PingAnLifeInsurance at SemEval-2023 Task 12: Sentiment Analysis for Low-resource African Languages with Multi-Model Fusion

MeiZhi Jin and **Cheng Chen** and **MengYuan Zhou** and **MengFei Yuan** and **XiaoLong Hou** and **XiYang Du** and **LianXin Jiang** and **JianYu Li**
Ping An Life Insurance Company of China, Ltd.
{JINMEIZHI005, CHENCHENG498, ZHOUMENGYUAN425, YUANMENGFEI854, HOUXIAOLONG430, DUXIYANG037, JIANGLIANXIN769, LIJIANYU002}
@pingan.com.cn

Abstract

This paper describes our system used in the SemEval-2023 Task12: Sentiment Analysis for Low-resource African Languages using Twitter Dataset. The AfriSenti-SemEval Shared Task 12 is based on a collection of Twitter datasets in 14 African languages for sentiment classification. It consists of three sub-tasks. Task A is a monolingual sentiment classification which covered 12 African languages. Task B is a multilingual sentiment classification which combined training data from Task A (12 African languages). Task C is a zero-shot sentiment classification. We utilized various strategies, including monolingual training, multilingual mixed training, and translation technology, and proposed a weighted voting method that combined the results of different strategies. Substantially, in the monolingual subtask, our system achieved Top-1 in two languages (Yoruba and Twi) and Top-2 in four languages (Nigerian Pidgin, Algerian Arabic, and Swahili, Multilingual). In the multilingual subtask, Our system achived Top-2 in publish leaderBoard.

1 Introduction

Sentiment Analysis (SA) refers to the identification of sentiment and opinion contained in the inpfut texts that are often user-generated comments. The problem of automatic sentiment analysis is growing research topic. Although sentiment analysis is an important area and already has a wide range of applications, it clearly is not a straightforward task and has many challenges related to Natural Language Processing (NLP).

In recent years, sentiment analysis has gained significant attention due to its numerous vital applications. The growth of social networks has resulted in the generation of more complex and interrelated information, enabling various artificial intelligence (AI) applications such as sentiment analysis, machine translation, and detection of harmful content. Social media platforms have become important sources of data for sentiment analysis. However, most sentiment analysis research focuses on high-resource languages such as English, while low-resource languages like African languages remain underrepresented. According to UNESCO (2003), around 2,058, or 30%, of all living languages are African languages. However, most of these languages lack curated datasets required for the development of AI applications, and this issue is not unique to sentiment analysis, but affects NLP research as a whole.

Sentiment analysis has garnered increasing interest due to its applicability across many domains, including public health, commerce/business, art and literature, social sciences, neuroscience, and psychology. Previous shared sentiment analysis tasks have been conducted, such as those by (Rosenthal et al., 2019), (Nakov et al., 2019), (Pontiki et al., 2016), and (Ghosh et al., 2015), among others. However, none of these tasks have included African languages, highlighting the need for concerted efforts to create resources for such languages (Alabi et al., 2020).

In this paper, we describe a system for sentiment analysis tasks in African low-resource languages on the Twitter dataset. Specifically, we introduce a novel ensemble system. For different languages, different strategies are proved effective in our system. To conduct monolingual training on each language data, we initially employed several pre-training models such as naija-twitter-sentiment-afriberta-large[1], TwHIN-BERT[2], and DeBERTa[3] model. We also utilized multilingual mixed training by using four folds of each language as the training dataset and one fold of each language as the validation dataset, and implemented 5-fold cross-

[1] https://huggingface.co/Davlan/naija-twitter-sentiment-afriberta-large
[2] https://huggingface.co/Twitter/twhin-bert-large
[3] https://huggingface.co/microsoft/deberta-v3-large

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 679–685

validation during training. Additionally, we used translation technology to translate all languages into English and used both single language training and multilingual mixed training methods. Finally, we combined the results of each approach through weighted voting to obtain the ultimate outcome.

2 Background

Sentiment analysis is the automated process of understanding and recognizing human sentiment from text, audio, or video. It uses natural language processing, text mining, and machine learning techniques to identify, extract, quantify, and study sentiment states from text. It can be used to detect and categorize positive, negative, and neutral sentiment in a given dataset. Some subcategories of research in sentiment analysis include: multimodal sentiment analysis, aspect-based sentiment analysis, fine-grained opinion analysis, language specific sentiment analysis.

Traditional methods such as lexicon-based approaches (Piryani et al., 2020), feature-based models (Pang and Lee, 2007) and rule-based models (Diwali et al., 2022) have been applied to sentiment analysis tasks. More recent approaches have focused on using machine learning algorithms to improve the accuracy and efficiency of sentiment analysis (Zhang and Zheng, 2016). In particular, deep learning methods are being applied to a wide range of natural language processing tasks, including sentiment analysis (Tang et al., 2015). These methods have been proven to be highly effective in extracting sentiment from text. Additionally, there has been research into utilizing transfer learning and ensemble methods for sentiment analysis tasks.

Recently, pre-training methods have shown their powerfulness in learning general semantic representations, and have remarkably improved most natural language processing (NLP) tasks like sentiment analysis. These methods build unsupervised objectives at word-level, such as masking strategy, next-word prediction or permutation. Such word prediction based objectives have shown great abilities to capture dependency between words and syntactic structures. However, due to the most of pre-training models only pre-trained in high-resource languages such as English, the African low-resource languages unable to perform semantic analysis well.

3 System overview

We utilized various approaches in this work, incorporating weighted voting to combine outcomes from different strategies. Initially, we employed several pre-training models, including naija-twitter-sentiment-afriberta-large, TwHIN-BERT and DeBERTa model, to conduct monolingual training on each language data. Additionally, we utilized multilingual mixed training to perform joint training, using four folds of each language as the training dataset and one fold of each language as the validation dataset, and implementing 5-fold cross-validation during training. Furthermore, we leveraged translation technology to translate all languages into English, and used both the single language training and multilingual mixed training methods for training. Finally, we combined the outcomes of each approach through weighted voting to obtain the ultimate result.

3.1 Strategy For Task A

3.1.1 Monolingual Training Technology

naija-twitter-sentiment-afriberta-large - The naija-twitter-sentiment-afriberta-large model is the first multilingual Twitter sentiment classification model for four Nigerian languages (Hausa, Igbo, Nigerian Pidgin, and Yorùbá). It is based on the fine-tuned castorini/afriberta_large model and achieves state-of-the-art performance on the Twitter sentiment classification task when trained on the NaijaSenti corpus. The model is capable of classifying tweets into three sentiment categories: negative, neutral, and positive. Specifically, the naija-twitter-sentiment-afriberta-large model is an xlm-roberta-large model that was fine-tuned using an aggregate of four Nigerian language datasets obtained from the NaijaSenti dataset.

TwHIN-BERT - TwHIN-BERT (Zhang et al., 2022) is a new multi-lingual language model for Tweets, which is trained on more than 7 billion tweets from over 100 different languages. Unlike previous pre-trained language models, it not only uses text-based self-supervision techniques like MLM but also uses a social objective that is based on the rich social engagements within a Twitter Heterogeneous Information Network (TwHIN). TwHIN-BERT can be used as a direct replacement for BERT in various natural language processing and recommendation tasks. It not only outperforms similar models in semantic understanding tasks such as text classification but also excels in social

recommendation tasks such as predicting user-to-tweet engagement.

3.1.2 Multilingual Mixed Training Technology

We utilized multilingual mixed training to perform joint training, using four folds of each language as the training datasets and one fold of each language as the validation datasets, and implementing 5-fold cross-validation during training. The TwHIN-BERT model as our base model for training. In addition, we also use DeBERTa as the base model after translating all languages into English. Multilingual mixed training is a powerful technique that can improve the performance of NLP models on low-resource languages, and it has been used successfully in various NLP tasks such as machine translation, language modeling, and named entity recognition.

3.1.3 Translation Technology

Due to the fact that most pre-trained models perform well on English datasets but not on African low-resource language datasets. We leveraged translation technology to translate all languages into English. We used the DeBERTa model as our base model for training, and the Monolingual training and multilingual mixed training methods are adopted.

3.1.4 Weighted voting fusion technology

We utilized a weighted voting approach to combine results, whereby we assigned scores to each strategy and used voting to merge them, taking the label with the highest weight as the final output. For example, if a sentence has labels "negative, negative, neutral, positive" across four strategies with scores of 0.67, 0.68, 0.7, and 0.69 respectively, the weight of "negative" would be 0.67+0.68=1.35, "neutral" would be 0.7, and "positive" would be 0.69. The final output would be "negative".

3.2 Strategy For Task B

3.2.1 Data Augmentation Technology

Due to the data of this task is low-resource languages, we decided to extend relevant datasets for these sentiment analysis tasks on same languages. We tried to search on GitHub, huggingface and relevant papers. Finally, we found the datasets in the paper "NaijaSenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis" (Muhammad et al., 2022), we can get the datasets of four languages, including Hausa (hau), Igbo (ibo), Naija (pcm) and Yorùbá (yor).

3.2.2 Translation Technology

Similar to TaskA's strategy, we translated data sets of all languages into English separately, then combine them together. We also use the pre-trained model DeBERTa as base model and get an English classification model after fine-tuning. In the inference stage, we also translate all the test sets into English before inference.

3.2.3 Pre-Classification Technology

Based on TaskA's classification models, we added a pre-classification model to identify which language the input is. This pre-classification model base on DeBERTa-large model's fine-tuning. On the other hand, we regarded the language names as labels for training. After identifying the language, we can use the relate classification model in TaskA. In this part, we can separate the TaskB into two parts: pre-classification model and single classification model.

4 Experimental setup

4.1 Dataset

The datasets for AfriSenti-SemEval (Muhammad et al., 2023a) competition are tweets collected from Twitter. Each tweet is annotated by three annotators following the annotation guidelines in (Muhammad et al., 2023b). The sentiments of the tweet was determined by majority vote method. The dataset involves tweets labelled with three sentiment classes (positive, negative, neutral) in 14 African languages. It consists of three sub-tasks:

Task A: Monolingual Sentiment Classification Task A is a monolingual sentiment classification which covered 12 African languages. Given training data in a target language, determine the polarity of a tweet in the target language (positive, negative, or neutral).

Task B: Multilingual Sentiment Classification Task B is a multilingual sentiment classification which combined training data from Task-A (12 African languages). Given combined training data from Task-A (Track 1 to 12), determine the polarity of a tweet in the target language (positive, negative, or neutral).

Task C: Zero-Shot Sentiment Classification Task C is a zero-shot sentiment classification. Given unlabelled tweets in two African languages

(Tigrinya and Oromo), leverage any or all of the available training datasets (in Task:A) to determine the sentiment of a tweet in the two target languages.

In this competition, the organizers provided 14 African languages, among which 12 languages had labels and the other 2 languages did not have labels. Figure 1 and Figure 2 respectively show the distribution of labels for task A and task B. From the figures, we can see the distribution of label numbers for each language. **we found the following conclusions:**

- Some languages have small datasets such as Xitsonga (ts).
- Some label distribution is unbalanced in some languages such as Nigerian Pidgin (pcm).
- The dataset of Task B is combined the Task A dataset across all languages.
- The size of testing dataset is twice the size of validation dataset in all languages.

4.2 Training Details

Learning rate initialization. To enhance the performance of our system, we implemented a learning rate scheduling technique for the different layers of the pre-trained text model. The model layers were grouped into three categories with distinct hyperparameters. The first group comprised layer-0 to layer-7, which were assigned learning rates of 1e-5/2.6. The second group consisted of layer-8 to layer-15, which were set to a learning rate of 1e-5. The third group included layer-16 to layer-23, which were assigned a learning rate of 1e-5 * 2.6. This approach of assigning different learning rates for different layers of the model proved to be effective in improving the overall performance of the system.

multi sample dropout. Our system utilized four dropout samples, and experimental results demonstrate a significant improvement in both performance and effectiveness.

Optimization. The AdamW optimizer to optimize the loss function was applied in our systems.

Learning rate decay. A method of cosine annealing to decay the learning rates for each batch to avoid falling into a local optimal solution was used in our system.

Adversarial Training. Adversarial Training was applied to improve modal robustness and generalization in our system.

5 Results

In this section, the results of the experiments for our runs will be discussed and compared to the results published in Table 1. We attempted to use various pre-trained models as the base model for training and employed both Monolingual language and multilingual mixed training methods. Additionally, we leveraged translation technology to translate all languages into English and utilized DeBERTa as the base model for training. Finally, we utilized a weighted voting approach to merge the results of multiple strategies. The result of Task B is mapping from Task A.

Through the experiment, we found the following conclusions:

- Using AfriBERTa (naija-twitter-sentiment-afriberta-large) model as the baseline for monolingual training produces the best results on the following four languages: Hausa (ha), Igbo (ig), Nigerian Pidgin (pcm), and Yoruba (yo).
- Using TwHIN-BERT model as the baseline for monolingual training produces the best results on the following languages: Twi (twi).
- Using multilingual mixed training based on TwHIN-BERT model produces the best results on the following languages: Algerian Arabic (dz), Mozambique Portuguese (pt), .
- Using DeBERTa model as the baseline for monolingual training after translating into English produces the best results on the following languages: Kinyarwanda (kr), Swahili (sw), Xitsonga (ts).
- Using multilingual mixed training based on DeBERTa model after the texts translated into English produces the best results on the following languages: Amharic (am), Darija (ma).
- Multilingual mixed training outperforms monolingual language training when training on raw data.
- When dealing with a small dataset such as Swahili (sw), training after translation into English can lead to a significant improvement.

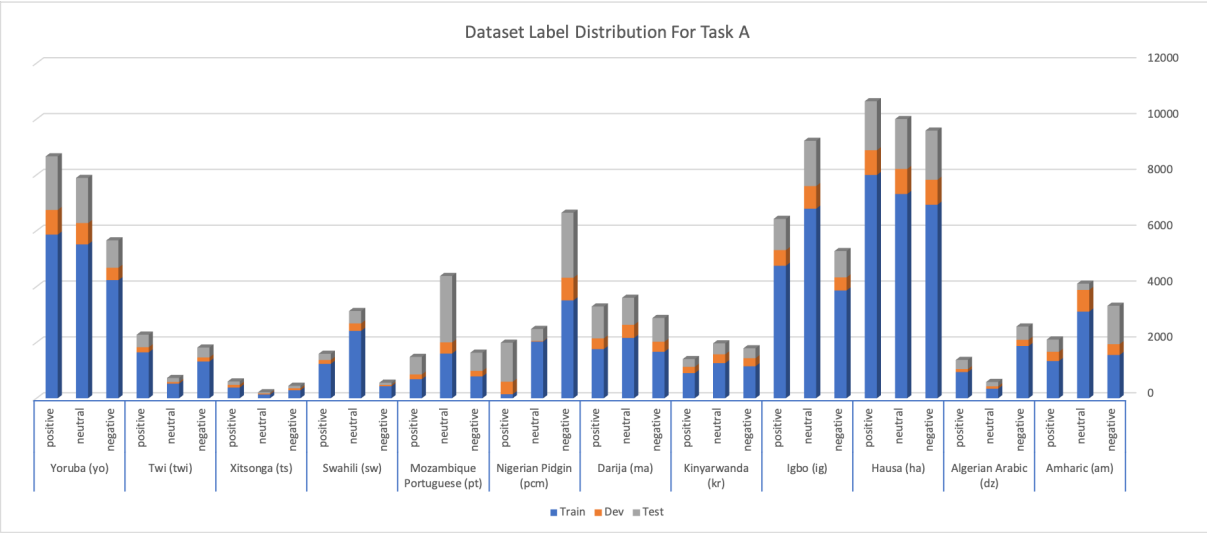

Figure 1: Dataset Label Distribution For Task A

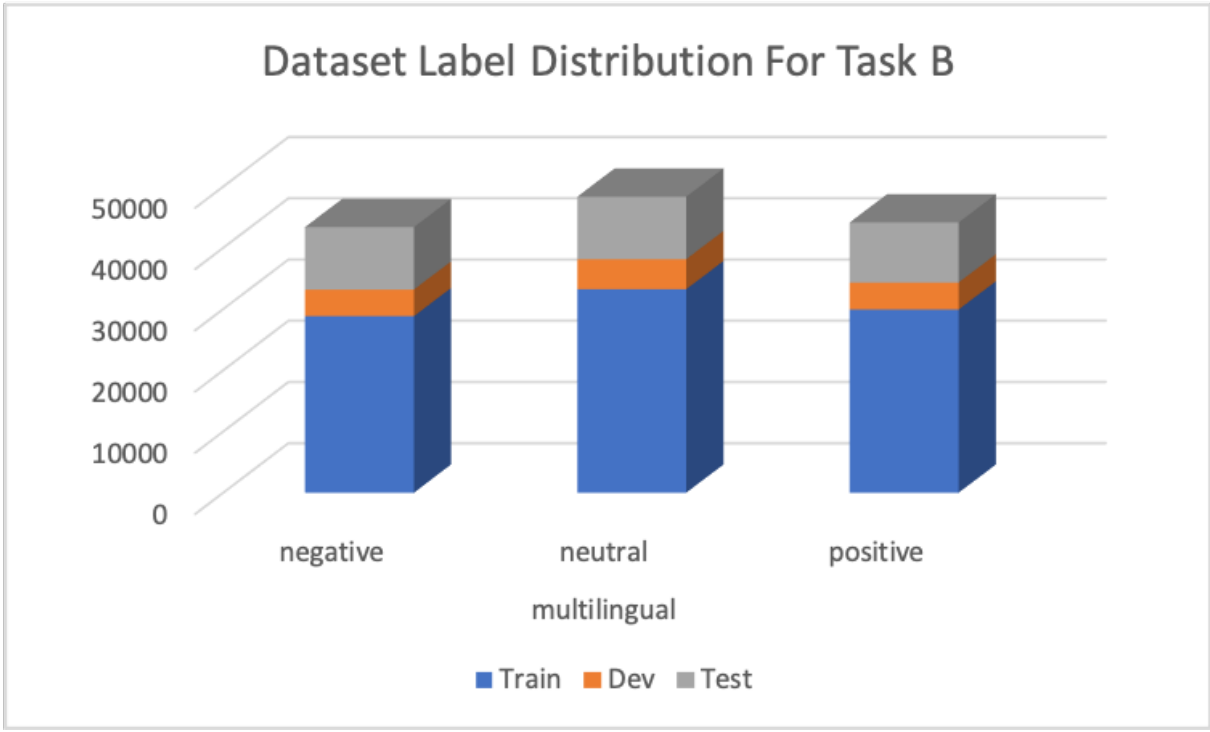

Figure 2: Dataset Label Distribution For Task B

Language	afriberta	sin+twhinbert	mul+twhinbert	sin+en+deberta	mul+en+deberta	merge
Amharic (am)	0.6733	0.6653	0.5982	0.6827	0.6983	0.6977
Algerian Arabic (dz)	0.4215	0.5295	0.7201	0.6732	0.6639	0.73
Hausa (ha)	0.81	0.7648	0.7539	0.7588	0.6427	0.811
Igbo (ig)	0.8114	0.7939	0.7791	0.7237	0.5726	0.8139
Kinyarwanda (kr)	0.6026	0.5893	0.6055	0.7088	0.6632	0.6026
Darija (ma)	0.3919	0.5295	0.5541	0.543	0.5739	0.5794
Nigerian Pidgin (pcm)	0.7553	0.7143	0.7243	0.7043	0.7264	0.7594
Mozambique Portuguese (pt)	0.565	0.7243	0.7498	0.7243	0.6823	0.7353
Swahili (sw)	0.6083	0.6006	0.615	0.6423	0.6149	0.6489
Xitsonga (ts)	0.4876	0.4808	0.5151	0.5538	0.5018	0.5626
Twi (twi)	0.5987	0.6688	0.6572	0.6347	0.6062	0.6828
Yoruba (yo)	0.8016	0.6565	0.6896	0.7666	0.6731	0.8016
multilingual	0.6885	0.6965	0.7017	0.7088	0.6535	0.75

Table 1: Experimental results of Task A and Task B (the multilingual results based on the single classification model after language identifying)

6 Conclusion

In this paper, we proposed a system to Sentiment Analysis for Low-resource African Languages using Twitter Dataset. To train the model, we tried different pre-trained models as the base model and used both Monolingual language and multilingual mixed training techniques. We also used translation technology to convert all languages into English and used DeBERTa as the base model for training. Ultimately, we merged the results of multiple strategies using a weighted voting method. The system performs well in AfriSenti-SemEval, obtained two Top-1 and three top-2 in Task A, achieved the Top-2 in Task B. Moving forward, we will continue to validate the effectiveness of more base models in different languages, and at the same time attempt to translate all languages into other African languages (such as Hausa, Igbo, etc.).

7 Acknowledgements

This research was supported by the PingAn Life Insurance. We thank the organizers of Bayero University, Kano, Masakhane, Universität Hamburg, Hamburg; Masakhane and Ahmadu Bello University Zaria, Masakhane for their support.

References

Jesujoba O. Alabi, Kwabena Amponsah-Kaakyire, David Ifeoluwa Adelani, and Cristina España-Bonet. 2020. Massive vs. curated embeddings for low-resourced languages: the case of yorùbá and twi. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 2754–2762. European Language Resources Association.

Arwa Diwali, Kia Dashtipour, Kawther Saeedi, Mandar Gogate, Erik Cambria, and Amir Hussain. 2022. Arabic sentiment analysis using dependency-based rules and deep neural networks. *Appl. Soft Comput.*, 127:109377.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John A. Barnden, and Antonio Reyes. 2015. Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, pages 470–478. The Association for Computer Linguistics.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Baye Messelle, Hailu Beshada Balcha,

Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku, and Steven Arthur. 2023a. AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Sa'id Ahmad, Nedjma Ousidhoum, Abinew Ali Ayele, Saif M. Mohammad, Meriem Beloucif, and Sebastian Ruder. 2023b. SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, David Adelani, Sebastian Ruder, Ibrahim Sa'id Ahmad, Idris Abdulmumin, Shehu Bello Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahuddeen Abdullahi, Anuoluwapo Aremu, Alipio Jeorge, and Pavel Brazdil. 2022. Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 590–602, Marseille, France. European Language Resources Association.

Preslav Nakov, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Theresa Wilson. 2019. Semeval-2013 task 2: Sentiment analysis in twitter. *CoRR*, abs/1912.06806.

Bo Pang and Lillian Lee. 2007. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135.

Rajesh Piryani, Bhawna Piryani, Vivek Kumar Singh, and David Pinto. 2020. Sentiment analysis in nepali: Exploring machine learning and lexicon-based approaches. *J. Intell. Fuzzy Syst.*, 39(2):2201–2212.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia V. Loukachevitch, Evgeniy V. Kotelnikov, Núria Bel, Salud María Jiménez Zafra, and Gülsen Eryigit. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 19–30. The Association for Computer Linguistics.

Sara Rosenthal, Saif M. Mohammad, Preslav Nakov, Alan Ritter, Svetlana Kiritchenko, and Veselin Stoyanov. 2019. Semeval-2015 task 10: Sentiment analysis in twitter. *CoRR*, abs/1912.02387.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Deep learning for sentiment analysis: successful approaches and future challenges. *WIREs Data Mining Knowl. Discov.*, 5(6):292–303.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2022. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations. *CoRR*, abs/2209.07562.

Xueying Zhang and Xianghan Zheng. 2016. Comparison of text sentiment analysis based on machine learning. In *15th International Symposium on Parallel and Distributed Computing, ISPDC 2016, Fuzhou, China, July 8-10, 2016*, pages 230–233. IEEE Computer Society.