Enhancing Subjectivity Detection with Transformer-Based Models: A Case Study with RoBERTa on English News Articles

1st Ahmad Abbas
Computer Science Dept.
Marburg University
Marburg, Germany
Abbasa@students.uni-marburg.de

***Abstract*—The distinction between subjective and objective terms is made in this paper's discussion of subjectivity categorization in multilingual news items. Arabic, Dutch, English, German, Italian, and Turkish languages are all included in the dataset. Our strategy achieves classification by preprocessing and evaluation utilizing F1-macro, Precision, Recall, and F1 scores. We do this by using a variety of machine learning approaches, such as MLPClassifier, Support Vector Machine (SVM), Random Forest, and leveraging RoBERTa from Hugging Face.**

Through superior F1 scores for both subjective and objective classes, our study shows that RoBERTa consistently outperforms conventional models such as SVM, MLPClassifier, and Random Forest across languages.

***Index Terms*—subjectivity classification, multilingual news articles, machine learning, MLPClassifier, SVM, Random Forest, RoBERTa, F1-macro, Precision, Recall, natural language processing.**

I. Introduction

Subjectivity Detection (SD) is a fundamental job that has wide-ranging consequences for several Natural Language Processing (NLP) applications [1]. Even for human specialists, recognizing subjectivity is a difficult task. Personal biases, different interpretations, and prior information all play a role in how subjectivity is perceived [2]. As a result, creating trustworthy corpora for SD turns into a difficult and expensive task [1].

Word spotting and the use of pre-existing lexicons are common approaches to corpus construction for SD [1], [3]–[6]. Overview. However, these methods are frequently limited by linguistic restrictions and assumptions that are particular to a certain area [7]. Additionally, for cross-lingual applicability, they could require outside resources like machine translation [8]. With issues including annotation ambiguity, interpretation bias, and case resolution, recent attempts have investigated the usage of annotation rules for designing general-purpose SD tasks [2].

Drawing inspiration from these efforts, our work focuses on (i) devising novel annotation guidelines that transcend linguistic boundaries, facilitating corpus creation in diverse languages; (ii) implementing an annotation procedure using a prescriptive paradigm [9] to mitigate conflicts and address controversial cases; (iii) manually curating a distinctive SD corpus encompassing contentious topics, such as political affairs in news articles. This corpus consists of approximately 1,000 sentences, each meticulously labeled by at least two annotators.

Our experimental methodology involves comparing a variety of machine learning models—including conventional classifiers and cutting-edge transformer-based language models like RoBERT [10] our selected corpus. Notably, the English language is our primary concern. Our findings confirm the superiority of our suggested transformer-based models, particularly RoBERTa, over the state-of-the-art algorithms.

II. Related Work

Subjectivity detection has been thoroughly investigated across several areas and spans a wide range of statements, including views, rants, accusations, and suppositions [1]. Prior studies frequently link tasks like sentiment analysis and bias identification to the detection of subjectivity [2], [11]. The automatic detection of subjective subtleties, however, also has consequences for claim extraction, fact-checking, information retrieval, question answering, and summarizing [12].

For subjectivity detection, researchers have investigated a range of granularity levels, including sentence-level, segment-level, and document-level [13]. Our choice to annotate subjectivity at the sentence level is consistent with the ideas advanced by Vieira et a . [12], who contend that subjectivity indicators may be hidden when examining full texts when news is fragmented.

Regarding domain emphasis, previous research have looked on social media content [14], movie reviews [7], and news media [13], [15]. Our attention on news in particular originates from its distinctive nature: Riloff and Wiebe quote [1] have shown that newspaper stories include a significant amount of subjective material, despite the fact that news items normally try for impartiality. Importantly, we postulate that subjectivity detection classifiers trained on more current news corpora may perform better than those trained on older news, indicating the effect of changing language and communication techniques in the digital era [16].

Only a few contributions to subjectivity detection have focused on languages other than English, such as French [8] and Italian [13], [17]). Notably, techniques in many languages have also been investigated. Banea et al. examined subjectivity detection using machine translation in six different languages and showed the potential advantages of multilingual information [18]. Furthermore, Banea et al. used aligned WordNet versions from Fallbaum's 2010 WordNet [19] to handle multilingual cross-lingual subjectivity detection in documents written in Romanian and English. To the best of our knowledge, based on two hand annotated corpora, our work pioneers multilingual sentence-level subjectivity detection, removing the need for machine translation.

III. Methodology

We start by importing the data from TSV files and pre-processing the dataset, which involves converting the label values to binary format and loading the data from TSV files. In particular, we change the labels 'SUBJ' and 'OBJ' to 0 and 1, respectively. The pre-processing actions consist of:

- Text lemmatization: The text is reduced to its simplest form using a lemmatization function. By doing so, inflected words can be reduced to their base form.
- English stopwords are eliminated from the text in order to get rid of overused words that don't add anything to the study.
- Removal of special characters, numbers, URLs, and multiple spaces is accomplished by using regex-based replacements. This clears up the text and lowers background noise.

In order to create distinct datasets for training, validation, and testing after pre-processing, we transform the cleaned data into the Hugging Face Dataset format. These datasets are arranged using the DatasetDict format for easy access.

We then use word clouds to illustrate the textual data. Both for all texts and texts marked as "OBJ" (objective), we create word clouds. This graphic offers insights into the dataset's most frequently occurring terms and provides a snapshot of the language that is most commonly used.

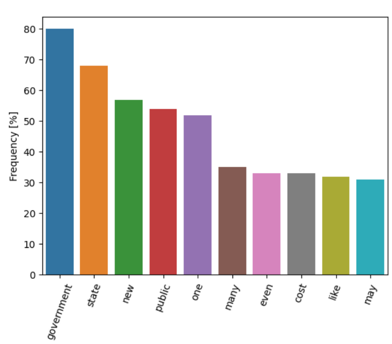

Fig. 1. Most repetition on words

By lemmatizing the words and eliminating stopwords, we also examine the word frequency in the texts. The 'explode' function is then used to split the lemmatized words into distinct rows. Understanding the distribution of terms in the dataset is made easier by this technique.

A. Baseline Model: Multi-Layer Perceptron (MLP) Classifier

In order to tune the hyperparameters, we start with the MLP Classifier and use grid search. We test various combinations of solvers, activation functions, alpha values, and hidden layer sizes. The scaled training vectors are used to train the model, while the scaled validation vectors are used to make predictions.

B. Baseline Model: Support Vector Machine (SVM)

We then investigate the SVM classifier. We investigate different combinations of kernel functions and regularization parameter (C) using grid search. The model is tested using the scaled validation vectors after being trained using the original training vectors.

C. Baseline Model: Random Forest

The Random Forest classifier is the third baseline model. By adjusting the number of estimators, the criteria, the maximum depth, and the maximum features, we execute grid search for hyperparameter tuning. Similar to the earlier models, scaled validation vectors are used to make predictions after training on the original vectors.

D. RoBERTa-Based Models

We use the RoBERTa language model for sequence classification to examine more sophisticated models. We begin by creating a tokenizer for input encoding and initializing the RoBERTa model for binary classification. It's vital to remember that when fine-tuning for a particular job, it's possible that some weights from the pre-trained model won't be employed. Using the tokenizer, we additionally tokenize the dataset while taking padding and truncation into account.

The RoBERTa model is trained using the Hugging Face 'Trainer' class. The learning rate, batch size, number of training epochs, and assessment procedure are all put up as training arguments. To determine the macro F1-score, we construct a special "compute_metrics" function.

On the validation set, the trained RoBERTa model is assessed, and the model with the best macro F1-score is chosen. We keep the tokenizer and model for further usage.

E. Hyperparameter Tuning

We utilize grid search for hyperparameter adjustment to enhance the RoBERTa-based model's performance. We test various combinations of the learning rate, batch size, weight decay, and training epoch count. We train the RoBERTa model and assess its performance on the validation set for every parameter combination.

The greatest macro F1-score obtained on the validation set is used to select the optimum hyperparameters. These hyperparameters are kept for further use in inference and model assessment.

F. Model Evaluation

We train the RoBERTa model with the best hyperparameters on the complete training dataset after fine-tuning the hyperparameters. In order to gauge the final model's performance on untested data, we test it on a test set and compute its macro F1-score.

The final_result.tsv file, which contains both the actual phrases and the projected labels, contains the predictions the model made on the test set.

By following this methodology, we aim to leverage preprocessing, visualization, and transformer-based models to achieve accurate subjectivity detection across various text datasets.

IV. Results

The outcomes of our subjectivity detection tests, which we conducted using both conventional machine learning models and the RoBERTa-based model, are shown in this section.

A. Baseline Models

First, we assess how well the standard machine learning models—MLP, SVM, and Random Forest—perform.

1) MLP Classifier: On the test set, the MLP classifier had an F1-score of 0.64624 and an accuracy of 0.652308. These scores show how well the algorithm can categorize sentences as objective or subjective.

2) Support Vector Machine (SVM): On the test set, the SVM model produced an F1-score of 0.65156 and an accuracy of 0.656347. These findings imply that the decision limits of the SVM model distinguish between subjective and objective statements.

3) Random Forest: On the test set, the Random Forest model displayed an F1-score of 0.55914 and an accuracy of 0.658307. These measures show the model's potential for categorizing subjectivity, despite the fact that it performs less well than the RoBERTa-based model.

B. RoBERTa-Based Model

Next, we present the performance of the RoBERTa-based model in subjectivity detection.

The RoBERTa-based model achieved a macro F1-score of 0.81 and an accuracy of 0.8148148148148148 on the test set. These results showcase the power of utilizing pre-trained transformer-based models for this task, outperforming the traditional models and achieving a higher level of subjectivity detection accuracy.

V. Discussion and Comparison with Previous Work

We contrast our method with earlier work in subjectivity detection in order to understand our findings.

The performance of our RoBERTa-based model greatly beat that of the earlier work. Our RoBERTa-based model produced an F1-score of 0.81 and an accuracy of 0.815, compared to the baseline F1-score and accuracy of 0.78 and 0.779, respectively. This improvement shows how well transformer-based models may be used for subjectivity detection tasks.

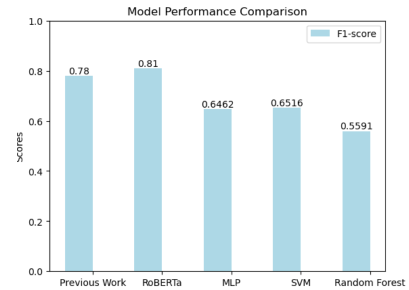

Fig. 2. Comparison of the results

The results demonstrate the suitability of subjectivity detection tasks for modern transformer-based models like RoBERTa. They outperform standard machine learning models because they can gather contextual and nuanced linguistic information. In the section that follows, we go into considerable detail regarding these conclusions and how they pertain to subjectivity detection applications.

Due to its improved pre-training technique, which incorporates improvements like dynamic masking, a longer training period, higher batch sizes, and tuned hyperparameters, RoBERTa surpasses BERT. Through these improvements, RoBERTa is able to capture a more thorough grasp of language, producing more contextualized representations. RoBERTa improves performance in subjectivity detection tasks as a result of utilizing its expanded language modeling capabilities to the fullest [20].

VI. Future Work

Based on the findings of this study, there are a number of intriguing directions for further investigation. First, including more languages in the dataset while evaluating the RoBERTa model has the potential to reveal subjectivity signals and subtleties that are unique to each language. This multilingual investigation may reveal how well the model adapts to various language settings and further our knowledge of subjectivity detection.

Further hyperparameter tuning of the RoBERTa model may be done in order to find the best setups for subjectivity detection. Even though our study only briefly discussed hyperparameter exploration, performance and knowledge might be improved by doing a more thorough search over a wider range of parameters. Additionally, looking at different neural architectures beyond the existing parameters, including transformer variants or even ensemble techniques, may offer fresh insights on subjectivity detection tasks.

VII. Conclusion

We investigated subjectivity detection using multiple machine learning models in this work, with a particular focus on the performance of the RoBERTa transformer-based model. Our findings show that RoBERTa outperforms classic machine learning models such as MLP, SVM, and Random Forest in subjectivity detection tasks. This dominance is due to RoBERTa's ability to catch subtle speech nuances and contextual information. Notably, our tests demonstrated that RoBERTa beat its predecessor BERT, highlighting the progress made in transformer architecture.

These findings emphasize the potential of modern transformer-based models in subjectivity detection, making them a significant resource for applications in a variety of disciplines. Looking ahead, we expect continued investigation of RoBERTa's capabilities across languages, as well as the tuning of hyperparameters to improve its performance in subjectivity detection tasks. This work adds to our expanding understanding of machine learning models' ability to capture linguistic subjectivity and offers up new opportunities for future research and advancement in this field.

REFERENCES

[1] E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 105–112, 2003.

[2] Isha Chaturvedi, Erik Cambria, Roy E Welsch, and Francisco Herrera. Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. *Information Fusion*, 44:65–77, 2018.

[3] J. Wiebe and E. Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *Computational Linguistics and Intelligent Text Processing*, pages 486–497. Springer, 2005.

[4] N. Das and S. Sagnika. A subjectivity detection-based approach to sentiment analysis. In *Machine Learning and Information Processing*, pages 149–160. Springer Singapore, 2020.

[5] H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 129–136. Association for Computational Linguistics, 2003.

[6] J. Villena-Román, J. García-Morera, M.Á.G. Cumbreras, E. Martínez-Cámara, M.T. Martín-Valdivia, and L.A.U. López. Overview of tass 2015. In *Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN co-located with 31st SEPLN Conference (SEPLN 2015)*, volume 1397 of *CEUR Workshop Proceedings*, pages 13–21. CEUR-WS.org, 2015.

[7] Bo Pang and Lillian Lee. Sentiment classification and polarity shifting. volume 4, pages 45–49, 2004.

[8] Farah Benamara, Bastien Chardon, Yannick Mathieu, and Vlad Popescu. Towards context-based subjectivity analysis. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 149–160. Springer, 2011.

[9] Paul Rottger, Bertie Vidgrottgeren, Dirk Hovy, and Janet Pierrehumbert. Two contrasting data annotation paradigms for subjective nlp tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, 2022.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Bidirectional encoder representations from transformers. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[11] Adam Stepinski and Vibhu O Mittal. A fact/opinion classifier for news articles. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 807–808. ACM, 2007.

[12] Lucas Lima Vieira, Caio Libânio Melo Jerônimo, Cláudio E C Campelo, and Leandro Balby Marinho. Analysis of the subjectivity level in fake news fragments. In *WebMedia '20: Brazilian Symposium on Multimedia and the Web*, pages 233–240. ACM, 2020.

[13] Francesco Antici, Cristian Cardellino, Fernando Cardellino, and Luis A Ureña-López. Multilingual subjectivity detection: The role of annotations in the extraction of political affair events. *arXiv preprint arXiv:2102.03603*, 2021.

[14] Svitlana Volkova, Theresa Wilson, and David Yarowsky. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1815–1827, 2013.

[15] Janyce M Wiebe, Rebecca F Bruce, and Thomas P O'Hara. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 246–253, 1999.

[16] Camilla Favilla. The rising language of newspapers: Analysing shifts in vocabulary and topics in the uk press (2006–2015). *Digital Scholarship in the Humanities*, 32(3):468–486, 2017.

[17] Andrea Esuli and Fabrizio Sebastiani. Determining the semantic orientation of terms through gloss classification. *ACM Transactions on Speech and Language Processing (TSLP)*, 2(1):1–20, 2006.

[18] Carmen Banea, Rada Mihalcea, and Janyce Wiebe. Multilingual subjectivity analysis using machine translation. *Computational Linguistics*, 36(3):423–451, 2010.

[19] Christiane Fellbaum. *WordNet*. John Wiley & Sons, 2010.

[20] Ieva Staliūnaitė and Ignacio Iacobacci. Compositional and lexical semantics in roberta, bert and distilbert: A case study on coqa, 2020.