Introduction

The exponential growth of Artificial Intelligence (AI), particularly in the domain of deep learning, has significantly transformed the landscape of natural language processing (NLP). Among its most impactful applications is sentiment analysis—an area where deep learning models have outperformed traditional approaches in deciphering the subjective tone embedded within human language. The proliferation of social media platforms such as X (formerly Twitter) and Facebook since the early 2000s has created an unprecedented corpus of real-time, unstructured public opinion data. According to Kemp, billions of users engage daily across these platforms, providing a dynamic, high-volume stream of sentiment-laden content. This immense data source has established sentiment analysis as a vital tool for sectors ranging from commerce and politics to public health, enabling timely insights into public opinion and behavioural trends.

To address the challenges inherent in sentiment classification from informal, context-rich social media posts, this study evaluates the performance of several state-of-the-art models: DistilBERT, BERTweet, VADER (Valence Aware Dictionary and Sentiment Reasoner), and a custom Feedforward Neural Network (FNN). Each model represents a unique approach to sentiment analysis—ranging from rule-based methods to transformer-based architectures and traditional neural networks. Transformer-based models, particularly BERT derivatives such as DistilBERT and BERTweet, have demonstrated exceptional capabilities in a range of NLP tasks due to their contextual understanding and language modelling depth (Devlin et al., 2019). Meanwhile, rule-based systems like VADER retain relevance in their computational efficiency and interpretability, especially for short-text classification (Hutto & Gilbert, 2014). The inclusion of a custom FNN aims to explore whether a domain-specific, lightweight neural architecture can provide improved performance on social sentiment datasets.

The core research question—"How successful are deep learning models like DistilBERT, BERTweet, VADER, and a custom FNN in accurately identifying sentiment in social media posts, and how do their results compare across sentiment categories?"—guides this comparative evaluation. Accurate sentiment classification is critical, as misclassifications can distort interpretations of public opinion, with significant downstream consequences for political decision-making, marketing strategy, and crisis management (Zimbra et al., 2018).

This investigation aims to systematically assess each model’s ability to capture the complex emotional undertones present in user-generated content, focusing on comparative accuracy across sentiment categories. By examining the linguistic challenges and contextual subtleties specific to social media, this study contributes to the broader effort of refining sentiment analysis techniques and enhancing the deployment of AI in real-world text analytics tasks.

Background

Sentiment analysis in the context of social media has emerged as a central research focus within AI, offering a unique window into real-time public discourse. Advances in natural language processing have made it possible to derive structured insights from vast volumes of unstructured text, allowing institutions to gauge public sentiment across domains such as politics, marketing, and education. Recent contributions to this field, such as those by Khan et al. (2023) and Rajan (2024), have showcased the growing sophistication of AI models in capturing emotional tone and intent from user-generated content.

Khan et al. (2023) proposed a Location-Based Election Prediction (LBEP) framework that integrated transformer models such as BERT, BERTweet, and ElecBERT with geospatial sentiment data derived from social media. Their model, combining VADER's rule-based scoring with transformer-based representations, accurately predicted election results in 41 US states, outperforming traditional polling mechanisms. This fusion of sentiment analysis and geographic metadata demonstrated the strategic advantage of AI in political forecasting.

In contrast, Rajan (2024) focused on the broader linguistic challenges of sentiment classification. By investigating the ability of AI models to manage informal language, ambiguity, and unconventional grammar, this study highlighted the robustness of transformer models in interpreting varied textual inputs. While it did not incorporate location-based data, Rajan's findings reinforced the importance of model adaptability in processing spontaneous digital communication, with implications extending to consumer analytics and public health monitoring.

Other notable applications of sentiment analysis include its use in tracking societal responses to complex issues. Chen, Sack, and Alam (2022) utilised AI models to extract sentiment around migration, demonstrating how computational methods can reveal latent societal tensions. Similarly, Chen, Vorvoreanu, and Madhavan (2014) applied sentiment analytics in educational contexts to interpret student feedback, showcasing the versatility of such tools across diverse fields.

Despite significant advancements, several gaps persist within existing literature. Few studies conduct side-by-side comparisons of rule-based models and transformer-based models within a unified evaluation framework. Additionally, while location-based sentiment modelling has shown promise in political domains, its application in areas like crisis response or personalised advertising remains underexplored. Persistent issues such as classification bias, failure to recognise sarcasm or irony, and the inability to differentiate between factual and subjective statements continue to constrain performance (Cambria et al., 2020).

The present work addresses these gaps by offering a multi-model evaluation of sentiment classification systems operating on a unified dataset. By comparing transformer-based models, a rule-based model, and a custom neural network under consistent conditions, this study seeks to advance the development of more robust, context-aware sentiment analysis frameworks tailored to the nuanced and evolving language of social media.

Methodology

This study undertakes a comparative evaluation of four sentiment classification models—DistilBERT, BERTweet, VADER, and a custom Feedforward Neural Network (FNN)—on the task of analysing sentiment from social media posts. The objective is to benchmark their performance across three sentiment categories: positive, negative, and neutral. A combination of quantitative metrics including precision, recall, F1 score, and Matthews Correlation Coefficient (MCC) is used to assess classification efficacy.

In addition to performance benchmarking, the methodology also investigates the impact of preprocessing strategies and dataset characteristics on model effectiveness. This dual focus enables a deeper understanding of model strengths, weaknesses, and applicability to real-world social data, particularly within the context of sentiment-rich public discourse.

The experimental pipeline follows a structured sequence: data preparation, model training, performance evaluation, and results comparison. Each model is evaluated under consistent conditions using the same dataset and split configuration to ensure fairness in comparison and validity of conclusions drawn.

Model Selection

The four models included in this study (DistilBERT, BERTweet, VADER, and a custom FNN) were selected to reflect diverse paradigms in sentiment analysis. Their architectural diversity allows for a thorough investigation into the trade-offs between accuracy, interpretability, and computational efficiency.

DistilBERT

DistilBERT is a distilled version of the BERT transformer model, optimised for computational efficiency while retaining approximately 95% of BERT's performance on downstream NLP tasks. It employs a multi-layer bidirectional transformer encoder and self-attention mechanism, which allows it to capture contextual dependencies across the input sequence. The scaled dot-product attention function used in DistilBERT is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt[2]{d_k}}\right)V$$

where Q, K, and V represent query, key, and value matrices, and dk is a scaling factor ensuring numerical stability.

BERTweet

BERTweet extends the BERT architecture by pretraining on a large corpus of Twitter data. This model is domain-tuned to handle social media-specific linguistic features such as abbreviations, emojis, hashtags, and informal grammar. The adaptation of BERT to social media syntax allows BERTweet to better represent the nuanced structure of tweets, making it particularly suitable for sentiment classification tasks in this domain.

VADER

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a rule-based model designed for sentiment analysis of social media content. It relies on a lexicon of sentiment-rich words and syntactic rules to compute a compound sentiment score, defined as:

$$\text{Compound Score} = \sum_{token \in tokens} \text{Valence(Token)}$$

This score is modified by contextual clues such as negation, intensifiers, and punctuation. Although not deep learning-based, VADER remains a baseline standard due to its efficiency and interpretability.

Feedforward Neural Network (FNN)

The custom model used in this study is a Feedforward Neural Network implemented in Keras with TensorFlow backend. Unlike transformer models, the FNN uses traditional dense layers to process numerical embeddings of input text. The architecture is composed of:

- Input Layer: Tokenised text is converted into vector embeddings.

- Hidden Layers: Dense layers with ReLU activations model non-linear relationships.

- Output Layer: A softmax function generates class probabilities for three sentiment categories.

The model is trained using categorical cross-entropy loss:

$$L = -\sum_{i=1} y_i \log(\hat{y_i})$$

where y_i represents the true label and $\hat{y_i}$ is the predicted probability for each class. The Adam optimiser is employed to enhance training efficiency and convergence speed.

Preprocessing and Evaluation Metrics

Text preprocessing transforms raw social media posts into a structured format suitable for model analysis. This process involves:

- Tokenisation: Breaking text into individual tokens.
- Normalisation: Standardising case and removing extraneous characters.
- Embedding: Converting tokens into numerical vectors.

Each model is trained and tested on a structured dataset, with performance evaluated using:

- **Precision**: $P = \frac{TP}{TP+FP}$

- **Recall**: $R = \frac{TP}{TP+FN}$

- **F1 Score**: $\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

These metrics provide insight into each model's ability to correctly classify sentiment while balancing false positives and false negatives. By applying these methodologies, the study offers a rigorous comparative analysis of state-of-the-art sentiment analysis models alongside a custom-built neural network, demonstrating their respective strengths and limitations in processing complex social media language.

Experiment

The experiments are conducted using the US_Presidential_Election_2020_Dem_Rep dataset, obtained from Hugging Face's datasets repository. A 5% subset of the dataset is utilised to ensure manageability while preserving label diversity. Text entries with missing content are replaced with empty strings, and TF-IDF vectorisation is applied to convert the corpus into a numerical format, limited to the top 1,000 terms for computational tractability.

Sentiment labels (positive, negative, neutral) are encoded using label encoding and transformed into categorical format. The dataset is partitioned into training and testing sets using an 80/20 stratified split to maintain class distribution.

The FNN is configured with two hidden layers comprising 64 and 32 neurons respectively, each followed by a dropout layer to prevent overfitting. ReLU activation is used in the hidden layers and softmax in the output layer. The model is trained using the Adam optimiser with a learning rate of 0.001 and categorical cross-entropy as the loss function. Early stopping with a patience of two epochs is employed to mitigate overfitting. Training is performed over a maximum of 10 epochs with a batch size of 64. GPU acceleration is enabled if available via TensorFlow.

In parallel, baseline models are tested using the following configurations:

DistilBERT: distilbert-base-uncased-finetuned-sst-2-english

BERTweet: vinai/bertweet-base

VADER: NLTK's SentimentIntensityAnalyzer

All models are evaluated on the same test set using F1 score, MCC, and confusion matrices. These results provide a comprehensive picture of model performance in analysing sentiment-rich political discourse.

Ethics

The deployment of AI-driven sentiment analysis systems, particularly in social media contexts, necessitates a rigorous consideration of ethical implications. These include concerns related to privacy, informed consent, algorithmic manipulation, and data sovereignty. The unstructured and highly personal nature of social media data raises questions about how user-generated content is collected, processed, and interpreted in large-scale computational frameworks.

As Nissenbaum (2010) argues, traditional notions of privacy are insufficient in the context of dynamic online ecosystems. Her framework of "contextual integrity" highlights the importance of preserving the expectations of information flow within specific contexts. The transformation of public social media content into training data for sentiment analysis models must, therefore, respect user expectations, particularly in terms of how their opinions are repurposed in political or commercial analyses.

Woolley and Howard (2018) identify the potential misuse of sentiment analysis within the realm of computational propaganda. Automated systems can be weaponised to influence public discourse through the amplification or suppression of certain sentiments, undermining democratic processes and civic trust. The risk of misapplication is especially salient when models are deployed to shape narratives around elections, crises, or polarising social events.

Data sovereignty further complicates the ethical landscape. As Kalathil and Boas (2010) note, the cross-border nature of data flows on social platforms challenges national jurisdictional authority. Sentiment analysis conducted across international datasets must navigate diverse regulatory regimes and cultural norms, requiring researchers to ensure that analysis respects both individual data rights and broader societal protections.

In light of these concerns, this study upholds principles of fairness, accountability, and transparency in model design and data handling. Data was sourced from a publicly available and ethically shared dataset, and all analyses were conducted with the aim of academic inquiry, not real-time intervention or prediction. Ethical oversight remains essential as AI systems scale, particularly in sensitive domains involving political discourse and public sentiment.

Results

The comparative evaluation of sentiment classification models demonstrated significant variability in performance across the tested architectures. The custom Feedforward Neural Network (FNN) substantially outperformed the transformer-based and rule-based baselines on all evaluated metrics.

The FNN achieved a weighted F1 score of 0.8725 and a Matthews Correlation Coefficient (MCC) of 0.8094, indicating a high degree of predictive accuracy and class discrimination. The confusion matrix revealed a balanced distribution of true positives across the positive, negative, and neutral sentiment classes, suggesting that the model was effective in handling class variability and ambiguity(.

In contrast, DistilBERT yielded an F1 score of 0.3118 and an MCC of 0.2012. Although the model was able to identify some negative sentiment correctly, the confusion matrix showed a marked bias toward the negative class. Misclassification of clearly positive inputs—for example, "The event was truly inspiring and left me hopeful!"—as negative, indicates a shortcoming in DistilBERT's ability to interpret sentiment-laden language with contextual optimism.

BERTweet performed notably worse, with an F1 score of 0.1728 and an MCC of 0.0000, approximating random guessing. The model demonstrated an overwhelming tendency to assign negative sentiment, regardless of textual content. For example, a clearly positive post such as "Had a great time catching up with old friends today!" was misclassified as negative, highlighting its inability to generalise from pretraining on informal language alone.

VADER, while computationally efficient, also underperformed, achieving an F1 score of 0.2437 and an MCC of 0.0498. The rule-based nature of VADER allowed it to identify strong polar sentiment but caused it to misclassify nuanced or ambiguous statements. For example, it interpreted the statement "Not the best experience, but it could have been worse" as entirely negative, despite its neutral connotation.

The confusion matrices associated with each model further underscore the custom FNN's superiority. While the transformer-based models struggled with domain adaptation and contextual subtleties, the FNN benefited from domain-specific tuning and architectural simplicity, enabling it to generalise effectively across sentiment categories.

These findings suggest that the integration of targeted pretraining, domain-specific embeddings, and balanced data representation can result in significant improvements in model performance. Future enhancements may involve hybrid architectures or ensemble methods to further improve robustness in handling complex sentiment expressions, including sarcasm and irony.

Conclusion

This study provides a comprehensive evaluation of four sentiment analysis models—DistilBERT, BERTweet, VADER, and a custom Feedforward Neural Network (FNN)—on a real-world dataset of politically charged social media posts. The results demonstrate the superior performance of the FNN, which achieved the highest scores across all evaluation metrics and demonstrated consistent accuracy in classifying sentiment across all three categories.

DistilBERT exhibited moderate success, highlighting the importance of fine-tuning transformer models on domain-specific content. BERTweet, despite being pre-trained on social media data, underperformed substantially, indicating that pretraining alone may not be sufficient for effective sentiment analysis without additional task-specific tuning. VADER, while simple and efficient, lacked the nuance required to correctly classify more ambiguous or context-sensitive posts.

A significant finding is the potential of hybrid or ensemble models that leverage the interpretability of rule-based systems and the contextual power of transformer architectures, combined with the targeted accuracy of neural networks. Such approaches could address the limitations observed in each individual model and enhance adaptability to evolving online language trends.

Ethical considerations were integral to this study, underscoring the importance of privacy, informed consent, and data governance in the use of AI for analysing public discourse. The study also calls for attention to data sovereignty and algorithmic accountability, particularly as sentiment analysis increasingly intersects with political, economic, and social decision-making.

Looking forward, expanding sentiment analysis tools to handle multi-lingual inputs, cultural variability, and domain-specific sentiment markers will be critical. Incorporating these advancements could significantly broaden the applicability and fairness of AI-driven sentiment classification.

Finally, this research supports the integration of sentiment analysis into predictive analytics pipelines. By correlating sentiment trends with real-world outcomes such as market fluctuations or electoral results, sentiment analysis can evolve into a powerful forecasting tool for both public and private sectors.

Appendices

Figure 1 (Neural Network Confusion Matrix)

Negative	**5100**	**858**	**410**
Neutral	**223**	**4560**	**161**
Positive	**259**	**548**	**6517**
	Negative	**Neutral**	**Positive**

Figure 2 (DistilBERT Confusion Matrix)

Negative	**5815**	**553**	**0**
Neutral	**2882**	**2142**	**0**
Positive	**3913**	**3393**	**0**
	Negative	**Neutral**	**Positive**

Figure 3(BERTweet Confusion Matrix)

Negative	**6368**	**0**	**0**
Neutral	**5024**	**0**	**0**
Positive	**7324**	**0**	**0**
	Negative	**Neutral**	**Positive**

Figure 4(VADER Confusion Matrix)

Negative	**6364**	**4**	**0**
Neutral	**5023**	**1**	**0**
Positive	**18**	**76306**	**0**
	Negative	**Neutral**	**Positive**

Figure 5 (F1 Scores)

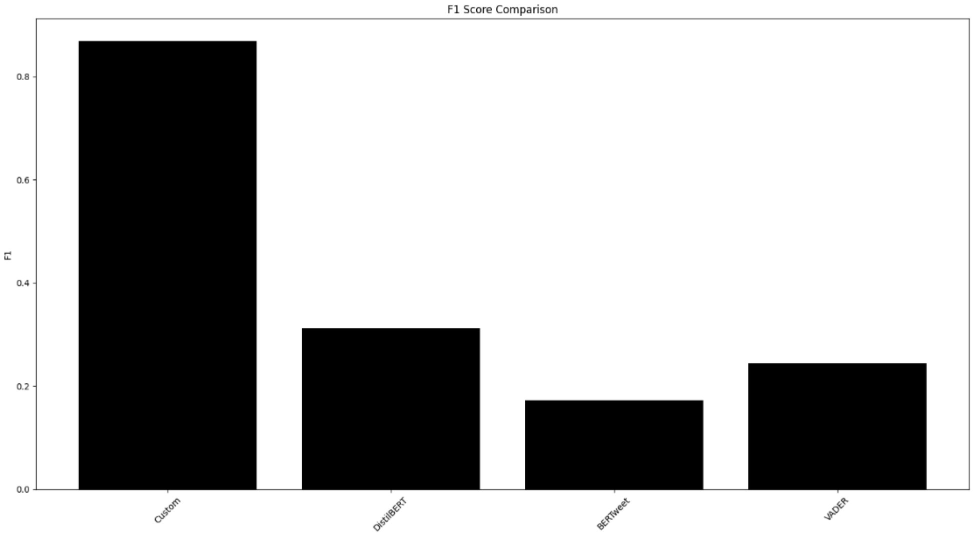
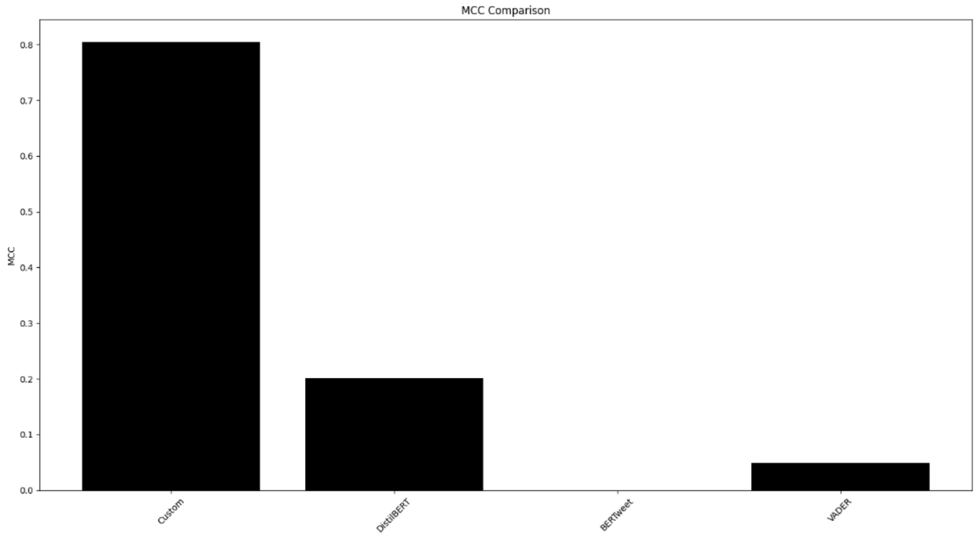


Figure 6 (MCC Scores)



References

[1] Khan, Asif & Boudjellal, Nada & Zhang, Huaping & Ahmad, Arshad & Khan, Maqbool. (2023). From Social Media to Ballot Box: Leveraging Location-Aware Sentiment Analysis for Election Predictions. *Computers, Materials & Continua*, 77, 3037-3055. DOI: 10.32604/cmc.2023.044403.

[2] Chen, Y., Sack, H., & Alam, M. (2022). Analysing social media for measuring public attitudes toward controversies and their driving factors: a case study of migration. *Social Network Analysis and Mining*, 12, 135. DOI: 10.1007/s13278-022-00896-4.

[3] Chen, Xin & Vorvoreanu, Mihaela & Madhavan, Krishna. (2014). Mining Social Media Data for Understanding Students' Learning Experiences. *IEEE Transactions on Learning Technologies*, 7, 246-259. DOI: 10.1109/TLT.2013.2296520.

[4] Rajan, K. (2024). Sentiment Analysis of Social Media Using Artificial Intelligence. *IntechOpen*. DOI: 10.5772/intechopen.113092.

[5] Nissenbaum, H. (2010). *Privacy in Context: Technology, Policy, and the Integrity of Social Life.* Stanford University Press.

[6] Woolley, S. C., & Howard, P. N. (2018). *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media.* Oxford University Press.

[7] Kalathil, S., & Boas, T. C. (2010). *Open Networks, Closed Regimes: The Impact of the Internet on Authoritarian Rule.* Carnegie Endowment.

[8] Kemp, S. (2023). *Digital 2023: Global Overview Report*. DataReportal. Retrieved from https://datareportal.com/reports/digital-2023-global-overview-report.

[9] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186. Retrieved from https://aclanthology.org/N19-1423.pdf.

[10] Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), 216-225.

[11] Zimbra, D., Abbasi, A., Chen, H., & Nunamaker, J. F. (2018). A Text Analytics Framework for Analyzing Online Public Opinions and Discussions. *Journal of Management Information Systems*, 35(1), 510-546. DOI: 10.1080/07421222.2018.1440753.