

A COMPARATIVE EVALUATION OF MACHINE LEARNING AND DEEP LEARNING MODELS FOR SENTIMENT ANALYSIS IN BRAND REPUTATION MONITORING ON SOCIAL MEDIA

Abhishen Raju

De Montfort University, United Kingdom

Abstract

Social media has become one of the main places where people talk about brands every day. Customers use platforms such as Twitter (X) to praise, complain and share their opinions, and these public messages can quickly change how a brand is seen. Because of this, many companies now rely on sentiment analysis to follow what people are saying and to react to problems and opportunities in real time (Liu, 2012). At the same time, the informal, sarcastic and context-dependent language used on these platforms makes sentiment analysis difficult for standard text classification methods (Cambria, 2016). This study compares six sentiment classification approaches—Logistic Regression, Naive Bayes, Support Vector Machines, Random Forest, a Long Short-Term Memory (LSTM) network and a stacked ensemble—on the widely used Sentiment140 dataset of 1.6 million labelled tweets (Go, Bhayani & Huang, 2009), with the specific goal of informing model choice for brand reputation monitoring. Under matched data and preprocessing conditions, LSTM achieved the highest individual accuracy (79.20%), with the ensemble model close behind (78.69%), while traditional models remained competitive around 75–77% at much lower computational cost. Error analysis shows that all approaches still struggle with sarcasm, context-dependent phrasing and ambiguous language. For practitioners, the results offer a concrete trade-off: where speed and low resource usage matter most, traditional models and Random Forest are reasonable choices; where accuracy is critical and computing power is available, LSTM is preferable. The findings also point to sarcasm detection and richer contextual representations as key areas for future work in brand reputation monitoring.

Keywords:

Sentiment Analysis, Brand Reputation Management, Machine Learning, Deep Learning, LSTM, Ensemble Learning, Social Media Analytics

1. INTRODUCTION

Social media has become one of the main places where people talk about brands every day. Customers use platforms such as Twitter (X) to praise, complain and share their opinions, and these public messages can quickly change how a brand is seen. Because of this, many companies now rely on sentiment analysis to follow what people are saying and to react to problems and opportunities in real time [15]. Sentiment analysis tries to classify text as positive, negative or neutral, but social media messages are difficult to handle with standard methods. Posts are often very short, full of slang and abbreviations, and they frequently use sarcasm. A tweet like “Oh great, another delay...” looks positive at first glance but clearly expresses frustration [18]. Traditional sentiment models such as Logistic Regression and Naive Bayes usually work with simple word-level features, for example bag-of-words or TF-IDF, and do not see the tone or context behind these expressions [20]. More advanced approaches, including Support Vector Machines, Random Forests and deep learning models such as Long Short-Term Memory (LSTM) networks, can

capture more complex patterns in text and often give better accuracy [22]. However, they also need more computing power and take longer to train and run. This raises a practical question for brand reputation monitoring systems that must process large volumes of data in real time: which model gives the best balance between accuracy and speed for everyday use in organisations? Although many studies have applied machine learning and deep learning to sentiment analysis, relatively few have compared these models side by side with a clear focus on brand reputation management [4]. In particular, there is limited work that discusses model selection as a trade-off between classification performance and computational cost under realistic social media conditions. This study addresses that gap by systematically comparing six models: Logistic Regression, Naive Bayes, Support Vector Machines, Random Forest, an LSTM network and an ensemble model on the widely used Sentiment140 dataset of 1.6 million labelled tweets [10]. The main objectives are: (1) to evaluate how traditional, advanced and deep learning models perform on binary sentiment classification of social media text; (2) to test whether an ensemble approach can improve on the best individual models; and (3) to analyse which types of errors, especially those related to sarcasm and context-dependent language, still occur, and what these errors mean for brands that depend on sentiment analysis to protect and manage their reputation.

2. RELATED WORK

Sentiment analysis is a branch of natural language processing that focuses on finding opinions and attitudes in text, such as whether a statement is positive or negative [15]. It usually starts with basic NLP steps like tokenisation, stemming and lemmatisation, which turn raw text into a cleaner and more consistent form that models can work with [13] [17]. These steps help reduce noise and the number of distinct word forms in the data. Even with these techniques, sentiment analysis still finds it hard to deal with sarcasm, idioms and meanings that depend heavily on context, because these patterns often cannot be captured by simple word counts or dictionaries [7].

2.1 TRADITIONAL MACHINE LEARNING APPROACHES.

Early work in sentiment analysis relied heavily on traditional supervised learning models. Logistic Regression and Naive Bayes are two of the most common baselines for text classification because they are simple, fast and easy to implement [20]. Both typically use bag-of-words or TF-IDF features, which represent documents by the words they contain and their relative importance. These methods often perform well when sentiment is expressed directly through clear positive or negative words. However, they struggle when sentiment depends on word order or

on how words interact in a sentence [19]. For example, Naive Bayes assumes that features are independent, which is convenient mathematically but does not match how language actually works [17]. This limits its ability to pick up subtle cues such as sarcasm or irony.

2.2 ADVANCED MACHINE LEARNING APPROACHES

To address some of these limits, later studies applied more powerful models such as Support Vector Machines (SVMs) and Random Forests. SVMs try to find a decision boundary that separates positive and negative examples with the largest possible margin in a high-dimensional feature space [8] [12]. This often gives better performance than simple linear models, especially when the data is not easily separated by a straight line. Random Forests are ensembles of many decision trees trained on different parts or versions of the data; their combined predictions tend to be more robust and less sensitive to noise [6]. In social media sentiment analysis, Random Forests have been used to handle messy, informal text more effectively. The drawback of these advanced models is that they often require more computation and careful tuning of parameters to reach their best performance.

2.3 DEEP LEARNING APPROACHES

Deep learning models have changed sentiment analysis by learning representations directly from sequences of words instead of relying only on hand-crafted features. Long Short-Term Memory (LSTM) networks, a type of recurrent neural network, are designed to keep useful information from earlier time steps and forget what is no longer needed [11]. This makes them well suited to modelling the flow of sentiment across a sentence or tweet, because they can take word order and context into account. Many studies report that LSTM-based models outperform traditional and advanced machine learning models on a range of sentiment tasks, including social media data, especially when enough training examples and computing resources are available [22] [1]. At the same time, LSTMs are more expensive to train and deploy, which can be a barrier for organisations that need to analyse large live streams of social media posts.

2.4 ENSEMBLE APPROACHES

Another line of work combines multiple models to improve overall performance. Ensemble methods use the idea that different models may make different mistakes, so combining them can reduce error and increase stability [9]. In sentiment analysis, researchers have built ensembles that mix traditional classifiers such as Logistic Regression and Naive Bayes with more complex models like SVMs, Random Forests or deep learning networks [21]. Some ensembles use voting schemes, while others stack models and train a meta-classifier on their outputs. These approaches often achieve higher accuracy and better generalisation than any single model [6] [14], but they also add complexity in terms of model selection, weighting and maintenance. For practical systems, it is important to balance these gains against extra computational and implementation cost.

Thus, the literature shows steady progress from simple, fast models towards more expressive but heavier approaches, including deep learning and ensembles. However, there is still no

clear consensus on which type of model is best suited for real-time brand reputation monitoring on social media, where both accuracy and speed matter [4]. This uncertainty motivates the comparative study presented in this paper.

3. METHODOLOGY

This section explains how the study was designed, which data was used, how the models were built and trained, and how their performance was evaluated. The goal is to compare six sentiment classification approaches under the same conditions so that differences in results can be linked to the models themselves rather than to changes in data or preprocessing.

4. RESEARCH DESIGN

The study uses comparative research design. Six models: Logistic Regression, Naive Bayes, Support Vector Machines (SVM), Random Forest, a Long Short-Term Memory (LSTM) network and a stacked ensemble are trained and tested on the same tweet data with the same preprocessing and train/validation/test splits [16]. This setup allows a direct comparison of accuracy and other metrics across traditional, advanced and deep learning approaches, instead of looking at any single model in isolation. The Fig.1 presents an overview of the research pipeline used in this study.

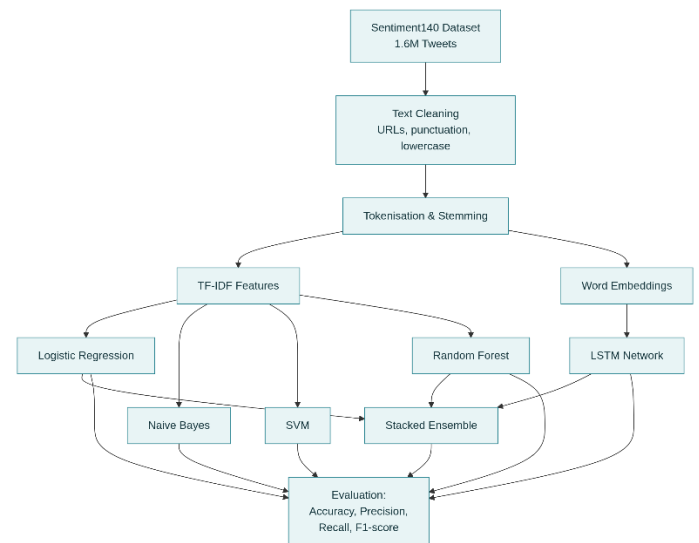


Fig.1. Overview of the research design and model evaluation pipeline.

4.1 DATASET AND PREPROCESSING

The experiments use the Sentiment140 dataset [10], which contains 1.6 million tweets automatically labelled as positive or negative based on emoticons. This dataset is widely used in sentiment analysis research and offers a large amount of real social media text. Before training, each tweet goes through several cleaning steps: removal of URLs, special characters and punctuation, conversion to lowercase, tokenisation, and stemming or lemmatisation to reduce different word forms to a common base [3]. These steps help reduce noise and the number of unique tokens, which in turn makes modelling easier. For the traditional

and advanced machine learning models, Term Frequency–Inverse Document Frequency (TF-IDF) is used to represent tweets as numerical feature vectors. TF-IDF emphasises words that are frequent in a tweet but not common across the whole dataset, which often improves classification compared to simple counts. Class imbalance between positive and negative tweets is handled using a mix of oversampling the minority class and adjusting class weights during training so that misclassifying minority examples is penalised more [2].

4.2 MODEL DEVELOPMENT

Logistic Regression and Naive Bayes are implemented as baseline classifiers using TF-IDF features. They provide a fast and interpretable starting point for comparison. SVM and Random Forest are then added to test whether more flexible decision boundaries and ensemble-of-trees structures can improve performance on the same data. Key hyperparameters for these models such as the SVM kernel type, the number of trees in the Random Forest and maximum tree depth are tuned using validation results rather than fixed by default values. The LSTM model is built to capture sequential and contextual patterns in tweet text. Tweets are mapped to word embeddings, which are then fed into the LSTM layers. The network is trained for a fixed number of epochs with early stopping based on validation loss to reduce overfitting. Finally, a stacked ensemble model is created by combining the outputs of Logistic Regression, Random Forest and LSTM. The ensemble uses a soft-voting mechanism where each model’s prediction is weighted by its validation accuracy, so that stronger models have more influence on the final decision.

4.3 TRAINING AND EVALUATION

The dataset is split into training (80%), validation (10%) and test (10%) subsets. The training set is used to fit the models and tune parameters, the validation set guides hyperparameter selection and early stopping, and the test set provides an unbiased estimate of final performance. For each model, accuracy, precision, recall and F1-score are calculated on the test data. Confusion matrices are also examined to see how often each model confuses positive and negative tweets. To check that results are not due to a single split, k-fold cross-validation is applied, and average performance across folds is compared with the main test-set scores.

4.4 IMPLEMENTATION

All models are implemented in Python. Traditional and advanced machine learning models (Logistic Regression, Naive Bayes, SVM, Random Forest) are built using the scikit-learn library, while the LSTM network is implemented in TensorFlow. Training is carried out on a multi-core machine equipped with a GPU to handle the higher computational demands of the deep learning model. This setup reflects a realistic environment for organisations that have access to standard data science infrastructure.

4.5 ETHICAL CONSIDERATIONS

The study uses the publicly available Sentiment140 dataset, which is pre-anonymised and does not include personally identifiable information. No attempt is made to link tweets back

to individual users. Potential demographic and topic biases within the dataset are recognised, and class-balancing techniques are applied as a partial mitigation. As the focus is on aggregate model performance for brand reputation monitoring, not on individual behaviour, the ethical risk is limited but still acknowledged.

5. RESULTS

This section presents the performance of all six models on the Sentiment140 test set. The Table.1 summarises test-set accuracy and the main strengths and weaknesses of each approach, and Figure 2 shows a visual comparison of accuracies.

Table.1. Comparative performance of sentiment classification models on the Sentiment140 test set

Model	Accuracy (%)	Strengths	Weaknesses
Logistic Regression	77.23	Simple, efficient	Struggles with sarcasm and context
Naive Bayes	75.50	Computationally efficient	Misclassifies context-sensitive text
SVM	72.74	Handles high-dimensional data	Computationally expensive on large data
Random Forest	77.07	Handles class imbalance well	Slower on large datasets
LSTM	79.20	Captures sequential context	Prone to overfitting, resource-intensive
Ensemble	78.69	Balanced performance	Only marginal gain over best single model

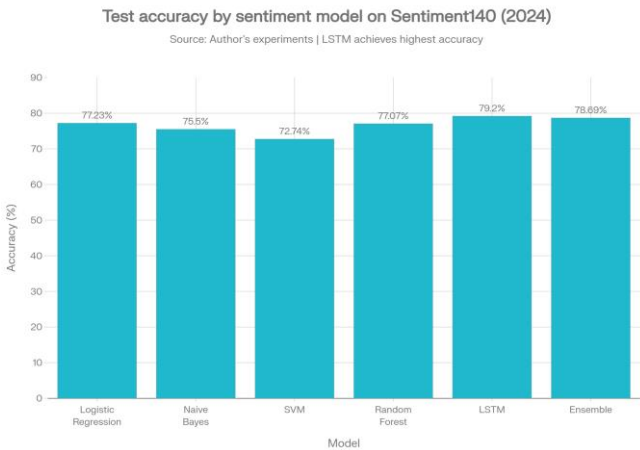


Fig.2. Test accuracy comparison across all six models.

Logistic Regression and Naive Bayes behaved as expected for linear, word-frequency-based classifiers. Logistic Regression reached 77.23% accuracy and Naive Bayes 75.50%. Both models were efficient to train and run, but their reliance on TF-IDF features meant they did not handle context or sarcasm particularly well. Random Forest achieved 77.07% accuracy, effectively

matching Logistic Regression while dealing better with class imbalance due to its ensemble-of-trees structure. SVM, evaluated on a reduced subset for computational reasons, achieved 72.74% accuracy and showed sensitivity to dataset size and parameter choices. The LSTM model produced the highest individual accuracy at 79.20%, reflecting its ability to capture sequential and contextual patterns in tweet text. Its performance was, however, affected by architecture decisions: deeper or more complex configurations tended to overfit, which required careful tuning and early stopping. The stacked ensemble, which combined Logistic Regression, Random Forest and LSTM with soft voting, achieved 78.69% accuracy. This represents a modest gain over the traditional and advanced models and slightly less than the best single model (LSTM), but with improved stability across classes.

Error analysis across all six models highlighted three recurring issues. First, sarcastic tweets such as “Oh great, another Monday...” were often misclassified as positive because of the literal polarity of words like “great”. Second, context-dependent phrases such as “It’s not bad” were sometimes misinterpreted when models focused on individual word sentiment rather than the full expression. Third, tweets containing slang or idiomatic language caused confusion for all models. Even the LSTM, despite its contextual capacity, struggled when sarcasm depended on information outside the tweet itself.

6. DISCUSSION

The results align with much of the existing work in sentiment analysis. Traditional models such as Logistic Regression and Naive Bayes offer solid baseline performance at low computational cost but remain limited in their ability to deal with non-literal and context-dependent language. Advanced models like Random Forest add robustness and slightly better handling of class imbalance, while SVM can perform well on carefully prepared subsets but may be expensive to scale. LSTM confirmed its expected advantages, delivering the highest accuracy by modelling word order and context rather than treating tweets as unordered bags of words. At the same time, training and tuning the LSTM required more time and resources, and the model was still not able to fully resolve sarcasm-related errors. This supports a growing view in literature that sequence modelling alone is not enough; further improvements likely depend on richer contextual information or architectures such as attention and transformer-based models.

From a practical brand-management perspective, these findings suggest a clear decision framework. In scenarios where organisations need to monitor very large volumes of social media data quickly and with limited computing resources for example, real-time dashboards tracking sentiment around product launches Logistic Regression or Random Forest can be sensible choices. They provide reasonably high accuracy with simple deployment and low computational cost. In settings where accuracy is more critical than speed, and where GPU or high-performance infrastructure is available, LSTM stands out as the most suitable option among the tested models. It offers the best individual performance and captures more nuanced patterns in text, despite its ongoing weaknesses around sarcasm and ambiguous language.

The ensemble model offers a middle path. By combining Logistic Regression, Random Forest and LSTM, it trades a small

amount of peak accuracy for increased robustness and more balanced performance across classes. For applications where consistent reliability matters more than maximising a single metric for instance, long-term monitoring of brand sentiment trends this balance may be attractive.

7. CONCLUSION

This study compared six sentiment classification models—Logistic Regression, Naive Bayes, SVM, Random Forest, LSTM and a stacked ensemble on the Sentiment140 dataset, with a specific focus on brand reputation monitoring on social media. Under matched data and preprocessing conditions, LSTM achieved the highest individual accuracy (79.20%), followed closely by the ensemble model (78.69%), while traditional models remained competitive around 75–77% at much lower computational cost. Across all models, the main sources of error were sarcastic tweets, context-dependent expressions and ambiguous or slang-heavy language. These errors appeared even in the best-performing LSTM model, suggesting that current approaches still have difficulty with sentiment that relies on external context or subtle shifts in meaning.

For practitioners responsible for brand reputation, the results provide an evidence-based guide to model choice. Linear models and Random Forest are suitable when speed, simplicity and low resource usage are the main priorities, while LSTM is preferable when organisations are prepared to invest in higher computing power to gain extra accuracy. The ensemble strategy can be used when a balance between robustness and performance is required.

8. LIMITATIONS

This work is limited to a single dataset (Sentiment140) and to binary sentiment classification, and it does not include transformer-based architectures such as BERT, which have recently shown strong results for contextual sentiment understanding. Computational constraints also restricted the range of LSTM and ensemble configurations tested.

9. FUTURE WORK

Future research should extend this comparison to transformer models and multi-class sentiment tasks and should evaluate performance across different social media platforms and languages. Such studies would help clarify how well the trade-offs observed here generalise beyond Twitter and binary sentiment and would support more refined model selection for brand reputation management in practice.

REFERENCES

- [1] S.M. Al-Selwi, M.F. Hassan, S.J. Abdulkadir and M.G. Ragab, “RNN-LSTM: From Applications to Modeling Techniques and Beyond-Systematic Review”, *Journal of King Saud University - Computer and Information Sciences*, Vol. 36, No. 5, pp. 102068-102088, 2024.
- [2] I.M. Alkhawaldeh, I. Albalkhi and A.J. Naswhan, “Challenges and Limitations of Synthetic Minority

- Oversampling Techniques in Machine Learning”, *World Journal of Methodology*, Vol. 13, No. 5, pp. 373-378, 2023.
- [3] V. Balakrishnan and L.Y. Ethel, “Stemming and Lemmatization: A Comparison of Retrieval Performances”, *Lecture Notes on Software Engineering*, Vol. 2, No. 3, pp. 262-267, 2014.
- [4] C. Bell, A. Olukemi and P. Brooklyn, “Social Media Sentiment Analysis for Brand Reputation Management”, Available at: <https://www.preprints.org/manuscript/202408.0029>, Accessed in 2024.
- [5] L. Breiman, “Bagging Predictors”, *Machine Learning*, Vol. 24, No. 2, pp. 123-140, 1996.
- [6] L. Breiman, “Random Forests”, *Machine Learning*, Vol. 45, No. 1, pp. 5-32, 2001.
- [7] E. Cambria, “Affective Computing and Sentiment Analysis”, *IEEE Intelligent Systems*, Vol. 31, No. 2, pp. 102-107, 2016.
- [8] C. Cortes and V. Vapnik, “Support-Vector Networks”, *Machine Learning*, Vol. 20, No. 3, pp. 273-297, 1995.
- [9] T.G. Dietterich, “Ensemble Methods in Machine Learning”, *Lecture Notes in Computer Science*, Vol. 1857, pp. 1-15, 2000.
- [10] A. Go, R. Bhayani and L. Huang, “Twitter Sentiment Classification using Distant Supervision”, Master’s Thesis, Department of Computer Science, Stanford University, pp. 1-89, 2009.
- [11] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory”, *Neural Computation*, Vol. 9, No. 8, pp. 1735-1780, 1997.
- [12] T. Joachims, “Text Categorization with Support Vector Machines: Learning with many Relevant Features”, *Proceedings of International Conference on Machine Learning*, pp. 137-142, 1998.
- [13] D. Jurafsky and J.H. Martin, “*Speech and Language Processing*”, 3rd Edition, Stanford University Press, 2018.
- [14] J. Kittler, M. Hatef and J. Matas, “On Combining Classifiers”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 3, pp. 226-239, 1998.
- [15] B. Liu, “*Sentiment Analysis and Opinion Mining*”, Morgan and Claypool Publishers, 2012.
- [16] Y. Liu, W. Zhang and J. Wang, “A Comparative Study of Machine Learning Methods for Text Classification”, *Journal of Data Science*, Vol. 9, No. 2, pp. 45-60, 2021.
- [17] C.D. Manning, P. Raghavan and H. Schütze, “*Introduction to Information Retrieval*”, Cambridge University Press, 2008.
- [18] D. Maynard and M.A. Greenwood, “Who Cares about Sarcastic Tweets? Investigating the Impact of Sarcasm on Sentiment Analysis”, *Proceedings of International Conference on Language Resources*, pp. 1-7, 2014.
- [19] T. Mullen and N. Collier, “Sentiment Analysis using Support Vector Machines with Diverse Information Sources”, *Proceedings of International Conference on Computational Linguistics*, pp. 1-5, 2004.
- [20] B. Pang and L. Lee, “Opinion Mining and Sentiment Analysis”, *Foundations and Trends in Information Retrieval*, Vol. 2, No. 1-2, pp. 1-35, 2008.
- [21] W. Yin, K. Kann, M. Yu and H. Schütze, “Comparative Study of CNN and RNN for Natural Language Processing”, *Proceedings of International Conference on Computational Linguistics*, pp. 1-5, 2017.
- [22] X. Zhou, X. Wan and J. Xiao, “Attention-Based LSTM Network for Cross-Lingual Sentiment Classification”, *Proceedings of International Conference on Natural Language Processing*, pp. 247-256, 2016.