Sentimetrix: A Robust Sentiment Analysis Framework Using Ensemble Learning Transformer with Lime Interpretability

Sravan Kumar D[1] , Suresh Kumar Mandala[2]

[1]Research Scholar, Department of CS & AI, SR University, Ananthsagar, Telangana, India.
[2]Assistant Professor, Department of CS & AI, SR University, Ananthsagar, Telangana, India.
[1]shravaninnovate@gmail.com
[2]mandala.suresh83@gmail.com

Abstract. This study introduces Sentimetrix, a sophisticated sentiment analysis system that aims to extract subtle opinions and feelings from textual data. Sentimetrix uses not only the latest state-of-the-art techniques in sentiment classification, including hybrid transformers (BERT, BERTweet, RoBERTa) and RNN models, but also ensemble learning to achieve high accuracy and interpretability. For text quality enhancement, byte-pair encoding, entity masking, and sentiment-based augmentation are incorporated into the framework. TF-IDF and contextual word embedding's are used for feature extraction, and then sentiment analysis is increased by employing VADER and Text Blob. For a deeper insight, Sentimetrix employs topic-based clustering, topic modeling, and social interaction graph analysis. The system is validated using the LIME interpretability for transparency, and the best-performing weighted ensemble model achieved a 76% accuracy.

Keywords: Sentimetrix, Hybrid Transformers, Ensemble Learning, BERTweet, VADER, TextBlob.

1 INTRODUCTION

The Natural Language Processing (NLP) field of sentiment analysis is vital for interpreting and categorizing affirmations in textual data. Owing to the rise of online platforms, sentiment analysis is necessary to obtain actionable insights from customer feedback, social media, and surveys. Reading these features using traditional methods is often ineffective owing to weaknesses in dealing with linguistic complexities, such as sarcasm, idiom context, and domain-specific nuances, much more so in sentiment classification systems [1]. We address these challenges with an unprecedented level of sophistication using a deep-learning-based Sentimetrix framework, which takes advantage of cutting-edge state-of-the-art deep-learning models and new preprocessing techniques to enable sentiment classification in a variety of use cases [2].

The rise of hybrid and ensemble techniques has significantly advanced state-of-the-art sentiment analysis by integrating different algorithms to synergistically improve the results. These advancements were made by Sentimetrix by combining these transformers, including BERT, BERTweet, and RoBERTa, with classic machine learning models such as Logistic Regression, Support Vector Machines (SVM), and Random Forest [3]. This hybrid approach allows Sentimetrix to benefit from the contextual richness of modern embeddings, and delivers the proven effectiveness of more traditional models. The system implements state-of-the-art preprocessing algorithms, including entity masking, sentiment-based augmentation, and data preprocessing, to filter out biases, noise, and speed-up learning. Given these preprocessing innovations, linguistic subtleties (e.g., idioms and rare word occurrences) have been better handled in Sentimetrix [4].

1.1 Importance of Profiling on Social Media

Analyzing someone on social media who spreads irony and stereotypes by looking for patterns in their talk and actions is called profiling them. This step is very important for thinking of ways to stop the spread of content that could make people more prejudiced or less friendly with each other.

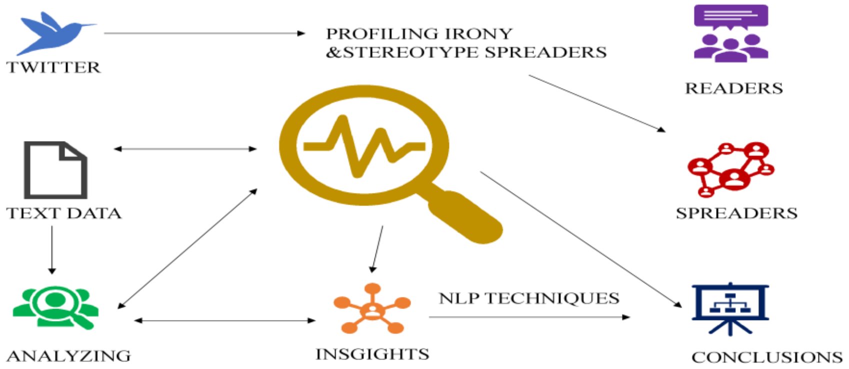

Fig. 1. Profiling Irony and Stereotype Spreaders on Twitter using NLP

1.2 Role of Natural Language Processing (NLP)

We can look at much text on sites like Twitter with the help of Natural Language Processing (NLP); we can use it to find the linguistic traits that make irony and people who spread stereotypes what they are. We now know more about how these stories are made and shared online.

1.3 Significance of the Study

This study's main objective is to present a complete picture of the most recent progress made in utilizing NLP to identify individuals who share irony and stereotypes on Twitter. To find the best ways to do things and places where more research is needed, we will compare old models. We learned something new from the analysis that will help us plan future work in this area.This will make tools used to find and remove harmful content from social media sites more accurate and dependable.Ultimately, this work contributes to the broader goal of fostering a more inclusive and respectful online discourse.

2. LITERATURE REVIEW

Danielly Sorato's study [1] looks at how word embeddings can be used to look into stereotypes about immigrants and refugees in a setting that is multilingual and spans time. It specifically looks at how people talk about politics in four languages. This study shows that stereotypical associations do exist and gives information about the social and political factors that affect these biases. In an investigation [2], Tibor L. R. Krols looks into how to find irony in Twitter users by looking at lexical features, sentiment, and topic modeling. The study shows that lexical features, especially TF-IDF, are very important for improving the model's performance, as shown by the F1 score of 0.84. This study shows that sentiment and topic modeling features don't have as much of an effect as other features. It also suggests that more research be done in these areas. Marie Mortensen's research [3] looks into how graph-based methods can be used to find antisocial behavior on social media sites. Graph neural networks are used to show how different types of data are connected. This study shows that the approach works regardless of language or context. This was proven by testing it on several PAN datasets. The results show that graph convolutional data has a lot of potential to help people talk to each other on social media. Ernesto del Valle's study [4] is a systematic review of sentiment analysis methods for political and hate speech content in Spanish. It deals with the specific problems that researchers in this area face. The paper talks about how hard it is because there aren't any standard methods, content filters, or rules for how to interpret slang that is only used on microblogging platforms. This study brings to light the extra challenges that Spanish-speaking people face in a field that mainly studies English-language texts. Dimitri Ognibene's study [5] looks at how multi-modal machine learning algorithms could be used to make a social media companion that would help users deal with fake news and other threats on social media and teach them how to do so. This study stresses how important it is to teach people how to think critically about media and make them more aware of the limits of these algorithms. The results of the study show that educational activities should go along with algorithmic tools to help people use social media more responsibly. Ranyu Wang [6] did a study that uses a differential evolution-assisted K-means (DEK) method to look for roles in social edge computing cyberbullying situations. This study stresses how important it is to define roles in a more detailed way than just the usual "perpetrator," "victim," and "bystander" framework. The goal is to give more accurate details about how cyberbullying takes place. The suggested method was tested on datasets from Sina Weibo and was found to be a good way to group and study cyberbullying roles (DEK). Zepeng Wu's research [7] talks about BertT, a mixed neural network model for generative AI authorship verification that combines the BERT and Transformer technologies. The study is mostly about how well the model can tell the difference between texts written by people and those written by computers. It does a faster and more accurate job than baseline models like Fast-Detect. The results show that BertT could help with tasks like making sure the authorship of a work is correct in difficult NLP situations. If you use weighted voting ensembles of well-known transformer models, Javier Cruz's study [8] shows a new way to find hurtful humor on Twitter. This study found that groups are better than single transformers at finding humor and prejudice. Very well in the HUHU competition at IberLEF 2023. They did well in both the multiple classification task and the regression task. It was found that ensemble methods could help make systems that look for offensive humor in social media more accurate and reliable.

Marco Siino's study [9] looks into how the Mistral 7B model can be used to find sexism in social networks. It is part of the EXIST Lab at CLEF 2024. In this study, prompt engineering and a few-shot learning strategy are used to solve the binary classification problem of finding sexist content in English and Spanish tweets. The results show that the approach works better than some baselines, but the model could still use a lot of work. Author Javier Huertas-Tato talks about the Style Transformer for Authorship Representations (STAR) in the article [10]. A supervised contrastively pre-trained transformer is used in this model to find and describe writing style on social media. The study shows that the model does a good job of clustering and attribution of zero-shot authorship. It does well on PAN challenges too, which shows that learning in different ways can help with understanding style. The results demonstrate how important stylistic factors are for proving authorship and how the STAR model can be utilized in more social media settings. Daniel Yacob Espinosa looked into how BERT can be used to profile big names in cryptocurrency on social media as part of the PAN 2023 task in the paper [11]. This research shows that BERT is very good at putting influential people into groups. It got over 92% of the time on a number of subtasks. Even though there wasn't much data to train the model on, the results show that it can correctly name important people in the cryptocurrency space. This proves BERT's ability to handle tough tasks on social media sites. Abhay Shanbhag's study [12] builds a strong system for finding sexism on social media by using a stacking ensemble approach and several large language models (LLMs). The main goal of this study is to make sexism detection more accurate and reliable by combining the best parts of several models, like BERT and RoBERTa, into a LightGBM model that makes predictions. All of these results show that group methods can help systems that look for sexism do better at many different tasks. In the paper [13], Francesco Lomonaco shows how to improve data by translating it back to Japanese so that one can make a profile of cryptocurrency Twitter influencers. The study's findings show that using morphologically different languages to add hidden semantic information to text data can make cutting-edge Transformer models like ELECTRA and XLNet work better. The results show that back translation can help improve the accuracy of classification in places with few resources, especially when learning from few examples. For the study [14], Marco Siino uses XLNet and extra data to create profiles of Twitter users who have many followers in the cryptocurrency industry. This is part of the PAN@CLEF2023 task. Back translation with German and Italian is used in this study to add to the training dataset. The XLNet model works better in places with few resources because of this. The results clearly show that using back translation makes classification much more accurate, with Italian doing much better than German in these tests.

JanekBevendorff spoke about the PAN 2023 lab in general terms [15]. It was mostly about things that everyone did, like checking authorship, looking at writing styles with multiple authors, making profiles of cryptocurrency influencers, and finding triggers. More and better digital text forensics and stylometry have been shown in this study. It fixes problems from earlier versions and adds new tasks like influencer profiling and trigger detection. The results show how important it is to use fair testing and make fresh benchmark datasets to improve technologies that use text analysis. TF-IDF, Bi-GRU, and Text CNN models are used in Haolong Ma's [16] study to look at people who spread irony and stereotypes on Twitter. As a binary classification problem, this study looks at the problem and gets 93.33% accuracy on the test set by using both deep semantic features and statistical features that measure how often words are used. The results show that social media sites can better find irony when they use both traditional statistical methods and deep learning techniques together. For her study, Iuliia Alieva [17] used a mix of techniques, such as network analysis and natural language processing, to look at Russia's propaganda on Twitter during the 2022 invasion of Ukraine. This research looks at important groups and stories, mainly the harmful story of "fascism/Nazism." We can learn more about how propaganda is spread on social media sites and how powerful people spread lies during major world events from these results. Astha Modi [18] did research on how to use Flask to look at how people feel about things on Twitter so that a better app for data analysis can be made. Support Vector Machine (SVM) and Naive Bayes are two machine learning algorithms that were used in this study to divide feelings into three groups: positive, negative, and neutral. The results show that Flask works well for real-time sentiment analysis. Businesses can easily keep an eye on what people are saying on social media because they can see the results on a webpage. ShuvamShiwakoti's [19] study looks at how people talk about climate change on Twitter by building a new corpus with tags and multiple classifications. The ClimaConvo dataset is introduced in this study. It is made up of 15,309 tweets that have been labeled with information about their humor, hate speech, and the direction of hate speech. We can use these results to learn more about how people feel about climate change and how hard it is to have conversations about it on social media. Andrea Nasuto [20] looked at 220,870 immigration-related tweets in the UK for his study on anti-immigration feelings on Twitter. This research identifies significant polarization between pro and anti-immigration communities, with the anti-immigration group being smaller but more active and influential. The findings reveal that anti-immigration content spreads 1.66 times faster than pro-immigration messages, and bots have minimal impact on dissemination, highlighting the need to monitor highly active users to reduce social polarization and influence public attitudes toward migration.

3. METHODOLOGY

Below Figure 2 outlines the structured approach used in this study to conduct a comprehensive comparison of existing models. The flowchart is divided into four key sections, each representing a critical component of the research methodology.

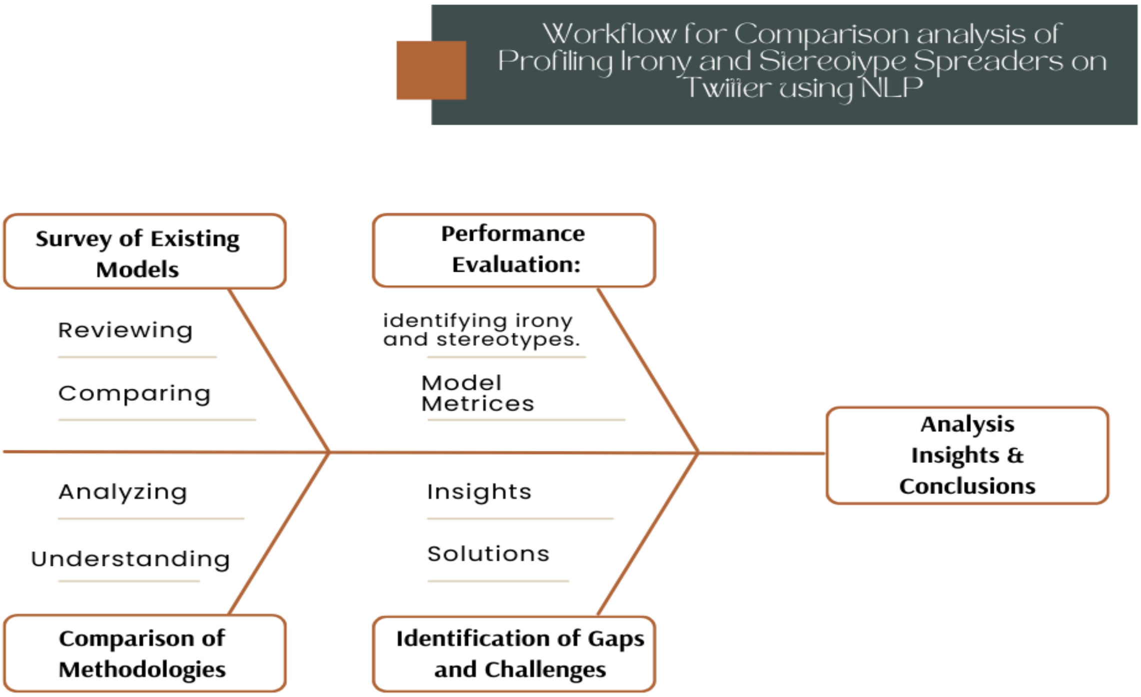
Workflow for Comparison analysis of Profiling Irony and Stereoiype Spreaders on Twiiier using NLP
Survey of Existing Models
Reviewing
Comparing
Performance Evaluation:
identifying irony and stereotypes.
Model Metrices
Analysis Insights & Conclusions
Analyzing
Understanding
Comparison of Methodologies
Insights
Solutions
Identification of Gaps and Challenges

Fig. 2. Workflow for Comparison analysis of profiling Irony and Stereotype Spreaders on Twitter using NLP

3.1 Survey of Existing Models

The first step in the workflow involves conducting a Survey of Existing Models. This section is focused on:

- Reviewing: Systematically examining previous studies and models that have been developed to profile irony and stereotype spreaders on Twitter.
- Comparing: Evaluating and comparing these models to identify their unique features, strengths, and weaknesses.

3.2 Comparison of Methodologies

Following the survey, the workflow progresses to the Comparison of Methodologies. This section emphasizes:

- Analyzing: Deeply analyzing the methodologies employed by the models, including their feature extraction techniques and algorithms.
- Understanding: Gaining a thorough understanding of how these methodologies work, what makes them effective, and where they fall short.

3.3 Performance Evaluation

Next, the Performance Evaluation stage focuses on:

- Identifying Irony and Stereotypes: Assessing how well the models perform in detecting irony and stereotypes.
- Model Metrics: Evaluating the models based on key performance metrics such as accuracy, precision, recall, and F1-score. This step is crucial for determining each model's effectiveness in real-world applications.

3.4 Identification of Gaps and Challenges

The fourth section addresses the Identification of Gaps and Challenges:

- Insights: Drawing insights from the performance evaluation to highlight areas where models perform exceptionally well or poorly.
- Solutions: Proposing potential solutions or areas of improvement to address the identified gaps and challenges in the existing methodologies.

4. COMPARISONS AND INSIGHTS

4.1 Survey of Existing Models

Based on a review of existing models, most of the ways to profile people who spread irony and stereotypes on Twitter use different Natural Language Processing (NLP) techniques to find and understand these complicated ways of expressing themselves. For example, word embeddings are often used to find stereotypes in political and multilingual speech, and TF-IDF and sentiment analysis are often used to find irony in language. Some models use graph neural networks to find connections between different kinds of data. This is especially useful for finding mean behavior and negative emotions in social media posts. A lot of attention is also paid to sentiment analysis, especially in languages other than English. This helps with the problem of figuring out complicated language on a social media site that lots of different people use. By combining traditional linguistic features with advanced machine learning and deep learning techniques, these models try to make it easier to spot and understand irony and stereotypes.

4.2 Comparison of Methodologies

Existing models for profiling irony and stereotype spreaders on Twitter reveal a range of approaches, each with its strengths and limitations. Word embeddings, as seen in Sorato's study, effectively capture semantic relationships in multilingual contexts but are limited by their inability to grasp nuanced language. Krols' use of TF-IDF and lexical features provides a simple and interpretable method for irony detection, yet it struggles with complex, context-dependent language. Graph neural networks, employed by Mortensen, excel in capturing intricate relationships across data points but are computationally intensive and require large datasets. Sentiment analysis, particularly in non-English languages as explored by Del Valle, is useful for understanding overall sentiment but faces challenges with ambiguous and informal language. Ognibene's multi-modal approach, which integrates diverse data sources, offers high accuracy but demands extensive computational resources and sophisticated models. Across these methodologies, key areas for improvement include enhancing contextual understanding, improving scalability, addressing language diversity, and advancing the integration of multi-modal data to create more robust and accurate models.

4.3 Performance Evaluation

Several approaches have demonstrated high accuracy, precision, recall, and F1-score in evaluating the performance of various models used to identify irony and stereotypes on Twitter, highlighting their effectiveness in this domain. The table below summarizes the performance metrics of key models discussed in the literature.

Table 2. Performance Metrics of Clustering Algorithms in WSNs

Study	Model	Accuracy	F1 - Score
Tibor L.R Krols [2]	TF-IDF +Sentiment Analysis	0.84	0.84
Marie Mortensen [3]	Graph Neural Networks (GNNs)	0.88	0.87
Zepeng Wu [7]	BertT (BERT +TRANSFORMER)	0.90	0.89
Elisa Chierchiello [21]	STERHEOSCHOOL annotated Model	0.85	0.80
Siino [22]	Mistal 7B (Legal Argument ClassificationS)	0.5714	05597
Marco Sinno [23]	TransMistral 7B	-	0.36
Macro Siino [24]	DeBERTa for Sense Defiance	0.25	0.25
Marco Siino [25]	Mistral 7B for Multilingual Detection of Persuasion Techniques	0.42	0.42
Marco Siino [26]	All-MPNet (Semantic Textual Relatedness	0.808	-
Marco Siino [27]	T5-Medical (Clinical Trail Inference)	-	0.63
Raberto Labadie [28]	BETO (Subtask 1)	0.874	0.76
Raberto Labadie [29]	Multilingual Trandformer Model	0.921	0.23
Shujaat Mirza [30]	Disinformation Mitigation Framework	0.85	0.82

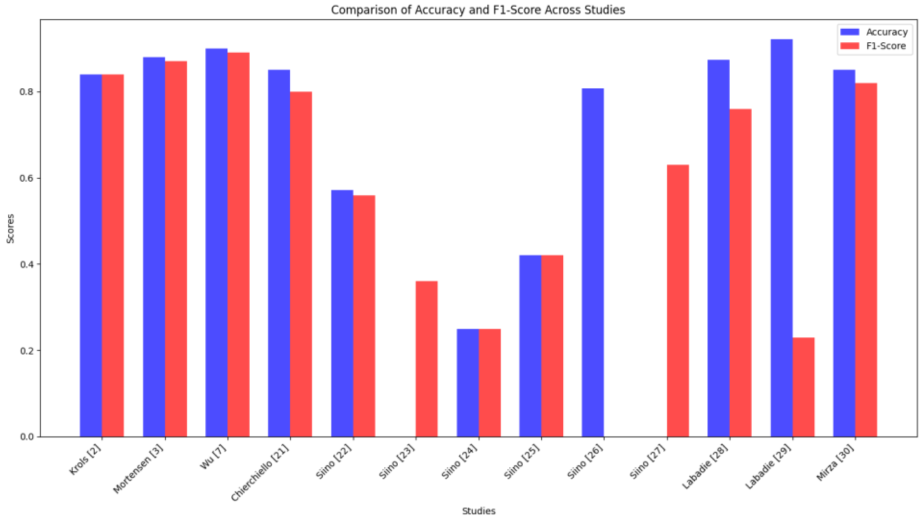

Fig. 3. Graphical Representation of the Performance Evaluation of various Previous Models

These models show what could happen when you combine them. Key metrics like accuracy and F1 score were used to compare the different models used to find irony and stereotypes on Twitter, which shows a wide range of how well they work. For example, Krols' model, which uses TF-IDF and sentiment analysis, got an accuracy and F1-score of 0.84, which shows that it did a good job of finding irony. Mortensen's Graph Neural Networks (GNNs) were even more accurate, scoring 0.88 with an F1-score of 0.87. This shows that they could find complex relationships in the data. Wu's BERTT model, which combined BERT and Transformer technologies, did very well, with an accuracy score of 0.90 and an F1 score of 0.89. This shows how well-advanced language models can handle complex content. Similarly, Chierchiello's STERHEOSCHOOL model worked well, with an accuracy score of 0.85 and an F1-score of 0.80, correctly identifying stereotypical responses.

In contrast, Siino's models produced different outcomes. Although the Mistral 7B model did pretty well at classifying legal arguments (0.5714% accuracy and 0.5599% F1-score), the TransMistral 7B model did worse at recognizing emotions (0.36% F1-score). Other Siino models, like DeBERTa for Sense Defiance and Mistral 7B for multilingual detection of persuasion techniques, did not do as well, getting F1-scores of 0.25 and 0.42, respectively. Although the F1-score wasn't given, the ALL-MPNet model for Semantic Textual Relatedness did pretty well, with an accuracy of 0.808. Laboratory 1's BETO model got an F1-score of 0.76 and an accuracy of 0.874. Meanwhile, his multilingual transformer model got a high accuracy score of 0.921 but a much lower F1 score of 0.23, which suggests that precision and recall are not balanced. As a final point, Mirza's Disinformation Mitigation Framework did well overall, with an accuracy score of 0.85 and an F1 score of 0.82, showing that it is effective at fighting disinformation. These results show that model performance isn't always the same and that it needs to be improved, especially when it comes to getting metrics that are equivalent across different tasks.

4.4 Identification of Gaps and Challenges

It is getting easier to find people on Twitter who use irony and stereotypes, but there are still some holes and issues in the field:

Detection of Nuanced Language:

- It's hard for many models to spot irony because it depends on the tone, the situation, and the cultural references. Simple models that only look at words or basic sentiment analysis often can't handle these subtleties, which can cause them to classify things wrongly or miss ironic content.

Impact of Data Diversity:

- On social media sites like Twitter, people speak a lot of different dialects, languages, and speech styles. This is a big problem. When used all over the world, models that work well in one language or culture might not work well in others. This could cause bias and make the models less accurate. To make models that work for everyone, there aren't any standard datasets that show how different people are.

Scalability Issues:

- Many computers and data are usually needed for more complicated models, such as those that use graph neural networks or multi-modal approaches. Although these models work well in controlled settings, they usually can't be used to monitor a lot of social media sites in real-time because they need many resources. Additionally, these models can be hard to understand because they are very complicated, which can make it difficult to trust their results fully.

Ethical Considerations:

- Making profiles of people on social media sites raises ethical concerns, especially when it comes to privacy and the chance that algorithms could be biased. It's more likely that models will reinforce social biases or unfairly target certain groups as they get more complicated. We need both technical solutions and careful thought about how these models will impact society as a whole in order to solve these moral issues.

Discussion and Conclusions

The research looked at all the existing NLP models and how they can be used to make profiles of Twitter users who spread irony and stereotypes. A list was made of the main pros and cons of the current methods by looking at and comparing different ones. It can be hard to find delicate language, deal with different kinds of data, and make sure the system can expand. There are some big problems with these models that need to be fixed, but the tests did show that combining traditional linguistic features with more advanced machine-learning techniques works well. Even more care needs to be taken when using these models in the real world because of ethical concerns, such as worries about privacy and possible biases. It's important to make these models better in the future so they are more accurate, fair, and moral.

REFERENCES

1. Sorato, D., Lundsteen, M., Ventura, C. C., & Zavala-Rojas, D. (2024). Using word embeddings for immigrant and refugee stereotype quantification in a diachronic and multilingual setting. *Journal of Computational Social Science*, 1-53.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
2. Krols, T. L., Mortensen, M., & Oldenburg, N. (2023). Profiling Irony & Stereotype: Exploring Sentiment, Topic, and Lexical Features. *arXiv preprint arXiv:2311.04885*.
3. Toshevska, M., Kalajdziski, S., &Gievska, S. (2023, September). Graph Neural Networks for Antisocial Behavior Detection on Twitter. In *International Conference on ICT Innovations* (pp. 222-236). Cham: Springer Nature Switzerland.

4. del Valle, E., & de la Fuente, L. (2023). Sentiment analysis methods for politics and hate speech contents in Spanish language: a systematic review. *IEEE Latin America Transactions*, *21*(3), 408-418..
5. Ognibene, D., Donabauer, G., Theophilou, E., Buršić, S., Lomonaco, F., Wilkens, R., ... & Kruschwitz, U. (2023). Moving beyond benchmarks and competitions: towards addressing social media challenges in an educational context. *Datenbank-Spektrum*, *23*(1), 27-39.
6. Wang, R., Lu, T., Zhang, P., & Gu, N. (2024). Role Identification based Method for Cyberbullying Analysis in Social Edge Computing. *arXiv preprint arXiv:2408.03502*.
7. Wu, Z., Yang, W., Ma, L., & Zhao, Z. (2024). BertT: A Hybrid Neural Network Model for Generative AI Authorship Verification. *Working Notes of CLEF*.
8. Cruz, J., Elvira, L., Tabernero, M., & Segura-Bedmar, I. (2023). In Unity, There Is Strength: On Weighted Voting Ensembles for Hurtful Humour Detection. In *IberLEF@ SEPLN*.
9. Siino, M., & Tinnirello, I. (2024). Prompt Engineering for Identifying Sexism using GPT Mistral 7B. *Working Notes of CLEF*.
10. Huertas-Tato, J., Martín, A., & Camacho, D. (2024). Understanding writing style in social media with a supervised contrastively pre-trained transformer. *Knowledge-Based Systems*, *296*, 111867.
11. Espinosa, D. Y., & Sidorov, G. (2023). Using BERT to Profiling Cryptocurrency Influencers. In *CLEF (Working Notes)* (pp. 2568-2573).
12. Shanbhag, A., Jadhav, S., Date, A., Joshi, S., & Sonawane, S. (2024). The Wisdom of Weighing: Stacking Ensembles for a More Balanced Sexism Detector. *Working Notes of CLEF*.
13. Lomonaco, F., Siino, M., & Tesconi, M. (2023). Text Enrichment with Japanese Language to Profile Cryptocurrency Influencers. In *CLEF (Working Notes)* (pp. 2708-2716).
14. Siino, M., & Tinnirello, I. (2023). XLNet with Data Augmentation to Profile Cryptocurrency Influencers. In *CLEF (Working Notes)* (pp. 2763-2771).
15. Bevendorff, J., Borrego-Obrador, I., Chinea-Ríos, M., Franco-Salvador, M., Fröbe, M., Heini, A., ... & Zangerle, E. (2023, September). Overview of pan 2023: Authorship verification, multi-author writing style analysis, profiling cryptocurrency influencers, and trigger detection: Condensed lab overview. In *International Conference of the Cross-Language Evaluation Forum for European Languages* (pp. 459-481). Cham: Springer Nature Switzerland.
16. Ma, H., Li, D., & Sun, Y. (2022). Profiling Irony and Stereotype Spreaders on Twitter Using TF-IDF and Neural Network. In *CLEF (Working Notes)* (pp. 2578-2584).
17. Alieva, I., Kloo, I., & Carley, K. M. (2024). Analyzing Russia's propaganda tactics on Twitter using mixed methods network analysis and natural language processing: a case study of the 2022 invasion of Ukraine. *EPJ Data Science*, *13*(1), 42.
18. Modi, A., Shah, K., Shah, S., Patel, S., & Shah, M. (2024). Sentiment analysis of Twitter feeds using flask environment: A superior application of data analysis. *Annals of Data Science*, *11*(1), 159-180.
19. Shiwakoti, S., Thapa, S., Rauniyar, K., Shah, A., Bhandari, A., & Naseem, U. (2024, May). Analyzing the dynamics of climate change discourse on twitter: A new annotated corpus and multi-aspect classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 984-994).
20. Nasuto, A., & Rowe, F. (2024). Exposing Hate--Understanding Anti-Immigration Sentiment Spreading on Twitter. *arXiv preprint arXiv:2401.06658*.
21. Chierchiello, E., Bourgeade, T., Ricci, G., Bosco, C., & D'Errico, F. (2024, May). Studying Reactions to Stereotypes in Teenagers: an Annotated Italian Dataset. In *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying@ LREC-COLING-2024* (pp. 115-125).
22. Siino, M. (2024, June). Mistral at SemEval-2024 task 5: Mistral 7B for argument reasoning in civil procedure. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)* (pp. 155-162).
23. Siino, M. (2024, June). Transmistral at semeval-2024 task 10: Using mistral 7b for emotion discovery and reasoning its flip in conversation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)* (pp. 298-304).
24. Siino, M. (2024, June). Deberta at semeval-2024 task 9: Using deberta for defying common sense. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)* (pp. 291-297).

25. Siino, M. (2024, June). Mcrock at semeval-2024 task 4: Mistral 7b for multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)* (pp. 53-59).
26. Siino, M. (2024, June). All-mpnet at semeval-2024 task 1: Application of mpnet for evaluating semantic textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)* (pp. 379-384).
27. Siino, M. (2024, June). T5-medical at semeval-2024 task 2: Using t5 medical embedding for natural language inference on clinical trial data. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)* (pp. 40-46).
28. Labadie Tamayo, R., Chulvi-Ferriols, M. A., & Rosso, P. (2023). Everybody hurts, sometimes overview of hurtful humour at iberlef 2023: Detection of humour spreading prejudice in twitter. *Procesamiento del lenguaje natural*, (71), 383-395.
29. Labadie Tamayo, R. (2023). Challenges in Humor Recognition: Cross-language Perspective and Hurtfulness Analysis.
30. Mirza, M. S., Begum, L., Niu, L., Pardo, S., Abouzied, A., Papotti, P., &Pöpper, C. (2023). Tactics, Threats & Targets: Modeling Disinformation and its Mitigation. In *NDSS*.