Journal of Computing and Communication Vol.4 , No.2 , PP. 100-112 , 2025

Analyzing Public opinion on Climate Change via Twitter: A Machine Learning Approach Using Historical Data

Mahmoud Mahdi [a], Ahmad Salah[b], Mohamed Omar[a]

[a]Department of Computer Science, Faculty of Computers and Informatics, Zagazig University, Zagazig, Egypt

[b]College of Computing and Information Sciences, University of Technology and Applied Science, Ibri, Oman

*Corresponding Author: Ahmad Salah [***ahmad.salah@utas.edu.om***]

ARTICLE DATA

Article history:
Received 26 April 2025
Revised 17 May 2025
Accepted 18 May 2025
Available online 30 July 2025

Keywords:
Climate Change
Tweets
Machine Learning
Classification
Sentiment Analysis

ABSTRACT

Examining public perspective regarding critical issues such as climate change in the context of a vast social media stream necessitates a computational approach. This study provides an initial benchmark of five classical machine learning classifiers (Multinomial Naive Bayes, Logistic Regression, Linear Support Vector classifier (SVM), Random Forest, Gradient Boosting) on multi-class categorization (Anti, Neutral, Pro, News) using the openly available Twitter Climate Change Sentiment dataset (2015-2018). Models were built using Term Frequency-Inverse Document Frequency (TF-IDF) for feature representation and performance was assessed using standard metrics (Accuracy, F1-score), as well as modeling generalizability by comparing training versus testing performance to identify overfitting. Experimental results showed that Linear SVC achieved the highest test F1-score (~70.7) but exhibited significant overfitting (≈28%), while Logistic Regression provided the best compromise producing a competitive F1-score (~67.8) and a notably higher degree of generalizability (≈11.5% drop in F1). Gradient Boosting showed remarkable robustness with minimal overfitting (≈1.4% drop in F1) but had less absolute performance (~58.3% F1). This study provides an important baseline for this classification task and highlights the importance of generalization in evaluating model performance in addition to predictions for reliable stance analysis, particularly in computational social science.

1. Introduction

Climate change represents one of the most fundamental and multifaceted challenges humanity faces today in the 21st century - a challenge that will require global action powered by science, understanding by policymakers, and engagement of people from all walks of life [1]. Online social media, particularly microblogging services such as Twitter (recently referred to as X), have emerged as powerful channels of public discourse pertaining to climate change, shaping public opinion, disseminating information (and misinformation), and galvanizing action [2, 3, 4]. The volume and velocity of user-generated content flowing through these platforms provides a significant opportunity, while also presenting challenges to understanding the dynamics relating to public sentiment, opinions, and narratives surrounding this important issue [5]. Therefore, the development and use of robust computational methods based mainly in Natural Language Processing (NLP) and Machine Learning (ML), to automate the analysis of public discourse at scale, is important [5].

The recent innovations in NLP have shaped our ability to capture nuanced semantic meaning, sentiment polarity, and argumentative opinion in short, informal texts that dominate social media platforms, largely, but not solely, featured in larger pre-trained transformer models and Large Language Models (LLMs) [7]. At the same time, user classification on a multilayered topic such as climate change is not simply a set of positive-negative sentiment classes; a consideration of the user relationship with the scientific consensus (pro-belief, anti-belief, or neutral to climate science, as an example), or type of post (factual news reporting) is necessary [8]. Accurately detecting opinions in a multi-class context remains an ongoing area of study, especially with multiple sociolinguistic phenomena, such as sarcasm, evolving language, and echo chambers [9,10].

Curated datasets are essential for advancing research in this area. In this study, we employ the publicly available "Twitter Climate Change Sentiment Dataset" [11], which consists of 43,943 tweets collected between April 27, 2015, and February 21, 2018; each tweet examined in this dataset was related to climate change and each tweet was coded with high reliability as determined by consensus of three independent reviewers. Tweets are categorized into four levels of support, labeling tweets as: News (factual news post), Pro (support of belief in manmade climate change), Neutral (neither supporting nor denying), or Anti (do not believe in manmade climate change). This dataset acts as a valuable historical snapshot of public discourse in a pivotal time, and a benchmark for building and evaluating computational models to help researchers know what or how the public does or does not believe around climate change [12, 13, 14,15].

This work will utilize this dataset to assess the state-of-the-art machine learning models performance for climate opinion classification as the problem will be classified as a classification problem. This research utilizes the Twitter Climate Change Sentiment Dataset to benchmark several traditional classifier models, including Naive Bayes [16], Logistic Regression [17], SVM [18], Random Forest [19], and Gradient Boosting [20]. In order to conduct feature extraction, we utilize TF-IDF vectorization and evaluate performance using a well-defined evaluation framework, where all classifiers have the same standard of evaluation. By implementing this procedural framework for evaluation of effectiveness of traditional classifiers, we intend to illustrate the performance of various classifiers in sentiment analysis, while illustrating strengths and limitations. We also consider challenges presented in sentiment analysis tasks, which include overfitting, class imbalance, and feature sparsity.

This paper adds to the growing climate change communication literature by:

1. We report on the performance of classical machine learning models for sentiment classification tasks on the Twitter Climate Change Sentiment Dataset.

2. The results may be serving as a baseline for work to be compared, such as pre-trained models, transformers, and hybrid models.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 exposes the utilized methodology. Section 4 lists the results and discusses the findings of this work. Section 5 concludes the paper.

2. Related Work

Techniques for sentiment analysis are extending beyond conventional product evaluations or social media, to explore more nuanced dimensions of perspectives within particular contexts, including within environmental policy. For example, in [21], the authors undertook a sentiment analysis on the word "change" within qualitative interviews of local government officials in Canada discussing climate change action plans and governance. The authors manually coded the data since they prioritized an accurate analysis within the context of the qualitative data rather than rely on computational methods (even on their qualitative data). They discovered "overwhelmingly positive emotional" as relating to institutional and/or strategic change, related to many factors such as leadership and planning features, and overwhelmingly negative emotional with respect to the slow or nonexistent pace of change, behavioral hurdles, and communicative framing issues etc. In addition to providing a working example of sentiment analysis being applied to gain insights from domain-specific qualitative data/attitudes, this also provides an example where the authors assert that sentiment analysis is important and be aware of the context (which is often a challenge for fully automated NLP systems with specialized corpora and nuanced concepts). Thus, it is a working use case of sentiment analysis as a tool for gaining insight from policy-related research to uncover drivers and barriers closely associated with specific keywords.

Elaborating upon applications of sentiment analysis in the field of climate change, the authors in [22] examined perceptions held within Malaysia by carrying out an analysis of editorial articles from The Sun Daily newspaper. The authors constructed a domain-specific corpus called the Malaysian Daily Climate Change Corpus (MyDCCC) as part of a mixed-method approach which included utilizing Azure Machine Learning that was completed as a preliminary first step of polarization by sentiment. A corpus-driven approach using AntConc was then used to identify salient terms of sentiment lexicon. The words were cross-referenced against the MPQA Subjectivity Lexicon to confirm significant findings. The authors indicated that there was a considerable prevalence of negative sentiment (90%) established based on sentiments related to climate change, identified via the editorials. The authors identified key negative salient terms that included long, critical, and serious, whereas the positive terms were not the most prevalent and included better, best, and hope. In [22], the author's discourse analysis demonstrated that this situated negative sentiment was fundamentally rooted in public frustration with governmental responses rather than denying climate change, thus the author suggested caution in interpreting sentiment beyond a binary value and in isolation of a context that is greater than simple sentiment scores. This study combines automated sentiment analytic resources with the analysis of corpus linguistics and discourse with the goal of identifying particular thoughts of public importance in a particular source of media.

In [23], the authors undertook a comparative analysis of lexicon-based, machine learning and hybrid methods of sentiment analysis of climate change Twitter data, utilizing seven popular sentiment lexicons (e.g., VADER, TextBlob, MPQA) and three machine learning classifiers (Logistic Regression, SVM, Naïve Bayes) in combination with Bag-of-Words and TF-IDF feature extractions. They observed that hybrid methods performed significantly outperformed the lexical and machine learning methods independently. The best performing approach combined the TextBlob lexicon with Logistic Regression classifier using TF-IDF features to provide the highest F1-score. They also noted that lemmatization was important because it generally provided better performance for machine learning and hybrid methods.

The authors in [24] investigated public sentiment on Twitter concerning climate change issues related to specific Sustainable Development Goals (SDGs). They applied and compared several NLP techniques, including the rule-based methods VADER and TextBlob, and a transfer-learning approach using BERT embeddings with a logistic regression classifier. Their results demonstrated that the BERT-based model achieved superior performance (69% accuracy) compared to the lexicon-based approaches for classifying sentiment in the collected climate-related tweets. While the overall sentiment detected was positive, the study highlighted challenges with data noise due to broad keyword selection and suggested domain-specific models like BERTweet could offer further improvements.

In [25], the authors explored climate change and energy discourse on Twitter in the UK and Spain using NLP methods in early-2019. Using the NRC Emotion Lexicon (EmoLex), they analyzed the data for positive/negative sentiment as well as the presence of eight discrete emotions (i.e., fear, trust, anticipation) within the tweets. The authors found the UK discourse to be less negative and to include more anticipation than the Spanish discourse, which was less negative overall and more dominated by fear. The authors used sentiment towards specific energy generators (e.g., positive for renewables, negative for coal) in a correlation with energy supply and public acceptance of energy sources — demonstrating the timeliness of discourse/sentiment analysis using social media for considering emotion and sentiment, complementing traditional surveys.

3. Proposed Methodology

This section explains the methodology employed in this study for the sentiment classification of climate change-related tweets derived from the Twitter Climate Change Sentiment Dataset. Figure 1 illustrates the complete workflow of the proposed methodology for processing text data, feature representation, and model evaluation. The workflow begins with the Raw Dataset, which goes through Data Pre-processing such as text cleaning, tokenization, stopword removal, and lemmatization or stemming. The raw text is converted to numerical features in the Feature Extraction phase, which consists of a feature vectorization step and a formalization of the vectorized into a Term Frequency-Inverse Document Frequency (TF-IDF) generated feature matrix. At this point, the pre-processed data is split between training and testing in the Train-Test Split phase, allowing model training on one subprocess, and model evaluation on another. The Model Training step refers to the machine learning model fit to the processed training data. Then the Model Evaluation step refers to applying various performance metrics such as accuracy, precision, recall, and F1-score to confirm the capability/modeling of the classifiers summarized as the final output, performance metrics. The workflow depicted in this pipeline is structured, systematic, and replicable for the evaluation of text data and classifier models.

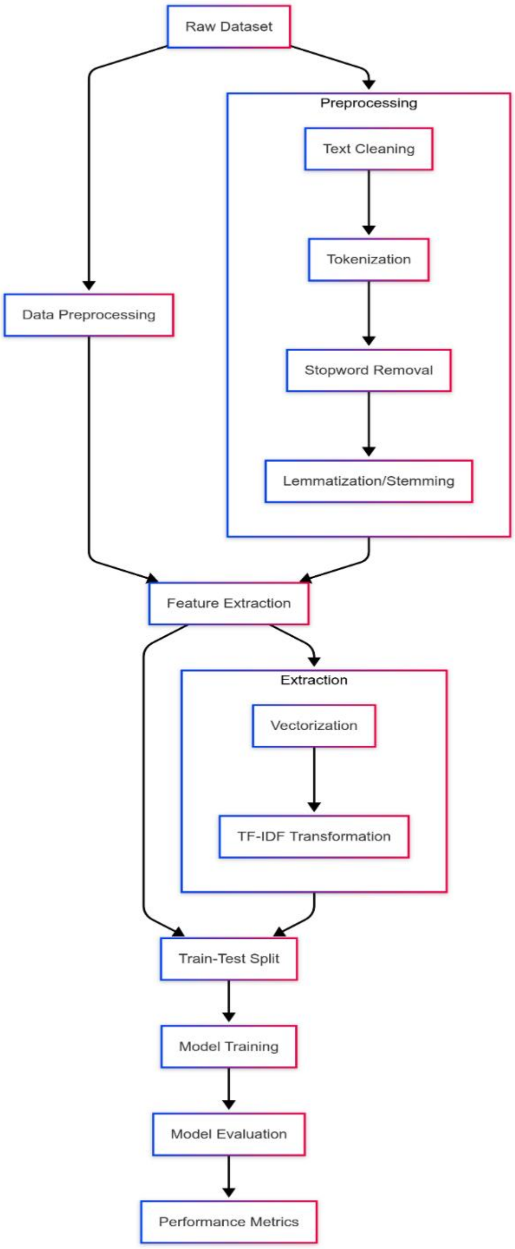

FIGURE 1. Workflow for Machine Learning-Based Text Classification.

3.1 The utilized dataset

The Twitter Climate Change Sentiment Dataset is a dataset of 43,943 annotated tweets that capture public sentiment about climate change. Tweets were collected between April 27, 2015 and February 21, 2018, and were assigned sentiment categories based on the consensus of three independent annotators. All tweets that received sentiment agreement from all three analysts were retained to ensure a high-quality dataset. The dataset organizes tweets into four sentiment classes: (2) News, for tweets linking to factual news articles; (1) Pro, for tweets representing support for the belief that climate change is man-made; (0) Neutral, for tweets that express no explicit sentiment about climate change; and (-1) Anti, for tweets denying that climate change was man-made. The class distribution is imbalanced, with fewer tweets in the Neutral and Anti classes— models trained on imbalanced datasets have the potential to incorporate that imbalance into its predictions when later evaluating or testing the models. Thus, we must deal with this class imbalance level fairly during the model training.

3.2. Data Preprocessing

To prepare the data for machine learning, a few preprocessing steps were performed. This includes missing value treatment, duplicate removal, and mapping sentiment labels to more descriptive names. To start, rows containing missing text values in the message column were deleted due to the detrimental impact of incomplete text data on the training of a model. Furthermore, duplicate tweets were removed from the dataset so it would not bias the results based solely on duplicate tweet counts. After this cleaning step, sentiment labels, which were originally denoted by numerical values (-1, 0, 1, 2), were mapped to more descriptive names (Anti, Neutral, Pro, and News) to further assist in data interpretation, analysis, and better understanding of the data and sentiments expressed within the data itself.

By this stage, the processed dataset was ready for a training and testing split, which was an 80-20 split. The training set was then used to train the machine learning models and the testing set was then used to test generalization capability of the models. In the interest of reproducibility, a random seed was assigned at the point of the training and testing split procedure. Figure 2 lists the main steps of the data proprocessing.

3.3. Feature extraction

The TF-IDF (Term Frequency-Inverse Document Frequency) approach was employed to convert textual data into a numerical representation that is amenable to machine learning. The TF-IDF approach is appropriate for representing text data because it models the importance of terms in a document relative to the corpus. With the TF-IDF approach, distinct terms, or terms that appear with low frequency across the corpus, are emphasized while common terms are down-rated. Distinct terms are generally more suited for the models to discover meaningful patterns in the text.

The TF-IDF Vectorizer was set to ignore, or remove, common English "stop words," which do not provide context to sentiment like "the" and "and." In addition, any term, or word, that appeared in more than 70% of tweets would also be removed ($max\_df = 0.7$). Common words that are so frequent, or prevalent, among all tweets aren't particularly helpful for distinguishing the classes from one another. Overall, the output of the TF-IDF matrix would be a sparse representation of the original text data and focused on the importance of terms (or words) found in each tweet.

TF-IDF is a statistical measure used to evaluate the importance of a term in a document relative to a collection of documents (corpus). It is calculated as the product of two components: TF and IDF. The formula for TF-IDF is expressed in Equation 1 as follows:

$$TF - IDF(t,d,D) = TF(t,d) \times IDF(t,D) \qquad (1)$$

where TF is calculated as $TF(t,d) = f(t,d) \,/\, \Sigma\, f(t',d)$, f(t,d) is the frequency of term t in document d, Σf(t′,d) is the total number of terms in document d and t' is any term in the document d. Also, IDF is calculated as $IDF(t,D) = log(|D| \,/\, (1 + |\{d \in D : t \in d\}|))$ where |D| represents the total number of documents in the corpus and | $d \in D : t \in d$ | is the number of documents containing term t.

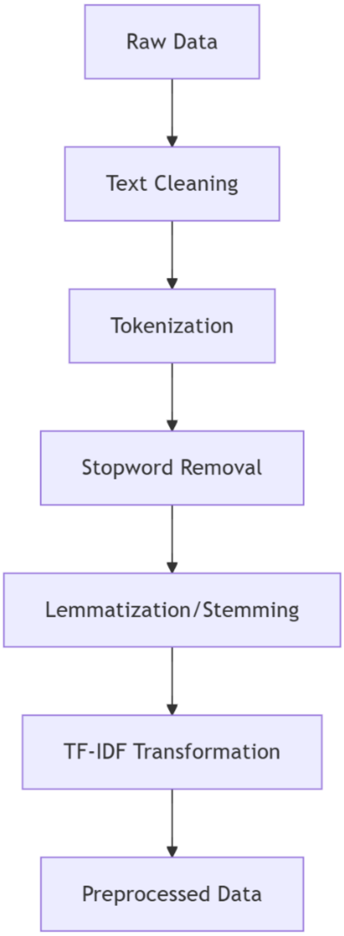

FIGURE 2. Pipeline for the data preprocessing.

3.4. Machine Learning Models

In this study, the sentiment of tweets was classified through five machine learning models namely Multinomial Naive Bayes, Logistic Regression, Linear Support Vector Machines (SVM), Random Forest, and Gradient Boosting Classifier. Collectively, these models span traditional supervised learning options and offer unique advantages to be deployed within the context of text classification.

1. The Multinomial Naive Bayes model assumes features are independent, resulting in a particularly valuable model for high-dimensional, sparse datasets as those generated through TF-IDF. The model is also computationally efficient and frequently has good predictive performance for text classification use cases.

2. As a linear model, Logistic Regression estimates class probabilities. The model is resistant to class imbalance and outputs are interpretable, providing a viable classification model for binary or multi-class problems.

3. In attempting to maximize the margin between classes, Linear Support Vector Machines (SVM) can be deployed as the model is known to perform well on high dimensional feature spaces, as those produced from TF-IDF. The SVM models are also known for their robustness and high predictive performance in text classification.

4. Random Forest is an ensemble model which generates multiple decision trees either through sampling dimensions or utilizing randomized decision-thresholds and then aggregates the predictions. Though Random Forest is often useful in describing non-linear relationships in data, its use for high-dimensional datasets may leave much to be desired.

5. Finally, Gradient Boosting Classifier is also an ensemble method. Like Random Forest, Gradient Boosting builds sequential decision-trees, to optimize across misclassified samples at each tree-generated stage. The model is computationally expensive; however, it is known to be an accurate predictive model.

3.5. Evaluation Metrics

The training and evaluation of the machine learning models was executed in a way that purposefully compared the models' performance on the sentiment classification. Each model was trained on the TF-IDF-transformed training data which led to a set of predictions for the test set to understand how well each model generalized. We evaluated each model by computing key performance metrics, including accuracy (Equation 2), precision (Equation 3), recall (Equation 4), and F1-score (Equation 5), all weighted to address the class imbalance in the data. The accuracy, precision, recall and F1-scores provided meaningful means for capturing the capability of each model to correctly classify tweets across the four sentiment classes (Anti, Neutral, Pro, and News). Additionally, confusion matrices were generated to examine common misclassifications for each model, providing information about class specific classification issues. We also compared training and testing F1 scores and flagged potential overfitting if there was a difference in the two means of at least 0.10. Each model's evaluation results were reported and visualized for the evaluation comparison, systematically achieving the goal of comparing the strengths and weaknesses of the various models. This comprehensive evaluation process assisted in the thorough examination of each models' performance, thereby providing useful information for their potential use to code and classify sentiment in tweets about climate change prior discourse.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \qquad (2)$$

where TP (True Positives) is the number of instances where the model correctly predicts the positive class, TN (True Negatives) is the number of instances where the model correctly predicts the negative class, FP (False Positives) is the number of instances where the model incorrectly predicts the positive class for a negative instance, and FN (False Negatives) is the number of instances where the model incorrectly predicts the negative class for a positive instance.

$$Precision = \frac{(TP)}{(TP + FP)} \qquad (3)$$

$$Precision = \frac{(TP)}{(TP + FN)} \qquad (4)$$

$$F1 - score = \frac{(Precision \cdot Recall)}{(Precision + Recall)} \qquad (5)$$

4. Experimental Results

4.1. Setup

The proposed work was written in the Python programming script. The Twitter sentiment dataset[1] was introduced and preprocessed utilizing pandas, including missing value removal and duplicate removal, alongside sentiment label mapping. Feature extraction was performed using Term Frequency-Inverse Document Frequency (TF-IDF), implemented via the TfidfVectorizer from sklearn.feature_extraction.text, configured to remove English stop words along with terms occurring in greater than 70% of tweets, after dividing the data into an 80% training and 20% testing set formed via sklearn.model_selection. Five different classifiers included as part of sklearn (MultinomialNB, LogisticRegression, LinearSVC, RandomForestClassifier, GradientBoostingClassifier) were trained and assessed via classification statistics (accuracy, precision, recall, F1-score) from sklearn.metrics on the TF-IDF features, including visualization in the form of a confusion matrix via matplotlib and seaborn.

4.2. Results

The performance measures for the five evaluated machine learning models using both the training and testing datasets are presented in tabular form in Table 1, and a graphical representation through confusion matrix in Figures 3-7 for convenience and summarizing purposes. These activities provide a measure of not only how well each model learned patterns in the training data but more importantly, how well it generalized to the unseen testing data, providing a measure of robustness. A meaningful difference in performance between training and testing performance, including the measure of F1-score (defined here as 0.10), can indicate overfitting, when the model has forgotten learning to simply state characteristics of the training dataset.

The training results for both Random Forest (99.99% F1-score) and Linear SVC (98.60% F1-score) produce superior performance levels indicating that they both learned patterns present within the training set. However, their measure of generalization, based upon performance on unseen test data, gives a very different indication of performance learning, which is evident in the confusion matrix for the respective models (Figure 4: Random Forest, Figure 5: Linear SVC). For Random Forest, the measure of F1-score dropped sharply to 65.27% testing data (almost a 34.7% difference), which can be considered as a classical and severe example of overfitting. For Linear SVC, it achieved the highest test F1-score of testing among the other models, which also measured an overfitting response, as the F1-score dropped almost 27.9% from training performance, as seen in Table 1; performance, and compare with knowledge of misclassification by testing in Figure 5: Linear SVC. Thus, while Linear SVC had the highest score for this actuality of the test dataset, the model demonstrates some reliability, albeit yielded no guarantees in generalization training data.

[1] https://www.kaggle.com/datasets/edqian/twitter-climate-change-sentiment-dataset/data

TABLE 1: Comparison of model performance metrics on training and test datasets.

Model	Accuracy		Precision		Recall		F1-score	
	Train	Test	Train	Test	Train	Test	Train	Test
Naive Bayes	64.75%	57.48%	78.25%	73.23%	64.75%	57.48%	58.07%	47.46%
Logistic Regression	80.41%	69.89%	81.84%	70.00%	80.41%	69.89%	79.32%	67.82%
Linear SVC	98.61%	71.61%	98.61%	70.77%	98.61%	71.61%	98.60%	70.67%
Random Forest	99.99%	67.41%	99.99%	68.20%	99.99%	67.41%	99.99%	65.27%
Gradient Boosting	**63.54%**	**61.73%**	**67.03%**	**63.79%**	**63.54%**	**61.73%**	**59.72%**	**58.31%**
Baseline [26]	-	-	-	83.00%	-	48.00%	-	50.00%
Logistic Regression [26]	-	-	-	67.00%	-	74.00%	-	69.00%

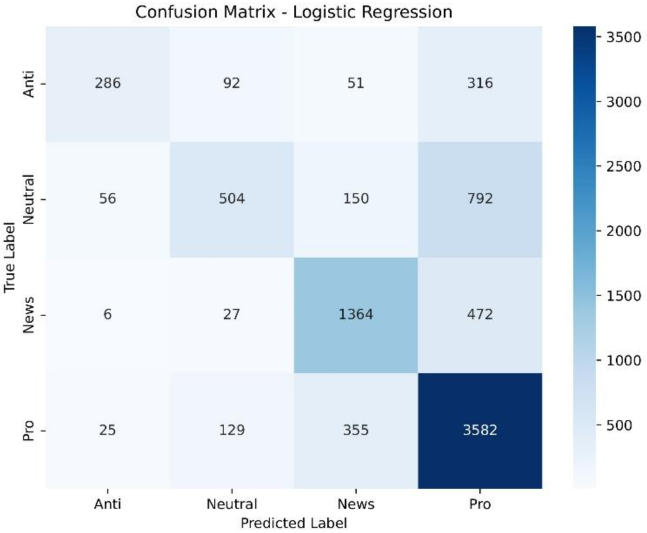

FIGURE 3. CM for the logistic regression model.

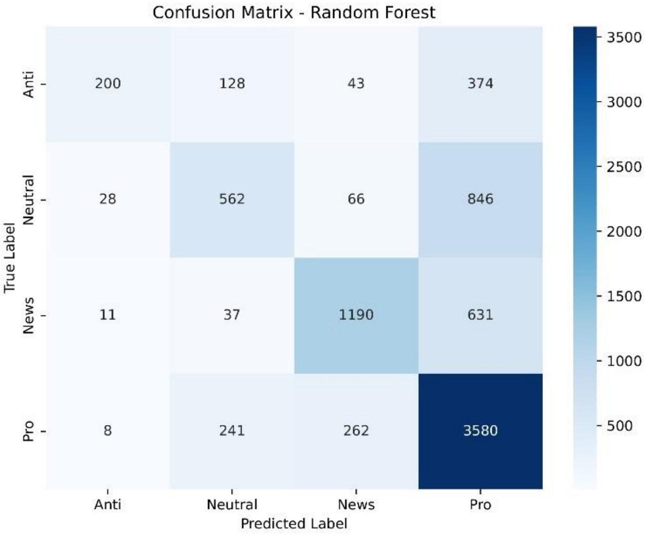

FIGURE 4. CM for the random forest model.

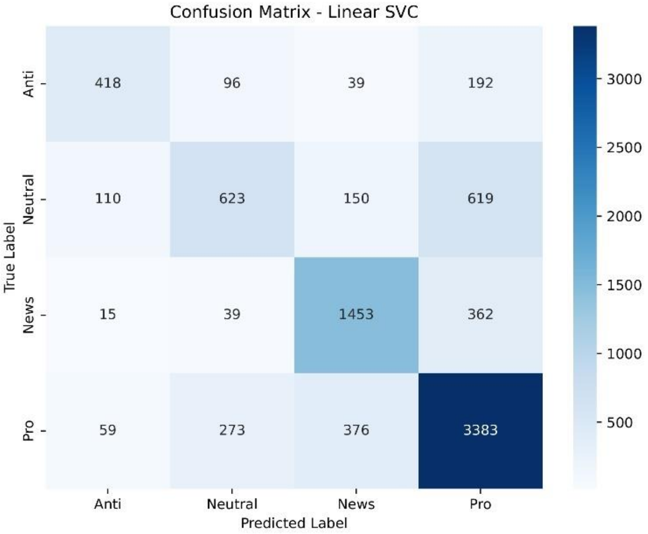

FIGURE 5. CM for the linear SVC model.

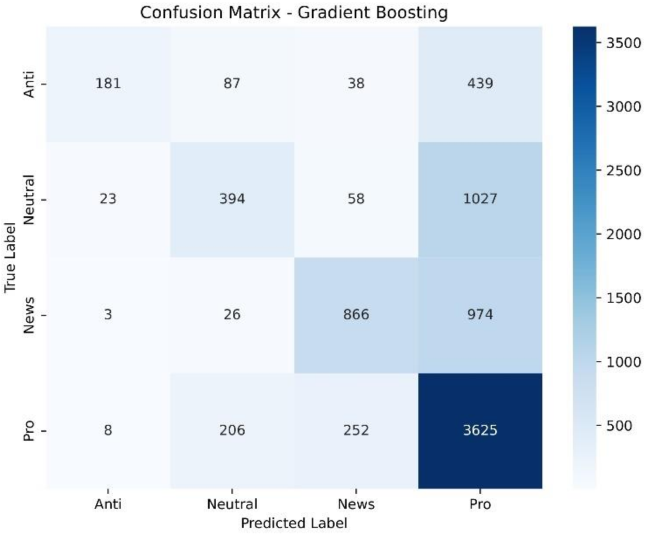

FIGURE 6. CM for the Gradient Boosting model.

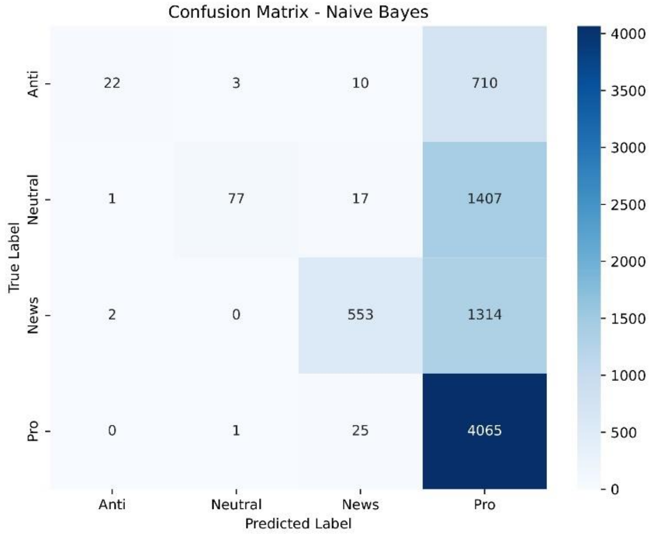

FIGURE 7. CM for the Naïve Bayes model.

4.3. Discussion

In contrast to the previously mentioned higher-performing models, both Naive Bayes and Gradient Boosting exhibited lower performance levels in their respective model evaluations from the test set. As indicated in Table 1 and observed in the confusion matrices presented in Figures 6 and 7, Naive Bayes had a test accuracy of 57.48% and an F1-score of 47.46%. The previous confusion checking discussion behaviorally does not have direct relevance to the study, but it is worth noting that the somewhat lower performance may be due to the core assumption of feature independence in class conditional probabilities that is often violated in more complex but real text data. Regarding Gradient Boosting model performance, a test accuracy of 61.73% and F1-score of 58.31% were found. While these values were lower compared to Linear SVC and the Logistic Regression model, Gradient Boosting provides more than just the absolute number presented in training performance; the model demonstrated excellent generalization, with training F1-score of 59.72% and testing F1-score of approximately 58.31%, indicating only a 1.4% performance score drop from training to testing data. In other words, the Gradient Boosting model performed best among all the models in relation to avoiding overfitting training data, and to provide some contextual, while the absolute predictive performance is limited on these performance figures and visualizations, as demonstrated in Figure 6. The Absolute testing performance metric scores for the Gradient Boosting and Naive Bayes models represented with Figures 6 and 7 both indicate that models applied with these configurations were performing "lower," indicating that both models could have benefited from hyperparameter tuning or that, simply, the models may function alternatively weaker on the complexity of the data represented.

Bringing the data evidence together in one condensation that considers the differences in the varied evaluation criteria is helpful in the context of future and current modelling. Strictly on understanding from absolute test performance, Linear SVC (Figure 5) was the highest performing classifying model. If the measure of evaluation is more of an indicator of generalization (avoiding overfitting training data), Linear SVC did poorly corresponding to a model published with equal paragraph efficiency or seeking generalization and dropped even significantly lower than the model in the logistic regression with substantial overfitting (See Figure 4). In practice, in this visualization construct that specifies combined reliance on absolute test performance and generalization parameters of Logistic Regression (Figure 3) was the second-best performing modelling choice that included dependency on each test measure ultimately providing an acceptable balance between two evaluation criteria overall. The Gradient Boosting model (Figure 6) demonstrates both favorable generalizations while did not have significant absolute test performance, while Naive Bayes showed little more than some worst absolute test performance while it started to show overlap of slight overfitting (although still weakest test performance). So, while any of the evaluations could be different relative to the perceived model "best," if the evaluation is solely about perceived maximum "score" on undiscovered unseen data, one could select Linear SVC (despite being advised not to consider overfitting). However, if absolute test performance is to be considered in addition to maximum model robustness, either of the earlier mentioned selections are a balance, or the even eventual superior generalizing power of data was Gradient Boosting classifier.

Linear models, like those used in the Linear SVC and the Logistic Regression model can be relatively effective with high dimensional, sparse datasets generated from TF-IDF, as evidenced in Figures 3 and 5. Although in this case, like most of the previous discussion, substantial use of overfitting at time was seen in the Random Forest model especially with the last variable into functionality of machine learning. The substantial overfitting of data seen especially from Random Forest (Figure 4) and equally concerning (if not concerning) in the Linear SVC (Figure 5) indicates how difficult it is to find the ultimate balance between model complexity and overfitting. All machine learning training frameworks need to consider regularization, or even careful hyperparameter tuning according to function. Thus, in this study either impression of previous combinations provides equal the most ideal and highest testing measures low scores for Linear SVC and Logistic Regression in combination on both measures such evaluation, while Logistic Regression demonstrated greater reliability in practice due to the checks for overfitting.

4.4. Limitations

While the proposed methodology performs favorably, it does exhibit minor weaknesses. First, as it relies on TF-IDF, it may not measure comprehensive semantic relationships or context of the text or text elements; utilizing some advanced techniques, e.g. word embeddings (BERT), may help to improve the results of model performance. Second, there was limited hyperparameter tuning, particularly for the models like Gradient Boosting, which could have improved model performance. Finally, stem, lemmatization and handling class imbalance during the preprocessing steps may have improved model performance as well. Integrating into these aspects would improve the robustness of the reports further.

5. Conclusion

The study explored the effectiveness of five established machine learning methods for categorizing public views on climate change as expressed in a historical Twitter dataset, using features extracted from the TF-IDF technique. The goal of the experiment was to produce performance baselines and assess the inherent trade–off of accuracy vs. generalizing to testing data on a multi-class classification task. There was a large variation in the obtained performance scores. The Linear SVC classifier obtained the highest F1–score (≈70.7%) on the held-out test set indicating strong overall prediction performance on similar training data distributions. However, this also came with a great deal of overfitting (≈28% drop in F1 performance from training to testing) raising questions about the model's performance for prediction on truly new/unseen data. Gradient Boosting achieved very good generalization with a limited performance drop in the held-out test data (≈1.4% drop in F1-score) but also had the much lower absolute F1-score of (≈58.3%). Logistic Regression presented as the most balanced mechanism providing the second highest F1–score performance (≈67.8%), while also presenting with much more controlled overfit (≈11.5%) over the Linear SVC model. random forest, like the SVC model, also severely overfit and naive bayes performed quite poorly across the board. The major conclusion drawn is that using the highest test score on its own as the criteria for selecting any of the models for use in practice can lead to poor practical implementation in noisy social media text. The amount of overfitting seen in the highest models shows that the readability of the models is highly dependent on generalization performance. Especially for work such as stance detection where robustness and reliability are vital elements, the commitment to models with better perceived balance sacrifices some test performance for improved generalization - as seen for the Logistic Regression model for this analysis. This work established a necessary baseline performance benchmark for classical methodologies with informed research and use for the various classical approaches for future and technique advancements in areas of deep learning technique or contextual embeddings (formatter embeddings - BERT or others). There is an opportunity for extended hyperparameter tuning to inform performance improvements and additional classes to be balanced/or equal representation levels to improve performance and reliability. Besides, the Friedman test will be applied to the performance differences to test if the reported results are significant.

References

[1] Chatterjee, Moumita, Piyush Kumar, and Dhrubasish Sarkar. "A novel framework for analyzing climate change tweets from online social media using supervised and unsupervised algorithms." In *2024 IEEE Calcutta Conference (CALCON)*, pp. 1-5. IEEE, 2024.

[2] Maynard D, Bontcheva K. Understanding climate change tweets: an open source toolkit for social media analysis. InEnviroInfo and ICT for Sustainability 2015 2015 Sep (pp. 242-250). Atlantis Press.

[3] Upadhyaya, Apoorva, Marco Fisichella, and Wolfgang Nejdl. "A multi-task model for sentiment aided stance detection of climate change tweets." In *Proceedings of the international AAAI conference on web and social media*, vol. 17, pp. 854-865. 2023.

[4] Toupin, Rémi, Florence Millerand, and Vincent Larivière. "Who tweets climate change papers? Investigating publics of research through users' descriptions." *Plos one* 17, no. 6 (2022): e0268999.

[5] Mumenthaler, Christian, O. Renaud, R. Gava, and Tobias Brosch. "The impact of local temperature volatility on attention to climate change: Evidence from Spanish tweets." *Global environmental change* 69 (2021): 102286.

[6] Fownes, Jennifer R., Chao Yu, and Drew B. Margolin. "Twitter and climate change." *Sociology Compass* 12, no. 6 (2018): e12587.

[7] Marcondes, Francisco S., Adelino Gala, Renata Magalhães, Fernando Perez de Britto, Dalila Durães, and Paulo Novais. "Case Study: LLM-Based Anxiety Climate Index." In *Natural Language Analytics with Generative Large-Language Models: A Practical Approach with Ollama and Open-Source LLMs*, pp. 53-73. Cham: Springer Nature Switzerland, 2025.

[8] Dahal, Biraj, Sathish AP Kumar, and Zhenlong Li. "Topic modeling and sentiment analysis of global climate change tweets." *Social network analysis and mining* 9 (2019): 1-20.

[9] Maynard, Diana G., and Mark A. Greenwood. "Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis." In *Lrec 2014 proceedings*. ELRA, 2014.

[10] Amendola, Miriam, Danilo Cavaliere, Carmen De Maio, Giuseppe Fenza, and Vincenzo Loia. "Towards echo chamber assessment by employing aspect-based sentiment analysis and gdm consensus metrics." *Online Social Networks and Media* 39 (2024): 100276.

[11] Qian, Edward. 2015. "Twitter Climate Change Sentiment Dataset." Kaggle.com. 2015. https://www.kaggle.com/datasets/edqian/twitter-climate-change-sentiment-dataset?resource=download.

[12] Mi, Zhewei, and Hongwei Zhan. "Text mining attitudes toward climate change: Emotion and sentiment analysis of the twitter corpus." *Weather, Climate, and Society* 15, no. 2 (2023): 277-287.

[13] Shyrokykh, Karina, Max Girnyk, and Lisa Dellmuth. "Short text classification with machine learning in the social sciences: The case of climate change on Twitter." *Plos one* 18, no. 9 (2023): e0290762.

[14] Uthirapathy, Samson Ebenezar, and Domnic Sandanam. "Topic Modelling and Opinion Analysis On Climate Change Twitter Data Using LDA And BERT Model." *Procedia Computer Science* 218 (2023): 908-917.

[15] Kvasničková Stanislavská, Lucie, Ladislav Pilař, Xhesilda Vogli, Tomas Hlavsa, Kateřina Kuralová, Abby Feenstra, Lucie Pilařová, Richard Hartman, and Joanna Rosak-Szyrocka. "Global analysis of Twitter communication in corporate social responsibility area: sustainability, climate change, and waste management." PeerJ Computer Science 9 (2023): e1390.

[16] Webb, Geoffrey I., Eamonn Keogh, and Risto Miikkulainen. "Naïve Bayes." Encyclopedia of machine learning 15, no. 1 (2010): 713-714.

[17] LaValley, Michael P. "Logistic regression." *Circulation* 117, no. 18 (2008): 2395-2399.

[18] Hearst, Marti A., Susan T. Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. "Support vector machines." *IEEE Intelligent Systems and their applications* 13, no. 4 (1998): 18-28.

[19] Rigatti, S.J., 2017. Random forest. *Journal of Insurance Medicine*, *47*(1), pp.31-39.

[20] Natekin, Alexey, and Alois Knoll. "Gradient boosting machines, a tutorial." *Frontiers in neurorobotics* 7 (2013): 21.

[21] Jost, François, Ann Dale, and Shoshana Schwebel. "How positive is "change" in climate change? A sentiment analysis." *Environmental Science & Policy* 96 (2019): 27-36.

[22] Taufek, Tasha Erina, Nor Fariza Mohd Nor, Azhar Jaludin, and Sabrina Tiun. "Public Perceptions on Climate Change: A Sentiment Analysis Approach." *GEMA Online Journal of Language Studies* 21, no. 4 (2021).

[23] Mohamad Sham, Nabila, and Azlinah Mohamed. "Climate change sentiment analysis using lexicon, machine learning and hybrid approaches." *Sustainability* 14, no. 8 (2022): 4723.

[24] Rosenberg, Emelie, Carlota Tarazona, Fermín Mallor, Hamidreza Eivazi, David Pastor-Escuredo, Francesco Fuso-Nerini, and Ricardo Vinuesa. "Sentiment analysis on Twitter data towards climate action." *Results in Engineering* 19 (2023): 101287.

[25] Loureiro, Maria L., and Maria Alló. "Sensing climate change and energy issues: Sentiment and emotion analysis with social media in the UK and Spain." ***Energy Policy*** **143 (2020): 111490.**

[26] Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. arXiv preprint arXiv:1701.02810.