Language-Agnostic method for sentiment analysis of twitter

Amir Reza Jafari, Reza Farahbakhsh, Mostafa Salehi, Noel Crespi

Language-Agnostic Method for Sentiment Analysis of Twitter

Amir Reza Jafari, Reza Farahbakhsh, Mostafa Salehi, Noel Crespi

Abstract With the different events and crises that we are witnessing these days, Twitter plays an essential role in sharing thoughts, opinions, and news worldwide in various languages. Understanding the sentiment of user-generated content has garnered much interest in both industrial and academic communities in recent studies. Due to the limited availability of data from low-resource languages, the focus on multilingual resources is a limiting and challenging issue of sentiment analysis task. Considering the importance of pre-processing in the implementation of a sentiment analysis system, we propose a method consisting of two steps for the pre-processing of tweets in different languages i) a language-agnostic step to replace or remove some elements in the Twitter data structure and ii) a text-normalization step based on the main high-resource language. In addition, we used machine translation techniques to translate low-resource language texts into the main language. We evaluated sentiment classification approaches based on four deep models: an RNN model and three BERT-based architectures namely vanilla-version, a language-specific and a large-scale pre-trained model for Twitter. The results show that our method had better accuracy when using a large-scale BERT-based pre-trained model.
Key words: BERT-based approaches, low resource languages, NLP, Sentiment analysis, Twitter

1 Introduction

Sentiment analysis as an application of natural language processing (NLP) is a prominent research area in which to identify and extract information from data, such as feelings, attitudes, emotions, and opinions. Twitter is one of the primary social media sources mainly used to share opinions and feelings about ongoing events worldwide. Twitter's Application Programming Interface (API) provides an easy-to-use interface for researchers to collect tweets on various subjects in different

A.R.Jafari, R.Farahbakhsh, N.Crespi
Samovar, Telecom SudParis, Institut Polytechnique de Paris, Palaiseau, France
e-mail: amir-reza.jafari_tehrani@telecom-sudparis.eu
e-mail: reza.farahbakhsh@telecom-sudparis.eu
e-mail: noel.crespi@telecom-sudparis.eu

M.Salehi
Faculty of New Sciences and Technology, University of Tehran, Tehran, Iran
e-mail: mostafa_salehi@ut.ac.ir

languages to perform multiple NLP tasks. As an example, after the COVID-19 outbreak, Twitter became a popular platform for users to share information about this pandemic, and many researchers concentrated on sentiment analysis of this subject on Twitter data to analyze users' opinions about ongoing issues [1] [2].

One of the challenges in analyzing such social media platforms is dealing with multilingual data as people usually discuss these trending topics in their native languages. As a result, having a general model that can be used for sentiment classification of social media data is essential to capturing worldwide input. With the emergence of transformers and attention-based models [3], many NLP downstream tasks, especially sentiment analysis, have proven their capability of state-of-the-art performance compared to previously proposed architectures [4]. However, these findings were mostly reported in languages with more available resources, such as English or Spanish [6]. With the introduction of the BERT model [5], many languages improved their performance on sentiment analysis task since it pre-trained on a wide range of languages but later studies showed language-specific BERT-based architectures outperform basic multilingual BERT.

On the one hand, such language-specific models need large corpora to have better performance on domain-specific data, while on the other hand, they need massive computational resources to train the models from scratch. Therefore, an improved, domain-specific sentiment approach that can be used for different languages without much effort in training on each language would be very appealing.

The main contributions of this paper are as follows:
(i) Propose a language-agnostic method for the sentiment analysis of domain-specific Twitter data using a machine translation technique to convert the low-resource language to the main high-resource one,
(ii) Investigate the impact of pre-processing on various classification architectures,
(iii) Compare sentiment classification of different BERT-based approaches on a low-resource language use case using the proposed method.

Our goal is to offer a method for pre-processing input tweets from different languages, translating them to a single high-resource language, and then analyzing different sentiment classification approaches for the proposed method to find the optimum approach. We compare transformer-based architecture with RNN architecture to evaluate our method for sentiment analysis of English domain-specific data and then compare the different transformer-based approaches to classify the sentiment of low-resource language data using a machine-translation tool.

2 Related work

Recently, research interest has increased significantly in sentiment analysis using new methods and algorithms based on deep learning, machine learning, and the use of transformers. Many algorithms, such as Naive Bayes, Decision Tree, KNN, SVM, and LSTM, as well as transformer-based algorithms, were applied to the IMDb dataset, which is a balanced sentiment dataset [7], [8], [9]. For various sentiment

classification approaches, Prajval et al.[10] provide a comparative study, and the technical and non-technical aspects and challenges of opinion mining and sentiment analysis are discussed in [11].

Ethem et al. [12] examined the possibility of training a model on a language with highly available resources from a multilingual perspective using the RNN algorithm, and reused it for the sentiment analysis of other languages such as Russian, Spanish, Turkish, and Dutch, for which fewer resources are available. Moreover, different machine learning algorithms are used for sentiment classification, along with a machine translation system for translating languages such as Telugu and Hindi to English in [13]. Also, Valentin et al. [14] analyzed the sentiment analysis of four other non-English languages such as French, Spanish, German and Italian in the tweet domain, by pre-training over English tweets and using automatic translation for non-English language adaptation. In our work, we used the machine translation technique as well to build a method for dynamically converting from any supported low-resource language to high resource one.

Twitter, is one of the main resources for determining users' emotions and feelings. Recent studies have concentrated on analyzing Twitter data on various subjects. Marco et al. [15] introduced a different approach for Twitter sentiment analysis based on the BERT language model by transforming tweet jargon to plain text as a way to avoid training directly on tweets from scratch. They used the Italian language as a case study but generally, their approach is applicable to different languages. Several papers have focused on pre-training on specific-language tweet corpora: Thakkar et al. [16] presented different pre-training strategies for the sentiment classification task in Latvian; Jose Angel et al. [17] proposed TWilBert, a BERT-based architecture that outperforms multilingual BERT on classification tasks; and Tuan Anh et al. [18] used CNN and LSTM architecture for sentiment analysis of informal Indonesian tweets. As a summary of the limitations and difficulties of sentiment analysis, a multi-class classification approach for Twitter is proposed in [19]. In our paper, we concentrated on more domain-specific sentiment analysis of Twitter data and examined the impact of pre-processing on two different classification methods and applied it to the Persian language as a low-resource language use case.

For the task of translating a low-resource language to a high-resource one, After the introduction of GNMT neural translation in 2016 as a hybrid model architecture based on RNN, the quality of translation has greatly improved for many languages [20]. Recent advances in Google translate show that the new model uses a transformer encoder and an RNN decoder instead of the original GNMT system [21]. After applying a variety of optimizations, this hybrid model is more stable in training and shows lower latency and better performance in translation. We used GNMT for in machine translation part of our proposed method.

3 Methodology

Here we present our approach for pre-processing Twitter data and the classification methods we used for analyzing the sentiment of tweets. Our approach can be used as

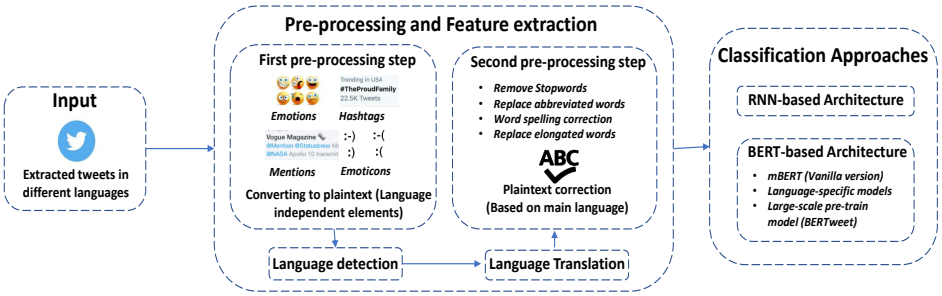

Fig. 1 Proposed methodology and flow of the sentiment classification

a language-agnostic approach, as any language, especially low-resource languages can cope with this system. The main idea is to choose a language with higher available resources as the main language for training, and then reuse that model for other low-resource languages using a machine translation to translate to the main language and evaluate the different popular approaches proposed for sentiment analysis.

As shown in Figure 1, the input is domain-specific tweets in different languages. A two-step pre-processing is then applied to the input data (Section 3.1).

After the language of tweets is detected, a translation model is used to translate data to the main language (Section 3.2) before applying the next step of pre-processing.

After the two pre-processing steps, different machine learning sentiment classification approaches for detecting the sentiment of tweets are introduced, and the architecture and fine-tuning are described (Section 3.3). The pseudocode for our method is presented in Algorithm 1.

Algorithm 1 Pre-processing steps for Twitter data

Require: Tweets from official Twitter API in different languages
MainLang ← Select main high resource language
for each Tweet T: **do** First step of pre-processing:
    Replace Emojies, Emoticons, time & date
    Remove URLs, emails, mentions
    if *Language*(T) in equal to *MainLang* **then**
        Continue;
    else
        Translate(T);
    end if
end for
for each Tweet T: **do** Second step of pre-processing:
    Remove stopwords
    Correct word spelling
    Replace abbreviated and elongated words
end for

3.1 Pre-processing steps

Tweets generally have a specific sentence structure that usually includes emojis, emoticons, hashtags, and mentions within the text along with other attributes which they usually consider noisy in analysis, as users often use colloquial language and abbreviations. To eliminate these noises, performing pre-processing on the data to convert tweets into a typical text structure is necessary. As a result, a two-step pre-processing method is proposed:

Step one: This pre-processing step is independent of the input language and is based on the general structure of the tweet, including its metadata such as emojis, emoticons, hashtags, and URLs as they are similar in all languages. Since most of the pre-trained models are trained on a normal plain text structure rather than Twitter-structured data, the objective is to remove these noisy elements and convert them to plain text, which can improve the result for the sentiment analysis task. Moreover, some elements such as email addresses, phone numbers, dates, etc, would not apply any special emotional load to the text, so they can be removed. However, emoticons and emojis affect the polarity of the text, so converting them to their equivalent meanings in the text would be ideal. We used the Ekphrasis[1] tool, which has several functionalities, to process social media data that can be used for various NLP tasks, such as understanding complex emoticons, emojis, and other unstructured expressions like dates, times, etc [22].

Step two: After completion of the first pre-processing step and the main language selection, the second stage of data pre-processing will be applied to prepare data for the classification model. For analysis, we consider English as the main language. At this stage of pre-processing, several basic actions are applied to the input data using NLTK[2] :

- Stopword removal: Words that may not be useful nor bring any valuable information to the text analysis are deleted. Some words, such as 'not', are excluded from this list, as they affect the sentiment classification.
- Abbreviated words: Abbreviated words are a shortened form of a written word or phrase that are replaced with the complete form. For example, "don't" is replaced with "do not", and "RU" with "are you".
- Correct word spelling: We used the spell correction functionality of Ekphrasis to correct misspelled words found in the input data of this step.
- Elongated words: Those words that use repeated alphabets to express an intense feeling about that word. For example, "Coooooool" is replaced with "cool"

3.2 Language labelling and translation

In this step, firstly the language of the tweet will be detected using "langdetect" library[3] and the language label will be added in the following forms:

[1] https://github.com/cbaziotis/ekphrasis

[2] https://www.nltk.org/

[3] https://pypi.org/project/langdetect/

- **Main language:** The language with high resource data. For example, we chose English so “en” would be added as a language label for English tweets; or
- **Subsidiary language:** Mainly low-resource languages that contain a smaller share of the input data like “fa” (for Farsi), and “ar” (for Arabic).

The main purpose of this step is to translate subsidiary language text into the main language using Google machine translation. We used this method for translating a low-resource language (Persian) dataset into the main language (English) in order to evaluate the classification approaches proposed in Section 3.3.

3.3 Classification approaches

In this paper, we use different language models and approaches to evaluate the sentiment analysis of Twitter data. Once the two pre-processing phases are complete, the input data is ready to enter the classifier system. For classification, we used several approaches:

The first approach uses a model based on RNN architecture, as it has proven reliable for classification tasks. We used the same structure as proposed in [12], with two bidirectional layers, each with 40 neurons and a dropout of 0.2. This approach is utilized to compare its accuracy with the mBERT model and thus to evaluate the performance of different architectures for the proposed analysis of Twitter data.

The second approach uses a pre-trained model based on transformer architecture. We used the mBERT model, which was pre-trained on Wikipedia data in 104 languages, and fine-tuned by using labelled tweets introduced as a training set in section 4.1.

The third approach is based on using a language-specific BERT-based model. Studies have shown that pre-training on large language-specific corpora with the same architecture as the BERT model may have better results on the downstream tasks on that specific language compared to the multilingual BERT model [5], and so we used this approach to evaluate this model on COVID-19 Twitter data in a low-resource language. We used the ParsBERT model, which is a language-specific model pre-trained and fine-tuned on Persian data [23]. To adapt this model for our evaluation, we used the hugging face model[4] , a sentiment analysis model fine-tuned on approximately 600,000 Persian tweets.

The fourth approach is based on using a large-scale pre-trained model for Twitter. We used the BERTweet model, which uses the same architecture as the BERT base, and the RoBERTa pre-training procedure, trained on a large-scale dataset of English tweets, including tweets related to the COVID-19 pandemic. The goal is to test this model on low-resource language tweets related to COVID-19. For the evaluation, we used the BERTweet model with the configuration introduced by [24].

[4] https://huggingface.co/nimaafshar/parsbert-fa-sentiment-twitter

Table 1 Classification results on BERT and RNN models with & without applying pre-processing

Model	Precision			Recall			F1-Score		
	Neg	*Neu*	*Pos*	*Neg*	*Neu*	*Pos*	*Neg*	*Neu*	*Pos*
RNN	**0.78**	0.58	0.81	**0.76**	0.74	0.82	**0.77**	0.65	0.81
Pre-Processing + RNN	0.77	**0.60**	**0.83**	0.75	**0.75**	**0.83**	0.76	**0.66**	**0.83**
BERT	0.89	0.61	0.87	0.82	0.80	**0.86**	0.85	0.70	0.86
Pre-Processing + BERT	**0.91**	**0.65**	**0.90**	**0.85**	**0.82**	**0.86**	**0.88**	**0.72**	**0.88**

4 Experiments and Results

4.1 Dataset

We used two sets of corpora containing tweets related to the COVID-19 pandemic in order to make the models more specialized in a single domain.

D1: This dataset is composed of nearly 45,000 English tweets available on Kaggle [5] containing original tweet texts and a sentiment label in five classes (very positive, positive, neutral, negative, and very negative). Since the data of each class was relatively imbalanced, and to consider three main categories of polarity in each classification model architecture, we combined the very negative and negative classes, as well as the very positive and positive classes, to have more balanced data. The overall distribution of each class is then: Positive- 43%, Neutral- 20%, and Negative- 37%.

D2: The D2 dataset consists of around 1000 tweets related to COVID-19 in Persian crawled from Twitter streaming API based on related hashtags. This is a human-annotated dataset annotated by computer science experts who are native speakers of Persian with a deep understanding of Persian linguistics. The overall distribution of each class is: Positive- 44%, Neutral- 22%, and Negative- 34%.

4.2 Results

Experiment 1: Our first experiment is the sentiment classification of the first dataset to compare BERT-based and RNN methods after applying pre-processing steps. The results of both the BERT-based and RNN models indicate that the BERT-based model has a better performance in sentiment analysis of Twitter data on COVID-19. Moreover, applying the two steps of pre-processing gives a slightly better performance in the BERT-based architecture, improving the accuracy of the model by about 2.5 percent, with better precision and recall for all three classes. However, for the RNN architecture, the results show that the two-step pre-processing method only improves the precision and recall of the "Neutral" class, with no significant difference observed in the other classes.

A deeper look at the results reveals that, as expected, the model has a higher error rate in detecting "Neutral" tweets compared to other categories, as the training set is unbalanced so that the number of tweets labelled "Neutral" is less than that of the

[5] www.kaggle.com/code/gauravduttakiit/covid-19-sentiment-analysis-on-complete-data/data

two other categories. It is also possible that the labelling of this category was not as accurate as that of the other categories.

Experiment II: To compare the different classification approaches introduced in Section 3.3 for a low-resource language, we conducted several experiments on each model with the Persian test set (D2). As presented in Table 2, the results show that there using a language-specific model offers improved performance over using a vanilla BERT-based model, since language-specific models are pre-trained and fine-tuned on larger language-specific corpora. More detailed observation shows that the error rate of incorrect classification for the "Negative" category increased in the first (BERT-based) approach compared to the language-specific model, most likely due to two main issues: 1- The loss of key information for negative sentences in the translation process due to the informal words used in tweets; and 2- Pre-training was performed on fewer data in that specific language in a domain-specific dataset. The other result obtained from this experiment is the confirmed effectiveness of using a pre-trained model with a large-scale in-domain corpus. Among the different approaches evaluated, the BERTweet model, which used large-scale Twitter data and domain-specific tweets related to COVID-19, outperformed the vanilla BERT-based approach and the language-specific approach by close to 6 and 2 percent, respectively.

Table 2 Classification results on low-resource language

Approach	**Model**	Precision	Recall	F1	Accuracy
BERT-based	mBERT	0.72	0.76	0.74	74.23
Lang Specific BERT-based	ParsBERT	0.78	0.79	0.78	78.49
Large-scale pre-trained model	BERTweet	**0.79**	**0.82**	**0.80**	**80.21**

5 Conclusion and future work

In this paper, we proposed a method for sentiment analysis of domain-specific Twitter data that can be used for a wide range of input languages. Although pre-training in each specific language would offer better accuracy, we used a machine translation system to translate a low-resource language to the main high-resource language, in this case, English. We first compared a BERT-based approach with one of the architectures based on RNN for multilingual data to show the possible improvements from using the BERT-based approach for English tweets related to COVID-19. Next, we evaluated two other approaches for the classification of a low-resource language; one based on a language-specific model and the other based on a large-scale pre-trained model on Twitter data. The results showed that the last approach had the best performance, attributed to the effectiveness of pre-training on a large-scale high-resource language and its usability for low-resource languages. For future work, we suggest the following directions: (i) Evaluate the proposed approach in additional low-resource languages; (ii) Analyze other transformer-based models for the sentiment classification of in-domain data, and (iii) Investigate the effect of linguistic and structural similarities in different languages. For example, languages like Persian, Arabic, and Urdu have similarities in typography and structure (as well as some common cultural references).

References

1. Xue, J., Chen, J., Hu, R., Chen, C., Zheng, C., Su, Y. & Zhu, T. Twitter discussions and emotions about the COVID-19 pandemic: Machine learning approach. *Journal Of Medical Internet Research*. **22**, e20550 (2020)
2. Chakraborty, A., Das, S. & Kolya, A. Sentiment analysis of covid-19 tweets using evolutionary classification-based LSTM model. *Proceedings Of Research And Applications In Artificial Intelligence: RAAI 2020*. pp. 75-86 (2021)
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., et al. Attention is all you need. *Advances In Neural Information Processing Systems*. **30** (2017)
4. Birjali, M., Kasri, M. & Beni-Hssane, A. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*. **226** pp. 107134 (2021)
5. Devlin, J., Chang, M., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*. (2018)
6. Pla, F. & Hurtado, L. Sentiment analysis in Twitter for Spanish. *International Conference On Applications Of Natural Language To Data Bases/information Systems*. pp. 208-213 (2014)
7. Sahu, T. & Ahuja, S. Sentiment analysis of movie reviews: A study on feature selection classification algorithms. *2016 International Conference On Microelectronics, Computing And Communications (MicroCom)*. pp. 1-6 (2016)
8. Rauf, S., Qiang, Y., Ali, S. & Ahmad, W. Using BERT for checking the polarity of movie reviews. *International Journal Of Computer Applications*. **975** pp. 8887 (2019)
9. Amulya, K., Swathi, S., Kamakshi, P. & Bhavani, Y. Sentiment Analysis on IMDB Movie Reviews using Machine Learning and Deep Learning Algorithms. *2022 4th International Conference On Smart Systems And Inventive Technology (ICSSIT)*. pp. 814-819 (2022)
10. Sudhir, P. & Suresh, V. Comparative study of various approaches, applications and classifiers for sentiment analysis. *Global Transitions Proceedings*. **2**, 205-211 (2021)
11. Shayaa, S., Jaafar, N., Bahri, S., Sulaiman, A., Wai, P., Chung, Y., et al., Sentiment analysis of big data: methods, applications, and open challenges. *IEEE Access*. **6** (2018)
12. Can, E., Ezen-Can, A. & Can, F. Multilingual sentiment analysis: An RNN-based framework for limited data. *ArXiv Preprint ArXiv:1806.04511*. (2018)
13. Arun, K. & Srinagesh, A. Multi-lingual Twitter sentiment analysis using machine learning.. *International Journal Of Electrical Computer Engineering (2088-8708)*. **10** (2020)
14. Barriere, V. & Balahur, A. Improving sentiment analysis over non-english tweets using multilingual transformers and automatic translation data-augmentation *ArXiv:2010.03486*. (2020)
15. Pota, M., Ventura, M., Catelli, R. & Esposito, M. An effective BERT-based pipeline for Twitter sentiment analysis: A case study in Italian. *Sensors*. **21**, 133 (2020)
16. Thakkr, G. & Pinnis, M. Pretraining and fine-tuning strategies for sentiment analysis of latvian tweets. *Human Language Technologies–the Baltic Perspective: HLT 2020*. **328** (2020)
17. Gonzalez, J., Hurtado, L. & Pla, F. TWilBert: Pre-trained deep bidirectional transformers for Spanish Twitter. *Neurocomputing*. **426** pp. 58-69 (2021)
18. Le, T., Moeljadi, D., Miura, Y. & Ohkuma, T. Sentiment analysis for low resource languages: A study on informal Indonesian tweets. *Proceedings Of The 12th Workshop On Asian Language Resources (ALR12)*. pp. 123-131 (2016)
19. Bouazizi, M. & Ohtsuki, T. Multi-class sentiment analysis on twitter: Classification performance and challenges. *Big Data Mining And Analytics*. **2**, 181-194 (2019)
20. Wu, Y., Schuster, M., Chen, Z., Le, Q., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K. & Others Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv Preprint ArXiv:1609.08144*. (2016)
21. Chen, M., Firat, O., Bapna, A., Johnson, M., Macherey, W., Foster, G., Jones, L., Parmar, N., Schuster, M., Chen, Z. & Others The best of both worlds: Combining recent advances in neural machine translation. *ArXiv Preprint ArXiv:1804.09849*. (2018)
22. Baziotis, C., Pelekis, N. & Doulkeridis, C. DataStories at SemEval-2017 Task 4 *Proceedings Of The 11th International Workshop On Semantic Evaluation (SemEval-2017)*. (2017,8)
23. Farahani, M., Gharachorloo, M., Farahani, M. & Manthouri, M. ParsBERT: Transformer-based Model for Persian Language Understanding. *Neural Processing Letters*. **53** (2021)
24. Pérez, J., Giudici, J. & Luque, F. pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks. (2021)