JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Sentiment Analysis on Social Media

[1]Devansh Choudhury, [2]Varun Sovani, [3]Mihir Parte, [4]Priyanka Shahane

[1,2,3] Student, Department of Computer Engineering, SCTR's Pune Institute of Computer Technology, Pune, India
[4]Assistant, Professor, Department of Computer Engineering, SCTR's Pune Institute of Computer Technology, Pune, India

Abstract : Sentiment analysis is the automated extraction of opinions of positive or negative nature from online text. It has gained considerable recognition from testers during this decade. Moreover, the exponential growth of internet users has been advancing rapidly parallel to rising technologies; that laboriously use online review spots, social networks and particular blogs to express their opinions. They harbor positive and negative opinions about people, associations, places, events, and ideas. The tools handed by natural language processing and machine literacy along with other approaches to work with large volumes of text, make it possible to begin rooting sentiments from social media. In this paper we talk over some of the challenges in sentiment extraction, some of the approaches that have been taken to address these challenges and our approach that analyzes sentiments from Twitter social media which not just gives the output beyond the polarity, but uses those polarities in product profiling, trend analysis and soothsaying. Promising results have shown that the approach can be further developed to feed business terrain needs through sentiment analysis in social media.

IndexTerms **- Sentiment Analysis, Natural Language Processing, Data Mining, Supervised Learning**

I. INTRODUCTION

People form conclusions about the world around them. They make positive and negative opinions about people, products, places and events. These kinds of opinions can be considered as sentiments. Sentiment analysis is the study of automated procedures for extracting sentiments from written languages. Growth of social media has redound in an explosion of publicly available, user generated text on the World Wide Web. These data and information can potentially be employed to provide real-time cognizance about the sentiments of people.

Blogs, online forums, comment sections on social media websites and social networking websites such as Facebook and Twitter can all be considered as social media. These social media websites can apprehend millions of peoples' views or word of mouth. Communication and the availability of these real time opinions from people around the world have made an innovation in computational linguistics and social network analysis. Social media is starting to become an increasingly important source of information for companies. People are more keen and happier to convey memories and accounts of their lives, knowledge, experiences and thoughts with the entire world through social media more than ever. They eagerly participate in social media discussions by expressing their opinions and comments. This way of sharing knowledge and emotions with the world drives businesses to collect more opinions of people on them, their products and to know how reputed they are and therefore take decisions accordingly. Therefore, it is clear that sentiment analysis is a principal element of leading innovative customer experience management and customer relationship marketing focused enterprises. Businesses are looking to automate the process of filtering out the noise, understanding the conversations, identifying the relevant content and taking appropriate action upon it. Many are now looking into the field of sentiment analysis. In the current era, having access to a surplus of information is no longer an issue. In this era, information has become the main trading object for many businesses. If we can create and employ mechanisms to search and retrieve relevant data and information and collect them to convert it to knowledge with accuracy and timeliness, that is where we get the exact usage of this large volume of information available to us.

However, in several occurrences, the required information is not found in a structured format but in unstructured format like documents written in human languages. Languages are indeterminate. A lot of words have different meanings for the same spellings or sounds of the words. Moreover, some people use different jargon, slang communications and short forms of the words for their ease. Consequently, it is strenuous to measure the sentiments accurately in terms of their polarity and the subjectivity of sentiments.

A lot of solutions rely on effortless dichotomous terms to convey an opinion about a post. To address the above-mentioned problems in the area of sentiment analysis, this adequate, and it will not generate precise and up-to-the-date knowledge for and assemblage of sentiments. In order to get a precise comprehension of the context after analyzing a sentiment, it should thoroughly solve the aforementioned problems. Other systems that try to offer solutions for these issues are still under the testing phase. A few systems have also inspected sentiments from multiple languages, which addressed the language barrier drawbacks.

This paper exhibits sentiment analysis as a tool that can analyze sentiments on Twitter social media, addressing the aforementioned problems and an application to gain business insights

II. LITERATURE SURVEY

Sentiment analysis is the study and identification of the views or opinions of people using tools like natural language processing, text analysis, document formatting and computational logistics. In this world where most people are on social media posting their thoughts, it is important to identify between positive, negative and neutral thoughts. Even among negative messages, we can differentiate between criticism and hate speech so that we only need to deal with the messages that really matter.

Line et. al. [1] collected COR posts from January 1, 2020 to February 1, 2020 from the Chinese social networking site- 'Weibo'. They established definitions for COR stigma, built deep learning classifiers and implemented a training process for COR stigma detection. They used 3 models, namely 'BERT', 'TextCNN', and 'BilSTM. The 'BERT' model prevailed with an accuracy of 0.986, precision of 0.955, recall of 0.970 and f1-score of 0.962.

Staphord et. al. [2] dealt with showing the differences in public opinion for the topic of discussion-'Monkeypox' on twitter. They use various techniques to handle the data. In the first stage of their project, they collect and translate data, and then preprocess it. For preprocessing, they removed unnecessary data like retweets, punctuation marks, user tags, emojis, hashtags, numbers, repeated words. The ones that could be changed into text format were replaced and the others were removed. For normalization, stemming and lemmatization were used. In the second stage, sentiment scores were calculated using VADER and TextBlob , vectorization of tokens was done. For the final stage, classification models were built using machine learning algorithms like Random Forest, Decision trees, Naïve Bayes, KNN, MLP, SVM, Logistic Regression. SVM model performed the best when compared to other models with accuracy of 0.9348.

[3] Bo et. al. proposed an experimental probability model 'CASA' to investigate the problem of social media sentiment analysis. Their model was tested on a twitter dataset. The model reduces the cost of manual data tagging by using untagged data to mine latent factors from the huge dataset. The model dominated its competing algorithms by getting an accuracy around 0.7, macro-precision of 0.7, macro-recall around 0.7 and macro-F1 around 0.7.

[4] Usman et. al. used the COVIDSENTI dataset to analyze user sentiments about the COVID -19 situation. To handle this, they took various approaches like machine learning classifiers, deep learning classifiers, hybrid models and transformer-based language models. Among these models and classifiers, 'BERT', 'DCNN with GloVe', 'fastText' performed the best in their respective categories, scoring 0.948, 0.869, 0.845 respectively.

[5] Mikolaj et. al. proposed solutions to analyze sentiments of cryptocurrency related social media posts using a corpus of data from Twitter, Facebook, Reddit, StockTwits. The first solution focused on training and finetuning a model built using a BERT architecture. The second solution classified emojis into either 'bullish' or 'boorish' using an SVM model. Among the models in the first solution, the CRYPTOBERT XL performed the best with an accuracy of 0.58, precision of 0.51, F1-score of 0.58 and recall of 0.61. Among the models used in the second solution, CRYPTOBERT performed the best with an accuracy of 0.60, precision of 0.46, F1-score of 0.55 and precision of 0.60.

[6] Ashima et. al. worked on improving sentiment analysis by using traditional methods and by shortening lengthened words. Experimentation was done on data taken from 'Facebook', 'Twitter', and personal chat messages. The proposed system did much better than the traditional system by an average of 22% increase in its performance metrics.

[7] Ernesto et. al. performed sentiment analysis on tweets related to the Taliban occupied Afghanistan situation. To accomplish this, they used three deep learning models called CNN, CNN-LSTM and GRU, then applied feature engineering techniques using machine learning algorithms like LR. The top performers were SVM and CNN-LSTM with an accuracy score of 0.97 for SVM.

[8] Ernesto et.al. built a racism detection system using the concepts of sentiment analysis. Tweets with negative sentiments are checked to see if the underlying message is racist or not. They took up an ensemble approach by combining CNN, GRU and RCNN to form a GCR-NN model. When compared with SVM and LR, GCR-NN does a better job at identifying racist tweets. The correct racism detection rates for SVM, LR and GCR-NN are 96, 95 and 97 respectively.

[9] Zhihua et. al. took up the task of improving detection of depression on Chinese social media networking sites. They collected the data from a 'Weibo' dataset to study the connection between human language and depression. It was observed that LightGBM performed the best at detecting depression with AUC score of 0.981, ACC score of 0.9614 and F1-Score of 0.9278 at k=300.

[10] Richard et.al. created a sentiment analysis system to identify discussion if HIV/AIDS through tweets and distinguish between positive, neutral and negative aspects of the tweets. They found out that the most dominant sentiment was negative (41.2%), tailed by positive (40.6%) and neutral (18.2%) coming in at last. 2 lexicon-based classifiers were used, namely VADER and TextBlob. VADER performed the best with the weighted average precision of 91%, weighted average recall of 90% and weighted average F1-Score of 90%.

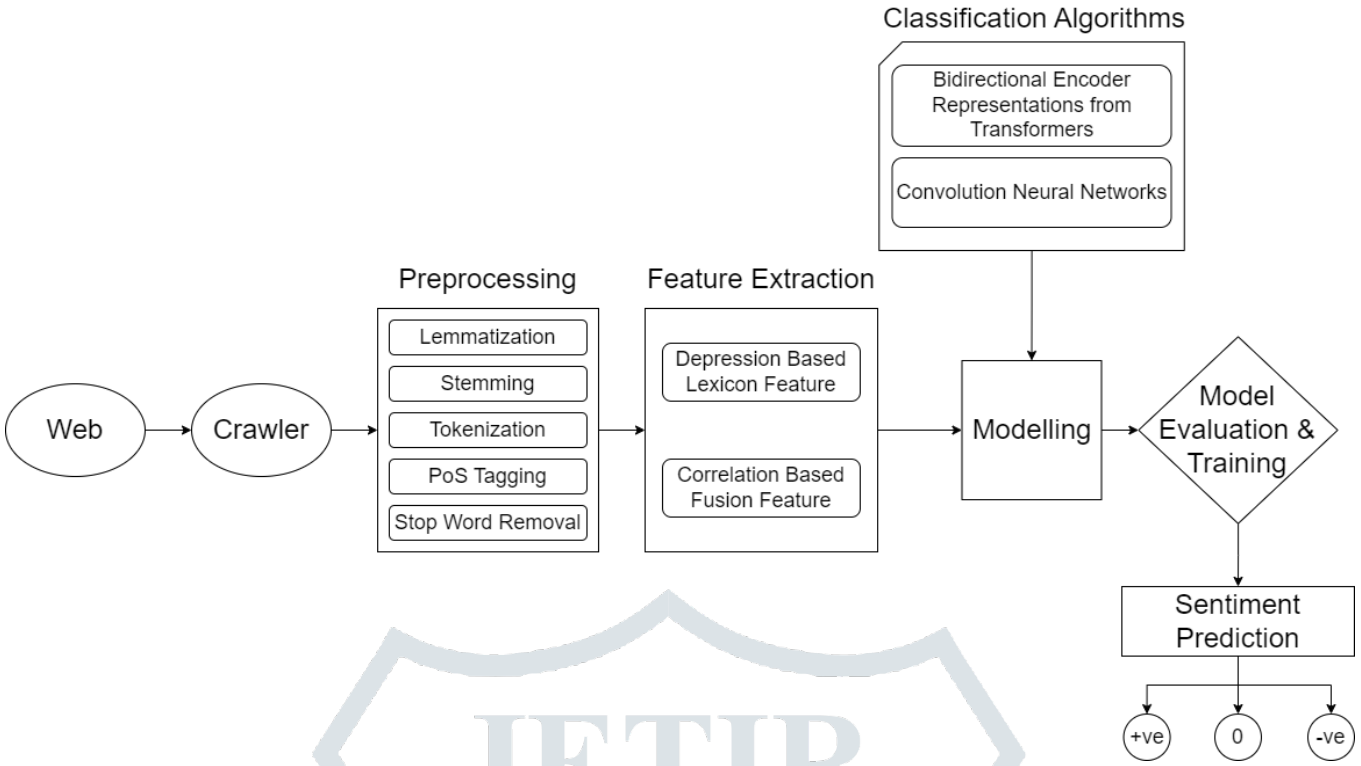
Sr No.	Paper	Best Classification Technique	Other Techniques Used	Dataset	Performance Parameter
1	A Deep Learning Approach for Semantic Analysis of COVID-19-Related Stigma on Social Media	BERT	TextCNN, BiLSTM	COR Weibo Posts dataset	TF-IDF, Accuracy, Precision, Recall, F1-Score
2.	A Machine Learning-Sentiment Analysis on Monkeypox Outbreak: An Extensive Dataset to Show the Polarity of Public Opinion From Twitter Tweets	SVM	Logistic Regression, Random Forest, Naive Bayes, KNN, XGBoost, MLP	Twitter Dataset	TF-IDF, Accuracy, Precision, Recall, F1-Score
3.	Context-Aware Social Media User Sentiment Analysis	CASA	CASA	Twitter Dataset	Accuracy, Macro-precision, Macro-recall, Macro-F
4.	COVIDSenti: A Large-Scale	BERT, DCNN	TF-IDF, SVM, RF,	COVIDSENTI dataset	TF-IDF, Accuracy,

	Benchmark Twitter Data Set for COVID-19 Sentiment Analysis	with GloVe, fastText	NB, DT, Word2Vec, GloVe, IWV, HyRank, distilBERT. XLNET, ALBERT		Precision, Recall, F1-Score
5.	Sentiment Classification of Cryptocurrency-Related Social Media Posts	CryptoBERT XL(StockTwits), BERTweet XL(emojis)	VADER, BERT, FinBERT, BERTweet, CryptoBERT, BERTweet XL, LUKE single emojis, LUKE with pairs	Cryptocurrency social media corpus(twitter, reddit, telegram, StockTwits)	Accuracy, Precision, Recall, F1-Score
6.	Improving Sentiment Analysis in Social Media by Handling Lengthened Words	Neural Network Approach	Traditional Approach- SOPMI, RF, DT, BN, LR, SVM, ME, EL, Neural Network Approach- BoW, Word2Vec, ANN, CNN, RNN, GRU, LSTM, Hybrid Neural Networks	Informal chats dataset	Precision, Recall, F-Measure, Senti-score
7.	Inquest of Current Situation in Afghanistan Under Taliban Rule Using Sentiment Analysis and Volume Analysis [7]	SVM, CNN-LSTMS	CNN, CNN-LSTM, GRU, SVM, LR, ETC, GNB, KNN, LSTMS	Afghanistan situation twitter dataset	Accuracy, Precision, Recall, F1-Score
8.	Racism Detection by Analyzing Differential Opinions Through Sentiment Analysis of Tweets Using Stacked Ensemble GCR-NN Model	GCR-NN	RF,DT, LR, KNN, GCR-NN(GRU, CNN & RNN)	Racist twitter posts dataset	Precision, Recall, F1-Score
9.	Leveraging Domain Knowledge to Improve Depression Detection on Chinese Social Media	LightGBM	LR, KNN, DT, RF, SVM, XGBoost, LightGBM	Weibo dataset	ACC, F1-Score
10.	Using Machine Learning to Establish the Concerns of Persons With HIV/AIDS During the COVID-19 Pandemic From Their Tweets	VADER	VADER, TextBlob,	Twitter dataset	Precision, Recall, F1-Score

Approach

To analyze sentiments and then come to a conclusion through them, we need to have enough sentiments in the correct format. There are thousands and millions of sentiment data on the web, especially in social media sites that can be used to get valuable conclusions. But they are too unstructured to get any valuable usage out of them. The text has to be converted into an interpretable way in order for us to proceed. It is the first part in our approach, which is developing a crawler to crawl data from Twitter social media. The Crawler should be able to crawl user sentiments from twitter and at the same time get user details in order to do product profiling for customers as the later part of the whole approach.

After gaining access to a plethora of data in structured format using a crawler and a database, the next thing on the agenda is to analyze sentiments. Sentiments come in various different languages. We will be covering the English language. Analyzing sentiments is a way of processing natural languages, therefore this part is about natural language processing. We will be using Natural Language Toolkit(NLTK), which is a widely used Python library that mainly deals with natural language processing tasks. There are different ways that we can use to analyze sentiment data using this toolkit, but none of them gives a hundred percent accuracy, because natural languages are used in many different ways by people. In our method, we will have implemented two supervised machine learning algorithms - Naïve Bayes Classification and Maximum Entropy Classification. They give the probability of a sentiment's polarity. Another method will have been used with SentiWordNet, which is a lexical database that assigns scores to words and as such, finds the sentiment score of an entire sentence. The first method, which is Naïve Bayes Classification, was chosen as the best method out of these three techniques after evaluating all the three methods. Using the selected sentiment analyzing method in this approach, it not only gives positive, negative and neutral sentiment scores to the user sentiments, but it can solve the issues of using short words, different jargon words and smileys in social media. To differentiate the senses of ambiguous words such as “apple”, it used word sense disambiguation techniques and was able to differentiate different senses for ambiguous words.



Architecture Diagram

For the third part of the project, we will have created a dashboard to show the results from the crawler and sentiment analysis. The dashboard will display how the sentiment polarity varies for a selected item against time using a graph. We can visualize from the dashboard how the user sentiment polarity of a specified brand or product is changing with time. For the final part, the output of the sentiment analysis module will be used as the input for the data mining module. It will use the sentiment scores of a particular product or service with the user information such as age, profession, area and gender to profile products, analyze the trend for that particular product or service and forecast. The result from data mining can be applicable in a most profitable way such as analyzing how people's sentiments change and will change for products and services with their age, location, profession and gender.

Businesses that are interested in knowing what people think about a particular product or service will be able to use this system. They can gain many insights from this system such as reception received by a product, getting genuine criticisms of their product from a stream of unfiltered comments, and much more.

III. Conclusion

It is a very important fact to analyze how people think in different contexts about different things. This becomes more important when it comes to the business world because business is dependent on their customers and they always try to make products or services in order to fulfill customer requirements. So knowing what they want, what they think and talk about existing products, services and brands is more useful for businesses to make decisions such as identifying competitors and analyzing trends. Both because people express their ideas on social media and they can access those data, it has been enabled in some way to do the above mentioned things by using those data. Using the sentiment scores for sentiments regarding a particular product or service with the user's information, it could successfully profile the products, analyze trends and forecasting. So, overall, the system is capable of saying how a set of people of a particular age range, particular area with a particular profession thinks about a particular product or service and how it will change the future, which is most useful information when it comes to the business world.

References

[1] L. Liu, Z. Cao, P. Zhao, P. J. -H. Hu, D. D. Zeng and Y. Luo, "A Deep Learning Approach for Semantic Analysis of COVID-19-Related Stigma on Social Media," in IEEE Transactions on Computational Social Systems, vol. 10, no. 1, pp. 246-254, Feb. 2023, doi:10.1109/TCSS.2022.3145404.

[2] S. Bengesi, T. Oladunni, R. Olusegun and H. Audu, "A Machine Learning-Sentiment Analysis on Monkeypox Outbreak: An Extensive Dataset to Show the Polarity of Public Opinion From Twitter Tweets," in IEEE Access, vol. 11, pp. 11811-11826, 2023, doi: 10.1109/ACCESS.2023.3242290.

[3] B. Liu et al., "Context-aware social media user sentiment analysis," in Tsinghua Science and Technology, vol. 25, no. 4, pp. 528-541, Aug. 2020, doi: 10.26599/TST.2019.9010021.

[4] U. Naseem, I. Razzak, M. Khushi, P. W. Eklund and J. Kim, "COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis," in IEEE Transactions on Computational Social Systems, vol. 8, no. 4, pp. 1003-1015, Aug. 2021, doi: 10.1109/TCSS.2021.3051189.

[5] M. Kulakowski and F. Frasincar, "Sentiment Classification of Cryptocurrency-Related Social Media Posts," in IEEE Intelligent Systems, vol. 38, no. 4, pp. 5-9, July-Aug. 2023, doi: 10.1109/MIS.2023.3283170.

[6] A. Kukkar, R. Mohana, A. Sharma, A. Nayyar and M. A. Shah, "Improving Sentiment Analysis in Social Media by Handling Lengthened Words," in IEEE Access, vol. 11, pp. 9775-9788, 2023, doi: 10.1109/ACCESS.2023.3238366.

[7] E. Lee, F. Rustam, I. Ashraf, P. B. Washington, M. Narra and R. Shafique, "Inquest of Current Situation in Afghanistan Under Taliban Rule Using Sentiment Analysis and Volume Analysis," in IEEE Access, vol. 10, pp. 10333-10348, 2022, doi: 10.1109/ACCESS.2022.3144659.

[8] E. Lee, F. Rustam, P. B. Washington, F. E. Barakaz, W. Aljedaani and I. Ashraf, "Racism Detection by Analyzing Differential Opinions Through Sentiment Analysis of Tweets Using Stacked Ensemble GCR-NN Model," in IEEE Access, vol. 10, pp. 9717-9728, 2022, doi: 10.1109/ACCESS.2022.3144266.

[9] Z. Guo, N. Ding, M. Zhai, Z. Zhang and Z. Li, "Leveraging Domain Knowledge to Improve Depression Detection on Chinese Social Media," in IEEE Transactions on Computational Social Systems, vol. 10, no. 4, pp. 1528-1536, Aug. 2023, doi: 10.1109/TCSS.2023.3267183.

[10] R. K. Lomotey, S. Kumi, M. Hilton, R. Orji and R. Deters, "Using Machine Learning to Establish the Concerns of Persons With HIV/AIDS During the COVID-19 Pandemic From Their Tweets," in IEEE Access, vol. 11, pp. 37570-37601, 2023, doi: 10.1109/ACCESS.2023.3267050.