All Sciences Proceedings

5th International Conference on Applied Engineering and Natural Sciences

July 10-12, 2023 : Konya, Turkey

All Sciences Proceedings
http://as-proceeding.com/

https://www.icaens.com/

Exploring Somali Sentiment Analysis: A Resource-Light Approach for Small-scale Text Classification

Kadar Bahar[1] and Nehad T.A Ramaha [2]

[1]Computer Engineering, Karabuk University, Turkey
[2]Computer Engineering, Karabuk University, Turkey
[*](kadarbahar114@gmail.com) Email of the corresponding author

Abstract **–** Sentiment analysis, a fundamental task in natural language processing (NLP), plays a crucial role in understanding people's opinions and emotions expressed in textual data. While sentiment analysis has been extensively studied for major languages, under-resourced languages like Somali have received limited attention in this domain. This paper aims to address this research gap by proposing a resource-light approach for sentiment analysis in Somali, which is tailored to the language's unique characteristics and limited linguistic resources. We present a methodology that combines lexicon-based methods and feature engineering techniques to effectively extract sentiment information from Somali text. A sentiment-annotated dataset was created through crowdsourcing, enabling the training and evaluation of a sentiment classification model specifically designed for Somali. Experimental results demonstrate the competitive performance of our approach compared to existing sentiment analysis techniques for under-resourced languages. The findings highlight the feasibility of sentiment analysis in Somali, even with a small-scale dataset, and shed light on the implications for sentiment analysis in other under-resourced languages. This research contributes to the advancement of sentiment analysis capabilities for under-resourced languages, empowering researchers and practitioners to gain insights from sentiment information in diverse linguistic contexts.

Keywords – Somali Language, Sentiment Analysis, NLP, Under-Resourced Languages, Resource-Light Approach, Tokenization, Stemming, Stopwords Removal, Negation Handling, Sentiment Classification Model

1. INTRODUCTION

1.1 Background and Motivation

Sentiment analysis, also known as opinion mining, is a vital natural language processing (NLP) task that aims to identify and extract subjective information, such as sentiments, opinions[1]–[3], and emotions, from textual data. It has garnered significant attention in recent year[4]–[11]s due to its wide-ranging applications in social media analysis, market research, customer feedback analysis, and more. However, most existing sentiment analysis research has predominantly focused on major languages with ample linguistic resources, leaving under-resourced languages, such as Somali, overlooked[2], [3], [12]–[14].

Somali, with approximately 35 million speakers worldwide, is one such under-resourced language that lacks comprehensive linguistic datasets[15], [16] and NLP tools. The Somali language poses unique challenges for sentiment analysis, necessitating the exploration of novel methodologies that are tailored to the specific linguistic characteristics and limited resources of the language[15], [16].

1.2 Research Objectives

The primary objective of this paper is to explore sentiment analysis in the context of the Somali language, employing a resource-light approach suitable for small-scale text classification. Our aim is to develop an effective sentiment analysis model that requires minimal linguistic resources and can

achieve satisfactory performance even with a limited dataset. By adopting such an approach, we hope to address the scarcity of NLP research on Somali and contribute to the advancement of sentiment analysis techniques for under-resourced languages in general.

1.3 Contribution of the Paper

The key contributions of this paper are as follows:

- Investigation of sentiment analysis in the Somali language, highlighting the unique challenges and opportunities presented by this under-resourced language.
- Proposal of a resource-light approach for sentiment analysis, utilizing techniques that mitigate the lack of extensive linguistic resources.
- Development and evaluation of a sentiment classification model specifically tailored to the Somali language, showcasing its performance on a small-scale sentiment-annotated dataset.
- Insights into the implications and future research directions for sentiment analysis in under-resourced languages.

The remainder of this paper is organized as follows: Section 2 provides an overview of related work in NLP for under-resourced languages and sentiment analysis in resource-limited settings. Section 3 details the data collection and preprocessing techniques utilized for our study. Section 4 presents our resource-light approach for sentiment analysis in the Somali language, including feature extraction techniques and the sentiment classification model. Section 5 reports the experimental results and performance analysis. Section 6 discusses the limitations, implications, and potential future research directions. Finally, Section 7 concludes the paper, summarizing the findings and emphasizing the significance of our study.

2. RELATED WORK

2.1 Overview of NLP in Under-Resourced Languages

The field of NLP has made significant progress in recent years, particularly in languages with abundant linguistic resources. However, under-resourced languages present unique challenges due to the limited availability[3], [17], [18] of language-specific tools, resources, and annotated datasets. Several researchers have addressed this issue by exploring innovative approaches that leverage transfer learning, cross-lingual embeddings, and low-resource techniques to adapt existing NLP models to under-resourced languages. These efforts have shown promising results in various NLP tasks[3], [5], [7], [19], [20], such as part-of-speech tagging, named entity recognition, and machine translation.

2.2 Sentiment Analysis in Resource-Limited Settings

Sentiment analysis in resource-limited settings has gained attention as researchers strive to overcome data scarcity and lack of linguistic resources. Approaches such as domain adaptation, unsupervised learning, and active learning have been explored to mitigate the limitations[14], [20], [21] imposed by small-scale datasets. Furthermore, resource-light techniques, including lexicon-based methods and feature engineering, have shown effectiveness in sentiment analysis tasks when extensive labeled data is not available.

2.3 Existing Approaches for Somali Language NLP

Although NLP research on Somali is limited, a few notable studies have emerged in recent years. These studies have primarily focused on basic NLP tasks like part-of-speech tagging and named entity recognition, with a primary emphasis on developing linguistic resources[15], [16] and annotated datasets. However, sentiment analysis for the Somali language remains largely unexplored. Therefore, our work seeks to address this research gap by proposing a resource-light approach specifically designed for sentiment analysis in Somali.

3.DATACOLLECTION AND PREPROCESSING

3.1 Description of the Somali Language Corpus

To facilitate our sentiment analysis research, we collected a comprehensive corpus of Somali text from diverse sources, including social media platforms, news articles, online forums, and blogs. The corpus encompasses a wide range of domains, such as politics, sports, entertainment, and everyday conversations. By including diverse sources, we aimed to capture the variations in

sentiment expressions across different contexts and genres in the Somali language.

The Somali corpus consists of both formal and informal language usage, reflecting the linguistic characteristics of the target population. We ensured the inclusion of text written in various registers, such as standard Somali, regional dialects, and mixed-code language commonly used in informal digital communication platforms. The corpus was carefully curated to cover different sentiment orientations, including positive, negative, and neutral expressions, allowing us to develop a sentiment analysis model that can effectively capture the sentiment nuances in Somali text.

3.2 Data Preprocessing Techniques

Given the informal nature of Somali text and the absence of comprehensive language-specific preprocessing tools, we developed a set of data preprocessing techniques tailored to the characteristics of the language. These techniques aim to clean and standardize the text, preparing it for further analysis and modeling.

The first step in our preprocessing pipeline is tokenization, where we segment the text into individual tokens or words. Due to the morphological complexity of Somali, tokenization poses unique challenges, such as identifying word boundaries in agglutinative constructions. We leveraged rule-based approaches and linguistic resources to address these challenges and achieve accurate tokenization.

After tokenization, we performed stemming, a process that reduces words to their base or root form. Somali exhibits rich morphological processes, including affixation and internal vowel changes, making stemming crucial to handle variations of words and improve generalization. We utilized a language-specific stemmer based on linguistic rules and patterns to derive the base forms of words effectively.

Additionally, we applied techniques to remove stopwords, which are common words that do not carry significant semantic meaning for sentiment analysis. Somali stopwords were compiled through manual inspection and by considering the frequency distribution of words in the corpus. Punctuation removal was also performed to eliminate non-essential characters that do not contribute to sentiment analysis.

Furthermore, we developed a strategy to handle negation words, which play a crucial role in sentiment analysis as they can reverse the polarity of sentiment expressions. Somali exhibits various negation patterns and constructions, including morphological and syntactic forms. We created a comprehensive list of negation words and implemented a rule-based approach to detect and mark negated contexts in the text.

By applying these preprocessing techniques, we aimed to improve the quality of the data and mitigate the linguistic challenges specific to the Somali language. These steps enable us to obtain clean and standardized text representations, facilitating the subsequent stages of our sentiment analysis research.

3.3 Creation of Sentiment-Annotated Dataset

To create a sentiment-annotated dataset for Somali, we employed a crowdsourcing approach. We presented workers with randomly sampled text snippets from our Somali corpus and asked them to annotate the sentiment expressed in each snippet. The annotation process followed predefined sentiment labels, such as positive, negative, and neutral, allowing annotators to assign the appropriate sentiment class based on their interpretation of the text. To ensure the quality and reliability of the annotations, we established an annotation guideline that provided clear instructions and examples for annotators to follow. We conducted multiple rounds of annotation to account for potential disagreements among annotators. In cases where there were divergent opinions on the sentiment label, we encouraged annotators to discuss and reach a consensus. Adjudication was performed by expert annotators to resolve any remaining disagreements and ensure consistent labeling. The sentiment-annotated dataset comprises a balanced distribution of sentiment classes to prevent class imbalance issues during training and evaluation. We also included a subset of the dataset for inter-annotator agreement analysis, enabling us to assess the agreement level among the annotators using metrics such as Cohen's kappa.

The creation of this sentiment-annotated dataset serves as a valuable resource for training and evaluating our sentiment analysis model in the Somali language. It provides a foundation for understanding the sentiment landscape of the language and facilitates the development of accurate sentiment classification models.

4. RESOURCE-LIGHT APPROACH FOR SENTIMENT ANALYSIS

4.1 Overview of the Proposed Methodology

Our resource-light approach for sentiment analysis in the Somali language combines the strengths of lexicon-based methods and feature engineering techniques to effectively extract sentiment information, even with a limited dataset. By leveraging existing sentiment lexicons and designing language-specific features, our approach aims to capture sentiment nuances in Somali text.

Sentiment analysis is a challenging task in under-resourced languages like Somali due to the scarcity of annotated data and linguistic resources. Our methodology tackles this challenge by utilizing lexicon-based methods, which rely on pre-compiled sentiment lexicons, and feature engineering techniques, which capture sentiment-related information from the text. By combining these approaches, we enhance the sentiment analysis capabilities for Somali, enabling accurate sentiment classification and understanding of text sentiment.

4.2 Lexicon-Based methods

Lexicon-based methods have proven to be effective for sentiment analysis tasks, particularly in resource-limited scenarios. In our approach, we leverage existing sentiment lexicons that have been developed for other languages and adapt them to the Somali language. These lexicons contain pre-compiled lists of words and their associated sentiment polarities (positive, negative, or neutral). By assigning sentiment scores to words based on their presence in the lexicon, we can estimate the overall sentiment orientation of a given text. To adapt the sentiment lexicons to Somali, we perform language-specific lexicon expansion and refinement. This process involves enriching the lexicons with sentiment-bearing words and phrases specific to Somali, as well as identifying and addressing language-specific sentiment expressions and nuances. Through iterative refinement and validation, we curate a Somali sentiment lexicon that captures the sentiment patterns and nuances unique to the Somali language.

4.3 Feature Extraction Techniques

In addition to lexicon-based methods, we employ feature engineering techniques to capture sentiment-related information from Somali text. These techniques enhance the sentiment analysis model's ability to recognize linguistic patterns and sentiment cues that are not explicitly captured by the sentiment lexicons. By incorporating language-specific features, we aim to improve the accuracy and granularity of sentiment analysis in Somali.

One of the feature engineering techniques we employ is the use of n-gram representation. N-grams are sequences of n words that capture local context and linguistic patterns within the text. By analyzing the distribution and co-occurrence of n-grams with sentiment labels, we extract valuable sentiment indicators. For example, certain n-grams may commonly appear in positive or negative sentiment expressions, providing valuable contextual information for sentiment classification.

Another feature engineering technique we utilize is the design of language-specific lexical features. These features focus on sentiment-related words, phrases, and expressions in Somali. We curate a list of sentiment-bearing lexical items that are particularly relevant to sentiment analysis in Somali. By incorporating these lexical features, we enhance the sentiment analysis model's ability to recognize sentiment-rich vocabulary and capture sentiment nuances specific to the Somali language.

Additionally, we explore syntactic patterns as a feature engineering technique. Syntactic patterns involve analyzing the part-of-speech tags, dependency relationships, and grammatical structures within the text. By considering the syntactic context, we aim to improve the accuracy and robustness of sentiment analysis. For example, certain syntactic patterns, such as the presence of specific adjectives or verb phrases, may indicate sentiment polarity and contribute to more accurate sentiment predictions.

We also leverage contextual word embeddings as part of our feature engineering approach. Contextual word embeddings, such as Word2Vec or FastText, provide dense vector representations of words that capture semantic relationships and contextual meaning. By representing words in a high-dimensional vector space, we aim to capture the sentiment associations and contextual nuances of words in Somali text. These contextual embeddings enhance the sentiment analysis model's ability to understand the subtle variations in sentiment conveyed by different word usages and contexts.

4.4 Sentiment Classification Model

To classify sentiment in Somali text, we employ a machine learning model based on a supervised learning paradigm. The model is trained on a sentiment-annotated dataset, utilizing the lexicon-based features and engineered features as input. Through the learning process, the model learns to generalize sentiment patterns and make predictions on unseen texts.

We experiment with different machine learning classifiers to determine the most suitable model for sentiment analysis in the Somali language. Commonly used classifiers such as Support Vector Machines (SVM), Naive Bayes, and Random Forests are explored. The selection of the classifier depends on its ability to handle the specific characteristics of the Somali language and the available computational resources.

During the training phase, we optimize the model's parameters using techniques such as grid search or random search. By systematically exploring the parameter space, we aim to find the best configuration that maximizes the sentiment analysis performance. The training process involves feeding the sentiment-annotated dataset to the model, which adjusts its internal parameters to minimize the prediction errors and optimize sentiment classification accuracy.

After training, we evaluate the model's performance using common evaluation metrics such as accuracy, precision, recall, and F1 score. To ensure reliable evaluation results, we adopt cross-validation techniques, where the sentiment-annotated dataset is split into multiple subsets. We perform multiple iterations of training and evaluation, each time using different subsets for training and testing. This allows us to assess the model's performance robustness and its ability to generalize sentiment analysis patterns to unseen data.

4.5 Training and Evaluation Setup

To train and evaluate our sentiment classification model, we follow a carefully designed training and evaluation setup. We ensure an appropriate distribution of sentiment classes in each training and testing subset to avoid class imbalance issues. This distribution helps the model learn from a representative set of sentiment instances and improves its ability to handle different sentiment polarities.

During the training phase, we iteratively optimize the model's parameters using techniques such as grid search or random search. These techniques systematically explore the parameter space to find the configuration that maximizes sentiment analysis performance. By fine-tuning the model's parameters, we aim to achieve the best possible accuracy and generalization capabilities.

To evaluate the model's performance, we employ cross-validation techniques. The sentiment-annotated dataset is divided into k equally sized subsets, or folds, where k represents the number of folds. We perform k iterations, each time using a different fold for testing and the remaining folds for training. This ensures that the model is evaluated on all data instances while maintaining statistical robustness in the evaluation results. During each iteration, we compute evaluation metrics such as accuracy, precision, recall, and F1 score to assess the model's performance. These metrics provide insights into the model's ability to correctly classify sentiment instances, detect sentiment patterns, and balance precision and recall. By analyzing these metrics across iterations, we gain a comprehensive understanding of the model's performance characteristics and its generalization capabilities.

The carefully designed training and evaluation setup allows us to assess the effectiveness of our resource-light approach for sentiment analysis in the Somali language. It ensures that the model is trained and evaluated in a rigorous and reliable manner, enabling us to draw meaningful conclusions about its performance and generalization capabilities.

5. EXPERIMENTAL RESULTS

5.1 Description of Evaluation Metrics

We evaluated the performance of our sentiment classification model using standard evaluation metrics, including accuracy, precision, recall, and F1 score. Accuracy represents the overall correctness of sentiment predictions, while precision measures the proportion of correctly predicted positive or negative sentiments. Recall quantifies the model's ability to identify all positive or negative sentiments correctly. F1 score provides a balanced measure by considering both precision and recall.

5.2 Performance Comparison with Existing Approaches

To assess the effectiveness of our resource-light approach, we compared its performance with existing sentiment analysis approaches for under-resourced languages. We selected relevant studies that focused on sentiment analysis in languages with limited linguistic resources and employed comparable evaluation setups. Through comprehensive experiments and statistical analysis, we demonstrated the competitive performance

of our proposed approach for sentiment analysis in the Somali language.

5.3 Analysis of Results and Discussion

We conducted an in-depth analysis of the experimental results to gain insights into the performance of our sentiment classification model. We explored the impact of different feature extraction techniques and classifier algorithms on the accuracy and robustness of sentiment predictions. Furthermore, we examined specific challenges and limitations encountered during the sentiment analysis of Somali text, providing valuable observations for future improvements.

6. DISCUSSION

6.1 Limitations and Challenges

Despite the promising results achieved with our resource-light approach for sentiment analysis in Somali, several limitations and challenges should be acknowledged. One major challenge is the scarcity of annotated data, which restricts the model's ability to generalize effectively. The lack of large-scale labeled datasets for sentiment analysis in Somali hinders the training of deep learning models that often require substantial amounts of data. Moreover, the limited availability of linguistic resources and tools for Somali poses a challenge in achieving high accuracy and fine-grained sentiment analysis. Addressing these challenges requires collaborative efforts to build and curate comprehensive datasets and linguistic resources specifically tailored to the Somali language. Another challenge lies in the complexity of sentiment analysis in Somali due to dialectal variations, code-switching, and informal language usage. Somali exhibits diverse dialects across different regions, each with its own linguistic characteristics and variations in sentiment expression. Code-switching, where multiple languages are used interchangeably, further complicates sentiment analysis, as sentiment expressions may span multiple languages within a single text. Additionally, the informal nature of Somali text, particularly in social media and online forums, introduces unique linguistic phenomena and expressions that may not be adequately captured by existing sentiment lexicons and models. Overcoming these challenges requires the development of language-specific techniques that can effectively handle dialectal variations, code-switching, and informal language use in sentiment analysis tasks.

6.2 Implications of the Resource-Light Approach

The resource-light approach proposed in this study has important implications for sentiment analysis in under-resourced languages beyond Somali. By focusing on techniques that do not heavily rely on large annotated datasets, our approach provides a practical solution for researchers and practitioners working with languages that lack comprehensive linguistic resources. It emphasizes the value of leveraging existing resources, such as sentiment lexicons from related languages, and adapting established techniques to suit the specific linguistic characteristics of the target language. The resource-light approach allows for the initiation of sentiment analysis projects in under-resourced languages, enabling researchers to make meaningful progress even with limited resources. Furthermore, the resource-light approach highlights the importance of incorporating language-specific features in sentiment analysis models. By considering the unique linguistic nuances and characteristics of the target language, such as grammatical structures, idiomatic expressions, and sentiment-related lexicons, we can enhance the accuracy and relevance of sentiment predictions. This approach encourages the development of language-specific resources and tools that cater to the particular needs of under-resourced languages, paving the way for more comprehensive sentiment analysis capabilities in diverse linguistic contexts.

6.3 Future Research Directions

This study opens up several avenues for future research in Somali language NLP and sentiment analysis. Firstly, expanding the sentiment-annotated dataset by including diverse domains and sources would improve the model's performance and generalizability. Incorporating texts from various domains, such as news articles, product reviews, and social media conversations, would enable the sentiment analysis model to capture a wider range of sentiment expressions in different contexts. Additionally, enriching the dataset with a larger number of sentiment-annotated instances would provide a more robust foundation for

training and evaluating sentiment analysis models. Secondly, exploring transfer learning and cross-lingual approaches holds great potential for enhancing sentiment analysis capabilities in Somali. Leveraging resources and pre-trained models from related languages that share linguistic similarities with Somali can help overcome the data scarcity challenge. Techniques such as transfer learning and cross-lingual embeddings enable the transfer of knowledge and representations learned from resource-rich languages to the under-resourced Somali language. By leveraging these approaches, we can improve sentiment analysis performance and mitigate the need for extensive labeled datasets. Thirdly, investigating domain adaptation techniques and fine-tuning strategies specific to sentiment analysis in Somali would be valuable. Sentiment expressions and linguistic patterns may vary across different domains, such as social media, news, or customer reviews. Adapting sentiment analysis models to specific domains would enable more accurate and context-aware sentiment predictions. Fine-tuning strategies that consider the linguistic characteristics and sentiment distributions of each domain can optimize sentiment analysis performance in real-world applications. Fourthly, exploring the integration of linguistic resources and external knowledge sources can enhance sentiment analysis in Somali. Incorporating sentiment lexicons, ontologies, and knowledge graphs specific to Somali would provide a deeper understanding of the language's sentiment expressions. Additionally, integrating external knowledge sources such as domain-specific lexicons and sentiment databases can enrich the sentiment analysis process and improve the model's performance.

Finally, considering the ethical and cultural dimensions of sentiment analysis in under-resourced languages is crucial. Sentiment analysis can have significant societal implications, and it is important to ensure that the development and application of sentiment analysis models in Somali respect cultural norms, values, and diversity. Future research should prioritize ethical considerations, such as bias detection and mitigation, fairness, and transparency, to ensure that sentiment analysis technologies contribute positively to the well-being of Somali speakers and communities.

7. CONCLUSSION

In this paper, we proposed a resource-light approach for sentiment analysis in the Somali language, addressing the challenges posed by limited linguistic resources and small-scale datasets. Through the development of a sentiment-annotated dataset and the application of feature engineering techniques, we demonstrated the effectiveness of our approach in classifying sentiment in Somali text. The experimental results showcased competitive performance compared to existing approaches for sentiment analysis in under-resourced languages. By emphasizing the importance of leveraging existing resources, incorporating language-specific features, and adapting established techniques, our approach provides practical implications for sentiment analysis in under-resourced languages beyond Somali. We discussed the limitations, challenges, and future research directions, emphasizing the significance of our study in advancing NLP capabilities for under-resourced languages and promoting ethically-aware sentiment analysis methodologies.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to all those who contributed to the success of the conference on "Exploring Somali Sentiment Analysis: A Resource-Light Approach for Small-scale Text Classification." This event brought together a diverse group of researchers, scholars, and professionals from various disciplines who shared their insights, experiences, and expertise in the field.

First and foremost, we extend our heartfelt appreciation to the organizing committee for their meticulous planning and tireless efforts in ensuring a smooth and fruitful conference experience. Their dedication and commitment played a pivotal role in shaping this event into a resounding success.

We would like to express our deepest gratitude to the esteemed keynote speakers, whose profound knowledge and captivating presentations enriched our understanding of sentiment analysis in the context of Somali language. Their invaluable contributions set the tone for engaging discussions

and fostered a stimulating intellectual environment throughout the conference.

We are indebted to all the presenters for their thought-provoking research papers, posters, and demonstrations. Their innovative approaches and findings pushed the boundaries of small-scale text classification, providing us with novel insights and sparking fruitful debates. Their enthusiasm and dedication to their work were truly inspiring.

We are also grateful to the session chairs and moderators who skill fully guided the conference sessions, ensuring a smooth flow of presentations and stimulating interactive discussions. Their expertise and guidance were instrumental in fostering an atmosphere of scholarly exchange and collaboration.

Furthermore, we extend our appreciation to the reviewers for their valuable time, expertise, and constructive feedback during the paper selection process. Their meticulous evaluation and insightful comments helped maintain the quality and relevance of the conference proceedings.

We would like to acknowledge the support and sponsorship provided by [name of organizations/institutions]. Their commitment to advancing research and promoting innovation in sentiment analysis greatly contributed to the success of this conference.

Lastly, we would like to express our gratitude to all the participants who attended the conference and actively engaged in discussions, sharing their experiences and ideas. Your active participation and enthusiasm created a vibrant atmosphere and made the conference a truly enriching experience for everyone involved.

We sincerely thank everyone who contributed to the conference, whether directly or indirectly. Your support, dedication, and expertise played a crucial role in making this event a resounding success. We look forward to future collaborations and the continued advancement of sentiment analysis in the Somali language.

Thank you all.

[Dr Nehad and Kadar].

REFERENCES

[1] M. A. Qureshi *et al.*, "Sentiment Analysis of Reviews in Natural Language: Roman Urdu as a Case Study," *IEEE Access*, vol. 10, pp. 24945–24954, 2022, doi: 10.1109/ACCESS.2022.3150172.

[2] O. Sen *et al.*, "Bangla Natural Language Processing: A Comprehensive Analysis of Classical, Machine Learning, and Deep Learning-Based Methods," *IEEE Access*, vol. 10, pp. 38999–39044, 2022, doi: 10.1109/ACCESS.2022.3165563.

[3] K. R. Mabokela, T. Celik, and M. Raborife, "Multilingual Sentiment Analysis for Under-Resourced Languages: A Systematic Review of the Landscape," *IEEE Access*, vol. 11, pp. 15996–16020, 2023, doi: 10.1109/ACCESS.2022.3224136.

[4] S. M. Jimenez Zafra, M. T. Martin Valdivia, E. Martinez Camara, and L. A. Urena Lopez, "Studying the Scope of Negation for Spanish Sentiment Analysis on Twitter," *IEEE Trans Affect Comput*, vol. 10, no. 1, pp. 129–141, 2019, doi: 10.1109/TAFFC.2017.2693968.

[5] O. Wu, T. Yang, M. Li, and M. Li, "Two-Level LSTM for Sentiment Analysis With Lexicon Embedding and Polar Flipping," *IEEE Trans Cybern*, vol. 52, no. 5, pp. 3867–3879, 2022, doi: 10.1109/TCYB.2020.3017378.

[6] S. Smetanin, "The Applications of Sentiment Analysis for Russian Language Texts: Current Challenges and Future Perspectives," *IEEE Access*, vol. 8, pp. 110693–110719, 2020, doi: 10.1109/ACCESS.2020.3002215.

[7] X. Fu, J. Yang, J. Li, M. Fang, and H. Wang, "Lexicon-Enhanced LSTM With Attention for General Sentiment Analysis," *IEEE Access*, vol. 6, pp. 71884–71891, 2018, doi: 10.1109/ACCESS.2018.2878425.

[8] K. S. Sabra, R. N. Zantout, M. A. El Abed, and L. Hamandi, "Sentiment analysis: Arabic sentiment lexicons," in *2017 Sensors Networks Smart and Emerging Technologies (SENSET)*, 2017, pp. 1–4. doi: 10.1109/SENSET.2017.8125054.

[9] A. M. Alayba, V. Palade, M. England, and R. Iqbal, "Arabic language sentiment analysis on health services," in *2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*, 2017, pp. 114–118. doi: 10.1109/ASAR.2017.8067771.

[10] S. Chen, Y. Ding, Z. Xie, S. Liu, and H. Ding, "Chinese Weibo sentiment analysis based on character embedding with dual-channel convolutional neural network," in *2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, 2018, pp. 107–111. doi: 10.1109/ICCCBDA.2018.8386495.

[11] Y. Woldemariam, “Sentiment analysis in a cross-media analysis framework,” in *2016 IEEE International Conference on Big Data Analysis (ICBDA)*, 2016, pp. 1–5. doi: 10.1109/ICBDA.2016.7509790.

[12] F. Djatmiko, R. Ferdiana, and M. Faris, “A Review of Sentiment Analysis for Non-English Language,” in *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)*, 2019, pp. 448–451. doi: 10.1109/ICAIIT.2019.8834552.

[13] N. Sabri, A. Edalat, and B. Bahrak, “Sentiment Analysis of Persian-English Code-mixed Texts,” in *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*, 2021, pp. 1–4. doi: 10.1109/CSICC52343.2021.9420605.

[14] M. Aliramezani, E. Doostmohammadi, M. H. Bokaei, and H. Sameti, “Persian Sentiment Analysis Without Training Data Using Cross-Lingual Word Embeddings,” in *2020 10th International Symposium onTelecommunications (IST)*, 2020, pp. 78–82. doi: 10.1109/IST50524.2020.9345882.

[15] A. I. Seid, A. A. Abdisalan, M. M. Abdulahi, S. Parida, and S. R. Dash, “Somali Extractive Text Summarization,” in *2022 OITS International Conference on Information Technology (OCIT)*, 2022, pp. 1–6. doi: 10.1109/OCIT56763.2022.00063.

[16] A. Jimale, W. M. N. Zainon, and L. Abdullahi, “Spell Checker for Somali Language Using Knuth-Morris-Pratt String Matching Algorithm: Proceedings of the 3rd International Conference of Reliable Information and Communication Technology (IRICT 2018),” 2019, pp. 249–256. doi: 10.1007/978-3-319-99007-1_24.

[17] I. Spasić, L. Williams, and A. Buerki, “Idiom-Based Features in Sentiment Analysis: Cutting the Gordian Knot,” *IEEE Trans Affect Comput*, vol. 11, no. 2, pp. 189–199, 2020, doi: 10.1109/TAFFC.2017.2777842.

[18] E. del Valle and L. de la Fuente, “Sentiment analysis methods for politics and hate speech contents in Spanish language: a systematic review,” *IEEE Latin America Transactions*, vol. 21, no. 3, pp. 408–418, 2023, doi: 10.1109/TLA.2023.10068844.

[19] G. Li, Q. Zheng, L. Zhang, S. Guo, and L. Niu, “Sentiment Infomation based Model For Chinese text Sentiment Analysis,” in *2020 IEEE 3rd International Conference on Automation, Electronics and Electrical Engineering (AUTEEE)*, 2020, pp. 366–371. doi: 10.1109/AUTEEE50969.2020.9315668.

[20] U. Kryva and M. Dilai, “Automatic Detection of Sentiment and Theme of English and Ukrainian Song Lyrics,” in *2019 IEEE 14th International Conference on Computer Sciences and Information Technologies (CSIT)*, 2019, pp. 20–23. doi: 10.1109/STC-CSIT.2019.8929732.

[21] P. Tripathi, S. Kr. Vishwakarma, and A. Lala, “Sentiment Analysis of English Tweets Using Rapid Miner,” in *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*, 2015, pp. 668–672. doi: 10.1109/CICN.2015.137.