Comparative Evaluation of Machine Translation Models Using Human-Translated Social Media Posts as References: Human-Translated Datasets

Shareefa Al Amer[1,2], Mark Lee[2], Phillip Smith[2]

[1]King Faisal University, [2]University of Birmingham

Correspondence: saalamer@kfu.edu.sa

Abstract

Machine translation (MT) of social media text presents unique challenges due to its informal nature, linguistic variations, and rapid evolution of language trends. In this paper, we propose a human-translated English dataset to Arabic, Italian, and Spanish, and a human-translated Arabic dataset to Modern Standard Arabic (MSA) and English. We also perform a comprehensive analysis of three publicly accessible MT models using human translations as a reference. We investigate the impact of social media informality on translation quality by translating the MSA version of the text and comparing BLEU and METEOR scores with the direct translation of the original social media posts. Our findings reveal that MarianMT provides the closest translations to human for Italian and Spanish among the three models, with METEOR scores of 0.583 and 0.640, respectively, while Google Translate provides the closest translations for Arabic, with a METEOR score of 0.354. By comparing the translation of the original social media posts with the MSA version, we confirm that the informality of social media text significantly impacts translation quality, with an increase of 12 percentage points in METEOR scores over the original posts. Additionally, we investigate inter-model alignment and the degree to which the output of these MT models align.

1 Introduction

The growing demand for multilingual content has driven the development of machine translation (MT) systems, which enable fast and cost-effective translation of large text volumes. Evaluating MT systems is critical to ensuring accuracy across diverse languages and text types, especially with the rise of social media as a rich, multilingual data source for tasks like sentiment analysis and topic detection. However, social media text's unstructured, error-prone nature—with linguistic variations, informal language, and abundant slang—poses unique challenges for MT.

This work contributes to the ongoing efforts to enhance MT systems for handling the intricacies of social media language, ultimately advancing the accessibility of multilingual content in an increasingly interconnected world. We propose a human-translated set of English X posts into Arabic, Italian, and Spanish, along with their corresponding translations by three distinct MT models (Google Translate, MarianMT, and Facebook M2M), to provide a comprehensive basis for assessing the efficacy of these systems in handling the complexities of social media language. Additionally, we have curated a dataset of 500 Arabic X posts translated by humans into both Modern Standard Arabic (MSA) and English.

These datasets serve as valuable resources for researchers in the field of natural language processing (NLP), providing diverse and representative samples of social media text for evaluating the performance of MT systems across multiple languages. Furthermore, to gain insights into the relative performance of the MT models, we evaluate each MT model against the other two by using reference translations from other models. This approach allows us to assess how similar the translations generated by each model are to those produced by the other models when using a specific reference. By conducting pairwise comparisons, we can identify strengths and weaknesses of individual MT models and gain a deeper understanding of their relative performance on social media text [1].

To explicitly illustrate our contributions, we list them below:

- **CrisisNLP-HMT-1K**: A Human and Machine Translated (HMT) version of the CrisisNLP dataset (Imran et al., 2016), compris-

[1]The datasets are available upon request by contacting the corresponding author via email.

Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025), pages 1–9

ing 1000 English X posts manually translated into three languages: Arabic, Italian, and Spanish. The dataset includes the same samples translated by three distinct machine translation (MT) models: Google Translate, MarianMT, and Facebook's M2M100.

- **Kawarith-HMT-500**: A Human and Machine Translated (HMT) version of the Kawarith dataset (Alharbi and Lee, 2021), consisting of 500 Arabic X posts manually translated by humans into both Modern Standard Arabic (MSA) and English. Additionally, the dataset includes translations generated by the same MT models into English, Italian & Spanish.

- A comprehensive evaluation of the several MT systems against the reference translations, as well as against one another, using BLEU and METEOR metrics.

We also aim to address the following question: Is there a correlation between the formality of X posts and the quality of their machine translations? We hypothesise that X posts written in a formal style would yield better translation outcomes. This question is explored in Section 5.3.1, where we systematically compare the output of MT models when translating dialectical Arabic versus Modern Standard Arabic (MSA) version of the same data.

These contributions are essential for tackling the challenges of evaluating machine translation systems, particularly in the context of social media language. The human-translated dataset serves as a valuable benchmark for assessing MT model performance in capturing the nuances and informalities of social media text. Through a comprehensive evaluation using established metrics such as BLEU and METEOR, we provide quantitative insights into the comparative performance of MT systems, thereby contributing to ongoing efforts to enhance the quality and accuracy of multilingual content translation, especially in the challenging domain of social media language.

The remainder of the paper is structured as follows: Section 2 provides an overview of related work, including existing parallel corpora and machine translation evaluation metrics. Section 3 explains how the data was sourced and processed. The methodology is described in Section 4. Section 5 covers the experimental setup, evaluation, and discusses the results. Finally, Section 6 concludes the paper.

2 Related Work

In recent years, the field of machine translation (MT) has witnessed significant advancements, driven by the growing demand for multilingual communication across various domains. Researchers have explored diverse approaches and techniques to improve the accuracy and effectiveness of MT systems, particularly in handling the challenges posed by social media text. In this section, we provide an overview of the related work in the literature, focusing on key contributions and trends in the field.

Machine translation systems are being used to translate social media posts, but their performance is not always perfect. Users have expressed concerns about potential inaccuracies in the translations of their posts and the lack of control over the translation process. The use of machine translation tools, such as Google Translate (Schuster et al., 2016), in social media is a relatively recent phenomenon, and it presents new challenges for translating informal language, dialects, and platform-specific language phenomena (Gupta et al., 2023). While these systems can help overcome language barriers, they also have limitations, and users often desire higher translation quality. Therefore, there is a need for machine translation systems specifically trained on social media text to improve their performance on informal language and address the concerns of users. Furthermore, machine translation has proven beneficial for creating parallel corpora to train Large Language Models (LLMs) for downstream tasks in natural language processing, reinforcing the importance of tailored translation systems for social media content (Al Amer et al., 2024, 2023).

2.1 Existing Parallel Corpora

The availability of resources for machine translation plays a crucial role in advancing research in the field. Large-scale parallel corpora, such as the Europarl (European Parliament Proceedings Parallel Corpus) Corpus for European languages (Koehn, 2005) and the United Nations Parallel Corpus (Ziemski et al., 2016) for multiple languages, have been instrumental in training and evaluating MT systems. Additionally, efforts such as the OPUS project provide open repositories of parallel corpora in various languages, facilitating research and development in machine translation (Tiedemann, 2012). The QT21 corpus contains parallel

source and human reference sentences from information technology and life sciences domains for the following language pairs: English to German (en-de), Latvian (en-lt) and Czech (en-cs), and German to English (de-en) (Specia et al., 2017).

The availability of parallel corpora extracted from social media platforms such as Twitter (now X) can significantly enrich research in machine translation and cross-lingual analysis. Several noteworthy Twitter parallel corpora have been developed, each contributing to our understanding of multilingual communication and the challenges of translating informal language on social media. Amid the global COVID-19 pandemic, the TwiConv-19 dataset emerged as a valuable resource for studying multilingual communication on Twitter during times of crisis. Curated to capture conversations related to the pandemic, TwiConv-19 provides parallel data of X posts in multiple languages, offering insights into how individuals across different linguistic communities engage with and respond to public health emergencies (Aktaş and Kohnert, 2020). Another existing dataset is the ParaCrawl project, a comprehensive multilingual parallel corpus collected from various online sources, including social media platforms like Twitter. With translations available for numerous language pairs, ParaCrawl facilitates research in machine translation and cross-lingual analysis by offering a diverse range of linguistic data sourced from the web (Bañón et al., 2020). Additionally, the Twitter Parallel Corpus by (Tiedemann, 2012) provides aligned tweet translations for several languages, enabling detailed investigations into cross-lingual phenomena and the adaptation of machine translation models to informal social media language. These datasets serve as invaluable resources for studying multilingual communication dynamics and advancing machine translation systems tailored for informal social media language.

Such resources are essential for evaluating different MT systems and for enabling various natural language processing (NLP) tasks. Evaluation of MT systems relies heavily on the availability of high-quality reference translations, which are essential for assessing the accuracy and fluency of machine-generated translations. Moreover, parallel corpora serve as valuable resources for training and fine-tuning MT models, enabling researchers to develop systems that perform well across different languages and domains. Beyond machine translation, parallel corpora are also used in other NLP tasks, including cross-lingual information retrieval, sentiment analysis, and language modeling (Schwenk and Li, 2019).

2.2 Evaluation Metrics

Evaluating machine translation (MT) systems on social media text is challenging due to the lack of standardised benchmarks and the subjective nature of translation quality. Common metrics include BLEU, which measures n-gram overlap between machine and reference translations (Papineni et al., 2002); METEOR, which accounts for unigram matches, stemming, and synonymy (Banerjee and Lavie, 2005); TER, which calculates the edit distance needed to transform the machine translation into the reference (Snover et al., 2006); ROUGE, which evaluates overlap between generated and reference summaries (Lin, 2004); and COMET, which leverages embeddings to compute similarity between outputs and references (Rei et al., 2020). While these metrics provide valuable insights, they may not fully capture the nuances of social media language, necessitating tailored evaluation approaches (Specia et al., 2010).

While these metrics offer valuable insights into MT system performance, they may not fully capture the nuances of social media language. Thus, exploring alternative evaluation methods tailored to the specific characteristics of social media text becomes imperative (Specia et al., 2010).

3 Data

The English data was sourced from the CrisisNLP dataset, a valuable resource containing English X posts collected and annotated specifically for crisis-related analysis and classification (Imran et al., 2016). This dataset provides a rich source of real-time social media data, offering insights into how individuals communicate and respond during crises. From this dataset, we selected a sample of 1000 X posts posted during various crisis events to enrich the diversity of content.

We sourced the Arabic data from the Kawarith dataset, another valuable resource containing Arabic X posts collected and annotated specifically for crisis-related analysis and classification (Alharbi and Lee, 2021). We sampled 500 X posts to generate parallel human translations.

In Table 1, we provide a detailed description of the selected data, which encompasses a wide range of topics, reflecting the varied nature of com-

munication on social media platforms during crisis situations. Through the careful selection of X posts, we aim to capture the nuances and complexities of language use in crisis-related contexts, facilitating a comprehensive analysis of machine translation performance and effectiveness.

	CrisisNLP-HMT-1K	Kawarith-HMT-500
# X posts	1,000	500
# disasters	27	7
# types	5	2
sum (words)	10,018	10,273
mean (words)	10.02	27.5
std	3.5	14.04
min (words)	2	4
25%	7	15.8
50%	10	27.5
75%	13	39.3
max (words)	19	52

Table 1: The statistics of the selected samples from CrisisNLP and Kawarith, respectively.

3.1 Data Pre-processing

Before providing the X posts to the translators, we conducted a pre-processing step to ensure consistency and minimise discrepancies between human and machine translations. Specifically, we removed user mentions (e.g., @username), retweet indicators (RT), URLs, and emojis from the X posts. This pre-processing step aims to streamline the translation process by presenting the translators with clean, focused text free from unnecessary information commonly found in social media posts, thereby reducing potential biases and variations in translation quality and enabling a more accurate comparison between human and machine translations.

4 Methodology

The data was sampled to cover diverse events worldwide, ensuring the inclusion of a variety of linguistic features, as each event took place in a different part of the world. The English dataset was sampled from the CrisisNLP dataset and cleaned following the process outlined in Section 3.1. Following this, human translators were engaged to translate the X posts into Arabic, Italian, and Spanish. Simultaneously, three distinct machine translation models—MarianMT (Helsinki-NLP/opus-mt), Google Translate (4.0.0-rc1), and Facebook's M2M (m2m100_418M)—were chosen for their popularity and accessibility to generate corresponding MT translations.

For the Arabic data sourced from the Kawarith dataset, human translators translated the X posts to English and converted the Arabic X posts to MSA. Similar to the English data, the Arabic data was translated by the same three MT models to English, Italian, and Spanish. Subsequently, an evaluation was conducted to assess these models against both human reference translations and each other.

In addition to evaluating the translation quality of the original X posts to the various languages, we introduce an additional dimension to our study by investigating the influence of the informal nature of X posts on translation quality. To assess this, we translated the Modern Standard Arabic (MSA) version of the X posts into English and used these translations as candidate translations against the human reference translations. Subsequently, we calculated BLEU and METEOR scores for both sets of translations. The comparison between translating the original Arabic X posts and their standardised counterpart can provide valuable insight into the impact of tweet informality on translation quality.

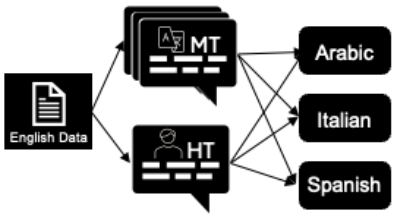

Figure 1: The process of generating the 12 different translations of the same English data. **MT** denotes Machine Translation, and **HT** denotes Human Translation.

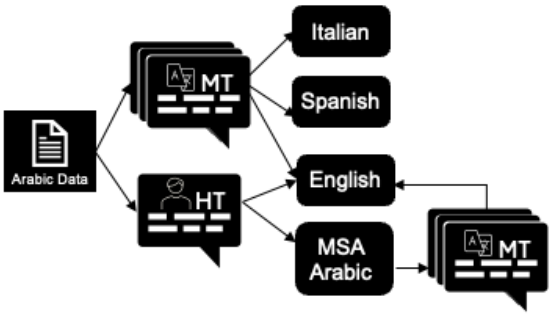

Figure 2: The process of generating the Modern Standard Arabic (MSA) version and different translations of the Arabic data. **MT** denotes Machine Translation, and **HT** denotes Human Translation.

5 Experiments

In this section, we outline the experimental setup designed to systematically evaluate and compare

machine translation models. We describe the process of data sampling, cleaning, and translation, followed by an overview of the evaluation metrics employed. Subsequently, we present the results of the evaluation, including BLEU and METEOR scores, and analyse the inter-model alignment to gain insights into the performance of the translation models.

5.1 Experimental Setup

The experimental setup involved several steps to ensure systematic evaluation and comparison of machine translation models. Initial data sampling from CrisisNLP, detailed in Section 3, was followed by thorough data cleaning as described in Section 3.1 to ensure consistency and quality. The translation process involved translating CrisisNLP-HMT-1K dataset into Arabic (Ar), Italian (It), and Spanish (Es) both manually by humans and using three distinct machine translation models: MarianMT, Google Translate, and Facebook M2M while the Kawarith-HMT-500 dataset was translated into English (En) and Modern Standard Arabic (MSA) by humans, and then translated into English (En), Italian (It), and Spanish (Es) using the same three machine translation models. For assessing the quality of the translated data, we utilise the Modern Standard Arabic (MSA) version of the Arabic X posts, translating them into English to serve as candidate translations for the MSA reference. The translation process is illustrated in Figure 1 and 2. BLEU and METEOR scores were then calculated against human reference translations and for inter-model comparison, with the results presented in Section 5.3 for analysis and interpretation.

5.2 Evaluation

The evaluation of machine translation quality is essential to ensure that machine translation systems produce accurate results comparable to human translations. In this section, we discuss the metrics used for evaluation and provide an overview of BLEU and METEOR, which are widely employed in the machine translation community for this purpose.

5.2.1 Metrics

Evaluating the quality of machine translations can be accomplished through various methods, all aimed at ensuring that machine translation systems produce accurate results comparable to human translations. While we have chosen BLEU and METEOR as the metrics for this purpose, which are widely used in the MT community, they provide unique insights into different aspects of translation quality, enabling a more comprehensive evaluation.

The evaluation of the translations involved comparison with human reference translations and inter-model comparison. This process entailed designating one machine translation as a reference and computing BLEU and METEOR scores against the other two translations to gain insights into their alignment.

Here, we briefly explain BLEU and METEOR, highlighting their similarities and differences.

BLEU (Bilingual Evaluation Understudy) quantifies the overlap between machine-generated and reference translations by comparing n-grams. Scores range from 0 to 1, with higher scores indicating better alignment. While widely used, BLEU has limitations, such as not accounting for fluency, coherence, or languages with complex grammar (Papineni et al., 2002).

METEOR (Metric for Evaluation of Translation with Explicit ORdering) evaluates translation quality using precision, recall, stemming, synonymy, and word order. It offers a more nuanced assessment than BLEU but still has limitations in capturing translation subtleties (Banerjee and Lavie, 2005).

5.3 Results & Discussion

Among the results illustrated in Table 2, MarianMT performed the best for translating the English data to Italian and Spanish, while Google Translate proved to be more effective for translating into Arabic. On the other hand, Kawarith-HMT-500 yielded the highest BLEU score when translated to English by Google Translate and the highest METEOR score when translated by MarianMT, with no significant difference between the two, as depicted in Table 3.

Given these considerations, the low BLEU scores observed might not necessarily reflect poor model performance due to the inherent complexity of the task. Translating social media content poses unique challenges in machine translation, and achieving high BLEU scores in this area is more difficult than in more formal types of text (Sabtan et al., 2021).

The METEOR scores are notably higher than the BLEU scores for the same models and languages. This difference could be due to METEOR's more comprehensive evaluation of translation quality,

	BLEU			METEOR		
	Ar	It	Es	Ar	It	Es
Google Translate	**0.144**	0.219	0.248	**0.354**	0.569	0.599
Facebook M2M	0.097	0.207	0.264	0.283	0.535	0.560
MarianMT	0.081	**0.239**	**0.325**	0.236	**0.583**	**0.640**

Table 2: BLEU and METEOR scores calculated for each Machine Translation model translating CrisisNLP-HMT-1K to Arabic, Italian, and Spanish using human translation as the reference.

	BLEU	METEOR
Google Translate	**0.177**	0.407
Facebook M2M	0.129	0.344
MarianMT	0.171	**0.411**

Table 3: BLEU and METEOR scores calculated for each Machine Translation model translating Kawarith-HMT-500 into English using human translation as the reference.

which accounts for factors like synonymy and sentence structure. This broader assessment approach is likely more forgiving than BLEU's strict n-gram matching, especially when applied to social media text.

It is also noteworthy that Arabic translation has the lowest scores among the other two languages. This difference may be attributed to the dissimilarities between the source (English) and target (Arabic) languages compared to translations within the same language family.

5.3.1 Impact of Text Informality on Translation Quality

As we hypothesise that the quality of the X posts being translated influences the quality of the generated machine translations, we pose a question: what if the X posts were written formally? How much would the translation quality improve? To investigate this, we use the Modern Standard Arabic (MSA) version instead of the raw X posts and translate it using the selected machine translation models. We then compare the BLEU and METEOR scores of the MSA-translated text with those of the raw tweet translations, both considering the human English translation as a reference. The results illustrated in Table 3 and 4 demonstrate a significant increase in both BLEU and METEOR scores, supporting the assumption that translated social media text tends to have relatively low scores. Specifically, there is an average increase of 8.7 percentage points in BLEU scores and an average increase of 12 percentage points in METEOR scores for the MSA-translated text compared to the raw X posts being translated. This increase suggests an improvement in translation quality when formal language is being translated. We also speculate that translating more contextual text could yield even better results considering the brief and context-limited nature of X posts. This can be investigated in future work.

	BLEU	METEOR
Google Translate	**0.263**	0.510
Facebook M2M	0.215	0.480
MarianMT	0.262	**0.526**

Table 4: BLEU and METEOR scores calculated for each Machine Translation model translating the standardised (MSA) version of Kawarith-HMT-500 into English using human translation as the reference.

5.3.2 Inter-Model Alignment

We were also curious about how closely the outputs of those MT models align. Conventional methods for evaluating machine translation quality often involve using reference human translations as a gold standard as we have done above. In this assessment, however, we used a variation of this method, where we use one machine translation as a reference for evaluating another machine translation. This can provide valuable insights into how different systems perform relative to each other. We show results of this assessment in Tables 5, 6, 7 and 8.

The BLEU scores in this analysis are slightly higher than the scores observed when evaluating the models against human translations. This suggests a higher degree of similarity between the outputs of these different models compared to their alignment with human translations. Specifically, the translations generated by MarianMT for Italian (It) and Spanish (Es) align more closely with the outputs of other models, as indicated by their higher scores. However, for Arabic (Ar), the lowest alignment scores are observed, emphasising the

		Reference								
		Google Translate			**Facebook M2M**			**MarianMT**		
Evaluated Model		Ar	It	Es	Ar	It	Es	Ar	It	Es
Google Translate	Ar				0.161			0.111		
	It					0.290			0.337	
	Es						0.258			0.372
Facebook M2M	Ar	**0.162**						0.081		
	It		0.290						0.360	
	Es			0.258						0.315
MarianMT	Ar	0.109			0.080					
	It		0.339			**0.362**				
	Es			**0.374**			0.315			

Table 5: BLEU score calculated for translating CrisisNLP-HMT-1K to demonstrate the similarity of the MT outputs in relation to each other. Best alignment for each language is highlighted in bold.

		Reference								
		Google Translate			**Facebook M2M**			**MarianMT**		
Evaluated Model		Ar	It	Es	Ar	It	Es	Ar	It	Es
Google Translate	Ar				0.379			0.351		
	It					0.664			**0.697**	
	Es						0.503			**0.719**
Facebook M2M	Ar	**0.472**						0.286		
	It		0.661						0.652	
	Es			0.605						0.620
MarianMT	Ar	0.333			0.221					
	It		0.679			0.638				
	Es			0.703			0.504			

Table 6: METEOR score calculated for translating CrisisNLP-HMT-1K to demonstrate the similarity of the MT outputs in relation to each other. Best alignment for each language is highlighted in bold.

notable differences in the generated translations for this specific language. This may be attributed to the complexity and structure of the Arabic language.

The BLEU results indicate that translations generated by Facebook's M2M100 model align best with those of MarianMT for translating English to Italian, while it aligns best with Google Translate for translating English to Arabic and Arabic to Italian. Moreover, Google Translate aligns best with MarianMT for translating English to Spanish and Arabic to English & Spanish. On the other hand, METEOR scores show the best alignment between Google Translate and MarianMT for translating English to Italian and Spanish, while it aligns more with Facebook's M2M100 for translating English to Arabic. When translating Arabic to English, Google Translate and MarianMT align the most with each other for translating Arabic to English, Italian, and Spanish.

Despite our initial anticipation of greater similarity among different translation models, the findings indicate that the translation outputs of these models may exhibit more similarity to each other than to the human reference translations. This analysis provides a distinctive viewpoint on the performance of these machine translation models, offering insights into their differences when compared to each other's outputs.

5.3.3 Key Findings and Recommendations

As we analyse our results, uncovering key insights, we highlight the outcomes of our experiments with the following findings:

- Google Translate provides the best translation

		Reference								
		Google Translate			**Facebook M2M**			**MarianMT**		
Evaluated Model		En	It	Es	En	It	Es	En	It	Es
Google Translate	En				0.176			**0.214**		
	It					0.143			0.139	
	Es						0.150			**0.183**
Facebook M2M	En	0.177						0.161		
	It		**0.144**						0.124	
	Es			0.154						0.167
MarianMT	En	**0.214**			0.161					
	It		0.139			0.124				
	Es			0.182			0.166			

Table 7: BLEU score calculated for translating Kawarith-HMT-500 to demonstrate the similarity of the MT outputs in relation to each other. Best alignment for each language is highlighted in bold.

		Reference								
		Google Translate			**Facebook M2M**			**MarianMT**		
Evaluated Model		En	It	Es	En	It	Es	En	It	Es
Google Translate	En				0.442			0.463		
	It					0.319			0.342	
	Es						0.359			0.411
Facebook M2M	En	0.442						0.427		
	It		0.335						0.334	
	Es			0.381						0.415
MarianMT	En	**0.495**			0.449					
	It		**0.358**			0.329				
	Es			**0.426**			0.402			

Table 8: METEOR score calculated for translating Kawarith-HMT-500 to demonstrate the similarity of the MT outputs in relation to each other. Best alignment for each language is highlighted in bold.

from English to Arabic among the other two tested models.

- MarianMT offers the best translation from English to Italian and Spanish.
- Google Translate and MarianMT demonstrate comparable translation quality from Arabic to English, both outperforming Facebook's M2M100.
- The formality of the Arabic text being translated significantly improves machine translation quality.
- Google Translate and MarianMT exhibit the highest level of similarity in generated translations from English to Italian & Spanish.
- Lower levels of alignment between MT models have been observed for Arabic translation from English.
- Translating Arabic to English yields better results than vice versa.

6 Conclusion

Our study examined the translation quality of social media text across multiple languages using human and machine translation. Formal language, such as Modern Standard Arabic, improved accuracy compared to informal text, and translation quality varied across languages and models, emphasising the role of formality and linguistic context.

The human-translated datasets we are publishing will support improvements in machine translation, cross-lingual research, and crisis communication. Future human evaluations of machine-generated translations will provide deeper insights into fluency and adequacy, refining our understanding of

translation performance in social media contexts.

References

Berfin Aktaş and Annalena Kohnert. 2020. TwiConv: A Coreference-annotated Corpus of Twitter Conversations. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*.

Shareefa Al Amer, Mark Lee, and Phillip Smith. 2023. Cross-lingual Classification of Crisis-related Tweets Using Machine Translation. In *Proceedings of Recent Advances in Natural Language Processing*, pages 22–31, Varna, Bulgaria. INCOMA Ltd.

Shareefa Al Amer, Mark Lee, and Phillip Smith. 2024. Adopting Ensemble Learning for Cross-lingual Classification of Crisis-related Text On Social Media. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 159–165, Bangkok,Thailand. Association for Computational Linguistics.

Alaa Alharbi and Mark Lee. 2021. Kawarith: An Arabic Twitter Corpus for Crisis Events. *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 42–52.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Proceedings of the Workshop ACL 2005*.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz-Rojas, Leopoldo Pla, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale Acquisition of Parallel Corpora. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Ananya Gupta, Jae D. Takeuchi, and Bart P. Knijnenburg. 2023. On The Real-world Performance of Machine Translation: Exploring Social Media Post-authors' Perspectives. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a Lifeline: Human-Annotated Twitter Corpora for NLP of Crisis-Related Messages. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*.

Philipp Koehn. 2005. Europarl : A Parallel Corpus for Statistical Machine Translation. *MT Summit*, 11.

C Y Lin. 2004. Rouge: A Package for Automatic Evaluation of Summaries. *Proceedings of the workshop on text summarization branches out (WAS 2004)*, (1).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, volume 2002-July.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*.

Yasser Muhammad Naguib Sabtan, Mohamed Saad Mahmoud Hussein, Hamza Ethelb, and Abdulfattah Omar. 2021. An Evaluation of the Accuracy of the Machine Translation Systems of Social Media Language. *International Journal of Advanced Computer Science and Applications*, 12(7).

Mike Schuster, Melvin Johnson, and Nikhil Thorat. 2016. Zero-Shot Translation with Google's Multilingual Neural Machine Translation System.

Holger Schwenk and Xian Li. 2019. A Corpus for Multilingual Document Classification in Eight Languages. In *LREC 2018 - 11th International Conference on Language Resources and Evaluation*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *AMTA 2006 - Proceedings of the 7th Conference of the Association for Machine Translation of the Americas: Visions for the Future of Machine Translation*.

Lucia Specia, Nicola Cancedda, and Marc Dymetman. 2010. A Dataset for Assessing Machine Translation Evaluation Metrics. In *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*.

Lucia Specia, Kim Harris, Frederic Blain, Aljoscha Burchardt, Vivien Macketanz, Inguna Skadin, Matteo Negri, and Marco Turchi. 2017. Translation Quality and Productivity: A Study on Rich Morphology Languages. In *Proceedings of Machine Translation Summit XVI: Research Track*.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*.

Michal Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations Parallel Corpus v1.0. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*.