مجلة الدلتا الدولية للعلوم التجارية ونظم المعلومات

https://djicsi. journals.eKb.eg

Hybrid Language Models for Improved Multilingual Sentiment Analysis

النماذج اللغوية الهجينة لتحسين تحليل المشاعر متعدد اللغات

Submitted by:

Mohammed Abd Elmoneim Al Salamony

Assistant Teacher in Delta higher institute for Management and accounting information system

ABSTRACT

The rapid evolution of social media has facilitated deep insights into user opinions. However, sentiment analysis, particularly for low-resource languages like Arabic, remains underexplored due to limited resources. This study addresses this gap by investigating sentiment analysis on tweet texts from SemEval-17, 2.5+ Million Rows Egyptian Datasets Collection and the Arabic Sentiment Tweet dataset. We evaluated four pretrained language models and introduced two ensemble models. Our results demonstrate that monolingual models showed superior performance, while ensemble models surpassed baseline results, with the majority voting ensemble achieving the best performance, even outperforming English language benchmarks.

KEYWORDS: Tweet dataset; Sentiment Analysis; Language Models; SemEval-17; Egyptian Datasets Collection; BERT

١. INTRODUCTION

In recent years, the field of Natural Language Processing (NLP) has witnessed significant advancements, particularly with sentiment analysis emerging as a pivotal subfield. This analytical tool finds widespread applications across diverse domains such as market research, customer feedback analysis, social media monitoring, and brand reputation management. The ability to interpret sentiments conveyed in textual data has become indispensable for businesses and researchers aiming to derive actionable insights from vast volumes of user-generated content.

This project focuses on sentiment analysis, specifically exploring the complex landscape of multilingual tweet texts. The objective is to identify and classify sentiments expressed in these concise and dynamic messages into three primary categories: positive, negative, or neutral. The significance of this task extends beyond linguistic diversity, encompassing a wide array of domains including business, finance, politics, education, and various services, highlighting its broad applicability and relevance in today's data-driven world[1]. To embark on this initiative, we utilize established datasets from2.5+ Million Rows Egyptian Datasets Collection and SemEval-2017 Task 4, focusing on Task 4A, which involves classifying message polarity. The dataset's depth, incorporating English data from SemEval 2013 to 2017, provides a robust basis for our sentiment analysis efforts. Nevertheless, challenges arise, notably linguistic disparities within the Arabic dataset, prompting the development of the Arabic Sentiment Tweet Dataset (ASTD) to address these specific nuances and enhance the scope of our study[2, 3]. By integrating ASTD with the SemEval-2017 Arabic sentiment data, our goal is to minimize biases and enrich the representation of sentiments within the Arabic language context. This approach allows for a more comprehensive analysis that captures the nuances and diversity of Arabic sentiment expressions, thereby advancing our understanding and applicability in this critical area of study[4].

Our proposed methodology encompasses two primary components. Firstly, we aim to harness the capabilities of pretrained language models such as AraBERTv2, RoBERTa, multilingual BERT, and XLM-RoBERTa. These models will undergo meticulous fine-tuning using diverse datasets in both English and Arabic, thereby enhancing their performance to accommodate the intricate linguistic nuances of each language.

In the second component, we introduce an ensemble approach. Following individual fine-tuning, these models will be integrated into a unified, language-agnostic ensemble model. This combined model, trained on a consolidated dataset, is anticipated to deliver a comprehensive and resilient solution for sentiment analysis, capable of transcending linguistic barriers.

To gauge the effectiveness of our approach, we plan to conduct a thorough evaluation against state-of-the-art deep learning models. This comparative analysis will employ the macro-average F1 measure as the evaluation metric, chosen for its suitability in assessing performance across multi-class imbalanced datasets. This comprehensive evaluation framework ensures robust validation of our proposed methodology's efficacy and performance capabilities[2, 5].

In conclusion, this project resides at the crossroads of Natural Language Processing (NLP), sentiment analysis, and multilingual comprehension. Our objective is to provide valuable insights and methodologies that not only propel advancements in these fields but also furnish practical solutions applicable to diverse real-world scenarios and challenges.

٢. BACKGROUND

In the contemporary era marked by prolific social media usage, sentiment analysis has emerged as a vital tool, particularly on microblogging platforms like Twitter. The vast amounts of user-generated data available on these platforms provide a rich and fertile ground for gaining deep insights into public opinions, feedback, and sentiments. By analyzing this data, researchers and businesses can uncover trends, gauge public sentiment, and make informed decisions based on the collective mood and opinions expressed by users. This capability is increasingly important as social media continues to influence public discourse and shape societal trends[6]. One prevalent approach in sentiment analysis involves leveraging machine learning algorithms for the classification of sentiments. This method has been demonstrated in numerous studies, such as those utilizing distant supervision techniques. In these studies, researchers have effectively employed noisy labels derived from emoticons and acronyms found in Twitter messages to train their models. By harnessing the vast and diverse data available on social media platforms, these algorithms can learn to accurately classify the sentiment expressed in user-generated content. This approach not only enhances the accuracy of sentiment analysis but also provides a scalable solution to analyze large volumes of data efficiently[6]. This methodology effectively addresses the inherent brevity and simplicity found in tweets, making it possible to perform a more nuanced analysis of sentiments. By focusing on the concise and often straightforward nature of Twitter messages, this approach allows for the extraction and interpretation of subtle emotional cues and sentiments that might otherwise be overlooked. Consequently, it enhances the ability to understand and classify the diverse range of sentiments expressed in these short bursts of text, thereby contributing to a more detailed and accurate sentiment analysis. This refined understanding is crucial for comprehending the broader context of public opinion and sentiment as conveyed through social media platforms.

Linguistic diversity, cultural nuances, and geographically specific trends on social media platforms add significant complexity to the field of sentiment analysis. To address these challenges, cross-lingual and multilingual approaches have emerged as innovative solutions. These approaches aim to effectively analyze sentiments across different languages and cultural contexts. A prime example of such efforts is the research utilizing XLM-RoBERTa for cross-lingual sentiment analysis. This research demonstrates how knowledge can be transferred from resource-rich languages like English to resource-poor languages such as Hindi. By leveraging the robust capabilities of models like XLM-RoBERTa, researchers can bridge the gap between languages with abundant linguistic resources and those with limited ones, ensuring a more comprehensive and accurate sentiment analysis across diverse linguistic landscapes[7]. Such adaptability is crucial for conducting effective sentiment analysis across a wide range of linguistic landscapes. The ability to accurately interpret and analyze sentiments in multiple languages is essential for capturing the true essence of user opinions on a global scale. Furthermore, sentiment analysis is not confined to English-centric platforms. There has been substantial research focused on sentiment detection in Arabic tweets, which highlights significant efforts in this area. These studies often combine sophisticated pre-processing strategies with advanced transformer-based models such as AraELECTRA and AraBERT. This approach is designed to address the unique linguistic characteristics and nuances of the Arabic language, including the complexities of sarcasm and varied sentiment expressions. By tackling these challenges, researchers are enhancing the precision and reliability of sentiment analysis in Arabic, thereby broadening the scope and applicability of these techniques to a more diverse set of languages and cultural contexts[8]. The surge in social media usage, especially on platforms like Twitter, has resulted in an enormous influx of user-generated data. This exponential growth necessitates the development of effective text categorization and sentiment analysis techniques, which are vital for a wide range of fields including healthcare, policy-making, marketing, and beyond. The multilingual nature of social media data introduces additional complexity, making it imperative to explore and develop domain-agnostic and multilingual solutions that can accurately process and analyze content across different languages and

cultural contexts. These advancements are crucial for harnessing the full potential of social media data to drive informed decisions and insights in various domains[9]. Recent efforts in multilingual text categorization and sentiment analysis have focused on leveraging BERT-based classifiers and zero-shot classification approaches. These methodologies have shown promising accuracy and efficiency in sentiment classification across various languages. Comparative studies highlight the strengths of multilingual BERT-based classifiers and the adaptability of zero-shot approaches in developing efficient and scalable multilingual solutions. The robustness of language models, especially transformer-based architectures such as RoBERTa, has been a significant research focus. Investigations delve into factors such as pretraining techniques, hyper parameter optimization, and the impact of training data size to enhance model performance. This research aims to refine these models for better multilingual text processing, ensuring more accurate and effective sentiment analysis across diverse linguistic landscapes[10].The use of RoBERTa in aspect-category sentiment analysis demonstrates its superiority over traditional LSTM-based methods. This highlights RoBERTa's potential in capturing and extracting nuanced sentiments more effectively[11].

This comprehensive background lays the foundation for our project, underscoring the importance of exploring and comparing various sentiment analysis methodologies. It emphasizes the need to leverage pretrained language models and tackle linguistic challenges across different languages and domains. Drawing inspiration from the insights provided by existing studies, our proposed research aims to advance the understanding and practical application of sentiment analysis techniques in real-world scenarios.

٣. METHODOLOGY

In this study, we concentrate on the sentiment analysis of tweet data. Sentiment analysis involves determining the polarity of a given text, which can indicate positive, negative, or neutral sentiments. This technique is commonly employed to assess the quality of products, evaluate customer service, and analyze social media content.

٣,١. Data Collection and Preprocessing

We primarily utilized datasets from three distinct sources. Our English dataset was sourced from SemEval-17[12], while our Arabic data was curated from the extensive 2.5+ Million Rows Egyptian Datasets Collection[13] and ASTD (Arabic Sentiment Tweet Dataset)[2]. The datasets were derived from X (formerly named Twitter) data. The ASTD dataset comprises 4 classes (Objective, Positive, Negative, and Neutral), whereas the SemEval-17 dataset includes 3 classes (Positive, Negative, and Neutral). To ensure consistency, we excluded the "Objective" class from the ASTD data. For the English training data, we aggregated data from SemEval 2013-2016, excluding the test data from 2013 and 2014. The test data from SemEval 2013 and 2014 served as the development set for our study, while the SemEval-17 test data was used to evaluate model performance. Regarding Arabic data, we utilized data from subtasks A, B, and D of SemEval-17 and integrated it with the ASTD dataset. From this merged dataset, 90% of the data was allocated for training and 10% for validation purposes. The official test set of SemEval-17 subtask A was used as our test set for evaluation.

Tweet data typically includes various symbols, URLs, usernames, and invisible characters. To preprocess the data, we removed symbols, URLs, and invisible characters. As part of this preprocessing step, we utilized Byte-pair encoding (BPE) tokenizers integrated with pretrained transformer models to tokenize the text.

٣،٢. Model Selection and Baseline

We employed four different transformer-based pretrained language models: ArabicBERT version 02[5], RoBERTa base[10], multilingual BERT[14], and XLM-RoBERTa base[15].

We opted for the multilingual BERT model as our baseline system for this study. The decision was based on its relatively smaller and simpler architecture compared to the other models considered. Additionally, multilingual BERT supports both Arabic and English languages, whereas ArabicBERT is specifically designed for Arabic and RoBERTa for English. We excluded XLM-RoBERTa due to its greater complexity and higher number of training parameters compared to multilingual BERT.

Although we initially selected four pretrained models for our study, we also introduced two ensemble models for further investigation. In the first proposed model, we utilized model-specific tokenizers tailored for each language along with language-specific models. Specifically, we employed the AraBERT model for Arabic and the RoBERTa model for English. Input IDs and attention masks were fed into both models, and we extracted the pooler output from each. These outputs, along with the pooler output from the multilingual BERT model, were concatenated. This concatenated output was passed through a fusion layer, followed by a feed-forward network and softmax function. Finally, we selected the output from the softmax function.

In the second proposed model, we integrated a multi-head attention layer between the fusion layer and the feed-forward network. This modification enhances the model's ability to capture complex relationships within the concatenated output. For a detailed architectural depiction, please refer to Figures 1 and 2.

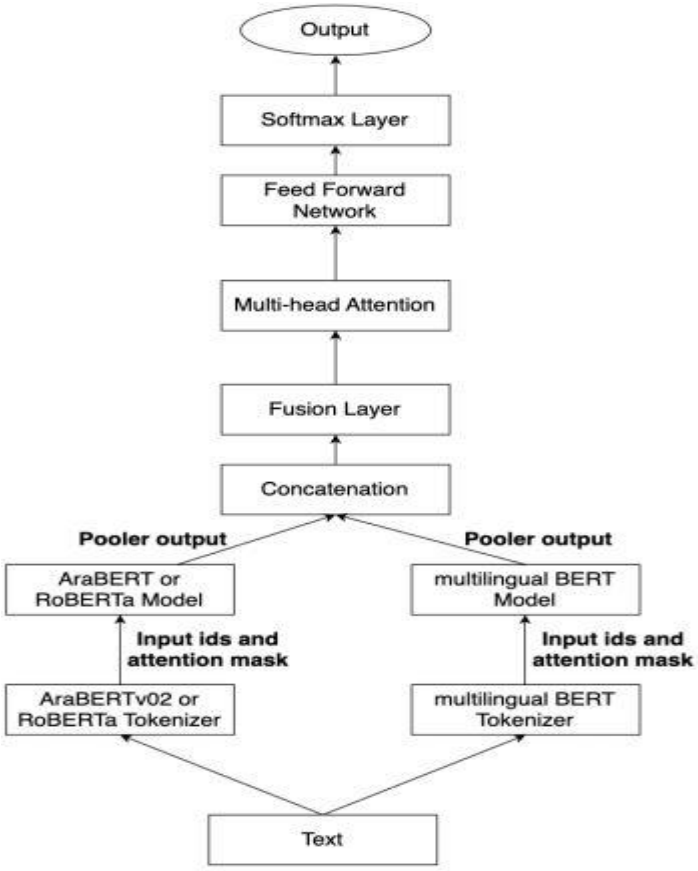

Figure 1: Ensemble of two pretrained language models followed by a Feed-Forward Network

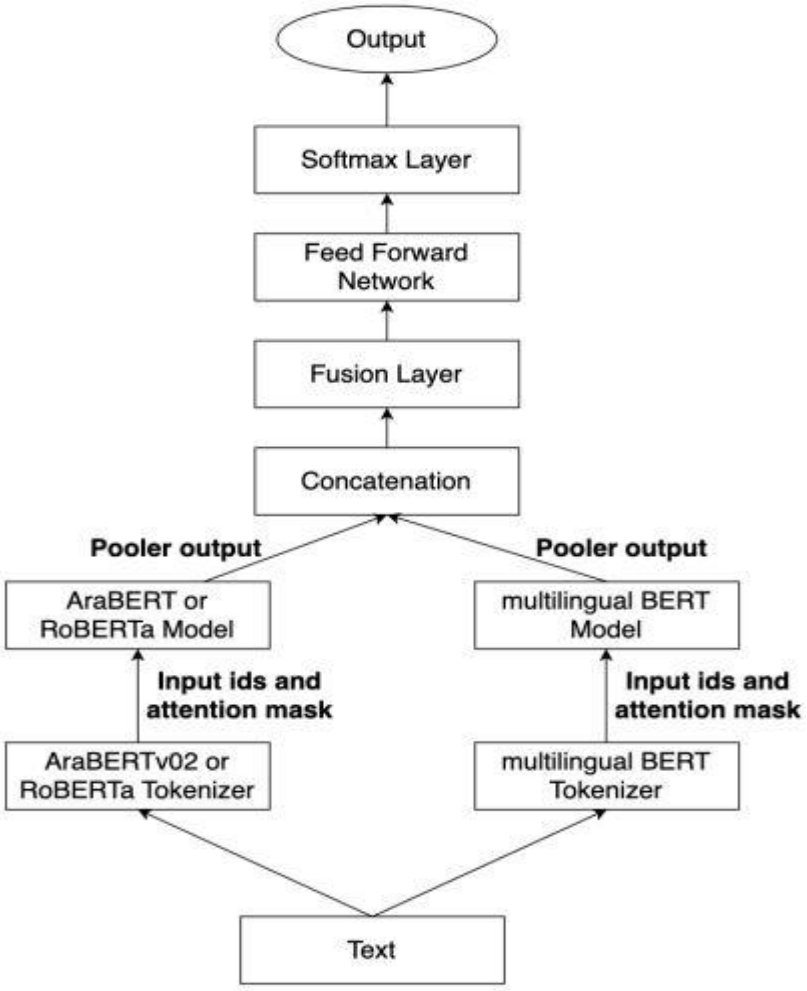

Figure 2: Ensemble of two pretrained language models followed by a multi-head attention and a Feed-Forward Network

٤. IMPLEMENTATION DETAILS

We began our exploration by reviewing two articles focused on sentiment analysis to understand the problem better. From these articles, we identified and selected datasets relevant to our study, focusing on two languages and combining multiple datasets for the Arabic language into a unified dataset. After completing data collection and preprocessing, we delved into examining state-of-the-art models.

Subsequently, we chose four pretrained language models for our investigation and proposed two ensemble models. We conducted several experiments using these models and datasets. Following this, we established a baseline model for comparison and curated evaluation metrics specific to our study. We then proceeded to evaluate and analyze the performance of each model. For a detailed depiction of our workflow, please refer to Figure 3.

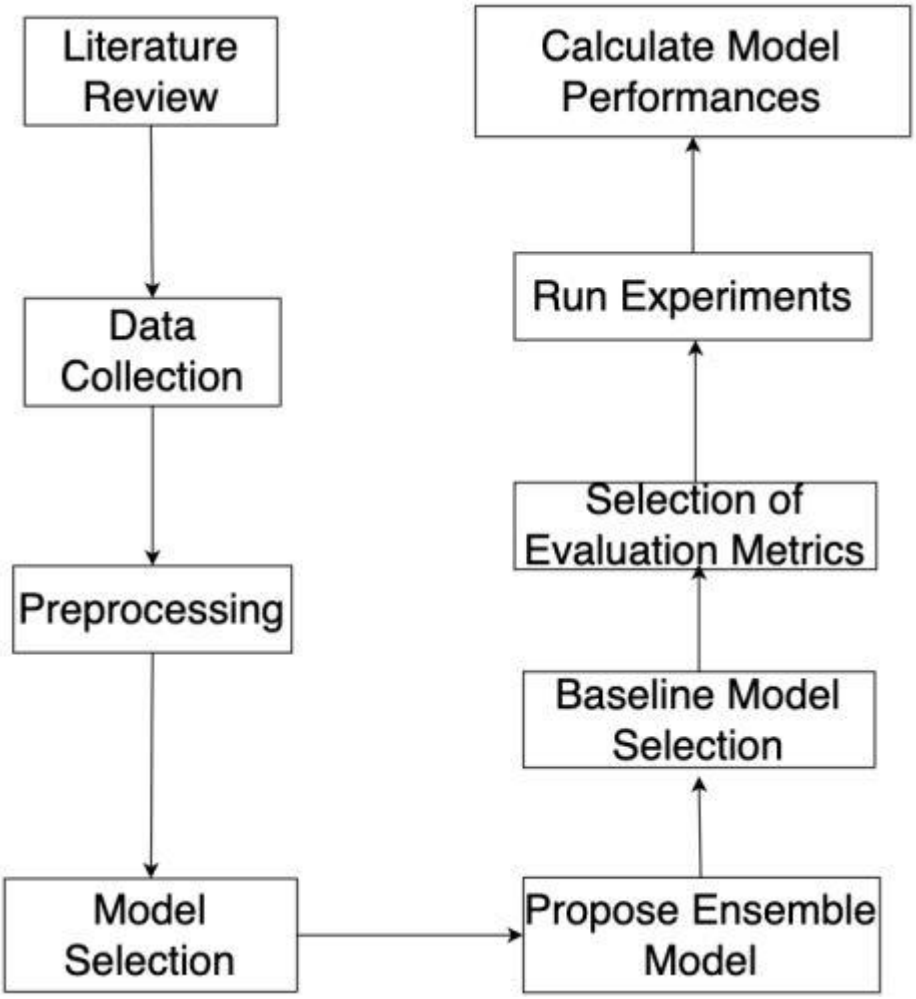

Figure 3: Detailed implementation workflow diagram of our study

We utilized several libraries and two machine learning frameworks to implement our models. Firstly, we opted for the PyTorch machine learning framework due to its scalability and the flexibility it offers in modifying network behavior at runtime. Additionally, we employed the Transformers library by Hugging Face, chosen for its comprehensive collection of pretrained language models, thus avoiding the time and resource-intensive task of developing models from scratch. The Transformers framework also seamlessly integrates with PyTorch, enhancing interoperability.

For data handling and preprocessing, we utilized the datasets library, which facilitated the fine-tuning of language models. Evaluation of our models was simplified using the Evaluate library. To compute evaluation metrics as outlined in section 5.2, we relied on scikit-learn. In managing data files, we employed libraries such as CSV, OS, random, and pandas. The re library was utilized for efficiently removing URLs using regular expressions as part of our preprocessing pipeline.

٥. EVALUATION

٥,١. Experimental Setup

We conducted our experiments across three distinct settings, employing the four selected pretrained language models alongside two proposed ensemble models. This approach allowed us to comprehensively evaluate and compare the performance of each model configuration under various conditions and datasets.

٥,١,١. Experiment Setup Configuration of the First Experiment

In this experimental setup, we categorized our experiments based on language-specific considerations. We fine-tuned two models each for Arabic and English languages, utilizing one language-specific model to compare the performances between multilingual and monolingual approaches. Specifically, for English, our monolingual model was RoBERTa base, and our multilingual baseline system was multilingual BERT, alongside the XLM-RoBERTa base model. In the case of Arabic, we employed AraBERT as the monolingual model, multilingual BERT as the baseline, and XLM-RoBERTa base model.

We employed binary cross entropy with logit loss as our loss function, utilized a learning rate (Adam) of 2e−5, set a maximum sequence length of 256, and utilized a batch size of 16. Training for all models was conducted over 3 epochs to ensure consistent evaluation across the experiments.

٥,١,٢. Experiment Setup Configuration of the First Experiment

In this experimental setup, we initially combined the English and Arabic training, validation, and test sets to train the model once. We first trained the ensemble model with a feed-forward network using the merged training data, validating the model at every epoch. Finally, we evaluated our proposed model using the merged test set and calculated the performance metrics individually. For training, we used cross-entropy loss as the loss function, set a learning rate (Adam) of 2e−5, a maximum sequence length of 256, and a batch size of 24, running the training for 2 epochs for both ensemble models.

٥,١,٣. Experiment Setup Configuration of the third Experiment

In this experimental setup, we initially train our proposed ensemble model with a feed-forward network using the English training set and validate it with the corresponding validation set. We then evaluate the model using the English test set. Subsequently, we train the same ensemble model with a feed-forward network using the Arabic training set, validate it with the Arabic validation set, and finally evaluate it with the Arabic test set. The model is validated at every epoch during training. The purpose of this third experimental setup is to identify any performance differences compared to the second experimental setup. For training, we use cross-entropy loss as the loss function, set a learning rate (Adam) of 2e−5, a maximum sequence length of 256, and a batch size of 24, running the training for 2 epochs for both ensemble models.

٥,٢. Evaluation Metrics

To measure performance across all experimental settings, we calculate accuracy, weighted precision, weighted recall, and macro F1 score. We specifically use the weighted and macro versions of these metrics to account for class imbalances.

٥,٣. Experimental Results

In Table 1, we presented the results of our experiments. From the results, the monolingual AraBERTv02 outperforms the Arabic language and the majority voting ensemble outperforms the English language. Our proposed model outperforms the baseline results for both languages.

Language	Training Data	Model	Accuracy	Precision	Recall	F1-macro
English	English	m-BERT (Baseline)	67.16	67.48	67.16	67.06
		RoBERTa	70.69	71.34	70.69	70.84
		XLM-RoBERTa	69.07	67.00	69.07	69.13
Arabic	Arabic	m-BERT (Baseline)	54.21	53.76	54.21	53.08
		AraBERTv02	69.79	69.96	69.79	**69.78**
		XLM-RoBERTa	63.89	63.63	63.89	63.74
English	English	Majority Voting Ensemble	70.95	71.55	70.95	**71.03**
Arabic	Arabic	Majority Voting Ensemble	66.69	66.37	66.69	66.42
English	English	Ensemble model with Feed Forward	68.91	69.26	68.91	68.59
Arabic	Arabic	Ensemble model with Feed Forward	67.67	69.01	67.67	67.82
English	English and Arabic	Ensemble model with multi-head attention Feed Forward	67.44	69.14	67.44	67.31
Arabic	English and Arabic	Ensemble model with multi-head attention Feed Forward	66.30	67.82	66.30	66.42
English	English and Arabic	Ensemble model with Feed Forward	70.03	70.50	70.03	69.88
Arabic	English and Arabic	Ensemble model with Feed Forward	67.61	68.01	67.61	67.12

Table 1: Performances on different sets of experiments including the baseline. **Bold** indicates the best system for languages.

٦. CONCLUSION

In this comprehensive sentiment analysis project, we navigated the intricate landscape of multilingual tweet texts, employing a diverse set of models, datasets, and ensemble strategies to effectively decipher sentiments. The evaluation results, as shown in the tabulated performance metrics, reflect the nuanced challenges and successes encountered across different linguistic and model scenarios.

Individual model assessments revealed notable disparities. For instance, RoBERTa exhibited robust performance in English sentiment analysis, outshining its counterparts. Conversely, m-BERT demonstrated varied efficacy, highlighting the sensitivity of model choice to language nuances. AraBERTv02 showcased commendable accuracy in Arabic sentiment analysis, though it faced challenges due to the language's unique characteristics.

The power of ensemble methods emerged prominently, with Majority Voting Ensemble and the proposed Ensemble multiple BERT approaches demonstrating improved performance, mitigating individual model limitations. We explored the intricacies of language-independent models using combined English and Arabic datasets, highlighting their potential for broader applicability.

Despite variations in model performance, the consistent use of macro-average F1 as an evaluation metric allowed for a balanced assessment across imbalanced multi-class datasets. This choice emphasized our commitment to precision, recall, and overall model effectiveness in handling diverse sentiments.

In conclusion, this project delved into the dynamic realm of sentiment analysis, offering insights into the effectiveness of state-of-the-art language models across different languages. The exploration of ensemble techniques and language-independent models underscored the adaptability and potential generalization of sentiment analysis systems. As sentiment analysis continues to evolve, this project contributes valuable perspectives, paving the way for further advancements in understanding and interpreting sentiments across multilingual contexts.

Refrances

1. Cui, J., et al., *Survey on sentiment analysis: evolution of research methods and topics.* Artificial Intelligence Review, 2023. **56**(8): p. 8469-8510.
2. Nabil, M., M. Aly, and A. Atiya. *Astd: Arabic sentiment tweets dataset*. in *Proceedings of the* 2015 conference on empirical methods in natural language processing. 2015.
3. Khalifa, Y. and A. Elnagar. *Colloquial arabic tweets: Collection, automatic annotation, and classification*. in *2020 International Conference on Asian Language Processing (IALP)*. 2020. IEEE.
4. *http://qufaculty.qu.edu.qa/telsayed/datasets/*.
5. Antoun, W., F. Baly, and H. Hajj, *Arabert: Transformer-based model for arabic language understanding.* arXiv preprint arXiv:2003.00104, 2020.
6. Gautam, G. and D. Yadav. *Sentiment analysis of twitter data using machine learning approaches and semantic analysis*. in *2014 Seventh international conference on contemporary computing (IC3)*. 2014. IEEE.
7. Kumar, A. and V.H.C. Albuquerque, *Sentiment analysis using XLM-R transformer and zero-shot transfer learning on resource-poor indian language.* Transactions on Asian and Low-Resource Language Information Processing, 2021. **20**(5): p. 1-13.
8. Wadhawan, A., *Arabert and farasa segmentation based approach for sarcasm and sentiment detection in arabic tweets.* arXiv preprint arXiv:2103.01679, 2021.
9. Manias, G., et al., *Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for classifying twitter data.* Neural Computing and Applications, 2023. **35**(29): p. 21415-21431.
10. Liu, Y., et al., *Roberta: A robustly optimized bert pretraining approach.* arXiv preprint arXiv:1907.11692, 2019.
11. Liao, W., et al., *An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa.* Applied Intelligence, 2021. **51**: p. 3522-3533.
12. Rosenthal, S., N. Farra, and P. Nakov, *SemEval-2017 task 4: Sentiment analysis in Twitter.* arXiv preprint arXiv:1912.00741, 2019.
13. El-Beltagy, S.R. *Nileulex: A phrase and word level sentiment lexicon for egyptian and modern standard arabic*. in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016.
14. Devlin, J., et al., *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805, 2018.
15. Conneau, A., et al., *Unsupervised cross-lingual representation learning at scale.* arXiv preprint arXiv:1911.02116, 2019.