International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307

OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT241022

Geographical Tweets Based Current Affairs Prediction Using Hybrid Features

Prof. Dhruvi Zala

Computer Engineering, Pacific School of Engineering, Surat, Gujarat, India

ARTICLEINFO

Article History:

Accepted: 20 Feb 2024

Published: 01 March 2024

Publication Issue

Volume 10, Issue 2

March-April-2024

Page Number

10-16

ABSTRACT

Twitter serves as a prominent and freely accessible social networking platform, enabling registered and authorized users to share their opinions and reviews through concise messages known as tweets. This research aims to conduct sentiment analysis on a Twitter dataset, seeking to predict and analyze current affairs based on user behaviour and opinions. The scope of current affairs encompasses a wide range of topics, including product reviews, political discourse, movie ratings, and more, expressed in both real-time streaming and offline data on the Twitter platform. During the feature extraction phase, a hybrid approach is employed, incorporating features such as acronyms, synonyms, and emoticons utilized by users in their tweets. Additionally, a hybrid dictionary is introduced, and instead of relying solely on unigrams and bigrams, a novel algorithm, n-gram, is implemented for more comprehensive analysis. This paper delves into various methodologies, specifically focusing on multiclass classification, and introduces a proposed system along with its noteworthy results. The research contributes to the understanding of user sentiments and behaviours on Twitter, offering valuable insights into diverse current affairs.

Keywords : Twitter, Sentiment Analysis, N-Gram, Opinion Mining, Hybrid Features, SVM, ANN, KNN, RF

I. INTRODUCTION

This research focuses on conducting sentiment analysis on a Twitter dataset to predict and assess the current state of affairs on the platform, aiming to survey and predict user behaviour and opinions. The scope of current affairs encompasses various tweet categories, including customer product reviews, political discourse, movie ratings, and more. Both real-time streaming and offline Twitter data in the form of tweets are subject to comprehensive analysis in this study.

Twitter, being a widely used and free social networking service, allows registered and authorized users to express their opinions through short messages,

commonly known as tweets. Organizations invest substantial resources in surveying their products and collecting feedback to identify system defects.

For instance, monitoring tweets related to a specific product can assist organizations in reducing costs. A negative tweet about a product, for instance, may prompt a cost reduction strategy. The analysis also includes a geographical perspective, conducting state-wise surveys to make predictions. This entails assessing reviews for a particular product in a specific area, understanding preferences or dislikes regarding movies in certain regions, gauging public opinion on political news in specific locales, and so forth.

Figure 1: Product Based on Location

In this research, a novel algorithm, referred to as n-gram, will be developed to enhance the prediction accuracy by encompassing features from unigram, 2-gram, and 3-gram. This integrated approach is expected to yield more precise predictions.

Furthermore, the analysis will incorporate hybrid features. Unlike the existing system, which removes emoticons, synonyms, and acronyms during pre-processing, the proposed system will utilize hybrid features, considering all these elements to enhance the accuracy of predictions.

The current system relies on the Afinn dictionary to calculate scores, but this dictionary has limitations due to its relatively small word coverage. The proposed system introduces a hybrid dictionary, combining lexicon and Afinn dictionaries, thus expanding the word pool and overcoming the limitations of the Afinn dictionary.

Moreover, the new system supports live tweet retrieval, facilitating real-time analysis of streaming data. This feature enhances the system's applicability and ensures its relevance in dynamic and evolving social media environments.

Figure 2: Tweets Example

II. RELATED WORK

In the study outlined in [1], sentiment analysis is conducted on Amazon product reviews. The experimentation involves employing various methods, such as Naïve Bayesian, Support Vector Machine Classifier, Stochastic Gradient Descent, Linear Regression, Random Forest, and Decision Tree, utilizing both unigram and, to a lesser extent, bigram models. Notably, the accuracy of the paper varies across different product categories, achieving 93.57% for Cell Phone & Accessories, 93.52% for Electronics, and 94.02% for Music Instruments.

In reference [2], the research focuses on sentiment analysis conducted on a Twitter dataset to discern public opinions regarding product reviews. The study employs a Support Vector Machine (SVM) classifier, incorporating multiclass classification. Notably, the feature extraction phase involves the utilization of emoticons, with each emoticon associated with its specific word meaning. This inclusion is aimed at enhancing the overall accuracy of the system.

In the study documented in [3], sentiment analysis is conducted to forecast the success of movies based on people's opinions and sentiments expressed on Twitter

through tweets. The analysis utilizes the Support Vector Machine (SVM) classification algorithm. Notably, the accuracy achieved in this research is recorded at 86%. This finding underscores the efficacy of employing the SVM algorithm for sentiment analysis in predicting the success of movies by leveraging insights from Twitter users' sentiments.

In the study documented in [4], the research delves into sentiment analysis of customer product reviews. The investigation employs three classification algorithms: Naïve Bayes, Support Vector Machine (SVM), and Decision Tree. Both unigram and bigram models are utilized for classification. Notably, the feature extraction process incorporates emoticons and synonyms, enhancing accuracy by substituting emoticons with their corresponding word meanings and replacing synonyms with their similar words.

In the investigation outlined in [5], sentiment analysis is carried out on reviews of smartphone products. The study specifically utilizes the Support Vector Machine (SVM) classification technique to categorize sentiments associated with positive smartphone product reviews. The analysis encompasses diverse datasets employed for sentiment classification and text analysis. Notably, the accuracy achieved in this research is noteworthy, reaching 90.99%. This result emphasizes the effectiveness of the SVM classification technique in discerning sentiments within the context of smartphone product reviews, showcasing its robust performance in sentiment analysis.

III.METHODS AND MATERIALS

A. API Interfacing: This phase involves obtaining user tweets or Twitter datasets through the Twitter API to conduct sentiment analysis on the dataset. The objective is to geographically predict sentiments related to political tweets, and movie ratings, and make predictions regarding product outcomes and client sentiment.

B. Pre-Processing: The pre-processing step focuses on refining the Twitter dataset. This includes the removal of unnecessary words and, the elimination of stop words, punctuations, and hashtags. The goal is to enhance the accuracy of the analysis results.

C. Features Extraction: During the feature extraction process, a comprehensive set of hybrid features is gathered, encompassing emoticons, acronyms, and synonyms. This step aims to enhance the accuracy of the system, employing the N-gram model as a novel algorithm. Unlike traditional Unigram and Bigram models, the N-gram model is expected to yield superior accuracy.

D. Multiclass Classification: Multiclass classification, also referred to as multinomial classification, is employed to categorize data into three or more classes. This approach facilitates the classification of sentiments related to political tweets, movies, and product reviews.

Artificial Neural Network [6]: Artificial neural networks represent electronic networks of neurons inspired by the brain's anatomical structure. These networks independently process and learn data by comparing their initial classification with the correct classification. Errors in the initial classification are fed back into the network, adjusting the network's formula for subsequent iterations.

Neurons within the network are organized in layers: input, hidden, and output. The input layer consists of input values for the subsequent layer of neurons, and there may be multiple hidden layers in a neural network. The output layer features one node for each category. A single pass through the network assigns a value to each node, ultimately categorizing the record based on the node with the highest value.

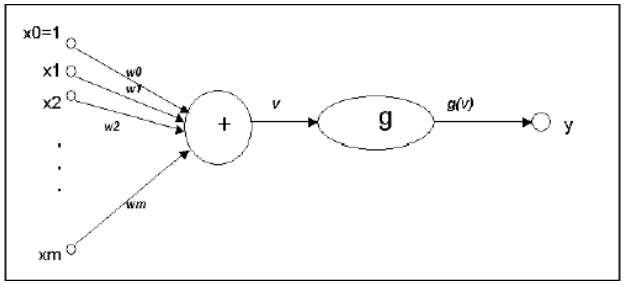

Figure 3: Artificial Neural Network

Support Vector Machine [6]: The Support Vector Machine is a supervised machine learning algorithm utilized for both classification and regression tasks, although its primary application is in classification. In this algorithm, each data point is graphically represented as a point in an n-dimensional space, where n corresponds to the number of features. The value of each feature is then depicted as the coordinate value within this space. The objective is to identify a hyperplane that effectively separates the two classes in this multi-dimensional space.

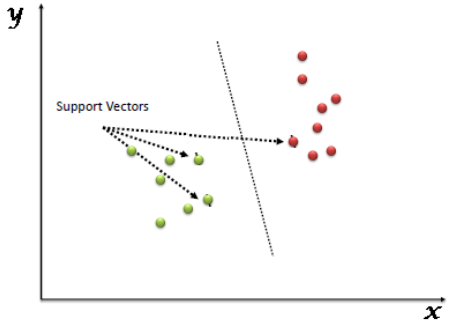

Figure 4: Support Vector Machine

K-Nearest Neighbors [6]: K-Nearest Neighbors is a non-parametric, instance-based machine learning algorithm designed to utilize a dataset with categorized data points to predict the classification of a new sample point.

The term "non-parametric" implies that no assumptions are made about the underlying data distribution. Additionally, KNN is characterized as a lazy algorithm, meaning it does not employ training data points to establish generalizations. In other words, there is no explicit training phase, or if present, it is minimal.

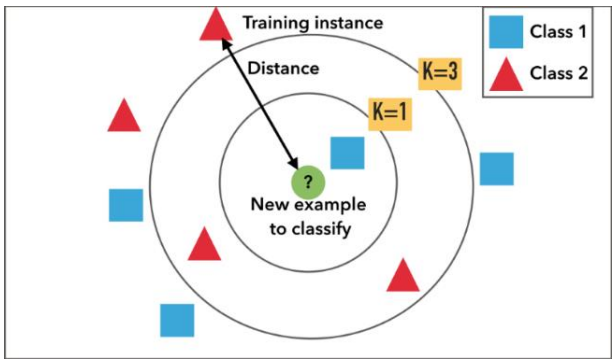

Figure 5: K- Nearest Neighbouring

Random Forest [6]: Random Forest is recognized as one of the most extensively employed and robust machines learning techniques, exhibiting a higher accuracy rate compared to recent algorithms. This method is particularly suitable for training with large datasets. Notably, Random Forest is straightforward to implement and demonstrates commendable performance across various challenges, including those characterized by non-linearity. The essence of Random Forest lies in its collection of tree-structured classifiers, with each tree relying on independently sampled values from a random vector. The collective distribution of all trees within the forest contributes to the algorithm's effectiveness.

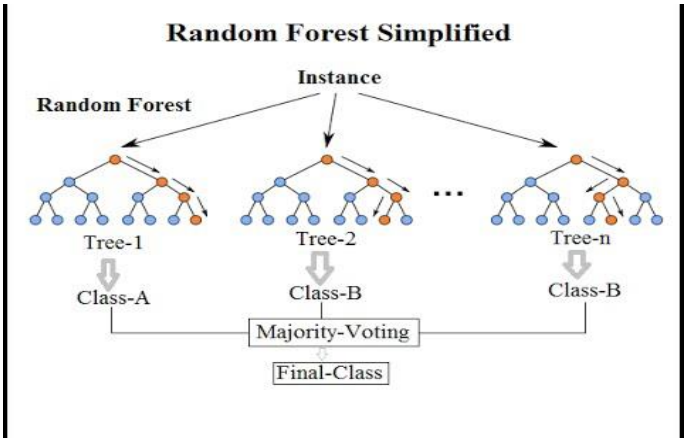

Figure 6: Random Forest [6]

IV.PROPOSED APPROACH

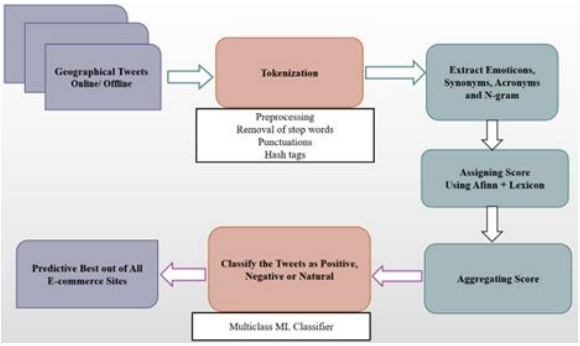

Figure 7: Proposed Flow

The research methodology involves retrieving and analyzing both live and offline Twitter datasets. Subsequently, a pre-processing step is undertaken to rectify inconsistencies, address incompleteness, and eliminate punctuations and hashtags. This ensures dataset uniformity, enhancing the accuracy of the analysis results.

In the subsequent phase, essential features for analysis are extracted from the tweets. This includes the incorporation of hybrid features such as emoticons, synonyms, and acronyms, as well as N-grams, which comprise a combination of Unigram, Bigram, and Trigram elements. The extraction of these features contributes to a comprehensive understanding of the content within the tweets.

Next, a scoring mechanism is implemented for all extracted features using dictionaries. This involves the utilization of a hybrid dictionary, formed by combining the Afinn dictionary and the lexicon dictionary. Assigning scores to features aids in quantifying their significance in the analysis process.
Finally, a multinomial classification is performed, employing various classification methods to conduct a survey. This phase involves categorizing and classifying the tweets based on the extracted features, providing insights into sentiment patterns and opinions expressed on Twitter across different classes.

V. PROPOSED ALGORITHM

Step 1: Retrieving Tweets using Tweeter API.
Step 2: Apply Preprocessing
Remove
Special Characters()
Hashtags()
URL()
Common English Words()
Repeated Words()
Step 3: Replace emoticons with word
:) – happy, etc.
Step 4: Replace synonyms words with their dictionary words
Glad – Happy, etc.
Step 5: Replace Acronyms words with their dictionary words
LOL– Laughing Out loud, etc.
Step 6: Apply N-gram Model
Sentence to Word break()
fori=1:all pair
If one_pair()=Ǿ or value else If two_pair()=Ǿ or value else If three_pair= Ǿ or value else score=0 or max(score)
end;
end
Step 7: Classify using Multi-Class Classification
Step 8:- End

VI.RESULTS & DISCUSSION

In Table 1, a comparison of the results is presented.

Table 1: Result Comparison

Classifier	Precision	Recall	Accuracy	Time
SVM	75%	75.26%	75%	0.277751 Sec.
ANN	68.2%	77.50%	68.2%	16.50096 Sec.
KNN	69%	75.45%	69%	0.051469 Sec.
RF	**88%**	**88.66%**	**88%**	**1.267681 Sec.**

Within this system, diverse classification methods are employed for multiclass classification, including Support Vector Machine (SVM), Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), and Random Forest (RF). Among these methods, SVM, ANN, and KNN constitute the existing implementation techniques, whereas RF represents the proposed implementation method. Notably, Random Forest exhibits the highest accuracy value, reaching 88%. This finding suggests the efficacy of the proposed RF method in achieving superior accuracy compared to the existing implementation techniques.

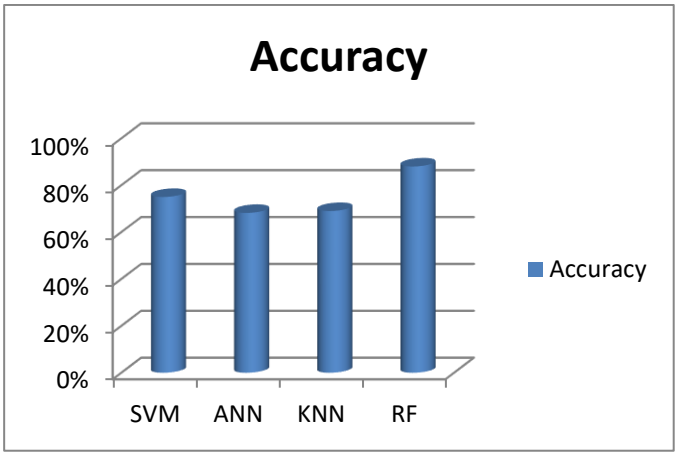

Figure 8: Analysis Chart

The chart above illustrates the accuracy lines of Support Vector Machine (SVM), Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), and Random Forest (RF) classifiers. Notably, the RF classifier exhibits the highest accuracy among all the classifiers.

VII. CONCLUSION & FUTURE PLAN

Recognizing the valuable insights that individual reviews provide for both consumers and company owners, I have introduced a novel geographical-based hybrid sentiment system. This system is designed to evaluate sentiments across various domains such as films, products, political parties, sports, etc., incorporating hybrid features like emoticons, synonyms, acronyms, and n-grams. Notably, these features are assigned scores using a hybrid dictionary, amalgamating the features of lexicon and Afinn dictionaries.

Following this comprehensive feature extraction and scoring process, the next step involves multiclass classification using various classification algorithms, including Random Forest (RF), Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). Remarkably, the Random Forest algorithm stands out with superior accuracy compared to other algorithms, achieving an accuracy rate of 88%. As a result, I have selected RF as the proposed algorithm for this work, recognizing its efficacy in sentiment analysis across diverse domains.

VIII. REFERENCES

[1] T. U. Haque, N. N. Saber and F. M. Shah, "Sentiment analysis on large scale Amazon product reviews," 2018 IEEE International Conference on Innovative Research and Development (ICIRD), Bangkok, Thailand, 2018, pp. 1-6, doi: 10.1109/ICIRD.2018.8376299.

[2] K. Lavanya and C. Deisy, "Twitter sentiment analysis using multi-class SVM," 2017 International Conference on Intelligent Computing and Control (I2C2), Coimbatore, India, 2017, pp. 1-6, doi: 10.1109/I2C2.2017.8321798.

[3] Q. I. Mahmud, A. Mohaimen, M. S. Islam and Marium-E-Jannat, "A support vector machine mixed with statistical reasoning approach to predict movie success by analyzing public sentiments," 2017 20th International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2017, pp. 1-6, doi: 10.1109/ICCITECHN.2017.8281803.

[4] Z. Singla, S. Randhawa and S. Jain, "Sentiment analysis of customer product reviews using machine learning," 2017 International Conference on Intelligent Computing and Control (I2C2), Coimbatore, India, 2017, pp. 1-5, doi: 10.1109/I2C2.2017.8321910.

[5] U. Kumari, A. K. Sharma and D. Soni, "Sentiment analysis of smart phone product review using SVM classification technique," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai,

India, 2017, pp. 1469-1474, doi: 10.1109/ICECDS.2017.8389689.

[6] A. Rane and A. Kumar, "Sentiment Classification System of Twitter Data for US Airline Service Analysis," 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), Tokyo, Japan, 2018, pp. 769-773, doi: 10.1109/COMPSAC.2018.00114.

[7] S. Ahuja and G. Dubey, "Clustering and sentiment analysis on Twitter data," 2017 2nd International Conference on Telecommunication and Networks (TEL-NET), Noida, India, 2017, pp. 1-5, doi: 10.1109/TEL-NET.2017.8343568.

[8] S. P. Algur and R. H. Patil, "Sentiment analysis by identifying the speaker's polarity in Twitter data," 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), Mysuru, India, 2017, pp. 1-5, doi: 10.1109/ICEECCOT.2017.8284629.

[9] Z. Jianqiang and G. Xiaolin, "Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis," in IEEE Access, vol. 5, pp. 2870-2879, 2017, doi: 10.1109/ACCESS.2017.2672677.

[10] N. Bhan and M. D'silva, "Sarcasmometer using sentiment analysis and topic modeling," 2017 International Conference on Advances in Computing, Communication and Control (ICAC3), Mumbai, India, 2017, pp. 1-7, doi: 10.1109/ICAC3.2017.8318782.

[11] Tahura Shaikh, Dr. Deepa Deshpande "Feature Selection Methods in Sentiment Analysis and Sentiment Classification of Amazon Product Reviews". International Journal of Computer Trends and Technology (IJCTT) V36(4):225-230 June 2016. ISSN:2231-2803.

[12] Z. Jianqiang, "Pre-processing Boosting Twitter Sentiment Analysis?" 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity), Chengdu, China, 2015, pp. 748-753, doi: 10.1109/SmartCity.2015.158.

[13] J. A. Banados and K. J. Espinosa, "Optimizing Support Vector Machine in classifying sentiments on product brands from Twitter," IISA 2014, The 5th International Conference on Information, Intelligence, Systems and Applications, Chania, Greece, 2014, pp. 75-80, doi: 10.1109/IISA.2014.6878768.

[14] M. A. Cabanlit and K. J. Espinosa, "Optimizing N-gram based text feature selection in sentiment analysis for commercial products in Twitter through polarity lexicons," IISA 2014, The 5th International Conference on Information, Intelligence, Systems and Applications, Chania, Greece, 2014, pp. 94-97, doi: 10.1109/IISA.2014.6878767.

[15] Kharde, V.A., & Sonawane, S.S. (2016). Sentiment Analysis of Twitter Data : A Survey of Techniques. *ArXiv, abs/1601.06971*.

[16] D. K. Zala and A. Gandhi, "A Review on Basic Methodology of Twitter Base Prediction System," 2018 3rd International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2018, pp. 447-451, doi: 10.1109/ICICT43934.2018.9034369.

[17] D. K. Zala and A. Gandhi, "A Twitter Based Opinion Mining to Perform Analysis Geographically," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019, pp. 59-63, doi: 10.1109/ICOEI.2019.8862548.

[18] D. K. Zala, "A Twitter Based Opinion Mining to Perform Analysis on Network Issues of Telecommunication Companies," 2018 3rd International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2018, pp. 437-441, doi: 10.1109/ICICT43934.2018.9034354.

Cite this article as :

Prof. Dhruvi Zala, "Geographical Tweets Based Current Affairs Prediction Using Hybrid Features", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 10, Issue 2, pp.10-16, March-April-2024. Available at doi : https://doi.org/10.32628/CSEIT241022

Journal URL : https://ijsrcseit.com/CSEIT241022