A Survey on Twitter Sentiment Analysis Using Machine Learning Techniques

G. Srikanth[1], K. Gangadhara Rao[1], Ramu Kuchipudi[1], Palamakula Ramesh Babu[1], R. Sai Venkat[1], T. Satyanarayana Murthy[1(✉)], and G. Venakata Kishore[2]

[1] Chaitanya Bharathi Institute of Technology, Hyderabad, India
{srikanthg_it,kgangadhar_it,ramukuchipudi_it,prameshbabu_it, saivenkatr_it,tsmurthy_it}@cbit.ac.in
[2] Srinidhi Institute of Technology, Hyderabad, India
kishore.g@sreenidhi.edu.in

Abstract. Twitter sentiment analysis involves various stages of the analysis process includes Text preprocessing techniques are applied to prepare the data, followed by examining the distribution of sentiment analysis. By utilizing the TF-IDF vectorizer, the textual data is transformed into numerical feature vectors. Three machine learning models, namely Bernoulli Naive Bayes (BernoulliNB), Linear Support Vector Classification (LinearSVC), and Logistic Regression, are created and evaluated using standard performance metrics like accuracy, precision, recall, and F1 score. The evaluation results effectively showcase the performance of each sentiment analysis model. The data is sourced from Twitter. The Logistic Regression model stands out in accurately classifying sentiments, while the LinearSVC and BernoulliNB models also exhibit high performance. The trained models are saved for future utilization, facilitating their integration into practical applications. This study presents a comprehensive approach to Twitter sentiment analysis, encompassing data preprocessing, model development, model evaluation, and storage for Twitter sentiment analysis tasks.

Keywords: twitter · sentiment analysis · BernoulliNB · LinearSVC · Logistic Regression

1 Introduction

In recent years, the younger generation has shown a growing interest in social media platforms such as Google Plus, WhatsApp, Facebook, and Twitter. These platforms have become an integral part of their lives, offering trending insights and current topics within seconds. People now openly express their social-related issues through comments, reviews, posts, hashtags, and emojis, which quickly gain popularity as they are followed by many. Moreover, social media has become an excellent opportunity for businesses to connect with consumers effortlessly.

O. Castillo et al. (Eds.): PerSOM 2023, LNICST 517, pp. 304–310, 2024.
https://doi.org/10.1007/978-3-031-66044-3_22

People heavily rely on user-generated content, such as comments and online reviews, to make decisions. For example, when considering purchasing a product, individuals search for online reviews and engage in discussions on social media platforms. The content displayed for a product and the discussions on social media significantly impact the success of a business. To automate the analysis of sentiment based on reviews or comments on social media, sentiment analysis (SA) has been introduced. SA aims to determine whether the information shared on social media is positive or negative in each scenario. By analyzing social media tags, we can gain insights into how people are currently reacting to various aspects of the world. Twitter sentiment analysis has become a popular area of research. Conducting sentiment analysis on Twitter data can be challenging due to certain characteristics unique to this platform. Tweets are limited to 140 characters, use informal English, contain irregular expressions, and include abbreviations and slang words. Researchers have focused on addressing these challenges through studies on sentiment analysis tweets. Twitter sentiment analysis approaches can generally be classified into two main categories: machine learning-based approaches and lexicon-based approaches. In this study, machine learning techniques are utilized to tackle Twitter sentiment analysis. Twitter sentiment analysis approaches can be generallycategorized into two main approaches, the machine learning approach, and a lexicon-based approach. In this study, we use machine learning techniques to tackle twitter sentiment analysis. Most classification algorithms are designed to predict nominal class data labels. However, predicting categories or labels on an ordinal scale presents additional pattern recognition challenges. This type of problem, known as ordinal classification or ordinal regression, has recently gained significant attention. This article focuses on performing sentiment analysis on Twitter data using machine learning. The process involves importing necessary dependencies and datasets, preprocessing the text to clean and transform it into a suitable format, analyzing the data, splitting it into training and testing sets, converting the text data into numerical features using the TF-IDF vectorizer, and creating and evaluating different models. Three popular machine learning models-Bernoulli Naive Bayes, Linear Support Vector Classification (LinearSVC), and Logistic Regression-are employed in this study. These models are trained on the training data and evaluated using appropriate metrics. Logistic Regression, for example, is a machine learning classifier used to classify different values based on specific attributes or features.

2 Literature Survey -Twitter Sentiment Analysis

This section explores prior investigations in the context of classification of disaster tweets and the performance of fundamental Machine Learning algorithms to advanced Deep Learning algorithms is discussed. A collection of traditional methods [1–9] is presented based on the information to detect the medical resource tweets during an emergency. Majority- Voting is considered for the

ensemble methods. The performance has been far better than the traditional algorithms on various parameters. It also combines this method performance over Bag Of Words. The performance of accuracy (82.4%). This analysis is restricted only to the Nepal and Italy earthquake datasets owing to the lack of labeled data. A model to identify NAR tweets [10–12] at the time of crisis. They suggest that the layering of a CNN with conventional classifiers (based on features) is effective for identifying the informative tweets. The authors also recommend that the combination of Convolutional Neural Networks, K-NN & SVM with specific features of domain exceeded the performance of diverse assemblages. This suggested model performs better than earlier methods on nepal and italy earthquake datasets. (Le, et al., proposed a comparison analysis of basic ML models with pre-trained model, BERT. They underlined that any DL algorithm, a LSTM or CNN with LSTM or a Convolutional Neural Networks (CNN) learns from one-hot encoding vectors of text,. If one-hot embedded its vector length equals to the word size, then this technique has dimensionality issues. A solution to this is to represent input as a low- dimensional space vector. This is implemented using many types of embedding techniques TF_IDF and Count Vector are considered for word representation and BERTLARGE architecture is used for implementing customized classification task. Disaster Tweets dataset from Kaggle repository is considered for this analysis. This paper substantiates that BERT (by tuning its parameters) is very constructive in classifying text [13–15]. This paper investigates the performance of text classification algorithm that uses TF-IDF Vectoriser, and a linear classification algorithm for a vector machine. The model predicts if a tweet is for a real emergency or not using a binary classifier.. The BERT layer's is followed by a drop out layer and then a dense layer. This study highlights the performance of BERT as a Text Classifier.- In this paper, the authors through their work emphasize on the significance of data processing, This paper throws light on importance of the information split affects efficiency of the classifier. They have used crisis_NLP and crisis_LexT26 datasets to make imbalanced and balanced datasets. And applied various information preprocessing methods to enhance the efficiency of classifier and further equate the performance of models (BERT, Default BERT, BERT with Non-Linear Layer, BERT with Long-Short Term Memory, BERT with Convolutional Neural Network.) on imbalanced and balanced datasets. They conclude that elimination of unessential data can lead to enhanced classifications. The models give better performance with balanced dataset. (Ma, n.d.) In this study, the authors have presented a comparision analysis on performance of the default BERT architecture and other custom based BERT architectures with the default Bi-LSTM for classification. Glove is used for embedding text into numerical vectors. The dataset used for the analysis is m CrisisLexT6 Hurricane Sandy dataset and n Crisis_NLP. The bidirectional LSTM with GloVe Twitter embeddings is set as baseline and several BERT based models (baseline BERT, BERT with NonLinearity, BERT with LSTM, BERT with CNN) are developed and they outperform the baseline performance [16]. This paper highlights on a comparison of word embeddings using CNN and Bi-LSTM as encoders for text classification.

This article explains the implementation of proposed methodology used for this study. This article is structured as follows. It starts with the introduction, which explains the overview of the execution, which is followed by the dataset and training specifications. Next to that, the implementation is detailed and finally the evaluation metrics used are presented. The datasets used and the challenges are elucidated [17–45]. The performance evaluation metrics used are F1-Score, Accuracy, Precision, Recall, AUC-ROC curve.

3 Conclusion

Analyzing emotions on Twitter involves multiple stages, starting with data collection and retrieval. The dataset is then preprocessed to clean the text data and make it suitable for analysis. Data analysis provides insights into the sentiments expressed by Twitter users, revealing diverse opinions in various contexts. To develop and evaluate a classification model, the data is divided into training and testing sets. The TF-IDF vectorizer is employed to convert the textual data into numerical representations used for model training and testing. Three models are utilized in this project: BernoulliNB, LinearSVC, and Logistic Regression. The BernoulliNB model, based on the Naive Bayes algorithm, is trained and evaluated to test distribution hypotheses. The LinearSVC model, utilizing a support vector machine, is also employed to validate the hypotheses. Additionally, the logistic regression model, commonly used for classification tasks, is utilized for sentiment prediction. The models are evaluated based on accuracy and their effectiveness in classifying emotions. Each model has its own strengths and limitations, but overall, they have proven effective in capturing sentiment expressed in tweets. The results demonstrate the feasibility of performing sentiment analysis on Twitter data using machine learning algorithms. Once the models are trained and evaluated, they can be saved for future use. Saved models can be easily reused and deployed in real-world applications requiring sentiment analysis. In summary, the Twitter Sentiment Analysis project leveraged machine learning models, including BernoulliNB, LinearSVC, and Logistic Regression, to identify and classify sentiment in tweets. This project enhances understanding of diverse opinions expressed by Twitter users and facilitates deeper insights into public sentiment on various topics and events. By employing these models, businesses and organizations can gain a better understanding of customer sentiment, leading to more informed decisions and tailored strategies. The study also emphasizes the significance of techniques like TF-IDF vectorization in preparing data for statistical analysis. Overall, this project contributes to the ongoing advancement of sentiment analysis in understanding social media sentiment.

References

1. Ameen, Y.A., Bahnasy, K., Elmahdy, A.: Classification of Arabic Tweets for damage event detection. Int. J. Sci. Eng. Res. **114**, 160–166 (2020)
2. Chanda, A.K.: Efficacy of BERT embeddings on predicting disaster from Twitter data. arXiv. Association for Computing Machinery **2021**, 1–14 (2021)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, vol. 1, pp.4171–4186 (2019)
4. Gao, H., Barbier, G.: Harnessing the crowdsourcing power of social media for disaster relief. IEEE Intell. Syst. 11–14 (2011)
5. Goswami, S., Raychaudhuri, D.: Identification of Disaster-Related Tweets Using Natural Language Processing (2020)
6. Irawan, R., Isa, S.M.: Social media disaster relevance classification for situation awareness during emergency response in Indonesia. Int. J. **87**, 3216–3222 (2020)
7. Kabir, Y.: A Deep Learning Approach for Tweet Classification and Rescue Scheduling for Effective Disaster Management, arXiv, pp. 1–14 (2019)
8. Kalyan, K.S., Sangeetha, S.: SECNLP?: a survey of embeddings in clinical natural language processing. J. Biomed. Inform. **101**, 103323 (2020)
9. Madichetty, S., Sridevi, M.: Improved classification of crisis-related data on Twitter using contextual representations. Procedia Comput. Sci. **1672019**, 962–968 (2020)
10. Madichetty, S., Sridevi, M.: A stacked convolutional neural network for detecting the resource tweets during a disaster. Multimedia Tools Appl. **803**, 3927–3949 (2021)
11. Malekzadeh, M., Hajibabaee, P., Heidari, M., Zad, S., Uzuner, O., Jones, J.H.: Review of graph neural network in text classification, pp. 0084–0091 (2022)
12. Messages, C., Imran, M., Mitra, P., Castillo, C.: Twitter as a Lifeline: Humanannotated Twitter Corpora for NLP of Crisis-related Messages, pp. 1638– 1643 (2016)
13. Naaz, S., Abedin, Z.U., Rizvi, D.R.: Sequence classification of tweets with transfer learning via BERT in the field of disaster management. EAI Endorsed Trans. Scalable Inf. Syst. **831**, 1–8 (2021)
14. Peters, M.E., et al.: Deep contextualized word representations. In: NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, vol. 1, pp. 2227–2237 (2018)
15. Weimar, B., Wiegmann, M., Kersten, J., Potthast, M.: Analysis of detection models for disaster-related tweets, pp. 872–880 (2020)
16. Wang, C., Nulty, P., Lillis, D.: A comparative study on word embeddings in deep learning for text classification, pp. 37–46 (2020)
17. Soh, W.T.: Text-based Graph Convolutional Network - Bible Book Classification - A semi-supervised graph-based approach for text classification and inference (2019)
18. Satyanarayana Murthy, T., Varma, M.K., Roy, S.: Improving the performance of association rules hiding using hybrid optimization algorithm. J. Appl. Secur. Res. **15**(3), 423–437 (2020). https://doi.org/10.1080/19361610.2020.1756155
19. Satyanarayana Murthy, T., Gopalan, N.P., Yakobu, D.: An efficient un-realization algorithm for privacy preserving decision tree learning using McDiarmid's Bound. Int. J. Innovative Technol. Exploring Eng. (IJITEE), **8**(4S2), 499–502 (2019)

20. Satyanarayana Murthy, T., Gopalan, N.P., Gunturu, S.: A novel optimization based algorithm to hide sensitive item-sets through sanitization approach. Int. J. Modern Educ. Comput. Sci. (IJMECS) **10**(10), 48–55 (2018). https://doi.org/10.5815/ijmecs.2018.10.06
21. Satyanarayana Murthy, T., Gopalan, N.P., Alla, D.S.K.: The power of anonymization and sensitive knowledge hiding using sanitization approach. Int. J. Modern Educ. Comput. Sci. (IJMECS), **10**(9), pp. 26–32 (2018). https://doi.org/10.5815/ijmecs.2018.09.04.
22. Satyanarayana Murthy, T., Gopalan, N.P.: A novel algorithm for association rule hiding. Int. J. Inf. Eng. Electron. Bus. (IJIEEB) **10**(3), 45–50 (2018). https://doi.org/10.5815/ijieeb.2018.03.06
23. SaiBabu, A., Murthy, T.S.N.: Security provision in publicly auditable secure cloud data storage services using SHA-1 algorithm. Int. J. Comput. Sci. Inf. Technol. (IJCSIT) **3**(3), 4084–4088 (2012)
24. Sathyanarayana Murthy, T., Mohan Krishna Varma, N., Ravuri, D., Kishore Babu, D., Nazeer, S.: Classification of Precious and Non-precious Tweets Using Deep Learning. In: Rout, R.R., Ghosh, S.K., Jana, P.K., Tripathy, A.K., Sahoo, J.P., Li, KC. (eds.) Advances in Distributed Computing and Machine Learning, vol. 427. LNNS, pp. 393–399. Springer, Singapore (2022). https://doi.org/10.1007/978-981-19-1018-0_33
25. Satyanarayana Murthy, T., Mohan Krishna Varma, N., Roy, S., Nazeer, S.: Effective classification of tweets using machine learning. In: Kumar, R., Ahn, C.W., Sharma, T.K., Verma, O.P., Agarwal, A. (eds.) Soft Computing: Theories and Applications. LNNS, vol. 425, pp. 439–446. Springer, Singapore (2022). https://doi.org/10.1007/978-981-19-0707-4_40
26. Murthy, T.S., Gopalan, N.P., Ramachandran, V.: A naive bayes classifier for detecting unusual customer consumption profiles in power distribution systems - APSPDCL. In: 2019 Third International Conference on Inventive Systems and Control (ICISC) at JCT College, Coimbatore, India, pp. 673–678 (2019)
27. Satyanarayana Murthy, T., Preethi, G., Gopalan, N.P.: An efficient way of anonymization without subjecting to attacks using secure matrix method. In: proceedings of the IEEE International Conference on Intelligent Computing and Control Systems at VAIGAI COLLEGE OF ENGG,MADURAI, pp 1462–1465 (2018)
28. Satyanarayana Murthy, T., Gopalan, N.P.: An efficient meta-heuristic chemical reaction based algorithm for association rule Hiding using an advanced perturbation approach. In: proceedings of the IEEE International Conference on Intelligent Computing and Control Systems, at VAIGAI COLLEGE OF ENGG,MADURAI. Indexed in IEEE (2018)
29. Gopalan, N.P., Satyanarayana Murthy, T.: Association rule Hiding using chemical reaction optimization. In: Presented a paper at 7th International Conference on Soft Computing for Problem Solving - SocProS 2017. IIT Bhubaneswar, ORISSA (2017)
30. Devarajan, D., et al.: Cervical cancer diagnosis using intelligent living behavior of artificial jellyfish optimized with artificial neural network. IEEE Access **10**, 126957–126968 (2022)
31. Maheswari, V.U., Aluvalu, R., Kantipudi, M.P., Chennam, K.K., Kotecha, K., Saini, J.R.: Driver drowsiness prediction based on multiple aspects using image processing techniques. IEEE Access **10**, 54980–54990 (2022)
32. Satyanarayana Murthy, T., Gopalan, N.P., Balaji, B.: A modified un-realization approach for effective data perturbation. Int. J Intell. Enterp. 408–421 (2023). https://doi.org/10.1504/IJIE.2023.10054103

33. Satyanarayana Murthy. T., Udayakumar, P., Alenezi, F., Laxmi Lydia, E., Ishak, M.K.: Coot optimization with deep learning-based false data injection attack recognition. Comput. Syst. Sci. Eng. **46**(1), 255–271 (2023)
34. Yonbawi, S., Alahmari, S., Satyanarayana Murthy, T., Maddala, P., Laxmi Lydia, E., et al.: Harris hawks optimizer with graph convolutional network-based weed detection in precision agriculture. Comput. Syst. Sci. Eng. **46**(2), 1533–1547 (2023)
35. Yonbawi, S., Alahmari, S., Murthy, T.S., Daniel, R., Lydia, E.L., et al.: Modified metaheuristics with transfer learning based insect pest classification for agricultural crops. Comput. Syst. Sci. Eng. **46**(3), 3847–3864 (2023)
36. Ahmed, M.A., Murthy, T.S., Alenezi, F., Lydia, E.L., Kadry, S., et al.: Design of evolutionary algorithm based unequal clustering for energy aware wireless sensor networks. Comput. Syst. Sci. Eng. **47**(1), 1283–1297 (2023)
37. Devaraj, F.S., Satyanarayana Murthy, T., Alenezi, F., Laxmi Lydia, E., Md Zawawi, M.A., et al.: Enhanced metaheuristics with trust aware route selection for wireless sensor networks. Comput. Syst. Sci. Eng. **46**(2), 1431–1445 (2023)
38. Kalyani, K., Parvathy, V.S., Abdeljaber, H.A.M., Murthy, T.S., Acharya, S., et al.: Effective return rate prediction of blockchain financial products using machine learning. Comput. Mater. Continua **74**(1), 2303–2316 (2023)
39. Satyanarayana Murthy, T.: An efficient diabetic prediction system for better diagnosis. Int. J. Intell. Enterp. 408–421 (2022). https://doi.org/10.1504/IJIE.2022.126397,
40. Satyanarayana Murthy, T., Gopalan, N.P., Athira, T.R.: Hiding critical transactions using modified un-realization approach". Int. J. Bus. Intell. **15**(3), 223–234 (2020)
41. Navaneetha Krishnan, S., Sundara Vadivel, P., Yuvaraj, D., Satyanarayana Murthy, T., Malla, S.J., et al.: Enhanced route optimization for wireless networks using meta-heuristic engineering. Comput. Syst. Sci. Eng. **43**(1), 17–26 (2022)
42. Shanmuga Priya, S., Yuvaraj, D., Satyanarayana Murthy, T., Chooralil, V.S., Navaneetha Krishnan, S., et al.: Secure key management based mobile authentication in cloud. Comput. Syst. Sci. Eng. **43**(3), 887–896 (2022)
43. Satyanarayana Murthy, T., Varma, M.K., Yadav, A.K.: A Diaetic Prediction System based on Mean Shift Clustering. ISI, IIETA Publisher,vol. 36, no. 2, pp. 231–235 (2021). https://doi.org/10.18280/isi.260210
44. Satyanarayana Murthy, T., Varma, M.K., Harsha.: Brain tumour segmentation using U-net based adversarial networks. Traitement du Signal, vol. 36, no. 4, pp. 353–359 (2021). https://doi.org/10.18280/ts.360408
45. Satyanarayana Murthy, T., Banothu, B., Varma, M.K.: An un-realization algorithm for effective privacy preservation using classification and regression trees, Revue d'Intelligence Artificielle,IIETA Publisher, vol. 33, no. 4, pp. 313–319 (2019). https://doi.org/10.18280/ria.330408.