Sentiment-based Candidate Selection For NMT

Alex Jones alexander.g.jones.23@dartmouth.edu
Dartmouth College, Hanover, NH, 03755, United States

Derry Tanti Wijaya wijaya@bu.edu
Department of Computer Science, Boston University, Boston, MA, 02215, United States

Abstract

The proliferation of user-generated content (UGC)—e.g. social media posts, comments, and reviews—has motivated the development of NLP applications tailored to these types of informal texts. Prevalent among these applications have been sentiment analysis and machine translation (MT). Grounded in the observation that UGC features highly idiomatic, sentiment-charged language, we propose a decoder-side approach that incorporates automatic sentiment scoring into the MT candidate selection process. We train monolingual sentiment classifiers in English and Spanish, in addition to a multilingual sentiment model, by fine-tuning BERT and XLM-RoBERTa. Using n-best candidates generated by a baseline MT model with beam search, we select the candidate that minimizes the absolute difference between the sentiment score of the source sentence and that of the translation, and perform two human evaluations to assess the produced translations. Unlike previous work, we select this minimally divergent translation by considering the sentiment scores of the source sentence and translation on a continuous interval, rather than using e.g. binary classification, allowing for more fine-grained selection of translation candidates. The results of human evaluations show that, in comparison to the open-source MT baseline model on top of which our sentiment-based pipeline is built, our pipeline produces more accurate translations of colloquial, sentiment-charged source texts[1].

1 Introduction

The Web, widespread internet access, and social media have transformed the way people create, consume, and share content, resulting in the proliferation of user-generated content (UGC). UGC—such as social media posts, comments, and reviews—has proven to be of paramount importance both for users and organizations/institutions (Pozzi et al., 2016). As users enjoy the freedoms of sharing their opinions in this relatively unconstrained environment, corporations can analyze user sentiments and extract insights for their decision-making processes, (Timoshenko and Hauser, 2019) or translate UGC to other languages to widen the company's scope and impact. For example, Hale (2016) shows that translating UGC between certain language pairs has beneficial effects on the overall ratings customers gave to attractions and shows on TripAdvisor, while the absence of translation hurts ratings. However, translating UGC comes with its own challenges that differ from those of translating well-formed documents like news articles. UGC is shorter and noisier, characterized by idiomatic and colloquial expressions (Pozzi et al., 2016). Translating idiomatic expressions is hard, as they often convey figurative meaning that cannot be reconstructed from the meaning of their parts (Wasow et al., 1983), and remains one of the open challenges in machine translation (MT) (Fadaee et al., 2018). Idiomatic expressions, however, typically carry an additional property: they imply an affective stance rather

[1] Code and reference materials are available at `https://github.com/AlexJonesNLP/SentimentMT`

than a neutral one (Wasow et al., 1983). The sentiment of an idiomatic expression, therefore, can be a useful signal for translation. In this paper, we hypothesize that a good translation of an idiomatic text, such as those prevalent in UGC, should be one that retains its underlying sentiment, and explore the use of textual sentiment analysis to improve translations.

Our motivation behind adding sentiment analysis model(s) to the NMT pipeline are several. First, with the sorts of texts prevalent in UGC (namely, idiomatic, sentiment-charged ones), the sentiment of a translated text is often arguably as important as the quality of the translation in other respects, such as adequacy, fluency, grammatical correctness, etc. Second, while a sentiment classifier can be trained particularly well to analyze the sentiment of various texts—including idiomatic expressions (Williams et al., 2015)—these idiomatic texts may be difficult for even state-of-the-art (SOTA) MT systems to handle consistently. This can be due to problems such as literal translation of figurative speech, but also to less obvious errors such as truncation (i.e. failing to translate crucial parts of the source sentence). Our assumption however, is that with open-source translation systems such as OPUS MT[2], the correct translation of a sentiment-laden, idiomatic text often lies somewhere lower among the predictions of the MT system, and that the sentiment analysis model can help signal the right translation by re-ranking candidates based on sentiment. Our contributions are as follows:

- We explore the idea of choosing translations that minimize source-target sentiment differences on a continuous scale (0-1). Previous works that addressed the integration of sentiment into the MT process have treated this difference as a simple polarity (i.e., positive, negative, or neutral) difference that does not account for the degree of difference between the source text and translation.
- We focus in particular on idiomatic, sentiment-charged texts sampled from real-world UGC, and show, both through human evaluation and qualitative examples, that our method improves a baseline MT model's ability to select sentiment-preserving *and* accurate translations in notable cases.
- We extend our method of using monolingual English and Spanish sentiment classifiers to aid in MT by substituting the classifiers for a single, multilingual sentiment classifier, and analyze the results of this second MT pipeline on the lower-resource English-Indonesian translation, illustrating the generalizability of our approach.

2 Related Work

Several papers in recent years have addressed the incorporation of sentiment into the MT process. Perhaps the earliest of these is Sennrich et al. (2016), which examined the effects of using honorific marking in training data to help MT systems pick up on the T-V distinction (e.g. informal *tu* vs. formal *vous* in French) that serves to convey formality or familiarity. Si et al. (2019) used sentiment-labeled sentences containing one of a fixed set of sentiment-ambiguous words, as well as valence-sensitive word embeddings for these words, to train models such that users could input the desired sentiment at translation time and receive the translation with the appropriate valence. Lastly, Lohar et al. (2017, 2018) experimented with training sentiment-isolated MT models—that is, MT models trained on only texts that had been pre-categorized into a set number of sentiment classes i.e., positive-only texts or negative-only texts. Our approach is novel in using sentiment to re-rank candidate translations of UGC in an MT pipeline and in using precise sentiment scores rather than simple polarity matching to aid the translation process.

In terms of sentiment analysis models of non-English languages, Can et al. (2018) experimented with using an RNN-based English sentiment model to analyze the sentiment of texts translated into English from other languages, while Balahur and Turchi (2012) used SMT to

[2] https://github.com/Helsinki-NLP/Opus-MT

generate sentiment training corpora in non-English languages. Dashtipour et al. (2016) provides an overview and comparison of various techniques used to tackle multilingual sentiment analysis.

As for MT candidate re-ranking, Hadj Ameur et al. (2019) provides an extensive overview of the various features and tools that have been used to aid in the candidate selection process, and also proposes a feature ensemble approach that doesn't rely on external NLP tools. Others who have used candidate selection or re-ranking to improve MT performance include Shen et al. (2004) and Yuan et al. (2016). To the best of our knowledge, however, no previous re-ranking methods have used sentiment for re-ranking despite findings that MT often alters sentiment, especially when ambiguous words or figurative language such as metaphors or idioms are present or when the translation exhibits incorrect word order (Mohammad et al., 2016).

3 Models and Data

3.1 Sentiment Classifiers

For the first portion of our experiments, we train monolingual sentiment classifiers, one for English and another for Spanish. For the English classifier, we fine-tune the BERT Base uncased model (Devlin et al., 2019), as it achieves SOTA or nearly SOTA results on various text classification tasks. We construct our BERT-based sentiment classifier model using BERTForSequenceClassification, following McCormick and Ryan (2019). For our English training and development data, we sample 50K positive and 50K negative tweets from the automatically annotated sentiment corpus described in Go et al. (2009) and use 90K tweets for training and the rest for development. For the English test set, we use the human-annotated sentiment corpus also described in Go et al. (2009), which consists of 359 total tweets after neutral-labeled tweets are removed. We use BertTokenizer with 'bert-base-uncased' as our vocabulary file and fine-tune a BERT model using one NVIDIA V100 GPU to classify the tweets into positive or negative labels for one epoch using the Adam optimizer (Kingma and Ba, 2014) with weight decay (AdamW in PyTorch) and a linear learning rate schedule with warmup. We use a batch size of 32, a learning rate of 2e-5, and an epsilon value of 1e-8 for Adam. We experiment with all hyperparameters manually, but find that the model converges very quickly (i.e. additional training after one epoch improves test accuracy negligibly, or causes overfitting). We achieve an accuracy of 85.2% on the English test set.

For the Spanish sentiment classifier, we fine-tune XLM-RoBERTa Large, a multilingual language model that has been shown to significantly outperform multilingual BERT (mBERT) on a variety of cross-lingual transfer tasks (Conneau et al., 2020), also using one NVIDIA V100 GPU. We construct our XLM-RoBERTa-based sentiment classifier model again following McCormick and Ryan (2019). The Spanish training and development data were collected from Mozetič et al. (2016). After removing neutral tweets, we obtain roughly 27.8K training tweets and 1.5K development tweets. The Spanish test set is a human-annotated sentiment corpus[3] containing 7.8K tweets, of which we use roughly 3K after removing neutral tweets and evening out the number of positive and negative tweets. We use the XLMRobertaTokenizer with vocabulary file 'xlm-roberta-large' and fine-tune the XLM-RoBERTa model to classify the tweets into positive or negative labels. The optimizer, epsilon value, number of epochs, learning rate, and batch size are the same as those of the English model, determined via experimentation (without grid search or a more regimented method). Unlike with the English model, we found that fine-tuning the Spanish model sometimes produced unreliable results, and so employ multiple random restarts and select the best model, a technique used in the original BERT paper (Devlin et al., 2019). The test accuracy on the Spanish model was 77.8%.

[3] https://www.kaggle.com/c/spanish-arilines-tweets-sentiment-analysis

3.2 Baseline MT Models

The baseline MT models we use for both English-Spanish and Spanish-English translation are the publicly available Helsinki-NLP/OPUS MT models released by Hugging Face and based on Marian NMT (Tiedemann and Thottingal, 2020; Junczys-Dowmunt et al., 2018; Wolf et al., 2019). Namely, we use both the en-ROMANCE and ROMANCE-en Transformer-based models, which were both trained using the OPUS dataset (Tiedemann, 2017)[4] with Sentence Piece tokenization and using training procedures and hyperparameters specified on the OPUS MT Github page[5] and in Tiedemann and Thottingal (2020).

4 Method: Sentiment-based Candidate Selection

We propose the use of two language-specific sentiment classifiers (which, as we will describe later in the paper, can be reduced to one multilingual sentiment model)—one applied to the input sentence in the source language and another to the candidate translation in the target language—to help an MT system select the candidate translation that diverges the least, in terms of sentiment, from the source sentence.

Using the baseline MT model described in Section 3.2, we first generate $n = 10$ best candidate translations using a beam size of 10 at decoding time. We decided on 10 as our candidate number based on the fact that one can expect a relatively low drop off in translation quality with this parameter choice (Hasan et al., 2007), while also maintaining a suitably high likelihood of getting variable translations. Additionally, decoding simply becomes too slow in practice beyond a certain beam size.

Once our model generates the 10 candidate translations for a given input sentence, we use the sentiment classifier trained in the appropriate language to score the sentiment of both the input sentence and each of the translations in the interval $[0, 1]$. To compute the sentiment score $S(x)$ for an input sentence x, we first compute a softmax over the array of logits returned by our sentiment model to get a probability distribution over all m possible classes (here, $m = 2$, since we only used positive- and negative-labeled tweets). Representing the negative and positive classes using the values 0 and 1, respectively, we define $S(x)$ to be the expected value of the class conditioned on x, namely $S(x) = \sum_{n=1}^{m} P(c_n \mid x)\ v_n$, where c_i is the ith class and v_i is the value corresponding to that class. In our case, since we have only two classes and the negative class is represented with value 0, $S(x) = P(\text{positive class} \mid x)$. After computing the sentiment scores, we take the absolute difference between the input sentence x's score and the candidate translation t_i's score for $i = 1, 2, ..., 10$ to obtain the *sentiment divergence* of each candidate. We select the candidate translation that minimizes the sentiment divergence, namely $y = \text{argmin}_{t_i} |S(t_i) - S(x)|$. Our method of selecting a translation differs from previous works in our use of the proposed sentiment divergence, which takes into account the degree of the sentiment difference (and not just polarity difference) between the input sentence and the candidate translation.

5 Experiments

5.1 English-Spanish Evaluation Data

The aim of our human evaluation was to discover how Spanish-English bilingual speakers assess both the quality and the degree of sentiment preservation of our proposed sentiment-sensitive MT model's translations in comparison to those of the human (a professional translator), the baseline MT model (Helsinki-NLP/OPUS MT), and a SOTA MT model, namely Google Translate.

[4] http://opus.nlpl.eu

[5] https://github.com/Helsinki-NLP/OPUS-MT-train

The human evaluation data consisted of 30 English (*en*) tweets, each translated using the above four methods to Spanish. We sample 30 English tweets from the English sentiment datasets that we do not use in training (Section 3.1) as well as from another English sentiment corpus (CrowdFlower, 2020)[6]. In assembling this evaluation set, we aimed to find a mix of texts that were highly idiomatic and sentiment-loaded—and thus presumably difficult to translate—but also ones that were more neutral in affect, less idiomatic, or some combination of the two.

5.2 English-Spanish Evaluation Setup

For the English-Spanish evaluation, we hired two fully bilingual professional translators using contracting site Freelancer [7]. Both evaluators were asked to provide proof of competency in both languages beforehand. The evaluation itself consisted of four translations (one generated by each method: human, baseline, sentiment-MT, Google Translate) for each of the 30 English tweets above, totaling 120 texts to be evaluated. For each of these texts, evaluators were asked to:

1. Rate the *accuracy* of the translation on a **0-5** scale, with 0 being the worst quality and 5 being the best
2. Rate the *sentiment divergence* of the translation on a **0-2** scale, with 0 indicating no sentiment change and 2 indicating sentiment reversal
3. Indicate the reasons for which they believe the sentiment changed in translation

5.3 English-Spanish Evaluation Results

As depicted in Table 1, the results of the English-Spanish human evaluation show improvements across the board for our modified pipeline over the vanilla baseline model. For the purposes of analysis, we divide the 30 English sentences (120 translations) into two categories: "all" (consisting of all 120 translations) and "idiomatic," consisting of 13 sentences (52 translations) deemed particularly idiomatic in nature. Although methods exist for identifying idiomatic texts systematically, e.g. Peng et al. (2014), we opt to hand-pick idiomatic texts ourselves. We do this in hopes of curating not only texts that contain idiomatic "multi-word" expressions, but also ones that are idiomatic in less concrete ways, which will enable us to gain more qualitative insights in the evaluation. Examples of such sentences are discussed in Section 7.

In the 'all' subset of the data, we see a +0.12 gain for our modified pipeline over the baseline in terms of accuracy (where higher accuracy is better), as well as a +0.11 reduction in sentiment divergence (where smaller divergence is better). On the idiomatic subset, the differences are more pronounced: we see a +0.80 gain over the baseline for accuracy and a +0.35 reduction in sentiment divergence. While our pipeline lags behind Google Translate in all metrics for English-Spanish—due to the superiority of Google Translate over OPUS MT in multiple regards (training data size, parameters, multilinguality, compute power, etc.)—our modification moves OPUS MT closer to this SOTA system. As a benchmark and to validate the soundness of our evaluation set, we include results for translations performed by a professional human translator, which, as expected, are vastly superior to those for any of the NMT systems used across all metrics and subsets of the data.

We also provide qualitative insights gained from the evaluations, in which evaluators were asked to identify *why* they believe the sentiment of the text *per se* changed in translation. The codes corresponding to these qualitative results are listed in the rightmost column of Table 1, and may be identified as follows:

- "MI" indicates the Mistranslation of Idiomatic/figurative language *per se*

[6] https://data.world/crowdflower/apple-twitter-sentiment

[7] https://www.freelancer.com/

	BLEU (Tatoeba)	BLEU (all tweets)	BLEU (idiom. tweets)	Accuracy (all tweets)	SentiDiff (all tweets)	Accuracy (idiom. tweets)	SentiDiff (idiom. tweets)	Top-3 Qual.
Baseline								
en→es	31.37	38.93	39.28	2.06	0.92	1.37	1.23	MI, O, MO
en→id	31.17	–	–	2.98	0.77	2.50	1.00	MO, O, MI
SentimentMT								
en→es	22.15	39.10	43.47	2.18	0.81	2.17	0.88	MO, IG, MI
en→id	20.85	–	–	3.31	0.65	3.20	0.64	MO, O, MI
Google Transl.								
en→es	51.39	56.76	57.98	3.08	0.43	2.31	0.79	MI, MO, O
en→id	33.93	–	–	3.57	0.55	3.00	0.94	MO, MI, O/IR
Human								
en→es	100	100	100	4.28	0.10	4.44	0.08	MO, O, IR

Table 1: The BLEU scores on the Tatoeba dataset, the accuracy and sentiment divergence scores on Twitter data, and the top 3 reasons given for sentiment divergence for each translation method, language pair, and chosen subset of the Twitter data: all vs. idiomatic. *en→es* represents English-Spanish, and *en→id* represents English-Indonesian. Note that ratings for each language are given by different sets of evaluators, and shouldn't be compared on a cross-lingual basis.

- "MO" indicates the Mistranslation of Other types of language
- "IG" indicates Incorrect Grammatical structure in the translation
- "IR" indicates IRrecoverability of the source text's meaning, i.e. even the gist of the sentence was gone
- "LT" indicates a Lack of Translatability of the source text to the language in question
- "O" indicates some Other reason for sentiment divergence

The top three most frequently cited causes of sentiment divergence for both the baseline and Google Translate were mistranslation of idiomatic language *per se*, mistranslation of other types of language, and other reasons not listed on the evaluation form. For our modified pipeline, the only distinctive top three cause of sentiment divergence was incorrect grammatical structure in the translation; additionally, one human translation was surprisingly flagged as rendering the source text's meaning "irrecoverable." However, the actual *frequency* of these error codes varied among models. For instance, 'MO' was given 5 times to human translations but 13 times to the baseline model's, and 'O' was given 3 times to Google Translate's translations and 7 times to our pipeline's. Some translations flagged with the '**O**ther' category are deemed to be of special interest and are discussed in Section 7.

We also noted strong and statistically significant ($p << 0.05$) negative correlations between accuracy and sentiment divergence for both the whole and idiomatic subsets of the data; the values of Pearson's *r* (Lewis-Beck et al., 2004) with their corresponding p-values are reported in Table 2.

Additionally, we measure agreement between the two English-Spanish evaluators using Krippendorff's inter-annotator agreement measure α (Krippendorff, 2011), which we choose as a metric in order to compare with previous work examining human agreement on sentiment judgments. In line with Provoost et al. (2019)'s findings of moderate agreement ($\alpha = 0.51$), we see α values ranging from 0.638 to 0.673 for the whole and idiomatic subsets of the data, respectively.

	Pearson's r (p-value) (all)	Pearson's r (p-value) (idiom.)
en→es	-0.764 (3.42e-47)	-0.759 (9.90e-21)
en→id	-0.570 (1.09e-15)	-0.756 (8.67e-14)

Table 2: Pearson's correlation coefficient and corresponding p-value with respect to accuracy and SentiDiff for each of the evaluations, broken down into the full (all) and idiomatic subsets.

In terms of automatic MT evaluation, we note that although our method causes a decrease in BLEU score on the Tatoeba test data for both languages (Table 1: SentimentMT vs. Baseline)—which is to be expected, as Tatoeba consists of "general" texts as opposed to UGC, and we select potentially non-optimal candidates during re-ranking—our method *improves* over the baseline for the Spanish tweets (and more so on the idiomatic tweets) on which the human evaluation was conducted. This result supports the efficacy of our model in the context of highly-idiomatic, affective UGC, and highlights the different challenges that UGC presents in comparison to more "formal" text.

Google Translate still outperforms the baseline and our method in terms of BLEU score on Tatoeba and the tweets. The explanation here is simply that the baseline model is not SOTA, which is to be expected given it's a free, flexible, open-source system. However, as our pipeline is orthogonal to any MT model, including SOTA, it could be used to improve a SOTA MT model for UGC.

6 Method Extension

6.1 Translation with Multilingual Sentiment Classifier

As highlighted in Hadj Ameur et al. (2019), one of the major criticisms of decoder-side re-ranking approaches for MT is their reliance on language-specific external NLP tools, such as the sentiment classifiers described in Section 3.1. To address the issue of language specificity and to develop a sentiment analysis model that can be used in tandem with MT between any two languages, we develop a multilingual sentiment classifier following Misra (2020). Specifically, we fine-tune the XLM-RoBERTa model using the training and development data used to train the English sentiment classifier, and the same tokenizer, vocabulary file, hyperparameters, and compute resources (GPU) used in training the Spanish classifier. We then use this multilingual language model fine-tuned on English sentiment data to perform zero-shot sentiment classification on various languages, and incorporate it into our beam search candidate selection pipeline for MT.

We test the model using the same test data used previously. On the English test data, this multilingual model achieves an accuracy of 83.8%, comparable to the accuracy score achieved using the BERT monolingual model (85.2%). On the Spanish test set, the multilingual model achieves a somewhat lower score of 73.6% (*cf.* 77.8% for the monolingual trained model), perhaps showing the limitations of this massively multilingual model on performing zero-shot downstream tasks.

6.2 English-Indonesian Evaluation Setup

We use the multilingual sentiment classifier in our sentiment-sensitive MT pipeline to perform translations on a handful of languages; examples from this experimentation are displayed in Tables 4 and 5 in the appendix.

We perform another human evaluation, this time involving English→Indonesian translations in place of English→Spanish. We choose Indonesian, as it is a medium-resource language (unlike Spanish, which is high-resource) (Joshi et al., 2020), and because we were able to obtain two truly bilingual annotators for this language pair.

The setup of the evaluation essentially mirrors that of the *en→es* evaluation, except we don't obtain professional human translations as a benchmark for Indonesian, due to the difficulty of obtaining the quality of translation required. Thus, the resulting evaluation set contains only $30 * 3 = 90$ translations instead of 120.

6.3 English-Indonesian Evaluation Results

The accuracy and sentiment divergence averages for different subsets of the *en-id* data are located in Table 1, and we direct readers to Section 5.3 for a qualitative discussion of these results. Quantitatively, we observe that our modified model outperforms the baseline in accuracy and sentiment divergence on every subset of the *en-id* data, while being comparable or better than Google Translate on the "all" and idiomatic subsets, respectively (Table 1). Specifically, on the "all" subset we see reductions of +0.33 and +0.12 over the baseline for accuracy and sentiment divergence, respectively, and on the idiomatic subset we see respective reductions of +0.70 and +0.36. Google Translate achieves slightly better accuracy and sentiment preservation overall (+0.26 and +0.10 over our pipeline for accuracy and sentiment divergence, respectively), but lags behind our pipeline in the idiomatic category (-0.20 and -0.30 for accuracy and sentiment divergence, respectively, compared to our pipeline).

Qualitatively, we see very similar reasons listed for sentiment divergence as we did for English-Spanish: each of the NMT systems we looked at had errors most frequently in the MI, MO, and O categories, denoting mistranslation of idiomatic language, mistranslation of other types of language, and other reasons for sentiment divergence, respectively; with MO being more frequent than MI in English-Indonesian evaluations, potentially due to lower MT performances for this language than Spanish (i.e., BLEU score for English-Indonesian modified model is 20.85 on the Tatoeba dataset compared to 22.15 for English-Spanish). However, as noted in the analysis of the previous evaluation, not all of these errors occurred with equal frequency across systems. For instance, Google Translate and the human translator produced less errors overall than the OPUS MT system, so the error codes should be interpreted as indicating the relative frequency and prevalence of certain translation errors that affect sentiment, not as markers to be compared on a system-to-system basis. As with the English-Spanish evaluation, certain qualitative observations made by our evaluators will be discussed further in Section 7. In line with results on the previous evaluation, accuracy and sentiment divergence are shown to be strongly negatively correlated, with Pearson's *r* values of -0.570 and -0.756 for the whole and idiomatic subsets of the data, respectively, both of which are statistically significant ($p << 0.05$) and are displayed in Table 2.

	acc. (all)	SentiDiff (all)	acc (idiom.)	SentiDiff (idiom.)
en→es	0.675	0.638	0.767	0.673
en→id	0.661	0.516	0.612	0.541

Table 3: Values of Krippendorff's alpha agreement measure α for both sets of evaluations with respect to accuracy ("acc.") and sentiment divergence ("SentiDiff") across different subsets.

Table 3 shows Krippendorff's alpha agreement measure (Krippendorff, 2011) for accuracy and sentiment divergence across both subsets, indicating moderate agreement, with higher agreement on accuracy. As was found with the English-Spanish evaluation, this is in line with previous findings of moderate human agreement on sentiment judgement (Krippendorff's α=0.51) (Provoost et al., 2019).

7 Discussion

Our experimentation with the various MT models generated a number of interesting example cases concerning the translation of idiomatic language. For example, given the tweet "Time Warner Road Runner customer support here absolutely blows," the baseline MT gives a literal translation of the word "blows" as "pukulan" (literally, "hits") in Indonesian; Google Translate gives a translation "hebat" ("awesome") that is opposite in sentiment to the idiomatic sense of the word "blows" ("sucks") in English; and our model gives a translation closest in meaning and sentiment to "blows," namely "kacau" (approx. "messed up" in Indonesian). There are also cases where our model gives a translation that is closer in degree of sentiment than what Google Translate produces. Given the source text "Yo @Apple fix your shitty iMessage," Google Translate produces "Yo @Apple perbaiki iMessage *buruk* Anda" ("Yo @Apple fix your *bad* iMessage"), which has roughly the same polarity as the source tweet. By contrast, our proposed model produces "Yo @Apple perbaiki imessage *menyebalkan* Anda," using the word "menyebalkan" ("annoying") instead of "buruk," which conveys a closer sentiment to "shitty" than simply "bad".

The evaluators of the English-Spanish translations provided us with rich qualitative commentary as well. For the sentence "Just broke my 3rd charger of the month. Get your shit together @apple," which is translated by the professional translator as "Se acaba de romper mi tercer cargador del mes. Sean más eficientes @apple," one evaluator acutely notes that "The expression 'Get your shit together' was translated in a more formal way (it loses the vulgarism). I would have translated it as 'Poneos las pilas, joder' to keep the same sentiment. We could say that this translation has a different diaphasic variation than the source text." This demonstrates that sentiment preservation is a problem not only for NMT systems, but for human translators as well. There are also problems attributed to challenges in machine translating informal texts. Acronyms such as "tbh" and "smh" made for another interesting case, as they weren't translated by any of the MT models for any language pairing, despite their common occurrence in UGC. The same evaluator also notes that "The acronym 'tbh' was not translated" in the sentence "@Apple tbh annoyed with Apple's shit at the moment," and says "this acronym is important for the sentiment because it expresses the modality of the speaker." In another example, we see our sentiment-sensitive pipeline helping the baseline distinguish between such a semantically fine-grained distinction as that between "hope" and "wish": the baseline translates the sentence "@Iberia Ojalá que encuentres pronto tu equipaje!!" as "@Iberia I *wish* you'd find your luggage soon!!," while our pipeline correctly chooses "@Iberia I *hope* you will find your luggage soon!!." We observe similar issues contribute to sentiment divergence in Spanish and Indonesian despite the fact that these are typologically disparate languages with different amounts of training data in the MT system.

In terms of automatic MT evaluation, our method improves over the baseline for the Spanish tweets on which the human evaluation was conducted. This result supports the efficacy of our model in the context of highly-idiomatic, affective UGC. And while Google Translate still outperforms the baseline and our pipeline in terms of BLEU score on Tatoeba (for both languages) and the tweets (for which only Spanish had a gold-standard benchmark)–given that the baseline model that we built our pipeline on is not SOTA–our pipeline can be added to any MT system and can also improve SOTA MT for UGC.

Furthermore, our approach also lends itself to many practical scenarios, e.g. companies who are interested in producing sentiment-preserving translations of large bodies of UGC but who lack the sufficient funds to use a subscription API like Google Cloud Translation. In these contexts, it may be beneficial—or even necessary—to improve free, open-source software in a way that is tailored to one's particular use case (thus the idea of "customized MT" that many companies now offer), instead of opting for the SOTA but more costly software.

More generally, since our approach shows that we can improve performance of an MT model for a particular use case i.e., UGC translation using signals beyond translation data that is relevant for the task at hand i.e., sentiment, it will be interesting to explore other signals that are relevant for improving MT performance in other use cases. It will also be interesting to explore the addition of these signals in a pipeline (our current method), as implicit feedback such as in Wijaya et al. (2017), or as explicit feedback in an end-to-end MT model for example, as additional loss terms in supervised (Wu et al., 2016), weakly-supervised (Kuwanto et al., 2021), or unsupervised (Artetxe et al., 2017) MT models. Beyond the potential engineering contribution for low-resource, budget-constrained settings, our experiments also offer rich qualitative insights regarding the causes of sentiment change in (machine) translation, opening up avenues to more disciplined efforts in mitigating and exploring these problems.

8 Conclusion

In this paper, we use several distinct sentiment classifiers trained on Twitter data to help machine translation models select sentiment-preserving translations of highly idiomatic source texts. Diverging from previous works, we use continuous (rather than binary or categorical) sentiment scores to select minimally divergent translations, and we test the performance of our pipeline with automated and human evaluations for English-Spanish and English-Indonesian translations.

Furthermore, we implement our sentiment-aware translation pipeline on free, open-source MT models available on Hugging Face[8]. Although many of these models are non-SOTA, our choice to use them represents a real-world scenario: Many users and companies do not have the resources or budget to subscribe to a SOTA translation API or train their own MT model from scratch. Our pipeline poses a lightweight solution for getting more with less, in a somewhat niche yet ubiquitous translation context (social media posts).

In future work, we would like to evaluate the effect of sentiment classifier performance on the downstream MT results, including the effects of classifier architecture, the number of sentiment categories and their distribution in the training data (e.g., UGCs with more informal words may contain more affective texts), etc. We would also like to investigate how continuous sentiment scoring compares with binary or categorical scoring for this task, using a larger evaluation set for idiomatic texts (e.g. in English (Michel and Neubig, 2018) or constructed in other languages (Wibowo et al., 2021)), or from a dataset we create ourselves. Finally, further work should establish benchmarks and put forth improvements for cross-lingual sentiment classification (i.e. the extent to which sentences that are translations of each other are assigned similar sentiments)—including the problem of zero-shot transfer—adding onto recent work in cross-lingual performance benchmarks (Hu et al., 2020; Liang et al., 2020).

References

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2017). Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

[8] https://huggingface.co/Helsinki-NLP

Balahur, A. and Turchi, M. (2012). Multilingual sentiment analysis using machine translation? In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 52–60, Jeju, Korea. Association for Computational Linguistics.

Can, E. F., Ezen-Can, A., and Can, F. (2018). Multilingual sentiment analysis: An RNN-based framework for limited data. *CoRR*, abs/1806.04511.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

CrowdFlower (2020). Apple Twitter sentiment. Online. Data.world dataset. Accessed 21 August 2020.

Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A. Y. A., Gelbukh, A., and Zhou, Q. (2016). Multilingual sentiment analysis: State of the art and independent comparison of techniques. *Cognitive Computation*, 8:757–771.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fadaee, M., Bisazza, A., and Monz, C. (2018). Examining the tip of the iceberg: A data set for idiom translation. *arXiv preprint arXiv:1802.04681*.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *Processing*, 150.

Hadj Ameur, M., Guessoum, A., and Meziane, F. (2019). Improving Arabic neural machine translation via n-best list re-ranking. *Machine Translation*, 33:1–36.

Hale, S. A. (2016). User reviews and language: how language influences ratings. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1208–1214.

Hasan, S., Zens, R., and Ney, H. (2007). Are very large N-best lists useful for SMT? In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 57–60, Rochester, New York. Association for Computational Linguistics.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020). XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization. *arXiv e-prints*, page arXiv:2003.11080.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. *arXiv preprint arXiv:2004.09095*.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Kingma, D. P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv e-prints*, page arXiv:1412.6980.

Krippendorff, K. (2011). Computing Krippendorff's Alpha-Reliability. *Annenberg School of Communication Scholarly Commons*.

Kuwanto, G., Akyürek, A. F., Tourni, I. C., Li, S., and Wijaya, D. (2021). Low-resource machine translation for low-resource languages: Leveraging comparable data, code-switching and compute resources. *arXiv preprint arXiv:2103.13272*.

Lewis-Beck, M. S., Bryman, A., and Liao, T. F. (2004). Pearson's correlation coefficient. *The SAGE encyclopedia of social science research methods*, 1(0).

Liang, Y., Duan, N., Gong, Y., Wu, N., Guo, F., Qi, W., Gong, M., Shou, L., Jiang, D., Cao, G., Fan, X., Zhang, R., Agrawal, R., Cui, E., Wei, S., Bharti, T., Qiao, Y., Chen, J.-H., Wu, W., Liu, S., Yang, F., Campos, D., Majumder, R., and Zhou, M. (2020). XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

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

Lohar, P., Afli, H., and Way, A. (2018). Balancing translation quality and sentiment preservation (non-archival extended abstract). In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 81–88, Boston, MA. Association for Machine Translation in the Americas.

McCormick, C. and Ryan, N. (2019). BERT fine-tuning tutorial with PyTorch. Online. Accessed 21 August 2020.

Michel, P. and Neubig, G. (2018). MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, Brussels, Belgium. Association for Computational Linguistics.

Misra, S. (2020). Guessing sentiment in 100 languages. Online. Accessed 21 August 2020. Media type: Blog.

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

Mozetič, I., Grčar, M., and Smailović, J. (2016). Twitter sentiment for 15 european languages. Slovenian language resource repository CLARIN.SI.

Peng, J., Feldman, A., and Vylomova, E. (2014). Classifying idiomatic and literal expressions using topic models and intensity of emotions. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2019–2027, Doha, Qatar. Association for Computational Linguistics.

Pozzi, F. A., Fersini, E., Messina, E., and Liu, B. (2016). *Sentiment analysis in social networks*. Morgan Kaufmann.

Provoost, S., Ruwaard, J., van Breda, W., Riper, H., and Bosse, T. (2019). Validating automated sentiment analysis of online cognitive behavioral therapy patient texts: An exploratory study. *Frontiers in Psychology*, 10:1065–1077.

Sennrich, R., Haddow, B., and Birch, A. (2016). Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.

Shen, L., Sarkar, A., and Och, F. J. (2004). Discriminative reranking for machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 177–184, Boston, Massachusetts, USA. Association for Computational Linguistics.

Si, C., Wu, K., Aw, A. T., and Kan, M.-Y. (2019). Sentiment aware neural machine translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 200–206, Hong Kong, China. Association for Computational Linguistics.

Tiedemann, J. (2017). OPUS. University of Helsinki.

Tiedemann, J. and Thottingal, S. (2020). OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisbon, Portugal. European Association for Machine Translation.

Timoshenko, A. and Hauser, J. R. (2019). Identifying customer needs from user-generated content. *Marketing Science*, 38(1):1–20.

Wasow, T., Sag, I., and Nunberg, G. (1983). Idioms: An interim report. In *Proceedings of the XIIIth International Congress of Linguists*, pages 102–115. CIPL Tokyo.

Wibowo, H. A., Nityasya, M. N., Akyurek, A. F., Fitriany, S., Aji, A. F., Prasojo, R. E., and Wijaya, D. T. (2021). Indocollex: A testbed for morphological transformation of indonesian word colloquialism. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics: Findings*.

Wijaya, D. T., Callahan, B., Hewitt, J., Gao, J., Ling, X., Apidianaki, M., and Callison-Burch, C. (2017). Learning translations via matrix completion. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1452–1463.

Williams, L., Bannister, C., Arribas-Ayllon, M., Preece, A., and Spasić, I. (2015). The role of idioms in sentiment analysis. *Expert Systems with Applications*, 42(21):7375–7385.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. *Computing Research Repository*, arXiv: 1910.03771. version 5.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yuan, Z., Briscoe, T., and Felice, M. (2016). Candidate re-ranking for SMT-based grammatical error correction. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 256–266, San Diego, CA. Association for Computational Linguistics.

A Appendix

A.1 Example Translations

French	
Original	Why are people such wankers these days?
Baseline	Pourquoi les gens sont-ils si branleurs ces jours-ci?
SentimentMT	Pourquoi les gens sont-ils si cons ces jours-ci?
Finnish	
Original	I'm sorry—I'm feeling kinda yucky myself—5am is going to come too quick.
Baseline	Olen pahoillani, olen itsekin aika naljaillen, että aamuviideltä tulee liian nopeasti.
SentimentMT	Olen pahoillani, että olen itse vähän kuvottava, mutta aamuviideltä tulee liian nopea.
Portuguese	
Original	Time Warner Road Runner customer support here absolutely blows.
Baseline	O suporte ao cliente do Time Warner Road Runner é absolutamente insuportável.
SentimentMT	O suporte ao cliente do Time Warner Road Runner aqui é absolutamente estragado.
Indonesian	
Original	Yo @Apple fix your shitty iMessage
Baseline	Yo @Apple perbaiki pesan menyebalkanmu
SentimentMT	Yo @Apple perbaiki imessage menyebalkan Anda

Table 4: Example texts exhibiting our MT pipeline's performance using the multilingual sentiment model fine-tuned with XLM-RoBERTa.

B Evaluation Instructions

The following are excerpts from the instructions given to evaluators for both the English-Spanish and English-Indonesian evaluations:
The document you are now looking at should contain prompts numbered up to 120. For each of these prompts, you will be asked to do three things:

1. Rate the *accuracy* of the translation. Please rate the **accuracy** of the translation on a **0 to 5** scale, where **0** indicates an "awful" translation, **2.5** indicates a "decent" translation, and **5** indicates a "flawless" translation . . .
2. Please rate the sentiment divergence on a **0 to 2** scale, where **0** indicates that the sentiment of the source sentence **perfectly matches** that of the translation and **2** indicates that the sentiment of the source sentence is the **opposite** of that of the translation . . .
3. Indicate the *reasons* for sentiment divergence . . .

C Sample Prompt for Human Evaluations

Below is an excerpt from a translation evaluation prompt that evaluators were asked to respond to:

- *Accuracy:*
- *Sentiment divergence:*
- *Please bold all of the below which had an effect on the sentiment of the translation:*
    1. *The translation contained literal translation(s) of figurative English language*
    2. *The translation contained other types of mistranslated words*
    3. *The original (English) sentence can't be properly translated to Spanish*

    . . .