International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

Social sensing with big data: Detecting hate speech in social media

Umar Ibrahim [1], Usman Lawal Gulma [2,*] and Ishaq Abdullahi Lawal [3]

[1] *Department of Computer Science Education, Adamu Augie College of Education, P.M.B 1012, Argungu, Kebbi State, Nigeria.*

[2] *Department of Geography, Adamu Augie College of Education, P.M.B 1012, Argungu, Kebbi State, Nigeria.*

[3] *Department of General Studies Education, Adamu Augie College of Education, P.M.B 1012, Argungu, Kebbi State, Nigeria.*

International Journal of Science and Research Archive, 2024, 11(02), 1146–1152

Publication history: Received on 22 February 2024; revised on 29 March 2024; accepted on 01 April 2024

Article DOI: https://doi.org/10.30574/ijsra.2024.11.2.0540

Abstract

The internet's accessibility and social media platforms, like Facebook and Twitter, have accelerated the spread of hate speech and fake news, both of which can be detrimental to society's overall well-being. Identifying and tracking hate speech is becoming increasingly difficult for the public, private citizens, legislators, and academics. Despite efforts to leverage automatic detection and monitoring techniques, their performances are still far from satisfactory. This study employs Natural Language Processing (NLP) and Machine Learning (ML) approaches to detect hate speech for decision-making. The result showed that the Support Vector Machine (SVM) algorithm has the best performance with an accuracy of 0.86 compared to the Random Forest with 0.8 accuracy. The manual evaluation of the performance of our algorithm yielded an inter-annotator agreement Cronbach's alpha (α = .775).

Keywords: Hate speech; Natural language processing; Machine learning; Social sensing; Big data

1. Introduction

Recent statistics have shown that Nigeria has about 90.48 million internet users (about 55.4% penetration) out of which 33 million are actively using social networking sites with WhatsApp, Facebook, and Twitter accounting for 95%, 89%, and 61% of the users [1]. A growing number of government agencies and companies also use social media networks for public interaction. The emergence of social networking sites such as Twitter and Facebook has democratized information, allowing users to share content about themselves and their social surroundings at an unprecedented scale. These sites have resulted in large volumes (big data) of unsolicited geographically and socially relevant data being created in real-time. These new forms of big spatial data sources contain lots of information (digital footprints) that can be harnessed for studying the dynamics of the human social environment known as *social sensing*. Social sensing is an emerging and novel research field, where unsolicited observations of real-world events are spontaneously mentioned in cyberspace and the users of such platforms act as human sensors [2]. In contrast to physical sensors that report events objectively, human sensors can recognize, summarize, and report perceptions of events differently.

However, the availability of the internet and social networking sites has propelled the spread of fake news and hate-related content in society. Hate speech includes any words, actions, writings, or displays that could incite others to engage in violent or discriminatory behavior [3]. In other words, hate speech is a form of verbal abuse that targets specific groups or individuals based on their identity, beliefs, or characteristics [4]. It can jeopardize social cohesiveness and democratic values, as well as negatively affect the victim's mental health, general well-being, and safety. Thus, it is crucial to identify and stop hate speech if one hopes to maintain social harmony and online safety. Opinion mining from social media content can be a very valuable tool in governance [5]. No study of contemporary society can ignore the rich potential afforded by these new dimensions of social life.

* Corresponding author: Usman Lawal Gulma

However, social media content is messy and unstructured and the task of extracting valuable insights from these emerging datasets presents a big computational research challenge [6]. People are essential components in social sensing, therefore collection and analysis of these new forms of data sources can help us better understand society and make decisions. In this paper, we propose to use social sensing techniques to enhance the performance and explain ability of hate speech detection models. We will analyze large Twitter datasets using Natural Language Processing (NLP) algorithm in order to identify and reveal patterns of hate speech on social media. Additionally, utilizing machine learning (ML) techniques to automate the text classification process will yield findings that are less subjective and more accurate [7]. NLP and ML are often combined to classify and detect hate speech in social media [8].

The rest of the paper is structured as follows: section two reviews the related literature, then material and method is described in section three. In section four, results and discussion are presented and conclusion is given in section five.

2. Review of Related Literature

The fact that hate speech is frequently subtle and context-dependent makes it challenging to identify and distinguish from acceptable expressions of humor or opinion. Furthermore, hate speech is subject to change throughout time, taking on new forms and tactics to evade detection through established means. Consequently, it might not be possible to fully capture the subtleties and dynamics of hate speech by depending only on linguistic elements. To overcome this difficulty, several researchers have suggested utilizing social sensing techniques to add user profile and social context data to the hate speech identification process.

Researchers have employed different approaches to quantify and detect hate-related speech especially those from social media. For example, Oriola and Kotzé [9] evaluated different machine-learning techniques on South African tweets to analyze and detect offensive and hate speech. They obtained the best results with a support vector machine (SVM). Mutanga, and Naicker [10] employed ensemble machine methods (decision trees and SVM) for automatic detection of hate speech in tweets. They found that these methods outperform classical machine learning methods that suffer from high variance. Plaza-Del-Arco, Molina-González [11] employed a combined approach using natural language processing and machine learning to detect hate speech. They found that a combination of NLP and ML helps to detect hate speech more accurately.

Khanday, and Rabani [12] used manually annotated tweets collected during the COVID-19 pandemic and ensemble machine-learning methods to detect hate speech. They concluded that the Decision Tree classifier is the most effective compared to other ML methods. Haider, and Dipty [13] also explored different ML algorithms in their attempt to classify hate comments from social media texts. Their experimental results showed that Random Forest produced the highest accuracy compared to other learning methods such as logistic regression, SVM and Naive Bayes.

Finally, in a survey on hate speech detection, despite the introduction of deep learning, ensemble and transformer-based approaches NLP and ML remain relevant in hate speech detection tasks [14]. Combining two or more machine learning algorithms can minimize variance and increase learning capacity greatly. This study will leverage from previous literature to develop NLP and ML models to automatically analyze and detect hateful comments on Twitter datasets.

3. Materials and Method

The proposed framework for this study is an improved classification model that aims to increase classification accuracy to assist decision-makers in the stock market sector. The data collection exercise for this model begins with financial tweets and stock prices. In the second phase, Twitter data is pre-processed to ensure that only essential data is retained, as well as all necessary formatting on the stock data. The polarity of Twitter data is determined in the analysis phase, and public emotions are classified as hateful, not hateful or abusive values using Natural Language Processing. In the analysis phase, we will explore relevant Machine Learning algorithms for hate speech detection. We will then evaluate the performance of the forecasting model. The proposed conceptual model for the research is presented in Figure 1.

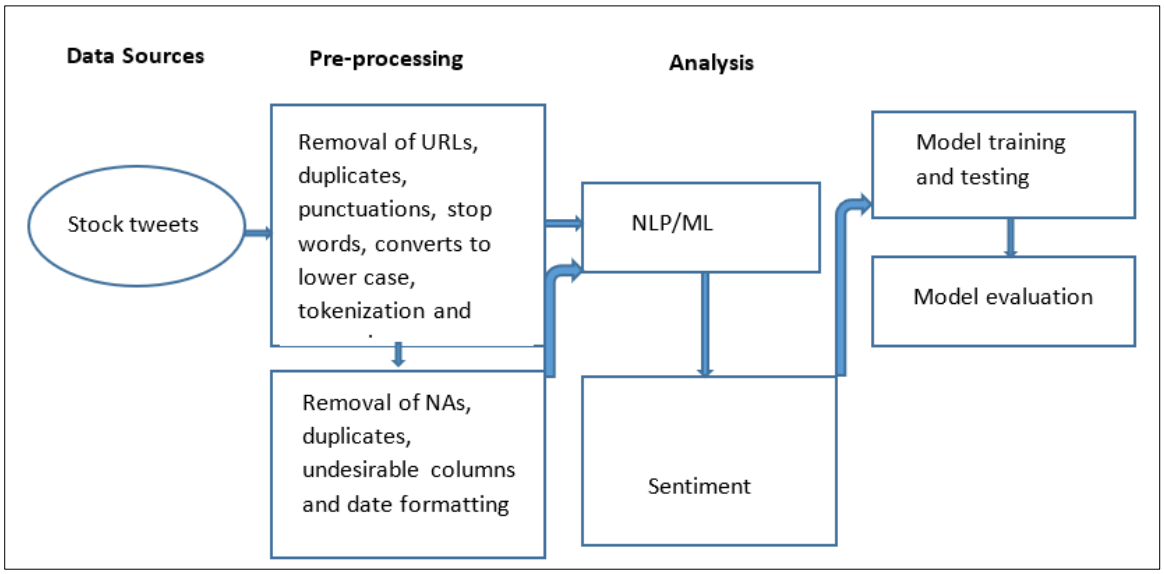

Figure 1 Conceptual model of the research proposal

3.1. Data

All the necessary credentials such as access key and access token will be obtained from the Twitter application programming interface (API). The API will then be utilized to stream data using Twitter handles of some pre-defined keywords within the Nigeria bounding box (top left 14°.5771N, 2°.6917E; lower right 4°.2405N, 15°.0000E). Streaming API can access real-time data of tweets using defined queries. Data streaming will be carried out for up to three months to acquire sufficient data for the study. Data has been described as an essential ingredient of the digital space [15].

3.2. Method of Data Analysis

Using social media texts to extract meaningful insights provides a challenging yet rich context for exploring computational models of natural language [16]. Because Twitter data is highly unstructured, it necessitates extensive data cleaning to prepare it for further analysis. This process will involve enormous tasks starting with the removal of abbreviations, white spaces, stop words, URLs, hashtags and user names. The process is concluded with feature extraction to filter relevant and irrelevant tweets. In the R programming environment, the NLP technique will be used to classify sentiments automatically. We then employ the Amazon Mechanical Turk (AMT) for manual validation. a commercial Human Intelligence Task (HTI) crowdsourcing platform to validate a sample of tweets. AMT is widely used by researchers for SA validation [17-22].

Estimating the reliability and accuracy of human annotation is implemented by quantifying the degree of agreement that exists between the annotators called *inter-rater reliability* (IRR) [23-27]. IRR is a measure of consistency among observation ratings provided by multiple annotators commonly evaluated using the intraclass correlation coefficient (ICC) [28]. ICC is a method used to measure the degree of correlation and agreement between raters [29]. The range of ICC is between 0 and 1. ICC value of 0 indicates random agreement and 1 indicates perfect agreement while negative ICC indicates disagreement between raters' scores [28]. ICC is usually expressed as Cronbach's alpha (α). Cronbach's alpha (α)provides a measure of the internal consistency of annotation scores [30], written as in Equation 1:

$$a = \frac{n}{n-1}\left(1 - \frac{\sum_i V_i}{V_t}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where α is a measure of consistency in agreement, V_i is the variance of scores in item i, V_t is the variance of test scores and n is the number of items.

Combining textual, social context, and user data to identify the subtleties and targets of hate speech on social media sites like Twitter is one approach to employing social sensing for hate speech identification. By taking into account the multi-modal data and the context-dependent nature of hate speech, this approach can increase the robustness and accuracy of hate speech identification [31].

4. Results and Discussion

In this study, we employ five different human annotators through Amazon Mechanical Turk (AMT), Overall the results obtained from AMT have been recognised to be of high quality, accurate and reliable [32].

In this study, we used irr Gamer, Lemon [33] and psych Revelle [34] R packages to measure Cronbach's (α) and Fleiss's (κ) respectively, in conformity with the literature [35]. Landis and Koch [36] provide a useful benchmark for interpreting Fleiss κ and is adopted in this study (Table 1).

Table 1 Interpretation of Fleiss's κ. Source: Landis and Koch [36]

Fleiss κ	Interpretation
< 0	Poor agreement
0.01 – 0.2	Slight agreement
0.21 – 0.4	Fair agreement
0.41 – 0.6	Moderate agreement
0.61 – 0.8	Substantial agreement
0.81 – 1.0	Almost perfect agreement

IRR results based on workers' scores are presented in Table 2. Before applying exclusion criteria, accuracy and reliability were assessed between the manually classified data and algorithm classification to evaluate the extent to which the results will differ after the exclusion criteria are applied. This decision was to measure the quality of inter-annotator results and to make comparisons between human annotations and algorithm classification simpler.

Table 2 Inter-rater reliability test between human annotators and algorithm

Inter-annotator			
Fleiss (κ)	**Interpretation**	**Cronbach's (α)**	**Interpretation**
.312	Fair agreement	.775	Good agreement
Algorithm and human average			
.445	Moderate agreement	.770	Good agreement

Overall, the performance of our algorithm compared to human annotation indicates good agreement (α = .770) and also good agreement for the inter-annotator agreement (α = .775). Our result is better compared to α = .55 value obtained in Provoost, Ruwaard [37] and α = .73 reported in Sloan and Morgan [38]. Nevertheless, the validation analysis in this study, indicates that the results of our hate speech algorithm is good and is potentially useful for further analysis.

Various machine learning and deep learning models have been used to tackle the problem of hate speech detection, as given in the literature review. Based on the analysis conducted, we have implemented the Naïve Bayes Classifier, Random Forrest (RF), Support Vector Machines (SVM) model and Recurrent Neural Network (RNN) model (Table 3).

Table 3 Machine learning algorithms implemented for hate speech detection

ML algorithm	Accuracy	Precision	Recall Score	F1 Score
Naïve Bayes	0.73	0.75	0.74	0.72
SVM	0.86	0.88	0.84	0.83
RF	0.80	0.85	0.83	0.80
RNN	0.81	0.78	0.74	0.72

Our proposal is a methodology for identifying hate speech and its targets that integrates linguistic, social, and emotional characteristics. Our framework consists of an emotion classifier that uses two separate approaches: lexicon-based and machine learning, to predict the emotions exhibited by the targets and consumers of hate speech. A lexicon-based method is a strategy that uses a pre-established set of terms or expressions that are connected to particular feelings, such hateful, neutral, and not hateful. Machine learning methods are used to extract complicated patterns from data and generate predictions from them.

5. Conclusion

In this study, we used a straightforward yet innovative method to identify hate speech in tweets. To find the best working model for this data, a variety of machine learning models are trained and assessed. Evaluation metrics like accuracy, precision score, recall score, and f1 score are computed. With an accuracy of 83%, the Support Vector Machine (SVM) Classifier model produced the best results, according to the data. The algorithm's performance was successfully manually validated with the help of Amazon Mechanical Turk (AMT).

In summary, social sensing is a potential method for tracking and identifying hate speech on social media sites. Social sensing enhances the accuracy and resilience of the detection algorithms by utilizing the textual, social, and emotional characteristics of users and their posts to identify the subtleties and settings of hate speech. The primary targets and origins of hate speech, as well as the chronological and spatial patterns of its dissemination, can all be found with the aid of social sensing. However, there are several difficulties that social sensing must overcome, including a lack of data, moral dilemmas, and cross-domain adaptability. Thus, future studies ought to concentrate on creating more efficient and morally sound procedures for gathering, marking, and examining social media data to identify hate speech.

Compliance with ethical standards

Acknowledgement

This research work was supported by the Tertiary Education Trust Fund (TetFund) under the Institution-Based Research (IBR) Intervention 2023.

Disclosure of Conflict of Interest

The authors have declared no conflict of interest.

References

[1] Internt Penetration in Nigeria. Available from: https://www.statista.com/statistics/1176092/internet-penetration-rate-nigeria// [Accessed 08/03/2024]. [Internet]. 2023.

[2] Arthur R, Boulton C, Shotton W. Social sensing of floods in the UK. PLoS ONE. 2018;13(1): e0189327.

[3] Gelber K. Differentiating hate speech: a systemic discrimination approach. Critical Review of International Social and Political Philosophy. 2021;24(4):393-414.

[4] Papcunová J, Martončik M, Fedáková D, Kentoš M, Bozogáňová M, Srba I, et al. Hate speech operationalization: a preliminary examination of hate speech indicators and their structure. Complex & intelligent systems. 2023;9(3):2827-42.

[5] Singh P, Dwivedi YK, Kahlon KS, Sawhney RS, Alalwan AA, Rana NP. Smart monitoring and controlling of government policies using social media and cloud computing. Information Systems Frontiers. 2020;22:315-37.

[6] Rahman MS, Reza H. A Systematic Review Towards Big Data Analytics in Social Media. Big Data Mining and Analytics. 2022;5(3):228-44.

[7] Mullah NS, Zainon WMNW. Advances in machine learning algorithms for hate speech detection in social media: a review. IEEE Access. 2021;9:88364-76.

[8] Khan U, Khan S, Rizwan A, Atteia G, Jamjoom MM, Samee NA. Aggression detection in social media from textual data using deep learning models. Applied Sciences. 2022;12(10):5083.

[9] Oriola O, Kotzé E. Evaluating machine learning techniques for detecting offensive and hate speech in South African tweets. IEEE Access. 2020;8:21496-509.

[10] Mutanga RT, Naicker N, Olugbara OO. Detecting hate speech on Twitter network using ensemble machine learning. International Journal of Advanced Computer Science and Applications. 2022;13(3).

[11] Plaza-Del-Arco FM, Molina-González MD, Ureña-López LA, Martín-Valdivia MT. A multi-task learning approach to hate speech detection leveraging sentiment analysis. IEEE Access. 2021;9:112478-89.

[12] Khanday AMUD, Rabani ST, Khan QR, Malik SH. Detecting twitter hate speech in COVID-19 era using machine learning and ensemble learning techniques. International Journal of Information Management Data Insights. 2022;2(2):100120.

[13] Haider F, Dipty I, Rahman F, Assaduzzaman M, Sohel A, editors. Social Media Hate Speech Detection Using Machine Learning Approach. International Conference on Computational Intelligence in Data Science; 2023: Springer.

[14] Subramanian M, Sathiskumar VE, Deepalakshmi G, Cho J, Manikandan G. A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. Alexandria Engineering Journal. 2023;80:110-21.

[15] Amoako G, Omari P, Kumi DK, Agbemabiase GC, Asamoah G. Conceptual Framework—Artificial Intelligence and Better Entrepreneurial Decision-Making: The Influence of Customer Preference, Industry Benchmark, and Employee Involvement in an Emerging Market. Journal of Risk and Financial Management. 2021;14(12):604.

[16] Boyd RL, Schwartz HA. Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. Journal of Language and Social Psychology. 2021;40(1):21-41.

[17] Snow R, O'Connor B, Jurafsky D, Ng AY. Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks. Proceedings of the conference on empirical methods in natural language processing. 2008:254-63.

[18] Akkaya C, Conrad A, Wiebe J, Mihalcea R. Amazon mechanical turk for subjectivity word sense disambiguation. Proceedings of the NAACL HLT. 2010:195-203.

[19] Taboada M, Brooke J, Tofiloski M, Voll K, Stede M. Lexicon-based methods for sentiment analysis. Computational Linguistics. 2011;37(2):267-307.

[20] Botchan N. Recognizing Meaning in the Crowd: Building Word Sense Inventories on Amazon Mechanical Turk.: Master Dissertation, University of Brandeis, Massachusetts, USA.; 2012.

[21] Shashidhar V, Pandey N, Aggarwal V. Spoken english grading: Machine learning with crowd intelligence. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015:2089-97.

[22] Burmania A, Parthasarathy S, Busso C. Increasing the reliability of crowdsourcing evaluations using online quality assessment. IEEE Transactions on Affective Computing. 2016;7(4):374-88.

[23] Remus R, Quasthoff U, Heyer G. SentiWS-A Publicly Available German-language Resource for Sentiment Analysis. LREC. 2010.

[24] Neviarouskaya A, Prendinger H, Ishizuka M. SentiFul: A lexicon for sentiment analysis. IEEE Transactions on Affective Computing. 2011;2(1):22-36.

[25] Sameki M, Gentil M, Mays KK, Guo L, Betke M. Dynamic Allocation of Crowd Contributions for Sentiment Analysis during the 2016 US Presidential Election. arXiv preprint arXiv:160808953. 2016.

[26] Benoit K, Conway D, Lauderdale BE, Laver M, Mikhaylov S. Crowd-sourced text analysis: reproducible and agile production of political data. American Political Science Review. 2016;110(2):278-95.

[27] Tosti-Kharas J, Conley C. Coding Psychological Constructs in Text Using Mechanical Turk: A Reliable, Accurate, and Efficient Alternative. Frontiers in Psychology. 2016;7.

[28] Hallgren KA. Computing inter-rater reliability for observational data: An overview and tutorial. Tutorials in Quantitative Methods for Psychology. 2012;8(1):23.

[29] Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine. 2016;15(2):155-63.

[30] Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16(3):297-334.

[31] Nagar S, Barbhuiya FA, Dey K. Towards more robust hate speech detection: using social context and user data. Social Network Analysis and Mining. 2023;13(1):47.

[32] Buhrmester M, Kwang T, Gosling SD. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? Perspectives on Psychological Science. 2011;6(1):3-5.

[33] Gamer M, Lemon J, Gamer MM, Robinson A, Kendall's W. Package 'irr'. Various coefficients of interrater reliability and agreement. 2012.

[34] Revelle W. Package 'psych'. The comprehensive R archive network. 2015;337:338.

[35] Liew JSY, Turtle HR, Liddy ED, editors. EmoTweet-28: a fine-grained emotion corpus for sentiment analysis. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16); 2016.

[36] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977:159-74.

[37] Provoost S, Ruwaard J, van Breda W, Riper H, Bosse T. Validating automated sentiment analysis of online cognitive behavioral therapy patient texts: an exploratory study. Frontiers in psychology. 2019;10:1065.

[38] Sloan L, Morgan J. Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. PloS one. 2015;10(11):e0142209.

Authors short Biography

Umar Ibrahim is an IT specialist and lecturer in the Department of Computer Science Education. He is also a director ICT of at Adamu Augie College of Education where he is employed. He has contributed to the data collection and analysis and manuscript writing. He has developed a lot of software for IT related tasks.

Dr **Usman Lawal Gulma** is a specialist in Geocomputation and spatial analysis, and he received his PhD from the University of Leeds, UK. He belongs to numerous professional associations, such as the Association of Nigerian Geographers (ANG), the Fellow Corporate Institute of Administration of Nigeria (FCAI), and the Geoinformation Society of Nigeria (GEOSON). Usman is employed with Adamu Augie College of Education in Argungu, Kebbi State. He conceived the research idea and contributed immensely to data analysis and methodology.

Ishaq Abdullahi Lawal studied English and has been a lecturer at Adamu Augie College of Education. He contributed to data cleaning and literature review as well as manuscript writing. He has co-authored in the publication of a number of peer-reviewed articles.