Analysing Mistranslation of Emotions in Multilingual Tweets by Online MT Tools

Hadeel Saadany *
Centre for Translation Studies
University of Surrey

Constantin Orăsan
Centre for Translation Studies
University of Surrey

Rocío Caro Quintana
RGCL
University of Wolverhampton

Félix do Carmo
Centre for Translation Studies
University of Surrey

Leonardo Zilio
Centre for Translation Studies
University of Surrey

Abstract

It is common for websites that contain User-Generated Text (UGT) to provide an automatic translation option to reach out to their linguistically diverse users. In such scenarios, the process of translating the users' emotions is entirely automatic with no human intervention, neither for post-editing, nor for accuracy checking. In this paper, we assess whether automatic translation tools can be a successful real-life utility in transferring emotion in multilingual tweets. Our analysis shows that the mistranslation of the source tweet can lead to critical errors where the emotion is either completely lost or flipped to an opposite sentiment. We identify linguistic phenomena specific to Twitter data which pose a challenge in translation of emotions and show how frequent these features are in different language pairs. We also show that commonly-used quality metrics can lend false confidence in the performance of online MT tools specifically when the source emotion is distorted in telegraphic messages such as tweets.

1 Introduction

Despite the tremendous improvement in the quality of automatic translation as a result of the use of Neural Machine Translation (NMT) systems, NMT output still contains errors. This is particularly noticeable with User-Generated Text (UGT) such as tweets which do not follow the common lexico-grammatical standards (Saadany et al., 2021b). In spite of this limitation, NMT systems are commonly used in multilingual platforms such as Twitter to provide its users with an idea of global views or emotions towards current events or public figures. In such scenarios, the component of the tweet that conveys emotions is often pivotal to the understanding of the tweet's message. There have been different studies which explored how far sentiment information can be captured from the machine-translated text (Demirtas and Pechenizkiy, 2013; Shalunts et al., 2016; Mohammad et al., 2016; Barhoumi et al., 2018). The objective of most research in this area, however, is from a sentiment classification perspective, rather than a translation accuracy perspective. It measures how far automatic translation of a language into English can help with the sentiment classification of that language by applying the available English sentiment resources on the target text. (Salameh et al., 2015; Araujo et al., 2016; Afli et al., 2017; Abdalla and Hirst, 2017).

The research presented in this paper, however, evaluates the preservation of the affect message from a user-related perspective. We assess how far NMT systems used in online platforms can be a successful real-life utility in transferring the user's fine-grained emotions such as anger and joy. Research has shown that NMT models are capable of producing an impressively fluent output that completely misses the correct meaning of the source (Koehn and Knowles, 2017). The problem exacerbates when the source text is a deliberately concise text that carries a strong sentiment message as is the case with tweets. Moreover, analysis of sentiment

*hadeel.saadany@surrey.ac.uk

mistranslations produced by online tools revealed typical errors related to linguistic phenomena such as contronyms, idiomatic expressions, and dialectical code-switching (Saadany and Orăsan, 2020). In this paper, we aim to investigate to what extent similar errors can be identified in the translation of tweets. To achieve this, we carry out an analysis of datasets of tweets automatically translated into different language pairs. At the end of this analysis, we attempt to provide answers to the following questions:

1. Are there specific linguistic features of tweets that can lead to mistranslation of emotions?
2. How far mistranslation can distort the affect message and whether different language pairs are equally affected?
3. Can traditional automatic quality measures adequately evaluate the mistranslation of sentiment?

To answer the above research questions, this paper is divided as follows: Section 2 presents our data compilation process and the approach used for evaluating the translation of emotion in tweets by MT systems. Section 3 analyses challenging features for the translation of emotions in multilingual tweets. It also provides a qualitative analysis of each feature based on its frequency in our compiled dataset and its prominence in each of the source languages explored. In section 4, we evaluate the efficacy of the MT automatic quality metrics in assessing the mistranlation of emotion within the multilingual UGT framework. Section 5 briefly reviews relevant research which addressed the challenges in the automatic translation of sentiment. Section 6 presents a conclusion on our experiment, limitations of the present study and recommendations for future research work.

2 Data Collection and Experiment Setup

In order to check how far automatic translation captures the specific emotion in tweets, we replicated a real-life scenario where MT systems are utilised spontaneously to translate the content of tweets. Twitter currently supports built-in translations, so users can click on a *Translate Tweet* prompt visible directly under the tweet text to translate it. Twitter mentions that it employs Google Translate for this service. To evaluate how far the MT system in this scenario can serve as a real-life tool, we used Google Translate API to automatically translate existing multilingual Twitter datasets previously annotated for four emotions (joy, fear, aggression, and anger). It is important to note that these four emotions were chosen as representative of the common fine-grained sentiments expressed in tweets. The authors of tweets are usually either happy, angry, or fearful of something or someone, and their anger can either be aggressive or passive. The datasets were collated from different emotion-detection and aggression-detection shared tasks (Mohammad and Bravo-Marquez, 2017; Mohammad and Kiritchenko, 2018; Basile et al., 2019; Zampieri et al., 2020). The source datasets amounted to approximately 30,000 tweets in three languages: 23,000 in English, 4000 in Arabic, and 3000 in Spanish.

We created two datasets from this source annotated data by using the Google Translate API. The first dataset was created by translating the Arabic and Spanish source datasets into English. The second dataset was created by translating part of the source English dataset into Romanian, Arabic, Spanish and Portuguese. These datasets were used to extract instances for our analysis. The next stage in our experiment was to extract instances in which the MT system *may have failed* to translate the emotion correctly. We call this failure "mistranslation of emotion" and it is identified by the discrepancy between the annotations of emotion in the source dataset and the emotions classified in the translated tweets. For example, if the original tweet is annotated as conveying 'anger' but a classifier predicted 'joy' for the translation, this pair was considered a potential mistranslation of emotion and was selected for manual analysis.

To get the classifications of emotions in the translated tweets, we used the standard methodology employed in emotion classification. To this end, we built a classifier by fine-tuning a Roberta XLM model (Liu et al., 2019) on the previously annotated 23,000 source English tweets. This data was pre-processed by deletion of punctuation, non-alphanumeric symbols, lemmatisation, and lower-casing. We also used the Demoji[1] Python library to transfer the emojis into their equivalent lexicon (e.g. 😕 is translated into

[1] https://pypi.org/project/demoji/

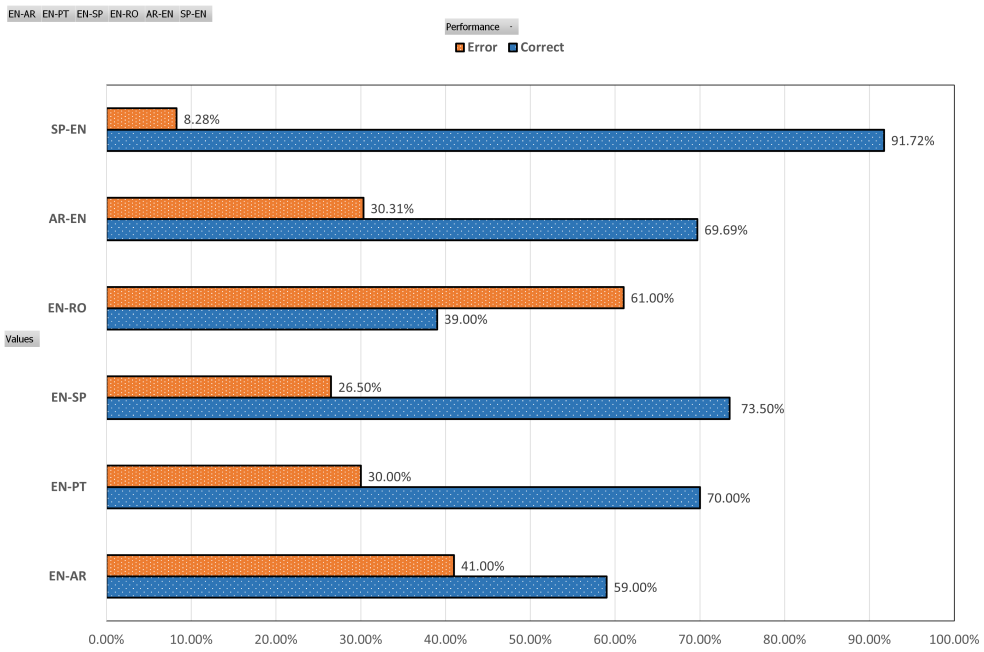


Figure 1: Frequency of Mistranslation of Emotion in the Analysed Dataset

"dislike"). The English emotion-detection model was trained on four epochs and fine-tuned with the following AdamW (Loshchilov and Hutter, 2017) optimiser hyperparameters: learning rate = $1 \times e^{-5}$, $\beta 1 = 0.9$, $\beta 2 = 0.98$ and $\epsilon = 1 \times e^{-8}$. We divided the dataset into 90% training and 10% validation set. The validation accuracy reached 92%. The English classifier was used to predict the emotion of the Google Translate output for the translation of the Arabic and Spanish dataset into English. For the experiments where English was the source language, we classified the back translation of the English tweets. Although the back translation may not be as accurate as the translated text, we opted for this compromise since ultimately the classifier's output will be manually compared to the source by human annotators. Thus, the classifier's predicted emotion was compared to the gold-standard emotion of the source text, instances of discrepancy were extracted as potential mistranslations of emotions.

3 Analysis of Challenging Features

To check the reasons for discrepancy between the predicted emotion and the emotion of the source text, a team of computational linguists who are native speakers of the analysed languages conducted a manual analysis on samples of the extracted potential mistranslations. The extracted samples for English to other languages amounted to 1600 tweets divided equally among the four target languages, and from the opposite direction, with English as a target language, they amounted to ≈3000 tweets divided between Arabic and Spanish as a source language. The disagreements in the dataset due to mistranslation of emotions are presented in Figure 1. Spanish has clearly fewer cases of discrepant emotions in tweets, both when these are translated into English (≈8%) and when they are translations from English (≈27%). Target languages like Romanian and Arabic show a much higher percentage of tweets with mistranslated emotions (61% and 41%, respectively). It is obvious from the analysed sample that some languages are more privileged than others in the real-life scenario we replicate for our experiment.

Next, we analysed in detail the linguistic features of instances where source tweet and translation have different emotion labels. The analysis showed that despite the unbalance in terms of MT accuracy among different language pairs as shown in Figure 1, there are common linguistic phenomena that cause distortion of emotion transfer among all the language pairs. Based on our analysis of the sample dataset, we selected the six features that the annotators found to be commonly constituting a challenge in transferring emotions by the MT engine for all the studied language directions. These linguistic features are: hashtags, slang, non-standard orthography, idiomatic expressions, polysemy,

Language Pair	Hashtags	Slang	Polysemy	Idiomatic Expressions	Grammar	Orthography
EN-ES	**44%**	14%	7.9%	6.3%	12.6%	14%
EN-PT	41.6%	16.6	2.7%	8.3%	**13.8%**	16.6%
EN-AR	25.6%	20.7%	**24.3%**	12%	6%	11%
EN-RO	24.6%	26%	18.6%	12%	6%	12.6%
AR-EN		**60%**	11%	7.9%	6.7%	13.9%
ES-EN		32.5%	16%	**16.5%**	12%	**22.6%**

Table 1: Frequency of Language Features per Language Pair

and grammar (especially negation structures). The following sections demonstrate the effect of these features on the translation of emotion with illustrative examples[2]. Table 1 presents a summary of our findings, which are discussed next. The following sections demonstrate these typical errors.

3.1 Hashtags

Emotions in tweets are expressed in a special style in line with Twitter's orthographic limitations and peculiarities. Thus, for example, authors of tweets frequently express their emotion as a trailing hashtag or a hashtagged non sequitur to a neutral or an ironic statement. The emotion of the tweets in such cases is retrieved solely from the hashtag. Our analysis has shown that this unique style of emotion transfer constitutes a challenge to the MT system. When the hashtags expressing emotion are either untranslated or mistranslated, the emotion expressed in the message is completely distorted. For example, the fear emotion in the English tweet "*Just waved daughter and her friend off to school, #terrifying!*" is entirely missed in the Arabic translation "لقد لوحت ابتها وصديقتها الصغيرة للتو إلى المدرسة #terrifying!" as the hashtag that carries the main emotional content is not translated.

Moreover, the hashtagged word in tweets is often written in non-standard orthography which causes the MT to output the hashtagged word as is without translation. For example, the anger emotion against customer service in the tweet "*I asked for my parcel to be delivered to a pickup store not my address #poorcustomerservice*" is missed in the Romanian translation "Am cerut livrarea coletului meu la un magazin de preluare, nu adresa mea #poorcustomerservice" as the hashtagged word is not translated. The MT treats such hashtags as out-of-vocabulary words and hence misses the affective message. The distortion of emotion is also caused by a wrong translation of the hashtagged word. The anger emotion of the English tweet "*CNN's Wolf Blitzer calls you an American astronaut and you don't correct him #disappointed*" is completely lost in the Spanish translation as the hashtag is mistranslated to "diseñado" meaning 'designed' instead of disappointed. The Spanish translation carries a neutral emotion. Almost 44% of the English hashtags in the dataset led to loss of the source emotion in the Spanish translations (see Table 1).

3.2 Slang and Dialectical Expressions

Research studies have shown that slang and dialectical expressions present several challenges to MT in general (Zbib et al., 2012; Saadany et al., 2022). Tweets are characterised by a wealth of slang expressions and code-switching between different dialects of one language based on the authors' demographics. It was observed from the manual analysis of the sample data that this stylistic quirk often distorts the translation of emotion in the source text. For example, the Spanish tweet "*Ni en pedo, bueno en pedo si*" is mistranslated in English as "*not even fart good fart yes*". The correct translation of the expression "*ni en pedo*" is "*no way*". The source tweet expressed a humorous comment which should read "*No way. Well, yes way*". In this example, the MT online engine provides an incomprehensible output due to a mistranslation of the dialectical version of the Spanish expression "*ni en pedo*" used mainly in Argentina, and therefore the emotion of the source text is completely lost.

Similarly, the MT system fails to detect the aggression in the English tweet "*The iconic **nigger** tweet*" when it is translated to Romanian as "*tweet-ul iconic negru*" (The iconic black tweet). The slang expression in the source tweet

[2]Due to space limitations, examples mentioned in this section are excerpts of tweets used for error analysis.

(nigger) carries the aggressive tone and hence the neutral translation (black) misses the aggressive emotion. By missing the racist slur, the Romanian translation wrongly transfers a positive/neutral emotion.

The amount of distortion of the affect message due to a mistranslation of slang or dialectical expressions varies from one language to another. It was observed that Arabic dialectical expressions posed a significant challenge to the MT system as it caused the flipping of the sentiment polarity of emotions in 60% of the Arabic tweets in the second dataset (see Table 1). For example, commenting on an event in the Middle East, a tweeter expresses joy "ايه كمية الانشكاح ديه" (What all this amount of happiness!). The MT system gives the exact opposite emotion "*What all this amount of anger!*". This owes to the fact that the dialectical expression "الانشكاح" (happiness) is mistranslated as "anger". The dialectical tweets were mostly mistranslated in aggressive Arabic tweets. For example, bullying a female football player, a tweet says "جايه بفستانها... تستلم جايزة افضل لاعبة خربوا الكورة الحريم" (She is coming with a dress to receive the best player prize..., women ruined football). The tweet is written in a Gulf dialect that was mistranslated by the MT engine as "*come to her dress and receive the prize for the best player who ruined the harem?*". The MT output misses the misogynist comment and transfers an overall 'joy' emotion despite the lack of semantic and grammatical coherence.

3.3 Non-standard Orthography

With its 280-character limit, Twitter users often have to resort to creative abbreviations and unconventional orthography. Moreover, linguists have observed that to encourage speed and immediacy of understanding, Twitter users type in the same way they speak (Ian, 2010). The manual analysis has shown that this specific linguistic phenomenon is a major culprit in a wrong transfer of the emotion within different language pairs. For example, the MT output of the English tweet "*watching sad bts video bc im sad. iwannacryy*" renders an incomprehensible affect message in Portuguese: "*assistindo ao vídeo do sad bts bc im sad. iwannacryy*". The reason is that the microblogging limitation causes the author of the tweet to use a creative word shortening by eliminating spaces "iwannacryy" as well as by texting in acronyms ("bc" meaning "because", "im" meaning "I am"). The affective message is missed in the Portuguese translation as all these emotional nuanced orthographic forms remain untranslated.

Another complication is that tweeters are more apt to use expressive lengthening to communicate strong emotion. These non-standard emotional expressions are usually treated as out-of-vocabulary by MT systems with all the language pairs the research team analysed. For instance, the anger in the Spanish tweet "*Por que sos re chantaaaa*" (Why are you such a liar?) is not transferred by the MT translation "*Why are you chantaaaa*" as the Spanish word "*chanta*" (liar) passes for out-of-vocabulary lexicon because of elongation. It is obvious that non-Spanish speakers would not understand the aggressive emotion in the Spanish tweet from the MT tool output.

3.4 Idiomatic Expressions

One of the challenging issues in the field of translation is the process of translating the different shades of meaning conveyed by an idiom (Al Mubarak, 2017). The reason is that translating idioms usually involves meta-linguistic information such as cultural and social norms. Because of their informal nature, conversational idioms are used extensively in tweets. The manual analysis has shown that a large number of idioms were literally translated, which did not only affect the sentiment preservation of the source text, but often produced nonsensical target text. For example, the Arabic tweet expressing joy in describing one particular public figure "والله دمه خفيف" has the idiomatic expression "دمه خفيف", meaning "funny". The tweet should read 'By God, he is so funny', but the MT output gives a literal translation, "*By God, his blood is so light*" which was predicted as having an 'anger' sentiment by our automatic classifier. The same problem also exists in language pairs with English as a source language. For example, an 'angry' tweet commenting on one of the candidates in the last American presidential elections – "*We have to keep u in line*" – has the idiomatic expressions "keep in line" meaning to discipline uncontrolled behaviour. This idiom was literally translated in Arabic as "في الطابور" (stay in the queue) and in Spanish as "*mantenerte en línea*" (stay fit). The

literal translation of the idiom in the two language pairs flips the emotion from anger to a neutral sentiment.

3.5 Polysemy

MT research has shown that polysemous words pose a challenge to MT systems when the contextual information is not clearly determined (Akhobadze, 2019). Due to the micro-blogging nature of tweets, polysemous words in tweets are usually lacking context. This adds to their ambiguity. The manual analysis of the translated data has revealed that this linguistic phenomenon distorts the tweeter's emotional message. One example is the aggressive English tweet "*the girl sitting in front of me is chewing her gum like a cow; I'm ready to snap*". The word snap here has the informal meaning of "burst in anger". The Romanian translation by the MT system, however, reflects a joy emotion as it gives the other meaning of snap "take pictures". Hence the MT Romanian output reads "*the girl in front of me chews her gum like a cow; I'm ready to take pictures*". Another more extreme example appears with the Arabic to English pair. Commenting on a Middle East political crisis, an aggressive tweeter threatens two Gulf countries "جايكم الدور خلينا نربيكم في اليمن ونربي قطر" (Your turn will come, Yemen and Qatar, we will teach you a lesson). The aggressive threat is lost in the MT output "*Come on let's educate Yemen and Qatar*". This is due to a mistranslation of the polysemous word "نربي" which could either positively mean "educate" or to negatively mean "teach a lesson" by inflicting punishment.

3.6 Grammar (Negation)

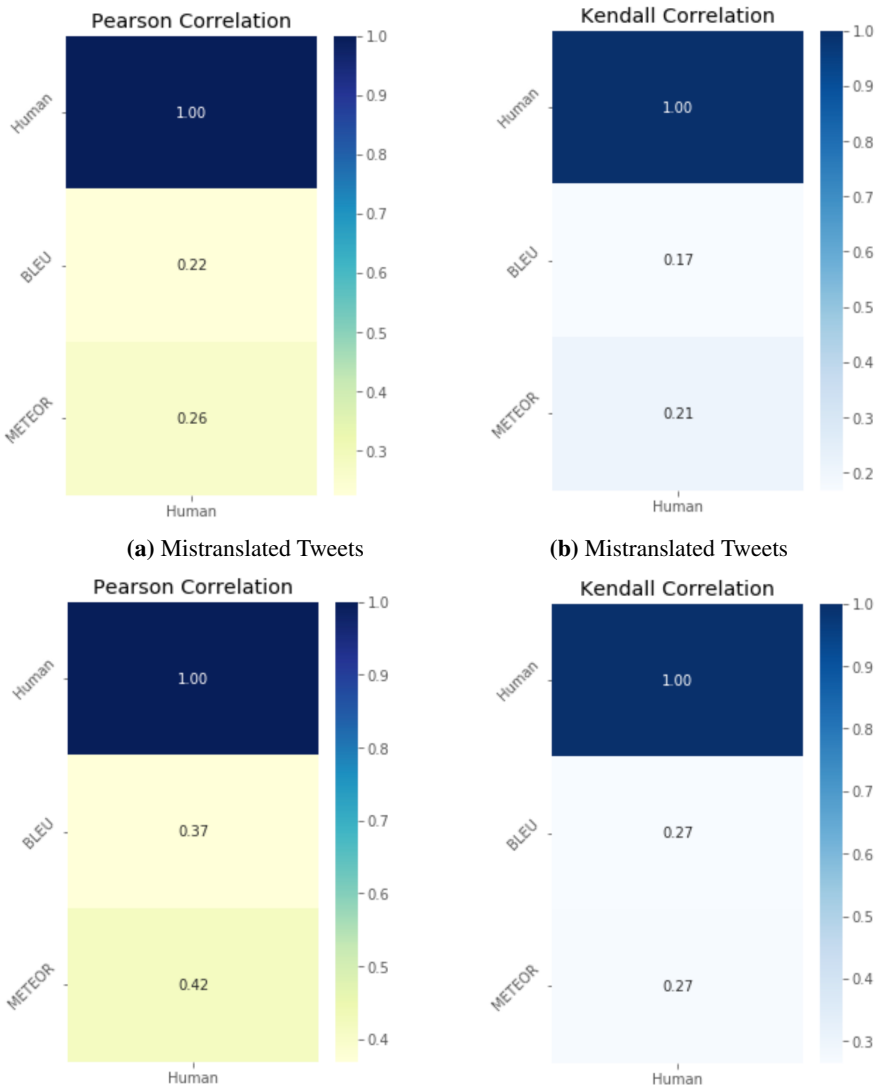
The analysis has shown that the distortion of the source emotion was also associated with a wrong translation of a negation marker between different language pairs. For example, the analysis has shown that missing negation structures in the English tweets distorts the emotion. The fear emotion in the English tweet "*A trip to the dentist never gets easier*" is flipped to joy in the Portuguese translation because of a wrong translation of the negative structure. The MT output in Portuguese is "*Uma ida ao dentista nunca foi tão fácil*" meaning "*A trip to the dentist has never been easier*". The emotion is not only distorted but the mistranslation fluently transfers the opposite affect message.

Moreover, negation was found to cause a problem when the source text is in dialectical Arabic. The lexico-grammatical realisation of negation differs between the Standard and dialectical Arabic as well as between its different dialects. Arabic dialects often treat negative particles as clitics, and hence a letter is added to the stem of the word to change it to negative (Mohamed et al., 2012; Mitkov and Angelova, 2021). The MT engine frequently either missed the Arabic dialectical negation and hence flips the phrase to an opposite sentiment pole or mistranslates the negated phrase altogether. For example, commenting on a terrorist attack, a tweeter angrily states "لا سامح الله من افتخر بقنبلة دمرت مئات المنازل" (May God not forgive (punish) the one who is proud of a bomb destroying hundreds of homes). The negation is missed and hence the online translation tool output reads "*May God forgive that one who is proud of a bomb destroying hundreds of homes*". The emotion of the tweeter in the Arabic translation is flipped from anger to sympathy towards a terrorist attack. If automatic translation is used to spot potential terrorist trends on social media platforms, this type of error would affect the accuracy of the algorithm and may bring dangerous consequences to users.

4 Measuring the Transfer of Sentiment

From the analysis of these language features, it can be observed that using automatic translation tools for translating emotion in multilingual UGT such as tweets involves several linguistic challenges. From our manual analysis, we found that such challenges can lead to a severe distortion of the source emotive message. However, despite these challenges, NMT systems such as Google Translate are extensively utilised by social media platforms without human post-editing. In the research environment, the reliability of MT systems is commonly determined by automatic quality metrics that are domain agnostic as they evaluate the translation accuracy regardless of the type of source text. In this section we assess how far the commonly used quality metrics are able to signal out critical mistranslation of emotions as the ones analysed in the previous sections.

The *de facto* standard for MT performance evaluation is the BLEU score with its different



(a) Mistranslated Tweets **(b)** Mistranslated Tweets

(c) Correctly Translated Tweets **(d)** Correctly Translated Tweets

Figure 2: Statistical correlation between BLEU and METEOR with segment-level human judgement

variations (Papineni et al., 2002). BLEU gives equal penalty weight to inaccurate translation of n-grams, which may lead to performance overestimation (or underestimation). For example, the "joy" emotion in the tweet from our Arabic dataset "*What is this amount of happiness!*" is flipped to anger by Twitter's Google Translate tool which outputs "*What is this amount of anger!*". Despite the distortion of the sentiment message, BLEU only mildly penalises the swapping of the two opposite emotive nouns 'happiness' and 'anger' and this translation receives a BLEU score of 0.76. The reason is that BLEU gauges the performance of an MT model by an indiscriminate n-gram matching, regardless of the semantic weight of each word. By human standards, the MT performance in such cases is highly over-estimated.

There have been numerous efforts to address the common pitfalls of n-gram-based metrics by incorporating semantic and contextual features in metrics specially when measuring the translation of sentiment (Saadany et al., 2021a). One very popular metric that has been introduced as a semantic-oriented metric is METEOR (Banerjee and Lavie, 2005). When it comes to evaluating an MT system performance in transferring emotion, even the semantically oriented automatic metrics do not give a penalty to a mistranslated sentiment proportional to the distortion it afflicts on the source message. For example, the negation in the Arabic tweet "*May God do not forgive those who put you in power*" is missed in the MT output: "*May God forgive the one who put you in*

power". The emotion is flipped from "anger" to "joy". Despite the distortion of the emotion, the mistranslation receives a METEOR score of 0.61.

To quantify the ability of the BLEU[3] and METEOR metrics to assess the transfer of emotion in translated tweets, we selected an evaluation dataset consisting of 300 tweets extracted from the Spanish and Arabic dataset that was classified as having a mistranslation of emotion in the English translations. The tweets in this dataset were chosen in a way where the main error in the translation is the distortion of emotion due to one of the six linguistic features discussed in the previous sections. We also created another evaluation dataset consisting of 100 tweet/translation pairs with the same language directions where the online MT tool transferred the correct emotion. The evaluation datasets were translated by native speakers in the research team. The translators were also asked to assign a score to each pair of source-translation tweet, where 1 is the poorest sentiment transfer and 10 is best sentiment transfer. The average scores of annotators were taken as the final human score. We compared the human scores of the mistranslated tweets and the correctly translated tweets with BLEU and METEOR scores of their translations. We followed the WMT standard methods for evaluating quality metrics and used absolute Pearson correlation coefficient r and the Kendall correlation coefficient $|\tau|$ to evaluate each metric's performance against the human judgement. Figures 2a and 2b show heatmaps visualising the Pearson and Kendall correlation coefficients for the mistranslated tweets, and Figures 2c and 2d show the coefficients of the studied metrics with the correctly translated tweets.

As seen from the Figures 2a, and 2b, with the mistranslated tweets BLEU score achieves only 0.22 and 0.17 Pearson and Kendall correlations, respectively. Similarly, METEOR records a Pearson correlation of 0.26 but a relatively lower Kendall correlation of 0.21. On the other hand, the correlation of the two metrics records (60%-68%) and (30%-60%) improvement on the correctly translated tweets for the Pearson and Kenall coefficients, respectively. Our results show that conventional metrics' performance seriously deteriorates with poor translation of emotions in tweets. Also, bearing in mind that the mistranslated tweets have critical translation errors that seriously distort the emotion, the low correlation results for the two metrics with the mistranslations dataset raise important doubts as to the reliability of these accepted metrics for ranking MT systems in terms of emotion transfer in UGT data such as tweets.

[3]We use the sacrebleu implementation of the BLEU score for all the experiments (Post, 2018).

5 Related Research

There has been a growing interest in analysing how far MT systems are capable of preserving the sentiment message, specifically in the automatic translation of tweets. Salameh et al.(2015) acknowledge the fact that aspects of sentiment may be lost in translation, especially in automatic translation of Arabic tweets. They show that the matching percentage between the manual sentiment annotation and an automatic sentiment annotation of the automatically translated dataset is 62.49% match as compared to the 68.65% match on a manually translated dataset.

Afli et al.(2017) propose a method to reduce the mistranslation of sentiment in Irish tweets. They manually expand the training data with an Irish-language sentiment lexicon when building an Irish-English MT system. The sentiment lexicon improves the sentiment accuracy of the translated text with an accuracy margin of 6%. Lohar et al.(2017) argue that machine translation of UGT becomes more difficult because of the level of noise it contains. Accordingly, the translation quality is affected in a way that may negatively impact sentiment preservation in the translation process. They show evidence of their analysis on a small dataset of 4000 English tweets and their translations in German. More recently, Saadany (2022) has shown that challenging features in tweets can lead to critical mistranslation of sentiment where the output of the MT system gives a deceivingly correct message that sometimes transfers a sentiment polarity opposite to the source tweet.

As for the evaluation of the output of online MT tools, there have been several studies that address the shortcomings of conventional quality metrics such as BLEU. For example, Mathur et al. (2020) points to the inconsistencies of BLEU as a parameterised metric since its score changes with a change of the parameters for tokenisation

and normalisation scheme. Saadany et al.(2021) demonstrate the inability of automatic metrics such as BLEU and METEOR to distinguish between a critical error that distorts the affect message in UGT data and a non-critical error where the MT affects the fluency of the source but still transfers the correct sentiment.

6 Conclusion

In this research, we evaluated the ability of the MT online system to translate fine-grained emotions in tweets between different language pairs. Our analysis has shown that there are linguistic features that are common among different language pairs which cause problems in translating tweets by NMT tools. More crucially, the manual analysis has shown that due to these linguistic challenges in tweets, the user of online MT tools may receive a fluent translation which deviates drastically from the sentiment of source in such a way that the reader would either understand the opposite sentiment or lose the sentiment all together. The error analysis presented in this paper, therefore, points to essential ethical issues that should be taken into consideration when adopting a fully automated translation technology to transfer users' stance on online platforms.

We also touched upon the reliability of automatic quality measures for evaluating MT systems performance in transferring emotion. We have shown that the standard evaluation measures were not able to give a penalty proportional to the incorrect translation of emotion in a sample dataset of mistranslated tweets. This points to the fact that critical mistranslation of emotions by online MT systems may go undetected if the performance is gauged by conventional metrics such as BLEU and METEOR. We believe that evaluating the performance of MT systems in translating sentiment-oriented text is an under-recognised problem in MT research. Future work should address the possibility of introducing a sentiment measure to reflect how far the MT system transfers the correct affective message in the source text as well as detect critical distortions of the source sentiment.

References

Abdalla, Mohamed and Graeme Hirst. 2017. Cross-lingual sentiment analysis without (good) translation. *arXiv preprint arXiv:1707.01626*.

Afli, Haithem, Sorcha Maguire, and Andy Way. 2017. Sentiment translation for low resourced languages: Experiments on Irish general election tweets. In *18th International Conference on Computational Linguistics and Intelligent Text Processing, Budapest, Hungry*, pages 17–21.

Akhobadze, Babulia. 2019. Polysemy in machine translation exemplified in English and Georgian. *Bull. Georg. Natl. Acad. Sci*, 13(1).

Al Mubarak, Amin Ali. 2017. The Challenges of Translating Idioms from Arabic into English A Closer Look at Al Imam AL Mahdi University–Sudan. *International Journal of Comparative Literature and Translation Studies*, 5(1):53–64.

Araujo, Matheus, Julio Reis, Adriano Pereira, and Fabricio Benevenuto. 2016. An evaluation of machine translation for multilingual sentence-level sentiment analysis. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, pages 1140–1145.

Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Barhoumi, Amira, Chafik Aloulou, Nathalie Camelin, Yannick Estève, and Lamia Belguith. 2018. Arabic sentiment analysis: an empirical study of machine translation's impact. In *Proceedings of the second Conference on Language Processing and Knowledge Management Kerkennah (Sfax), Tunisia*.

Basile, Valerio, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, Manuela Sanguinetti, et al. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics.

Demirtas, Erkin and Mykola Pechenizkiy. 2013. Cross-lingual polarity detection with machine translation. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, pages 1–8.

Ian, Tucker. 2010. Twitter spreads regional slang. `https://www.theguardian.com/science/2010/sep/05/tv-not-twitter-spreads-slang`.

Koehn, Philipp and Rebecca Knowles. 2017. Six challenges for neural machine translation. *ACL 2017*, page 28.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lohar, Pintu, Haithem Afli, and Andy Way. 2017. Maintaining sentiment polarity in translation of user-generated content. *The Prague Bulletin of Mathematical Linguistics*, 108(1):73–84.

Loshchilov, Ilya and Frank Hutter. 2017. Decoupled weight decay regularization. *In: 7th International Conference on Learning Representations (2017)*. http://arxiv.org/abs/1711.05101.

Mathur, Nitika, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997.

Mitkov, Ruslan and Galia Angelova. 2021. Proceedings of the international conference on recent advances in natural language processing (ranlp 2021). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*.

Mohamed, Emad, Behrang Mohit, and Kemal Oflazer. 2012. Transforming standard Arabic to colloquial Arabic. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 176–180.

Mohammad, Saif M and Felipe Bravo-Marquez. 2017. Wassa-2017 shared task on emotion intensity. *arXiv preprint arXiv:1708.03700*.

Mohammad, Saif and Svetlana Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Mohammad, Saif M, Mohammad Salameh, and Svetlana Kiritchenko. 2016. How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55:95–130.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Post, Matt. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Saadany, Hadeel and Constantin Orăsan. 2020. Is it Great or Terrible? Preserving Sentiment in Neural Machine Translation of Arabic Reviews. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 24–37.

Saadany, Hadeel and Constantin Orăsan. 2021. BLEU, METEOR, BERTScore: Evaluation of Metrics Performance in Assessing Critical Translation Errors in Sentiment-oriented Text. *TRITON 2021*, page 48.

Saadany, Hadeel, Constantin Orăsan, Emad Mohamed, and Ashraf Tantavy. 2021a. Sentiment-aware measure (SAM) for evaluating sentiment transfer by machine translation systems. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1217–1226.

Saadany, Hadeel, Constantin Orăsan, Rocio Caro Quintana, Felix do Carmo, and Leonardo Zilio. 2021b. Challenges in translation of emotions in multilingual user-generated content: Twitter as a case study. *arXiv preprint arXiv:2106.10719*.

Saadany, Hadeel, Constantin Orăsan, Emad Mohamed, and Ashraf Tantawy. 2022. A semi-supervised approach for a better translation of sentiment in dialectical arabic ugt. *arXiv preprint arXiv:2210.11899*.

Saadany, Hadeel. 2022. *A study of the translation of sentiment in User-Generated Text*. PhD dissertation, University of Wolverhampton.

Salameh, Mohammad, Saif Mohammad, and Svetlana Kiritchenko. 2015. Sentiment after translation: A case-study on Arabic social media posts. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 767–777.

Shalunts, Gayane, Gerhard Backfried, and Nicolas Commeignes. 2016. The impact of machine translation on sentiment analysis. *Data Analytics*, 63:51–56.

Zampieri, Marcos, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *arXiv preprint arXiv:2006.07235*.

Zbib, Rabih, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar Zaidan, and Chris Callison-Burch. 2012. Machine translation of arabic dialects. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 49–59.