United International Journal for Research & Technology

Volume 04, Issue 07, 2023 | Open Access | ISSN: 2582-6832

Named Entity Recognition on E-Mono: An Algorithm Enhancement Applied in Sentiment Analysis

Stephen Kent A. Malagday[1], Mark Raphael V. Sto. Domingo[2], Raymund M. Dioses[3], and Vivien A. Agustin[4]

[1,2,3,4]Computer Science Department, Pamantasan ng Lungsod ng Maynila, Manila, Philippines

***Abstract*—** Sentiment analysis plays a crucial role in natural language processing, aiming to categorize emotions expressed in text. One commonly used approach in sentiment analysis is the Extended Max-Occurrence with Normalized Non-Occurrence (EMONO) term weighting scheme. The EMONO scheme extends the Max-Occurrence with Normalized Non-Occurrence (MONO) approach by considering both the occurrence and non-occurrence frequencies of terms in sentiment classes. However, the original E-MONO approach still overlooks the significance of named entities and their associated sentiment. To address this limitation, we propose an enhanced approach that combines the E-MONO term weighting scheme with Named Entity Recognition (NER). By incorporating NER, we aim to improve the accuracy of sentiment analysis by accurately identifying named entities and analyzing sentiment towards specific entities. Additionally, we enhance the EMONO term weighting scheme by introducing extended max-occurrence groups and normalizing non-occurrences, providing a more comprehensive representation of term significance. The combination of NER and the enhanced EMONO term weighting scheme aims to capture the nuanced sentiment expressed towards named entities, leading to improved sentiment analysis results.

***Keywords*—** Sentiment Analysis, Entity-Aware Sentiment, Named Entity Recognition, Term Weighting, Natural Language Preprocessing.

I. INTRODUCTION

Text classification plays a vital role in various applications, and an essential aspect of this process term weighting. Term weighting assigns importance to individual features, highlighting their contribution to the distinctiveness of each document within a collection. In the Vector Space Model, which is a mathematical representation of documents, feature vectors are used to depict documents by assigning numerical values to selected or extracted features. Proper term weighting is crucial as it has been shown to significantly impact classification accuracy [3].

One specific application of text classification is sentiment analysis, also known as opinion mining, which aims to categorize the emotions expressed in textual data [2]. With the widespread sharing of opinions through social media, blog posts, articles, and other online platforms, sentiment analysis has become an invaluable tool for understanding public sentiment and opinion.

[1] introduced an enhanced term weighting algorithm called Extended Max-Occurrence with Normalized Non-Occurrence (E-MONO), building upon the original MONO algorithm. E-MONO expands the max-occurrence group of document frequency to include classes with higher values, allows the setting of the class-distinguishing power of a term, and addresses the uneven distribution of document frequency across classes through normalization of non-occurrences. The authors also applied variants of E-MONO to multi-class sentiment analysis, particularly addressing the short-distance frequency of a term in interclass sentiment analysis. The results achieved an accuracy of 74% on the K-Nearest Neighbors (KNN) classifier and 82% on the Support Vector Machine (SVM) classifier [1].

However, the study by [1] focused primarily on document-level analysis and did not incorporate feature selection techniques to improve classification results. To address these limitations, our research aims to enhance the existing results by incorporating named entity recognition (NER) and feature selection in sentiment analysis.

Named entity recognition (NER) plays a crucial role in identifying and extracting specific entities, such as people, organizations, locations, and more, from text [13]. By integrating NER into sentiment analysis, we can capture sentiment towards specific entities and achieve a more fine-grained analysis. This incorporation of NER is expected to improve the classification accuracy by considering entity-specific sentiment information.

In addition to NER, feature selection techniques are employed to identify the most informative and relevant features for sentiment analysis [6]. These techniques

help to reduce dimensionality, remove noise, and enhance the classification performance by selecting features that are most discriminative for sentiment classification. By incorporating feature selection methods, we aim to improve the overall accuracy of sentiment analysis and enhance the performance of the E-MONO algorithm.

In this paper, we build upon the work of [1], by incorporating NER and feature selection techniques into sentiment analysis to enhance the existing results. By leveraging the entity-specific sentiment information and selecting informative features, we anticipate achieving higher classification accuracies and more accurate sentiment analysis results.

Statement of the Problem

The traditional sentiment analysis algorithms often overlook the contextual information surrounding terms, which can impact their accuracy. In the study conducted by [1], they employed the EMONO term weighting algorithm for multiclass sentiment analysis but did not address the contextual issue. However [16] emphasized the significance of incorporating context to improve sentiment classification accuracy. Another limitation in Abalario's study was the lack of feature selection, resulting in the inclusion of irrelevant information that affected the algorithm's precision. Feature selection techniques like Chi-Square and Count Difference, as suggested by [19], have proven effective in sentiment analysis. Lastly, optimizing the dataset and considering feature redundancy can further enhance algorithm efficiency. Unfortunately, [1] did not implement a feature selection strategy that addresses feature redundancy, potentially hindering model performance.

Objectives of the study

The objective of this research is to enhance the accuracy and classification performance of the Extended Max Occurrence Non-Occurrence (E-MONO) algorithm through the use of Named Entity Recognition (NER). In this paper the researchers will:

1. Use Named Entity Recognition (NER) to identify words that correlate with sentiment and provide context to improve the analysis of the text.
2. Use Chi-square feature selection to generate the best subset from the dataset.
3. Use named entities with EMONO for sentiment classification.

II. RELATED WORKS

Named Entity Recognition

Named entity recognition (NER) is a crucial natural language processing (NLP) task that involves identifying and classifying named entities, such as specific people, organizations, places, or events, in a given text. These named entities are important because they often contain valuable information and context, and they can be used to improve the performance of other NLP tasks, such as information retrieval and text classification [23][17].

In recent years, deep learning algorithms that utilize transfer learning to create contextual representations of text have had a major impact on the field of NLP, including NER. These algorithms have led to the development of state-of-the-art NER systems that differ based on the method used to create contextual representations [9][2][7]. Techniques such as bidirectional LSTM, attention mechanisms, and transformers have been employed to improve NER performance. The use of context has been shown to improve classification results in NER and other NLP tasks [18]. For example, [18] found that using cross-sentence information consistently improved NER results when using BERT models. [22] proposed a method for improving NER by retrieving context from sentences and using cooperative learning, which outperformed baseline models. [13] explored generating relevant context for named entities by introducing the task CONTEXT-NER. Despite the complexity of the problem, their approach achieved a score of 39% on a one-shot evaluation with GPT-3. However, NER still faces challenges, such as handling ambiguous or novel entities, dealing with variations in entity names, or coping with imbalanced datasets. Named Entity Recognition is a vital aspect of NLP that involves identifying and classifying named entities in text. The use of context and advanced algorithms, such as deep learning and transfer learning, has significantly impacted the development of NER and improved its performance in various NLP tasks.

Extended Max Occurrence Non-Occurrence

Extended Max Occurrence Non-Occurrence (E-MONO) is an enhancement of MONO categorized into two major processes: extended max occurrence and normalized non-occurrence. [1] stated that MONO suffers setbacks in weighting terms with non-uniformity values. As it focuses more on class document frequency for non-occurrence members. It neglects to utilize the occurrence of interclass distinguishing power to give

good term weighting ability that gives great results for classification. E-MONO utilizes the top classes with the high-class occurrence and tackles the imbalance distribution of non-occurrence through normalization. EMONO utilizes an EMO number, which is the number of classes the MO group is composed of.

Similar to MONO, a sorted df collection is divided into two groups:

$$sorted\_df_{t_i} = \left\{ \overbrace{d_{i3}, d_{i1}}^{MO} \middle| \overbrace{d_{i4}, d_{i5}, \ldots, d_{ij-1}, d_{ij}}^{NO} \right\}$$

An EMO of 2, however sets the MO group to consist of 2 classes.

As the number of NO members is reduced, each NO member's percentage is calculated individually (assuming EMO = 2).

$$Normalized\_NO_{t_i} = \frac{NO_{t_{i1}} + NO_{t_{i2}}}{200}$$

Calculations of EMONO Local and EMONO Global are as follows:

$$EMONO_{Local}(t_i) = EMO_{t_i} * Normalized\_NO_{t_i}$$

$$EMONO_{Global}(t_i) = [1 + \alpha * EMONO_{Local}(t_i)]$$

EMONO Variants TF-EMONO and SRTF-EMONO are calculated as follows:

$$TF - EMONO = TF(t_i, d_k) * [EMONO_{Global}(t_i)]$$

$$SRTF - EMONO = SRTF(t_i, d_k) * [EMONO_{Global}(t_i)]$$

$$EMO_{t_i} = \frac{MO_{t_{i1}} + MO_{t_{i2}}}{D_{total}\left(MO_{t_{i1}} + MO_{t_{i2}}\right)}$$

Feature Selection Strategies for Sentiment Analysis

Feature selection plays a crucial role in sentiment analysis by choosing a subset of relevant features from a large dataset to build a useful model. Diverse feature types in text documents cause high dimensionality, and feature selection improves classifier performance [15]. According to [19], Chi-Square and Count Difference are among the best feature selection techniques for sentiment analysis. Chi-Square is a statistical test used to measure the independence and relationship between two events. A higher Chi-Square score indicates a more significant relationship between a feature and the target variable in sentiment analysis. [19] compared various feature selection techniques combined with different machine learning algorithms and found that Chi-Square performed best on Multinomial Naive Bayes with an accuracy score of 82.33%.

III. PROPOSED METHODS

A. Existing Algorithm

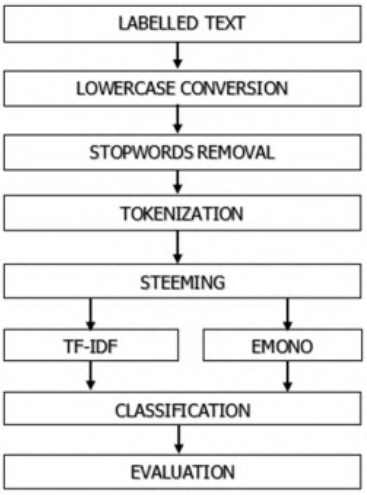

Fig.1 E-MONO Algorithm

Fig. 1 presents the existing algorithm by [1], EMONO and TF-IDF were used in their framework to compare the results of the two-term weighting schemes. This utilized steps for preprocessing and creating different branches for TF-IDF and EMONO.

A. Theoretical Framework of E-MONO with NER

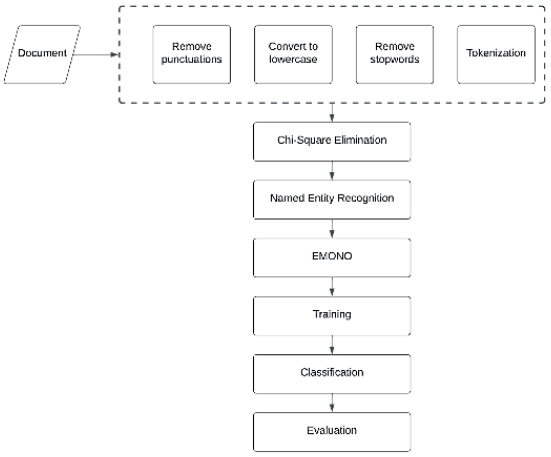

Fig.2 Theoretical Framework of E-MONO with NER

Fig. 2 presents the addition of Named Entity Recognition for improving E-MONO. The process starts by preprocessing the data by removing symbols,

converting to lowercase, removing stop words, and tokenization. Secondly is the extraction of NER entities and computations of the E-MONO values to be used as features. Next, for feature selection, the researchers used Chi-square to select the best features. Lastly, training the model, classification, and evaluation.

IV. METHODOLOGY

The present study endeavors to enhance sentiment analysis by incorporating Named Entity Recognition (NER) for contextual identification, employing the Chi-square feature selection technique. This section describes how to put these techniques into action.

Dataset

The dataset for this study was obtained from Kaggle, a platform renowned for hosting a diverse range of datasets relevant to various research domains. The dataset in focus is titled "Sentiment Analysis in Commodity Market: Gold". The dataset contains a total of 10,000+ news headlines across multiple dimensions into various classes. It has been sampled from a period of 20+ years ranging from 2000-2021. This is a news dataset for the commodity market where we have manually annotated 10,000+ news headlines across multiple dimensions into various classes. The dataset has been sampled from a period of 20+ years (2000-2021).

Pre-Processing

The initial stage of our study involved meticulous pre-processing of the dataset to ensure the data was in a suitable format for subsequent analysis. Firstly, all text data underwent a cleaning process, which involved removing punctuation marks. These marks often introduce noise that could potentially interfere with the accurate interpretation of the textual information. Next, all text data was transformed to lowercase to maintain consistency and eliminate duplication due to case differences. This normalization step is crucial in text mining as it standardizes the text, allowing the subsequent processes to treat similar words uniformly.

Furthermore, the researchers implemented a *stopword* removal process, eliminating commonly used words (like 'and', 'the', 'is', etc.) that generally do not contribute meaningful context to the sentiment conveyed. This step significantly reduces the dimensionality of our data while retaining the core sentiment-bearing phrases. The final stage of our pre-processing was tokenization, which involved breaking down the text into individual words or tokens, making it feasible to analyze the text data at a granular level.

Context Identification through Named Entity Recognition

The first step involves using NER to identify words that correlate with sentiment and provide context to improve the analysis of the text. To achieve this, the researchers will use a pre-trained NER model, such as *spaCy*, to extract named entities from the text dataset. The extracted entities will include people, organizations, places, and events.

Once the named entities are identified, the researchers will analyze their relationship with sentiment by calculating the correlation between their presence and the dataset's sentiment labels. This analysis will help the researchers determine which named entities are most relevant to sentiment analysis.

Extended Max-Occurrence with Normalized Non-Occurrence

E-MONO expands the max-occurrence group to include classes with higher document frequency values, allows the class-distinguishing power of a term to be set, and addresses the uneven distribution of document frequency across classes by normalizing non-occurrences. E-MONO is used as the term weighting scheme for text classification.

Chi-square Feature Selection Technique

The researchers will utilize the Chi-square feature selection technique to filter the features into a concise subset containing only germane information. Chi-square is a statistical method employed to discern the independence between each feature and the target variable, which in this case was the sentiment label. Each feature was assigned a Chi-square score, with higher scores indicating higher relevance to the sentiment analysis task. By focusing on these pertinent features and discarding the less relevant ones, the dimensionality of the dataset was reduced. This made the model more computationally efficient and enhanced the performance of the machine learning algorithms by reducing the risk of overfitting and improving generalization.

Training and Testing

The researchers employed the Support Vector Machine (SVM) model for the purpose of sentiment classification. SVM was chosen due to its efficacy in handling high-dimensional data and its robustness in

classifying linearly inseparable data through the use of kernel trick.

The SVM model was trained on a portion of the preprocessed dataset, wherein it learned to distinguish between different sentiment labels based on the selected features. The model was then tested on a separate portion of the data, which allowed the researchers to evaluate its performance and generalizability to unseen data.

Performance Metrics

The model's performance was evaluated using various metrics, each providing a unique perspective on the model's predictive capability. The chosen metrics were accuracy, precision, recall, and F1-score. By assessing the model using these comprehensive metrics, the authors aimed to gain a detailed understanding of its strengths and weaknesses, thereby identifying potential areas for improvement.

- Accuracy measures the overall correctness of the model, i.e., the proportion of true results (both true positives and true negatives) in the total number of cases examined.
- Precision, also known as the positive predictive value, measures the proportion of correctly predicted positive observations to the total predicted positives.
- Recall, also known as sensitivity, measures the proportion of correctly predicted positive observations to all observations in the actual class.
- F1-score is the harmonic mean of precision and recall and is an effective measure when the data has imbalanced classes.

IV. RESULTS

The initial problem statement of the research identified several key issues in the application of the E-MONO algorithm in sentiment analysis. It was noted that the algorithm might not sufficiently consider the context in which a term occurs. Furthermore, the researchers observed that the algorithm's overall precision could be compromised by irrelevant information in the dataset, and they also noticed a lack of feature selection strategy in previous studies, specifically the work by [1].

In addressing these problems, the researchers' work was guided by several objectives. The general objective was to enhance the accuracy and classification performance of the Extended Max Occurrence Non-Occurrence (E-MONO) algorithm. The researcher's approach was to incorporate Named Entity Recognition (NER) into the algorithm to identify words that correlate with sentiment and provide context to improve the text analysis.

Additionally, the researchers aimed to use the Chi-square feature selection technique to generate the most suitable subset from the dataset.

The results presented herein aim to discuss the extent to which these objectives were met and the implications of the findings in improving the efficiency and accuracy of the E-MONO algorithm in sentiment analysis**.**

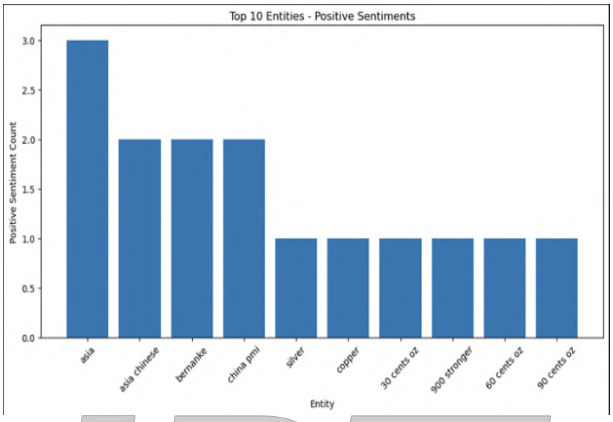

Fig. 1 Top 10 Entities – Positive Sentiments

Figure 1. illustrates the entities that are most frequently identified in the positive class.

The entities, listed by frequency, include asia, asia chinese, bernanke, chinese pmi, silver, copper, 30 cents oz, 900 stronger, 60 cents oz, and 90 cents oz. Notably, "asia" appears to be the most frequent entity in the positive sentiment class.

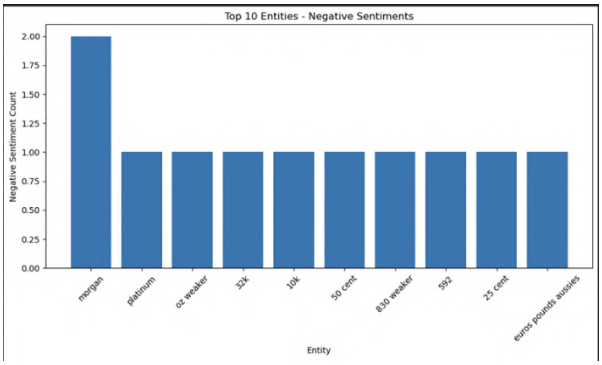

Fig 2*. Top 10 Entities - Negative Sentiments*

Figure 2. illustrates the entities that are most frequently identified in the negative class. The entities listed by frequency include morgan, platinum, oz weather, 32k, 10k, 50 cent, 830 weaker, 592, 25 cents, euros pounds aussies. Notably, "morgan" appears to be the most frequent entity in the negative sentiment class.

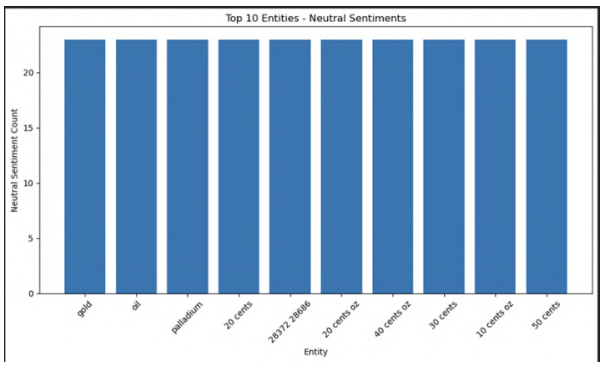

Fig 3. *Top 10 Entities - Neutral Sentiments*

Figure 3. illustrates the entities that are most frequently identified in the neutral class. The entities listed by frequency include gold, oil, palladium, 20 cents, 28372 28656 ,20 cents oz, 40 cents oz, 30 cents, 10 cents oz, and 50 cents. Notably, "gold" is the most frequent entity in the neutral sentiment class.

Named Entity Recognition (NER) technique has been utilized to identify words or phrases, also known as entities, that carry significant sentiment. These entities serve to provide context in the sentiment analysis, improving the overall understanding of the text and leading to more accurate classifications. This is in line with the objectives of the study, which aim to enhance the accuracy and performance of the E-MONO algorithm, particularly in the context of sentiment classification.

In terms of results, a notable pattern emerged among the top entities identified in each sentiment class. For the positive class, the most frequent entity was "asia", while for the negative class it was "morgan", and for the neutral class, it was "gold".

These entities provide valuable insights into the context of the sentiments expressed in the analyzed texts, potentially improving the precision of sentiment classification.

Recent studies have highlighted the importance and challenges of NER in various contexts, especially in social media, where the dynamic nature of language adds complexity to the task. For instance, the impact of temporal shifts in the semantics of language, particularly on social media platforms like Twitter, has been explored. This involves how the meaning of words can change or evolve over time, adding another layer of complexity to the task of NER. This underscores the importance of having an up-to-date language model to handle such shifts [21]

Table 1. *Chi-Square Feature Selection Result*

No. of features	Mean Score
971 features	0.847271
1295 features	0.846966
647 features	0.846048
323 features	0.840849
1619 features	0.845590

The findings from Table 1 demonstrate that Chi-Square feature selection played a crucial role in optimizing the feature set for sentiment classification. The fact that the highest mean score was achieved with 971 features indicates that this might be the optimal number of features for our dataset using the E-MONO algorithm.

This result contributes to the ongoing conversation on the significance of feature selection in machine learning algorithms, especially in sentiment analysis. As illustrated in the table, simply increasing the number of features does not necessarily lead to an improved performance. In fact, an excessive number of features can lead to lower mean scores, possibly due to the introduction of irrelevant or redundant features. This validates our initial assertion that feature selection, particularly using the Chi-Square method, can greatly enhance the performance of the E-MONO algorithm.

However, it should be noted that while 971 features yielded the best results in our study, this might not always be the case for other datasets or algorithms. Further studies could explore the variability of optimal feature numbers in different contexts or datasets.

Table 2. *Global MONO weights of terms*

Term	Weight
old	0.125593
futures	0.086971
oz	0.070556
prices	0.043706
dollar	0.034710
lower	0.044172
week	0.039531
higher	0.039187
trade	0.023000

Table 2 shows the entities with the highest MONO weights, with "gold" standing out with the highest weight of 0.125593, followed by "futures" with a weight of 0.086971, and 'oz.' with a weight of 0.070556. The other terms show lower weights, with 'trade' having the lowest weight among the top 10 entities at 0.023000.

The terms listed above represent the entities that have the most significant impact on sentiment classification using the E-MONO algorithm in the dataset. The entity 'gold', having the highest weight, can be considered the

most impactful term, likely contributing significantly to the sentiment analysis. The presence of terms like 'gold', 'futures.', and 'oz.' suggests that the algorithm effectively identifies important terms related to financial and trading contexts, given their frequent usage and relevance in such conversations.

Table 3: *Comparison Results of Different Features Vectorization on SVM*

	E-MONO		NER with E-MONO	
Vectorizer	**Accuracy**	**F1-weighted**	**Accuracy**	**F1-weighted**
TF-IDF	81%	81%	79%	79%
TF-EMONO EMO = 2	82%	82%	75%	75%
STRF-EMONO EMO = 2	82%	82%	**84%**	**84%**

Table 3. presents the comparison of different term weighting schemes when applied with Support Vector Machine (SVM). The data reveals that the existing E-MONO algorithm yields an accuracy and an f1-weighted score of 82%. However, an interesting improvement is noted when Named Entity Recognition (NER) is incorporated with the E-MONO algorithm. It can be seen that the SRTF-EMONO with NER, the accuracy and the f1-weigthed is 84% interestingly then NER is applied to the model TF-EMONO with EMO = 2 the accuracy and f1-weighted lowered to 75%.

The observed enhancement in the performance metrics implies that the incorporation of NER with the E-MONO algorithm successfully boosts the classification performance in sentiment analysis. These findings, thus, confirm our research hypothesis that the inclusion of NER could lead to an improvement in the accuracy of the sentiment classification process.

VII. CONCLUSION

The general objective of this research is to enhance the accuracy and classification performance of the Extended Max Occurrence Non-Occurrence (E-MONO) algorithm by addressing the limitations of the existing model, particularly its tendency not to consider the context in which a term occurs and its lack of a feature selection strategy. By integrating Named Entity Recognition (NER) and Chi-square feature selection, we aimed to improve the algorithm's accuracy and classification performance.

Our findings demonstrate that incorporating NER successfully identifies entities that strongly correlate with specific sentiments, thus providing the much-needed context for sentiment analysis. Notably, "Asia," "Morgan," and "Gold" were the most frequently identified entities in positive, negative, and neutral sentiment classes, respectively.

Furthermore, through comparative analysis using different term weighting schemes in conjunction with SVM, we found that the integration of NER and the adjustment of the EMO parameter in the STRF-EMONO algorithm significantly improved the accuracy and f1-weighted score from 82% to 84%. This underscores the effectiveness of our approach in enhancing sentiment classification performance.

These results not only address the limitations identified in previous studies, such as that of [1], but also pave the way for future research. Future work could explore further optimization strategies or consider other machine-learning techniques to increase the precision of sentiment classification. Despite the noted improvements, it's important to continuously advance these algorithms as language usage and sentiment expression continue to evolve.

In conclusion, our research demonstrates that the combination of NER and Chi-square feature selection can significantly enhance the accuracy and classification performance of the E-MONO algorithm in sentiment analysis. This advancement is a vital step in the pursuit of more accurate and efficient sentiment analysis techniques.

ACKNOWLEDGMENT

We would like to express our heartfelt gratitude to Abalorio for their invaluable inspiration and contribution to our research. Their pioneering work has served as a significant catalyst in shaping our research interests and motivating us to pursue our thesis.

Their groundbreaking research in the field of Natural Language Processing has been a constant source of

inspiration and has laid a solid foundation for our own investigation. Their dedication to advancing knowledge and their insightful findings have greatly influenced our research direction and methodologies.

We are particularly grateful for their willingness to share their expertise and provide valuable guidance during our thesis journey. Their valuable insights, constructive feedback, and thought-provoking discussions have been instrumental in shaping the conceptualization and execution of our research.

Furthermore, we extend our gratitude to our advisors and mentors who have provided us with continuous support, guidance, and encouragement throughout the research process. Their expertise and wisdom have played a pivotal role in the development and refinement of our research objectives and methodology.

We would also like to thank our colleagues and friends for their unwavering support and encouragement. Their constructive feedback, stimulating discussions, and moral support have been invaluable in enhancing the quality of our research.

Lastly, we are indebted to our families for their unconditional love, understanding, and encouragement throughout this challenging journey. Their usnwavering support and belief in our abilities have been a constant source of motivation and strength.

REFERENCES

[1] Abalorio, A. Sison, R. Medina, and G. Dalaorao, "Applying EMONO variants to multi-class sentiment analysis for short-distance inter-class frequency of term," Mathematical Statistician and Engineering Applications, 2022. [Online]. Available: https://www.philstat.org/index.php/MSEA/article/view/721

[2] Akbik, T. Bergmann, and R. Vollgraf, "Pooled contextualized embeddings for named entity recognition," in Proceedings of the 2019 Conference of the North, 2019. [Online]. Available: https://doi.org/10.18653/v1/n19-1078

[3] R. Alroobaea, "Comparative study of sentiment analysis on Amazon product reviews using recurrent neural network (RNN)," International Journal of Advanced Trends in Computer Science and Engineering, vol. 11, no. 3, pp. 141-146, 2022. [Online]. Available: https://doi.org/10.30534/ijatcse/2022/111132022

[4] Alsmadi and G. K. Hoon, "Term weighting scheme for short-text classification: Twitter corpuses," Neural Computing and Applications, vol. 31, no. 8, pp. 3819-3831, 2018. [Online]. Available: https://doi.org/10.1007/s00521-017-3298-8

[5] G. Ansari, T. Ahmad, and M. N. Doja, "Hybrid filter-wrapper feature selection method for sentiment classification," Arabian Journal for Science and Engineering, vol. 44, no. 11, pp. 9191-9208, 2019. [Online]. Available: https://doi.org/10.1007/s13369-019-04064-6

[6] Aretove, "Importance of feature selection in machine learning," Aretove Technologies, Dec. 23, 2020. [Online]. Available: https://www.aretove.com/importance-of-feature-selection-in-machine-learning

[7] Baevski, S. Edunov, Y. Liu, L. Zettlemoyer, and M. Auli, "Cloze-driven Pretraining of self-attention networks," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019. [Online]. Available: https://doi.org/10.18653/v1/d19-1539

[8] K. Chen, Z. Zhang, J. Long, and H. Zhang, "Turning from TF-IDF to TF-IGM for term weighting in text classification," Expert Systems with Applications, vol. 66, pp. 245-260, 2016. [Online]. Available: https://doi.org/10.1016/j.eswa.2016.09.009

[9] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018. [Online]. Available: https://doi.org/10.48550/arXiv.1810.04805

[10] T. Dogan and A. K Uysal, "Improved inverse gravity moment term weighting for text classification," Expert Systems with Applications, vol. 130, pp. 45-59,[Online]. Available: https://doi.org/10.1016/j.eswa.2019.04.015

[11] T. Dogan and A. K. Uysal, "A novel term weighting scheme for text classification: TF-MONO," Journal of Informetrics, vol. 14, no. 4, 101076, 2020. [Online]. Available: https://doi.org/10.1016/j.joi.2020.101076

[12] G. Domeniconi, G. Moro, R. Pasolini, and C. Sartori, "A comparison of term weighting schemes

for text classification and sentiment analysis with a supervised variant of tf.idf," Communications in Computer and Information Science, pp. 39-58, 2016. [Online]. Available: https://doi.org/10.1007/978-3-319-30162-4_4

[13] S. Gupta, H. Verma, T. Kumar, S. Mishra, T. Agrawal, A. Bagudu, and H. Bhatt, "Context-NER : Contextual Phrase Generation at Scale," 2021. [Online]. Available: https://arxiv.org/abs/2109.08079

[14] Z. Jiang, B. Gao, Y. He, Y. Han, P. Doyle, and Q. Zhu, "Text classification using novel term weighting scheme-based improved TF-IDF for internet media reports," Mathematical Problems in Engineering, vol. 2021, pp. 1-30, 2021. [Online]. Available: https://doi.org/10.1155/2021/6619088

[15] N. I. Khairi, A. Mohamed, and N. N. Yusof, "Feature selection methods in sentiment analysis," in Proceedings of the 3rd International Conference on Networking, Information Systems & Security, 2020. [Online]. Available: https://doi.org/10.1145/3386723.3387840

[16] Kumar and G. Garg, "The multifaceted concept of context in sentiment analysis," in Cognitive Informatics and Soft Computing, pp. 413-421, 2020. [Online]. Available: https://doi.org/10.1007/978-981-15-1451-7_44

[17] J. Li, A. Sun, J. Han, and C. Li, "A Survey on Deep Learning for Named Entity Recognition," IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 1, pp. 50 - 70, 2020. [Online]. Available: https://doi.org/10.1109/TKDE.2020.2981314

[18] J. Luoma and S. Pyysalo, "Exploring cross-sentence contexts for named entity recognition with BERT," in Proceedings of the 28th International Conference on Computational Linguistics, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.coling-main.78

[19] Madusu and S. E, "Efficient Feature Selection techniques for Sentiment Analysis," 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1911.00288

[20] G. Miner IV, J. E., A. Fast, T. Hill, R. Nisbet, and D. Delen, "Chapter 7 - Text Classification and Categorization," in Practical text mining and statistical analysis for non-structured text data applications21. M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018. [Online]. Available: https://doi.org/10.48550/arXiv.1802.05365

[21] Ushio, L. Neves, F. Barbieri, J. Collados, and V. Silva, "Named entity recognition in Twitter: A dataset and analysis on short-term temporal shifts," arXiv.org, 2022, November 15. [Online]. Available: https://arxiv.org/abs/2210.03797

[22] Wu, F. Wu, T. Qi, and Y. Huang, "Named entity recognition with context-aware dictionary knowledge," in Lecture Notes in Computer Science, pp. 129-143, 2020. [Online]. Available: https://doi.org/10.1007/978-3-030-63031-7_10

[23] J. Xie, Z. Yang, G. Neubig, N. A. Smith, and J. Carbonell, "Neural cross-lingual named entity recognition with minimal resources," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018. [Online]. Available: https://doi.org/10.18653/v1/d18-1034