A study of extracting sentiments for non-english languages with LLMs

Anonymous ACL submission

Abstract

Social media platforms like X (formerly Twitter), Facebook, Reddit are top platform to share feedback and raise concerns about product and services of particular company. Understanding of social data in various languages, but not limited to English is important to meet the ever-changing need of customers. Recently, large language models (LLMs) have shown significant performance for various NLP tasks. However, There are very limited labelled dataset and pre-trained models for non-english languages like Spanish, French and German. In this paper, we did extensive study of exploring various open pre-trained transformer based machine translation and sentiment models and evaluating and fine-tuning these models on open datasets. Our approach incorporates two strategies, Firstly we leverage sentiment models trained in non-english languages and improve it further by fine-tuning on domain data. Second, We fine-tuned machine translation models to translate text from non-english languages to english and use state of the art sentiment models in english language. Machine translation model after fine tuning achieves BLEU score between 50 and 60 that is considered as very high quality translation. We find that extracting sentiments by translating to english and then using english model with domain understanding yields better performance, which achieves 61.43 f1-score vs 55.26 with first approach.

1 Introduction

These days, people spend lot of time on social media, consuming content, sharing feedback and discuss about product and services in various languages. It is believed that people tends to get influenced and take decisions based on the feedback shared by others. Companies has to continuously analyzed the social data, understand the issues, monitor the performance of competitors and improve customer sentiment. Hence, Social listening becomes very critical for companies to improve its product and digitize the customer service. X (formerly Twitter) has around 50% data in non-english language (PARC, 2010) (Yeung, 2013) so we decided to pick top 3 popular languages (Spanish, German, French) other than english to understand sentiment from posts. It's a challenging task to train language model from scratch as it requires lot of labelled data and computations that subsequently requires more cost. We leveraged already pre-trained transformer based model using transfer learning and datasets that are freely available. We explored 5 different machine translation models from HuggingFace (Wolf et al., 2019) that can translate text from non-english languages (Spanish, French and German) to English and in opposite direction. Helsinki model (Tiedemann and Thottingal, 2020) performed better when compared using (Papineni et al., 2002) BLEU score. Total 6 Helsinki models (es-en, de-en, fr-en, en-es, en-de, en-fr) is further fine tuned on domain data from X (formerly Twitter), OPUS (Tiedemann, 2009) and Linguistic Data Consortium (LDC) (UPenn, 1992). Machine translation models (es-en, de-en, fr-en) are used to translate text from non-english language to english and leveraging state of the art sentiment models in english to extract sentiments. Other three models (en-es, en-de, en-fr) are used to generate training samples from english to non-english languages, since there is very limited data available in these languages.

Pre-trained sentiment models trained in non-english languages as well as english are leveraged to extract sentiment from social posts. There are very limited models available in languages like Spanish, German and French. To Further boost the performance of multilingual models we generated more training samples using machine translation models and fine-tuned models on domain data. We also explored state of the art sentiment models trained specifically on social data in english languages. Here, first we have to translate text from

non-english languages to english using machine translation models and then perform sentiment classification. Experimental results shown that sentiment extracted by translating text to english using machine translation and leveraging english model performs better.

2 Datasets

2.1 Machine Translation

We have used two sources i.e Linguistic Data Consortium(LDC) and OPUS to collect parallel corpus containing three Non-english (Spanish, French and German) and English language. Also used tweets from X (formerly Twitter) in english language using snscrape (JustAnotherArchivist, 2018) library in python. We translated tweets from english to three Non-english languages using googletrans library (Han, 2015) that is based on google translate.

OPUS is open dataset containing translated text from the web and for 300+ languages. We identified three news datasets that are as close as possible to Twitter data because 49% of posts shared on twitter is related to news. Following are the three OPUS datasets: News-commentary (WMT, 2019a), Global voices (CASMACAT, 2018) and WMT-News (WMT, 2019b). After further evaluation, News-commentary data is selected for fine-tuning of models. From LDC, ATIS - Seven Languages (Saab Mansour, 2021) dataset that was developed by Amazon Web Services, Inc. was selected. It consists of 5,871 English utterances from ATIS (Air Travel Information Services) corpora, translated into six languages: Spanish, German, French, Portuguese, Chinese, and Japanese. We combined LDC, OPUS and tweets data containing parallel corpus for three languages and distribution of samples is shown in Table 1.

Source	Number of samples	% Distribution
Tweets	56112	78%
OPUS	10000	14%
LDC	5871	8%
Total	71983	100%

Table 1: Data source and number of samples from each source.

Total samples of 71,983 was further split into three parts: Training has 50388 (70%), validation has 7195 (10%) and testing has 14400 (20%) samples.

2.2 Sentiment Classification

There are very limited labelled dataset for sentiment classification task in non-english languages (Spanish, German, French). We found cardiffnlp/tweet_sentiment_multilingual (Barbieri et al., 2022) dataset from HuggingFace that best suite our use-case of extracting sentiment from tweets. It contains 3033 tweets with 1011 samples each for negative, neutral and positive class for 8 languages (arabic, english, french, german, hindi, italian, portuguese and spanish) and has domain knowledge for our task. We have used it for evaluation of all the experiments done with respect to sentiment models.

For fine-tuning of multilingual models, we leveraged labelled mteb/tweet_sentiment_extraction (Muennighoff et al., 2022) dataset from HuggingFace in english language and translated english text to non-english using machine translation models. Total 18000 samples were used for fine-tuning having 6000 samples each for three non-english languages. Samples are split further into training (14400, 80%) and validation (3600, 20%) dataset.

3 Method and Experiments

To build machine translation model, we leveraged multiple transformer based pre-trained models available from HuggingFace and improved performance further by fine-tuning to adapt the domain knowledge. These models are implemented and fine-tuned using transformers library. There are two model components used: machine translation and sentiment classification. Architecture for all the machine translation models are encoder-decoder based, because it's a sequence-to-sequence task i.e. input and output both are text sequences. Machine translation model is fine-tuned on OPUS, LDC and tweet dataset for four languages like English, Spanish, German, French to further improve result.

Five state-of-the-art pre-trained translation models of various types and architectures are explored. Following are the models: mBART (Meta) (Tang et al., 2020), NLLB-200 (Meta) (Team et al., 2022), Helsinki machine translation model (Tiedemann and Thottingal, 2020), M2M-100 (Fan et al., 2020) and SMaLL-100 (Mohammadshahi et al., 2022). These models are classified into SDMT (Single Direction Machine Translation) and MDMT (Multi Direction Machine Translation). SDMT models can translate text in single direction only. For

Example, University of Helsinki model (opus-mt-es-en) that is pretrained on spanish and engish parallel data and translate spanish text to english. MDMT models can translate text in both directions. For Example M2M-100 can translatate text from any of the supported 100 languages to the rest 99 languages. MDMT models can further classified into English and Non-english centric. mBART-large is the english centric model where atleast one language in parallel data should be english. To overcome the lower performance for english centric models in non-english translation directions, non-english centric models like M2M100 and NLLB200 are developed. The parallel data used to pre-train these models are non-english centric i.e., one of the sentences in the sentence pair need not be english. These five pre-trained models are evaluated on OPUS test set using BLEU score and Helsiniki model performed better than others and finally it is selected to fine-tune further on domain specific dataset. Results for both pre-trained and fine-tuned model is shown in the results section below.

To extract sentiments from non-english language, We follow two strategies: First directly extracting sentiments using models trained in non-english languages. We explored 3 different pre-trained sentiment models: nlptown/bert-base-multilingual-uncased-sentiment (NLP Town, 2023), lxyuan/distilbert-base-multilingual-cased-sentiments-student(Lik Xun Yuan, 2023) and pysentimiento/robertuito-sentiment-analysis (Pérez et al., 2022). The last model was only trained for Spanish and English language and hence not considered for further experimentation. Also generated more training data for non-english languages using translation models (en-es, en-fr, en-de) for fine-tuning BERT base models: google-bert/bert-base-multilingual-uncased (Devlin et al., 2018) and distilbert/distilbert-base-multilingual-cased (Sanh et al., 2019). Second approach incorporates translating text present in non-english language to english using translation models (es-en, fr-en, de-en). Then use english models like cardiffnlp/twitter-roberta-base-sentiment-latest (Camacho-collados et al., 2022) and finiteautomata/bertweet-base-sentiment-analysis (Pérez et al., 2021) with domain understanding of social data. These models are evaluated on tweet_sentiment_multilingual dataset described in Datasets section.

Fine-tuned MT Model (HuggingFace)	BLEU Score
Helsinki-NLP/opus-mt-es-en	0.577
Helsinki-NLP/opus-mt-fr-en	0.582
Helsinki-NLP/opus-mt-de-en	0.539
Helsinki-NLP/opus-mt-en-es	0.644
Helsinki-NLP/opus-mt-en-fr	0.633
Helsinki-NLP/opus-mt-en-de	0.549

Table 2: It represents BLEU score for Fine-tuned Helsinki models like Helsinki-NLP/opus-mt-es-en that can translate text from spanish to english. Here es in the model name represent Spanish language (source language) and en represent English language (target language). Similarly fr represents French language and de represents German language.

4 Results

Pre-trained machine translation models are evaluated on spanish to english translation tasks for 2000 samples from OPUS dataset. We found the Helsinki MT models has better BLEU score and less inference time than other models. These models are single direction MT models and light weight. Inference time is also important metrics as it involves computation cost. Inference time is calculated on AWS GPU SageMaker instance ml.p3.16xlarge (64vCPU + 488 GB + 8 GPU) for 2000 samples. See Table 3 for comparison of pre-trained models using BLEU score and inference time.

Six Fine-tuned Helsinki models are evaluated using BLEU score on test data (Tweets, OPUS and LDC) containing 14,400 (20%) samples. All the models achieves state of the art BLEU score that is more than 0.5 that considered to be very high quality translation. It took 50 minutes on AWS GPU ml.p3.16xlarge instance (64vCPU + 488 GB + 8 GPU) for 14,400 samples. Table 2 shows the results.

We evaluated two pre-trained multilingual sentiment models using labelled test data containing 3033 tweets for all three languages. nlptown model achieves F1-score of 47.46 and results can be seen in the Table 4. We further improved the results by fine-tuning multilingual BERT and DistilBERT based models on domain data generated from translation model. The best model achieves F1-score of 55.26 and results can be seen in Table 5.

Non-english data is translated to english using fine-tuned machine translation models like (es-en, fr-en, de-en). Pre-trained state of the art sentiment models that are trained on tweets data are used to evaluate on same labelled test data containing 3033

Pre-trained MT Model (HuggingFace)	Model Type	BLEU Score	Inference Time (sec)
Helsinki-NLP/opus-mt-es-en	SDMT	**0.411**	**0.21**
facebook/m2m100_418M	MDMT and non-english centric	0.351	1.2
alirezamsh/small100	MDMT and non-english centric	0.353	0.54
facebook/nllb-200-distilled-600M	MDMT and non-english centric	0.317	0.54
mrm8488/mbart-large-finetuned-opus-es-en-translation	MDMT and english centric	0.241	0.63

Table 3: Performance comparison for various state of the art pre-trained machine translation model for translating Spanish text to English using BLEU score. Inference time is measured on ml.p3.16xlarge AWS SageMaker instance.

Pre-trained Multilingual Sentiment Model	Spanish	French	German	Overall
nlptown/bert-base-multilingual-uncased-sentiment	46.49	50.53	45.36	**47.46**
lxyuan/distilbert-base-multilingual-cased-sentiments-student	47.76	46.52	40.33	44.87

Table 4: F1-score of Pre-trained sentiment models in non-english languages (Spanish, French German)

tweets. Table 6 provides result and both the models achieves F1-score around 61 and performance is better than multilingual sentiment models.

5 Conclusion

In this paper, we explored various types of open source pre-trained sentiment models along with machine translation models from HuggingFace. We have evaluated these models on test data and identify the best machine translation model. Further, Fine-tuned the best model to improve on data that is closer to the actual use-case i.e. social tweets data. Fine-tuned machine translation models achieved BLEU score of more than 0.5 that is the state-of-the-art results when compared with other benchmarks. We experimented with two strategies: First directly extracting sentiments from models trained in non-english languages. Since there is limited labelled dataset for these languages, we generated more data from english labelled sentiment dataset using translation models and improved multilingual models by fine-tuning. Second approach incorporated translating text from non-english languages to english and leverage state of the art sentiment models in english language. We conclude that translating text from non-english languages to english and then leveraging models trained on domain data in english language performs better than first approach.

Limitations

All experiments are conducted for three popular non-english languages (Spanish, French and German) and similar results may not generalize for other non-english languages. Although, machine translation models have good BLEU score but any improper translation may affects performance of sentiment classification model. We have used same test data for evaluating sentiment models in non-english and english language. If quality of labels in test data is compromised then evaluation results might be different.

References

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Jose Camacho-collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa Anke, Fangyu Liu, et al. 2022. TweetNLP: Cutting-edge natural language processing for social media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–49, Abu Dhabi, UAE. Association for Computational Linguistics.

CASMACAT. 2018. Globalvoices.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

Suhun Han. 2015. Googletrans: Free and unlimited google translate api for python. translates totally free of charge. `https://github.com/ssut/py-googletrans`.

Fine-tuned Multilingual Sentiment Model (HuggingFace)	Spanish	French	German	Overall
google-bert/bert-base-multilingual-uncased	58.42	54.31	53.07	**55.26**
distilbert/distilbert-base-multilingual-cased	54.18	45.84	47.98	49.33

Table 5: F1-score of fine-tuned BERT and DistilBERT based sentiment models in multilingual languages.

Pre-trained Sentiment Model for English (HuggingFace)	Spanish	French	German	Overall
lxyuan/distilbert-base-multilingual-cased-sentiments-student	63.89	56.47	63.94	**61.43**
cardiffnlp/twitter-roberta-base-sentiment-latest	63.80	52.52	67.00	61.10

Table 6: F1-score of Pre-trained sentiment models in english language. Here non-english text first translated to english and then models trained in english language are used to extract sentiment

JustAnotherArchivist. 2018. A social networking service scraper in python. https://github.com/JustAnotherArchivist/snscrape/.

Lik Xun Yuan. 2023. distilbert-base-multilingual-cased-sentiments-student (revision 2e33845).

Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson, and Laurent Besacier. 2022. SMaLL-100: Introducing shallow multilingual machine translation model for low-resource languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8348–8359, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

NLP Town. 2023. bert-base-multilingual-uncased-sentiment (revision edd66ab).

Kishore Papineni, 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

PARC. 2010. Languages and social network behaviors: Top 10 languages on twitter.

Juan Manuel Pérez, Damián Ariel Furman, Laura Alonso Alemany, and Franco M. Luque. 2022. RoBERTuito: a pre-trained language model for social media text in Spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7235–7243, Marseille, France. European Language Resources Association.

Juan Manuel Pérez, Juan Carlos Giudici, and Franco Luque. 2021. pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks.

Batool Haider Saab Mansour. 2021. Atis - seven languages.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Jörg Tiedemann. 2009. *News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces*, volume V, pages 237–248.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

UPenn. 1992. Linguistic data consortium.

WMT. 2019a. News-commentary.

WMT. 2019b. Wmt-news.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Ken Yeung. 2013. 61 languages are found on twitter. here's how they rank in popularity.