Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.778103

TOPIC BASED FEATURE EXTRACTION MESSAGE CLASSIFICATION AND VISUALIZATION OF SENTIMENT POLARITY

[1,2]Kolluri David Raju [3]Dr. Bipin Bihari Jayasingh
[1] Research Scholar, CSE, Rayalaseema University, Kurnool, AP.
[2]Associate Professor, Dept. of CSE, Hitam, Hyderabad, TS
[3]Professor, Dept. of IT, CVR College of Engineering, Hyderabad, TS

Abstract:
This paper introduces a novel classification system designed for Topic-Based Message Polarity classification, a crucial task in sentiment analysis for determining the positive or negative sentiment of a given message or topic. The system prioritizes the stronger sentiment when both positive and negative sentiments are expressed within the same context. The experimentation phase utilizes a Twitter dataset to assess the effectiveness of the proposed system. The classification model is trained using a Support Vector Machine (SVM) classifier, leveraging insights from a training dataset of tweets. The trained model is then applied to ascertain the sentiment of an anonymous tweet. Various feature sets, including Brown Dictionary Features, Semantic features, linguistic features, word embedding features, and Sentiment Lexicon features, are employed as inputs to the SVM classifier to capture patterns within the training dataset. Notably, this work introduces a set of new features based on word embeddings, supplementing the existing feature set. These novel features are incorporated into the experimentation, demonstrating their efficacy in improving sentiment prediction for topic-based messages. he results indicate that the proposed approach, particularly with the introduction of new features, achieves commendable performance in topic-based sentiment prediction. Comparative analysis against existing solutions in sentiment classification reveals that the proposed features enhance sentiment identification capacity significantly. Furthermore, the study identifies the superiority of the proposed features over existing ones, showcasing their ability to elevate sentiment classification accuracy. This research contributes valuable insights and advancements to the field of sentiment analysis, specifically in the domain of topic-based message polarity classification. The findings highlight the effectiveness of the proposed system and underscore the significance of incorporating innovative features, such as those based on word embeddings, to enhance sentiment prediction models.

Keywords: Semantics, Twitter, Post, Word Embedding, Message, Classification

1. Introduction

In the era of social media dominance, platforms like Twitter have become invaluable sources of real-time information and user-generated content. Understanding and categorizing the sentiments expressed in Twitter posts, commonly known as tweets, is a critical aspect of sentiment analysis. This paper delves into the intricate task of Topic-Based Message Classification of Twitter posts, employing advanced techniques such as word embedding and semantic analysis to unravel the nuances of sentiment within the dynamic and concise nature of tweets [1].

Twitter, with its character-limited messages, presents a unique challenge for sentiment analysis. Traditional methods often struggle to capture the subtleties and context-dependent nature of sentiment in such short texts. As a result, there is a growing need for more sophisticated approaches that go beyond bag-of-words models. This study focuses on leveraging the power of word embedding and semantic analysis to enhance the accuracy and depth of sentiment classification in the context of Twitter posts [2].

Sentiment analysis holds immense significance in various domains, ranging from marketing and business intelligence to political analysis and public opinion tracking. The ability to discern sentiment in Twitter posts allows for a deeper understanding of user attitudes, preferences, and trends, which can be instrumental in making informed decisions and devising targeted strategies [3].

Word embedding, a technique that represents words as dense vectors in a continuous vector space, has emerged as a powerful tool in natural language processing tasks. In the context of Twitter sentiment analysis, word embedding helps capture semantic relationships between words and contextual nuances that traditional methods might overlook. This paper explores the integration of word embedding techniques to enhance the representation of words in Twitter posts and improve the overall accuracy of sentiment classification.

Semantic analysis involves understanding the meaning of words and their relationships in a given context. In the context of Twitter sentiment analysis, semantic analysis allows for a more profound comprehension of the underlying sentiment expressed in tweets. This study investigates how semantic features contribute to the task of Topic-Based Message Classification, shedding light on the contextual intricacies that influence sentiment in Twitter posts.

The primary objectives of this research are to develop an effective Topic-Based Message Classification system for Twitter posts and to assess the impact of incorporating word embedding and semantic analysis. The study aims to address the limitations of existing approaches by providing a more nuanced and context-aware sentiment analysis model tailored for the unique characteristics of Twitter content.

While this study focuses specifically on Twitter posts, the methodologies and insights gained could potentially be extended to other social media platforms. However, it's essential to acknowledge the limitations inherent in sentiment analysis, particularly in the inherent subjectivity and dynamic nature of user-generated content.

In essence, this research endeavours to unravel the intricacies of sentiment within the concise and dynamic world of Twitter posts, offering a novel perspective on Topic-Based Message Classification through the lens of word embedding and semantic analysis.

2. Literature Review

The landscape of sentiment analysis, particularly in the realm of Twitter posts, has evolved significantly in response to the dynamic nature of social media communication. This section reviews key developments in sentiment analysis methodologies, with a focus on Topic-Based Message Classification and the incorporation of word embedding and semantic analysis.

Early sentiment analysis predominantly relied on traditional approaches such as bag-of-words models and sentiment lexicons. While effective to some extent, these methods struggled to

capture the contextual nuances and brevity inherent in Twitter posts. The challenge of discerning sentiment in short and often ambiguous messages led researchers to seek more sophisticated solutions [4].

Twitter, with its 280-character limit, presented a unique challenge for sentiment analysis. Expressing sentiment concisely required algorithms to consider not only the explicit language used but also the implicit meaning and context surrounding the tweets. Traditional models faced limitations in coping with the fast-paced and evolving nature of Twitter conversations [5]-[6].

Word embedding, introduced as a breakthrough in natural language processing, gained traction in sentiment analysis due to its ability to capture semantic relationships between words. In the context of Twitter sentiment analysis, where brevity often results in ambiguity, word embedding techniques like Word2Vec and GloVe became pivotal in enhancing the representation of words and, consequently, sentiment [7].

Recognizing the importance of context in sentiment analysis, researchers delved into semantic analysis to understand the underlying meaning of words in specific contexts. This shift from a purely syntactic approach to a more semantic one enabled sentiment analysis models to grasp the subtle nuances of sentiment expressed in Twitter posts [8]. To improve sentiment prediction models, researchers began incorporating diverse feature sets beyond traditional linguistic features. Brown Dictionary Features, Semantic features, and Sentiment Lexicon features were introduced to capture deeper contextual information. This paved the way for a more comprehensive understanding of sentiment beyond the surface-level analysis.

Recent studies have explored ensemble learning techniques, deep learning architectures, and hybrid models that combine various features for sentiment prediction. The integration of these advanced methodologies showcases a trend toward building more robust and context-aware sentiment analysis systems [9]-[10].

While advancements have been made, challenges persist, particularly in handling sarcasm, irony, and cultural nuances in sentiment expression. The need for models that can adapt to evolving linguistic trends on Twitter remains an ongoing area of exploration.

Topic-Based Message Classification emerged as a crucial task within sentiment analysis, acknowledging that sentiment can vary significantly depending on the context of a conversation. This led to the development of systems like the one proposed in this paper, which prioritizes stronger sentiments in a given context [11]-[12].

In conclusion, the literature review underscores the evolution of sentiment analysis methodologies, especially in the context of Twitter posts. The integration of word embedding and semantic analysis represents a paradigm shift in understanding sentiment nuances. While advancements have been substantial, there is an ongoing need for models that can adapt to the ever-evolving landscape of social media communication. The proposed Topic-Based Message Classification system, with its emphasis on innovative features, adds a valuable dimension to this evolving field. This paper contributes to this ongoing discourse by addressing the limitations of existing approaches and offering a nuanced perspective on sentiment analysis in the dynamic realm of Twitter.

3. Proposed Method

In this study, a novel approach to sentiment polarity classification is introduced, illustrated in Figure 1. This approach encompasses three distinct steps for effective sentiment analysis:

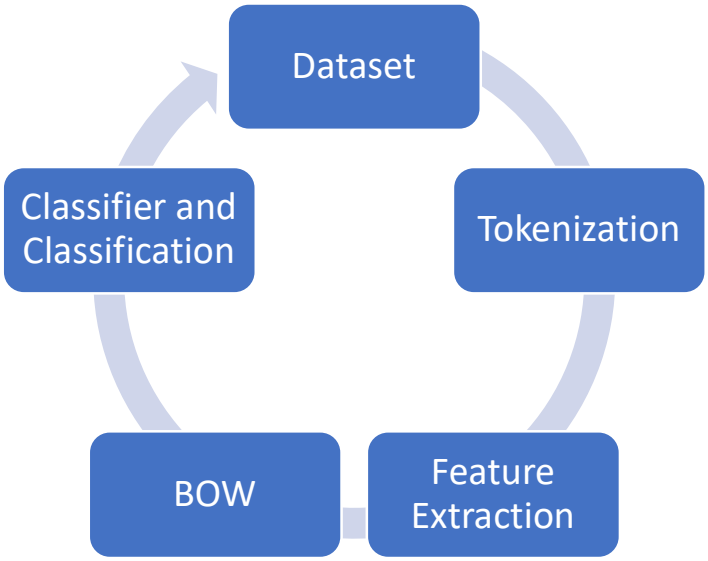

Figure 1: Proposed Model

Tokenization and Preprocessing:

In the initial step, the text undergoes tokenization and preprocessing. Tokenization involves the segmentation of the text into punctuations, words, and emoticons present in the tweet. Simultaneously, preprocessing is executed to eliminate words devoid of significant relevance for analysis, such as stop words. This initial phase lays the foundation for a refined and focused dataset for subsequent analysis.

Feature Extraction:

By adopting this systematic approach, the study aims to enhance the accuracy and efficiency of sentiment polarity classification. The integration of tokenization, preprocessing, feature identification, and SVM-based classification into a cohesive framework represents a holistic strategy to address the intricacies of sentiment analysis in Twitter data. This proposed methodology stands as an innovative contribution to the field, showcasing a structured and effective way to analyze sentiments in social media content.

The sentiment classification process for test tweets. The procedure involves the following steps:

1. **Input Test Tweets:** Test tweets, representing real-world data for sentiment evaluation, are provided as input to the learned model.

2. **Learned Model from SVM Classifier:** The SVM classifier, having been previously trained on a dataset during the learning phase, produces a learned model. This model encapsulates the discernment of positive or negative sentiments based on the patterns identified during training.

3. **Sentiment Detection:** The learned model is then applied to the input test tweets. By leveraging the knowledge acquired during training, the SVM classifier identifies and

categorizes the sentiments expressed in the tweets. The output of this process determines whether the sentiment of each tweet is positive or negative.

This classification process represents the culmination of the proposed approach. By employing a learned model and the capabilities of the SVM classifier, the system achieves the task of sentiment detection for test tweets. This step is pivotal for evaluating the efficacy of the proposed methodology and validating its performance in accurately classifying sentiments within the context of real-world Twitter data.

Feature Identification and Bag Of Words (BOW) Model Construction:

In this study, a novel feature, the Word2Vec representation of tweets, is introduced. The Continuous Bag of Words (CBOW) model is employed to generate Word2Vec representations for the words present in the tweets. This model utilizes 300 dimensions and considers five terms in each window for representation. Each tweet in both the training and test datasets is transformed into a vector by calculating the standard deviation and mean of the words in the Word2Vec vectors. This process results in a final vector containing 600 features.

The second step involves the identification of features based on semantic analysis and various linguistic resources. These extracted features are then utilized to construct a Bag Of Words (BOW) model. Within this model, tweets are represented as vectors, capturing the essential semantic elements of the text. This process contributes to a comprehensive understanding of the underlying sentiments expressed in the tweets.

Classification Model Construction:

The final step entails the utilization of tweet vectors as inputs for the classification algorithm. The chosen classification algorithm in this work is the Support Vector Machine (SVM) classifier. The SVM classifier is employed to train a model that discerns the polarity of test tweets. This phase is crucial for the development of a robust sentiment polarity classification model, leveraging the semantic features and linguistic resources identified earlier.

Having identified various types of features, the tweets in both the training and test datasets are represented as feature vectors. The ultimate step in the sentiment analysis process involves the classification of tweet messages using dedicated classification algorithms. These algorithms are trained to develop a model based on the generated feature vectors. Subsequently, the sentiment of an anonymous text is predicted by applying this model. The feature vectors of the test dataset are provided to the learned model to assess the performance of the classifier. The same learned model is also utilized to predict the sentiment class of an anonymous tweet. The implementation of the Support Vector Machine (SVM) is employed, utilizing LibSVM provided by the Scikit-learn library. SVM functions by identifying appropriate hyperplanes that effectively separate vectors of different classes based on the specified feature set. This robust classification approach contributes to the accurate prediction of sentiment polarity in Twitter data, ensuring the effectiveness of the proposed methodology.

4. Results and Discussion

The SVM classifier is employed in a two-stage process encompassing learning and classification. During the learning stage, the SVM classifier generates classification rules specific to each topic by utilizing training data as input. Subsequently, in the classification stage, these rules are applied to categorize test tweets from the topic-wise test dataset into either

negative or positive sentiments. The effectiveness of the proposed approach, incorporating a fusion of both newly introduced features and existing ones, is assessed through various metrics, including accuracy, recall, and F1 scores. The evaluation encompasses negative recall, positive recall, and macro-averaged recall values for all feature sets, employing a leave-one-out mechanism where one feature set is excluded in each experiment. The results of these assessments are presented in Figure 3. Additionally, the proposed approach introduces a novel combination of features, emphasizing not only the incorporation of existing feature sets but also the integration of newly proposed ones based on advanced word embedding techniques. This amalgamation aims to capture a more comprehensive set of patterns and semantic nuances, contributing to the refinement of sentiment classification. Furthermore, the evaluation metrics, such as accuracy, recall, and F1, serve as robust indicators of the proposed approach's efficacy. These metrics provide a comprehensive understanding of the model's performance across different sentiment categories and highlight its ability to discern sentiments accurately in the context of specific topics. In conclusion, the proposed two-stage SVM classification approach, leveraging a combination of new and existing features, demonstrates a holistic strategy for sentiment analysis in Twitter posts.

	All features	Linguistic Features	Semantic Features	Sentiment Lexicons	Proposed Feature
Negative Recall (RNegative)	0.885	0.882	0.878	0.861	0.857
Positive Recall (RPositive)	0.912	0.893	0.898	0.876	0.872
Macro Averaged Recall (RMacro)	0.898	0.887	0.888	0.868	0.864

Table 1: The Leave-one-out Experiments for four categories of features to measure macro-averaged recall

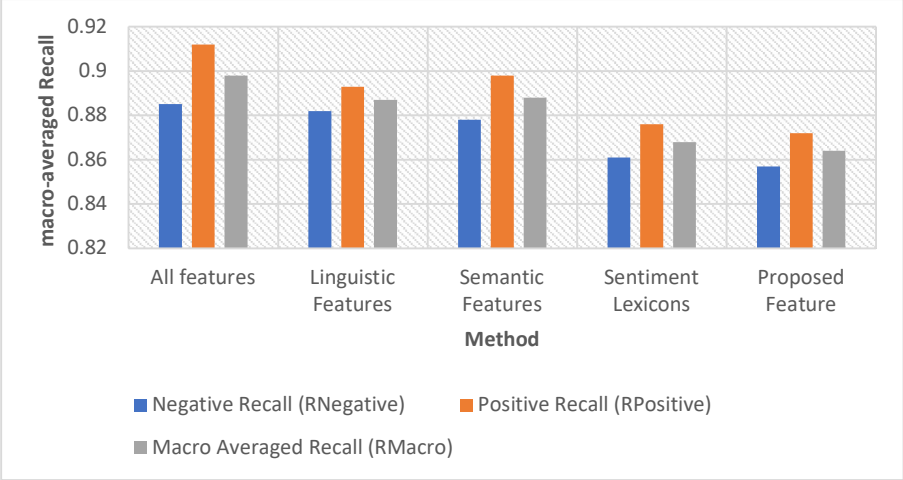

Figure 3: The Leave-one-out Experiments for four categories of features to measure macro-averaged recall

Figure 3 and Table 1 displays the computed values of Negative F1 and Positive F1, obtained by calculating both recall and precision for tweets expressing negative and positive sentiments,

respectively. The macro-averaged F1 is subsequently derived by incorporating the Negative F1 and Positive F1 values. This holistic evaluation approach provides a comprehensive assessment of the model's performance, considering both negative and positive sentiment predictions (Figure 4 and Table 2).

	All features	Linguistic Features	Semantic Features	Sentiment Lexicons	Proposed Feature
Negative Recall (RNegative)	0.876	0.857	0.863	0.854	0.848
Positive Recall (RPositive)	0.894	0.882	0.879	0.872	0.868
Macro Averaged Recall (RMacro)	0.885	0.869	0.871	0.863	0.858

Table 2: The Leave-one-out Experiments for four categories of features to measure the macro-averaged F1

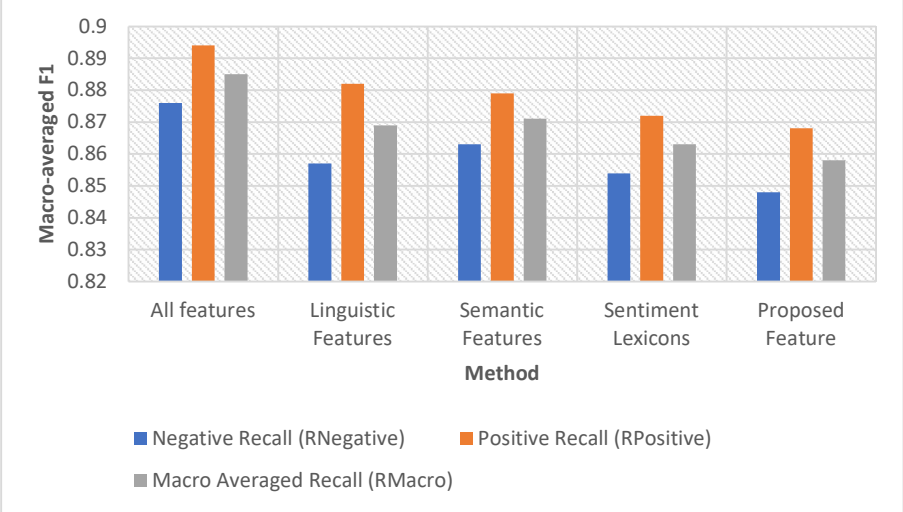

Figure 4: The Leave-one-out Experiments for four categories of features to measure the macro-averaged F1

The Accuracy values for negative tweets, positive Tweets and average accuracy is presented in Figure 5 and Table 3.

	All features	Linguistic Features	Semantic Features	Sentiment Lexicons	Proposed Feature
Negative Recall (RNegative)	0.874	0.862	0.859	0.847	0.837
Positive Recall (RPositive)	0.884	0.873	0.866	0.859	0.852
Macro Averaged Recall (RMacro)	0.879	0.867	0.862	0.853	0.844

Table 3: The Leave-one-out Experiments for four categories of features to measure average accuracy

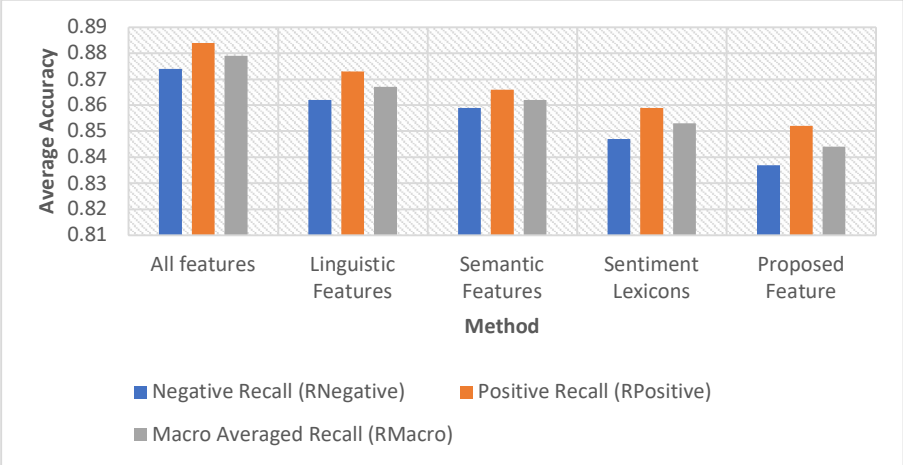

Figure 5: The Leave-one-out Experiments for four categories of features to measure average accuracy

5. Conclusions and Future Scope

In this study, we undertook the sentiment classification task on a dataset of tweets, exploring various feature sets. Notably, leveraging word embedding for tweet words yielded the most promising results in terms of sentiment prediction. Additionally, the integration of sentiment lexicon-based features showcased optimal outcomes, particularly when combined with other feature sets. Looking ahead, a significant avenue for future exploration lies in the application of deep learning methods to sentiment classification. The prospect of allowing models to autonomously learn features from text, rather than relying on manual extraction, represents a substantial leap towards enhancing the adaptability and effectiveness of sentiment classification systems. Furthermore, the continual evolution of social media communication demands ongoing research to stay abreast of emerging trends and linguistic nuances. Future studies could delve into the challenges posed by sarcasm, irony, and cultural context within tweets, seeking innovative methodologies that robustly handle these complexities. The integration of multi-modal data, such as images and videos accompanying tweets, could also be explored to enrich the contextual understanding of sentiments expressed. This holistic approach could offer a more comprehensive and nuanced sentiment analysis in the era of multimedia-rich social media platforms. Lastly, the deployment of sentiment classification models in real-time applications, such as customer feedback analysis or political sentiment tracking during elections, presents an exciting area for practical implementation. Integrating these models into decision-making processes could have far-reaching implications across diverse sectors. In conclusion, this study not only sheds light on effective features for sentiment classification in tweets but also identifies promising directions for future research, emphasizing the need for continuous adaptation to the dynamic landscape of social media and the exploration of cutting-edge technologies like deep learning in sentiment analysis.

REFERENCES

1. Pang, Bo and Lillian Lee (2008). "Opinion mining and sentiment analysis." In:Foundations and trends in information retrieval 2 (1-2), pp. 1–135.
2. Medhat, Walaa, Ahmed Hassan, and HodaKorashy (2014), Sentiment analysis algorithms and applications: A survey, Ain Shams Engineering Journal 5 (4), pp. 1093{1113, Elsevier.
3. Comito, C., Forestiero, A., & Pizzuti, C. (2019, October). Word embedding based clustering to detect topics in social media. In *IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 192-199).
4. Lagrari, F. E., & Elkettani, Y. (2021). Traditional and deep learning approaches for sentiment analysis: A survey. *Advances in Science, Technology and Engineering Systems Journal*, *6*(4), 1-7.
5. Ebrahimi, M., Yazdavar, A. H., & Sheth, A. (2017). Challenges of sentiment analysis for dynamic events. *IEEE Intelligent Systems*, *32*(5), 70-75.
6. Mohammad, S. M. (2017). Challenges in sentiment analysis. *A practical guide to sentiment analysis*, 61-83.
7. Zhao, L., & Zhao, A. (2019). Sentiment analysis based requirement evolution prediction. *Future Internet*, *11*(2), 52.
8. Saif, H., He, Y., & Alani, H. (2012). Semantic sentiment analysis of twitter. In *The Semantic Web–ISWC 2012: 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I 11* (pp. 508-524). Springer Berlin Heidelberg.
9. Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), 5731-5780.
10. Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (Eds.). (2017). *A practical guide to sentiment analysis* (Vol. 5). Cham: Springer International Publishing.
11. Hussein, D. M. E. D. M. (2018). A survey on sentiment analysis challenges. *Journal of King Saud University-Engineering Sciences*, *30*(4), 330-338.
12. Madhoushi, Z., Hamdan, A. R., & Zainudin, S. (2015, July). Sentiment analysis techniques in recent works. In *2015 science and information conference (SAI)* (pp. 288-291). IEEE.