                                          |                                                  | Başlangıç                                       |
| <b>Eğitim Durumu</b>                             | <b>Başlama - Bitirme Yılı</b>                                                                                                                                                                                                                                                                                                                                                                                                  | <b>Kurum Adı</b>                                 |                                                 |
| <b>Lise</b>                                      | 2013                                                                                                                                                                                                                                                                                                                                                                                                                           | 2017                                             | Devrek Anadolu Lisesi                           |
| <b>Lisans</b>                                    | 2017                                                                                                                                                                                                                                                                                                                                                                                                                           | 2022                                             | Karadeniz Teknik Üniversitesi                   |
| <b>Yüksek Lisans</b>                             | 2022                                                                                                                                                                                                                                                                                                                                                                                                                           | 2025                                             | İstanbul 29 Mayıs Üniversitesi                  |
| <b>Doktora</b>                                   |                                                                                                                                                                                                                                                                                                                                                                                                                                |                                                  |                                                 |
| <b>Çalıştığı Kurum/lar</b>                       | <b>Başlama - Ayrılma Yılı</b>                                                                                                                                                                                                                                                                                                                                                                                                  | <b>Çalışılan Kurumun Adı</b>                     |                                                 |
| <b>1.</b>                                        | 2022                                                                                                                                                                                                                                                                                                                                                                                                                           | 2024                                             | Altınbaş Üniversitesi                           |
| <b>2.</b>                                        | 2024                                                                                                                                                                                                                                                                                                                                                                                                                           | Devam Ediyor                                     | Bahçeşehir Üniversitesi Mentora Education Group |
| <b>Üye Olduğu Bilimsel ve Mesleki Kuruluşlar</b> |                                                                                                                                                                                                                                                                                                                                                                                                                                |                                                  |                                                 |
| <b>Katıldığı Proje ve Toplantılar</b>            | Yürütücü: Aziz Berkan KARADELİ<br>Danışman: Muhammed BAYDERE<br><i>Makine Çevirisi Motorlarının İngilizce-Türkçe ve Türkçe-İngilizce Dil Çiftlerindeki Performanslarının İncelenmesi: Sosyal Medya Metinleri Bağlamında Google Translate, Microsoft Translator ve Yandex. Translate Üzerine Karşılaştırmalı Bir Analiz.</i> TÜBİTAK Projesi 2209-A, Üniversite Öğrencileri Araştırma Projeleri Destekleme Programı, 2022–2023. |                                                  |                                                 |
| <b>Yayınlar:</b>                                 | Baydere, Muhammed ve Aziz Berkan Karadeli. “İngilizce–Türkçe ve Türkçe–İngilizce Sosyal Medya Yemek Tarifi Metinlerinde Makine Çevirisi Kalitesi: Bir İnsan Değerlendirme Çalışması.” 9. Uluslararası KTUDELL Dil, Edebiyat ve Çeviri Konferansı, Trabzon, Türkiye, 2023.                                                                                                                                                      |                                                  |                                                 |
| <b>Diğer:</b>                                    |                                                                                                                                                                                                                                                                                                                                                                                                                                |                                                  |                                                 |
| <b>İletişim (e-posta):</b>                       |                                                                                                                                                                                                                                                                                                                                                                                                                                |                                                  |                                                 |
|                                                  |                                                                                                                                                                                                                                                                                                                                                                                                                                | <b>Tarih</b><br><b>İmza</b><br><b>Adı Soyadı</b> | 09.09.2025<br><br>Aziz Berkan KARADELİ          |

