International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Twitter Data Radicalization Using Sentiment Analysis Techniques

Vivek Kumar[1], Mr. Satish Singh Verma[2] , Rohit Yadav[3], Chandan singh[4], Karamveer kumar[5]

[1,3,4,5]Students,[2]Assistant Professor

[1,2,3,4,5]Computer Science & Engineering, Babu Banarasi Das Institute of Technology & Management Lucknow, 226028, India

ABSTRACT

Social networks are the main resources to gather information about people spend hours daily on social media and share their opinion. Twitter is one of the special offers organizations a fast and effective way to analyze customers' perspectives toward Developing a program for sentiment analysis is an approach to be used to computation use natural language processing and machine learning concepts to create a model to how we can create a model for analysis of tweets which is trained by various Approach.

Keywords: Machine Learning, Anaconda, python, positive, negative, Social media,Natural Language Processing.

1. INTRODUCTION

What do we do when we want to express ourselves of reach out to a large audience? We log on to one of our favourite social media sites. Social Media has taken over in today's world and Twitter is one of the biggest platform where we express our views, emotions, opinion about a specific point. However, it's important to note that as a public platform, Twitter can also be a space where disagreements and conflicts arise. Due to the ease of sharing opinions, diverse perspectives can lead to debates and discussions, sometimes turning contentious. Nonetheless, Twitter continues to be a significant platform for individuals to express themselves and engage with others on a wide array of subjects.

(1) These emotions are used in various analytics for better understanding of humans.. In contrast, consumers have all the power when it comes to what consumers want to see and how consumers respond. By this, the company’s success and failure is publicly shared. However, the social network can change the behaviour and decision making of consumers.

(2) In this paper, we have attempted to create and trained a model for analysis on "Tweets" using machine learning and natural language processing . We try to simplify the polarity of the tweet where it is either radical or non-radical .If the tweet has both radical and non-radical elements, the more strong sentiment should be picked as the final label.

A. Natural Language Processing

Natural Language Processing (NLP) is a branch of Artificial Intelligence (AI) that focuses on the interaction between computers and human language, with the goal of enabling machines to understand, interpret, and generate human language in a meaningful way. Natural Language can be in form of test or sound, which are used for humans to communicate each other.NLP can enable humans to communicate to machines in a natural way. Modern NLP algorithms heavily rely on machine learning techniques Machine learning is indeed a subfield of artificial intelligence that focuses on the development of algorithms and models capable of learning from data without being explicitly programmed. Instead of relying on explicit instructions, machine learning algorithms use statistical techniques to automatically identify patterns, make predictions, and improve performance over time through exposure to relevant data.Tasks typically involved the direct hand coding of large sets of rules 7 instead for using general learning algorithms - often, although not always, grounded in static learn such rules through the analysis of large corpora of typical real-world examplesA corpus is a collection of documents or sentences that have been annotated with the correct values to be learned. examples in the corpus, allowing them to make predictions or perform tasks on unseen data based on the patterns they have learned.We are proposing a model which will take data from this social media platform. It will identify potential people who are on the edge of being radicalized and will associate a radicalization quotient with them to know their degree of radicalization, so that appropriate measures can be accordingly initiated. The degree of radicalization (0 – non radicalized and 1 – highly radicalized) will depict the extent to which a person might be vulnerable to being radicalized. We will also have an effective User Interface that will take data from Twitter API as input and will return a radicalization quotient. The User Interface will itself send an alert to the respective By analyzing such corpora, machine learning algorithms can identify statistical patterns and relationships within the data. These algorithms learn from the authorities (counselors / police) in the nearby area, based on the radicalization quotient of the user id, using the GPS. If it finds generous number of user belonging to a particular area, it might declare the area as a red zone and send an alert to the higher authorities as well.

2. LITERATURE REVIEW

(1) Analysis for Social Media : Analysis is a problem of text based analysis, but there are son as compared to traditional text based analysis This clearly shows that there are several opportunities for future research for negations, hidden polarity.

(2) Analysis Prototype System for Social Network Data: This paper discusses a prototype bused analysis of social network data for evaluation of the public social conscience of user provided topics and events.

(3) Analysis of Tweets Using Machine Learning Approach : The research of sentiment analysis in different aspects. This paper shows analysis types and techniques used to perform extraction survey paper, we trained a model using machine learning and nip concepts.

(4) Opinion Mining on Social Media Data 6: Po-Wei Liang etal used Twitter API to collect data from twitter. Tweets which contain opinions were filtered out. for polarity identification. They also worked for deletion of unwanted features by using the Mutual Information and Chi square feature extraction method. Finally, the approach for predicting the tweets as positive or negative did not give better accuracy by this method.

(5) Analysis for Using Twitter Data Set: The research of sentiment analysis of Twitter data can be performed in different aspects. This paper shows how we can create a model and trained it for sentiment analysis t used to perform extraction of sentiment from tweets.

3. METHODOLOGY

The collected tweets will be subjected to preprocessing. Applying a supervised algorithm like Support Vector Machine (SVM) to the stored data for sentiment analysis is a common approach in NLP tasks. SVM is known for its effectiveness in handling classification problems, including sentiment analysis. Using SVM, you can train a model on the labeled data, where the sentiment (positive, negative, or neutral) serves as the target variable. The trained model can then be used to predict the sentiment of new, unseen tweets. Anaconda is a popular distribution of the Python and R programming languages that is widely used in data science and machine learning applications. It provides a convenient platform for managing and working with the libraries and tools commonly used in these fields. With Anaconda, you can easily set up and manage Python environments for your projects. Environments allow you to isolate and control the packages and dependencies required for different projects, ensuring reproducibility and avoiding conflicts between different versions of libraries. We try to simplify the polarity of the tweet where it is either radical or non-radical .If the tweet has both radical and non-radical elements, the more strong sentiment should be picked as the final label.

We use the dataset from twitter which was label radical /non radical. The analysis of social media is being done for a number of purposes today. We found that work has been done in the field of identification of radicalization phenomenon as well. Our goal now was to create a model, by adding some new features to it so that it has a better accuracy and that it can process data in real time using the Twitter API. We will begin by extracting a generous amount of data from the Twitter API that will consist of both radicalized and normal tweets. We will extract multiple tweets for each of the different user ids and would perform text summarization on the tweets originating from a single id for generating a precise and concise summary of voluminous texts while at the same time preserving the key informational elements. We will then perform refinement and preprocessing of the collected data which would involve data labeling, data pruning etc. Further we will go for feature extraction and selection and preparation of training and testing dataset. Next we will build a model using deep learning methods and we are also willing to add some handcrafted features such as entropies, parts of speech, N-gram etc. which will enhance the performance of our model. We are targeting the deep learning models along with some embedding techniques like GloVe, FastText etc. Our model will go on to predict the radicalization quotient associated with a particular user id on Twitter, based on certain keywords, which would range between 0 and 1 (0 – non radicalized and 1 – highly radicalized). Upon detection of such a user id, it will automatically raise an alarm to the associated authorities based on the degree of radicalization given by the radicalization quotient using the GPS.

DATA DESCRIPTION

The data given from the dataset is in the form of comma separated values. The training dataset is a CSV (comma separated value) file of type tweet-id, sentiment, tweet where the tweet-id is a unique integer identifying the tweet, sentiment is either 1 (positive) or O (negative), and tweet is the tweet enclosed in ''. Similarly, the test dataset is a CSV (Comma-Separated Values) format and consists of tweet data, including tweet IDs and the corresponding tweets themselves. It is common for Twitter datasets to contain a mixture of words, emoticons, symbols, URLs, and references to people, reflecting the nature of tweets on the platform. Words and emoticons contribute to predicting the sentiment, but URLs and references to people don't. Therefore, URLs and references are being ignored. The words are also a mixture of misspelled words / incorrect, extra punctuations, and words with many repeated letters. The "tweets", therefore, must be pre-processed to standardize the dataset. The provided training and test dataset have 32000 tweets respectively. Preliminary statistical analysis of the contents of datasets, after pre processing.

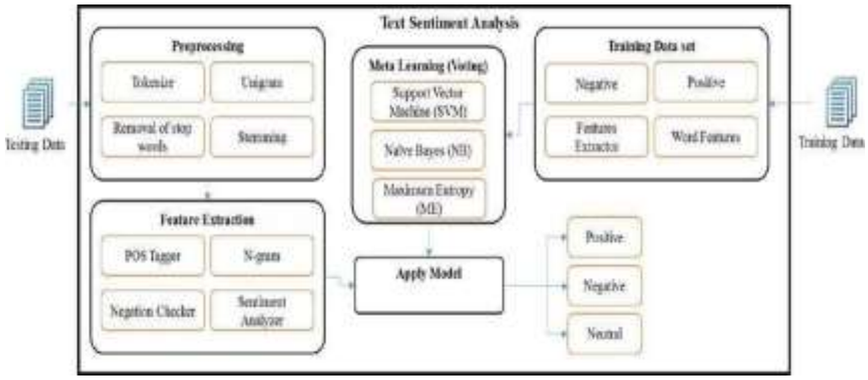


Fig. 1 – Methodology

3.1 Problem Statement

The machine learning algorithm uses linguistic features by an objective of system's performance optimization with the example data. The big data models like Pentaho and Mahout consists of plug-ins and library related to the machine learning algorithm that is evaluated for performing the classification of sentiment. In evaluation of big data, the user must define the method type, which should be given to the data and that method has been performed using big data analytics tools in order to solve the certain problem like predictive analytics. In most of the cases, two document sets are needed for performing the machine learning-based categorization. These sets are considered as the training as well as testing sets. In order to learn the features of the document, the training set has been given to the machine learning algorithm, and for evaluating the performance of the classifier, testing set is employed.

By using the machine learning algorithms, the text classification models are split into unsupervised as well as supervised learning algorithms. The unsupervised learning models are employed in finding the training documents, which are quite complex.

The supervised learning models employ more amounts of label training documents. Moreover, these supervised algorithms attain satisfactory efficiency but those are generally language dependent and domain specific. Moreover, these algorithms need label information that is frequently labour intensive. In the mean time, the unsupervised algorithms have more demand as the freely accessible information is frequently unlabelled and therefore best solutions are required. At that time, semi-supervised learning algorithm is developed and produces best results in classifying the sentiments. In order to construct best learning methods in unsupervised learning algorithms, it requires more amount of label as well as unlabelled data.

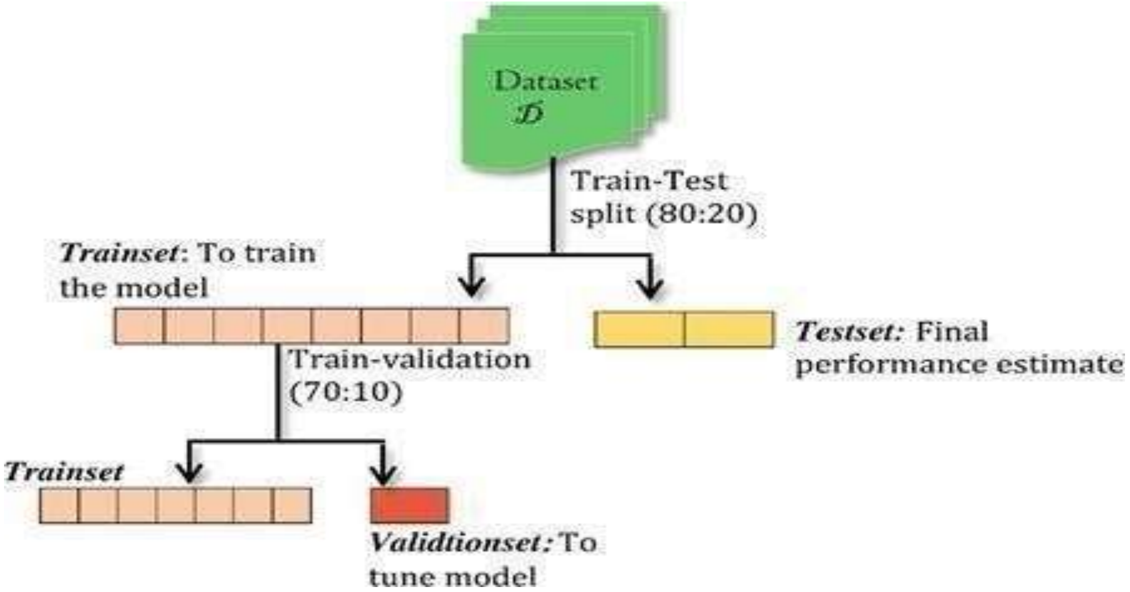
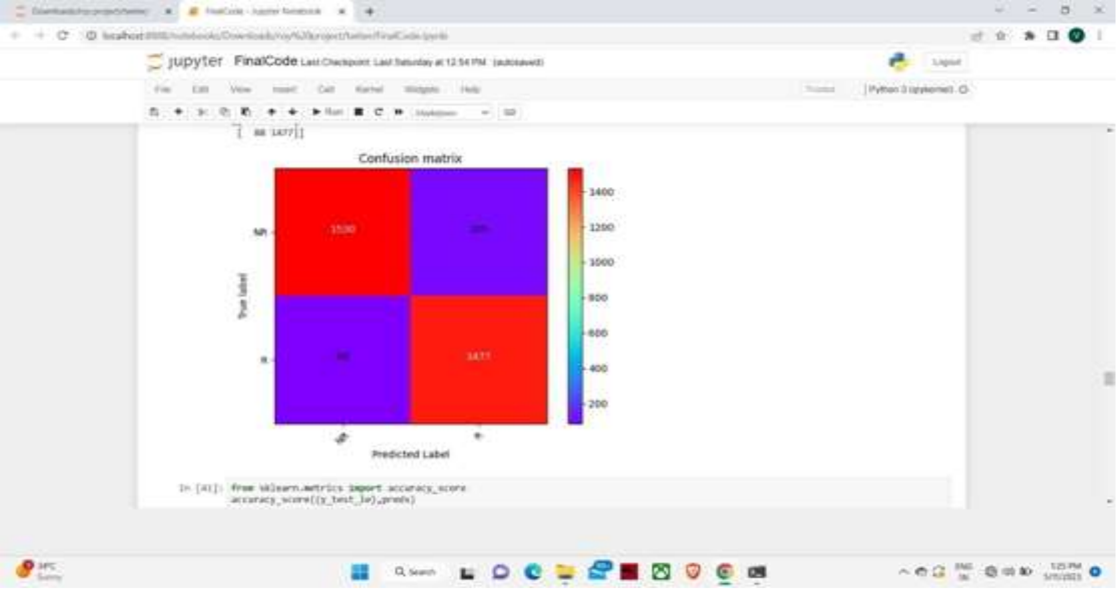

Dataset
$\mathcal{D}$
Train-Test split (80:20)
Trainset: To train the model
Testset: Final performance estimate
Train-validation (70:10)
Trainset
Validtionset: To tune model

Fig 2- Validation

3.2 Proposed Approach

In sentiment analysis, many machine learning algorithms have been employed for classification. The famous machine learning algorithms, which have attained more successful in classifying the text are NB, ME, and SVM. Many of the conventional models are associated with the public-related sentiments from social network, and text applications..However, there is less amount of work, in which the ontology and semantics are seemed to be more important research works in sentiment analysis. Now-a-days, the conventional research models are experimented using the public review datasets.

However, this kind of review has not been evaluated keenly by concerning the sentiments .By categorizing the sentiments as positive or negative, it will not provide the original and the concealed information beyond the actual concepts of sentiments. In addition, there are some specific sentences that are quite complicating and accurate classification cannot be performed. There have been many constraints such as assessment of sentiment during the review, and document exploration using many subjects. Moreover, traditional models concentrated on major problems rather than minor ones, in which the accuracy is not seemed to be optimized. It is observed that there are few research methodologies, which have been recognized as standard models. There is only little number of researches other than standard models, whose results are seemed to be less efficiency over a suitable approach. The evaluation of less dimension text might utilize less number of resources. However, the collection of sentiments from the collaborative environment will utilize more amounts of resources. It is unfortunate that in the conventional researches, the authors didn't found several confirmations related to the computational expenses of efficient techniques for performing huge data sentiment analysis.



```
In [41]: from sklearn.metrics import accuracy_score
         accuracy_score((y_test_lw),preds)
```

4. CONCLUSION

The present paper has developed the review of earlier contributions with various machine learning models using discrete information. The present review has explored several of research works that covered various implementations employed for sentiment analysis. Initially, the assessment has concentrated on clarifying the contribution of every task and observed the type of machine learning algorithm utilized. The evaluation also focused in recognizing the type of data employed. Later, the environment utilized and the performance metrics covered in each contribution was analyzed. In this technical paper, we discussed the methods of creating model for tweets analysis. We showed the res model ... The model were built can easy be scaled for new algorithms be it in Machine Learning, De Language Processing. Sentiment analysis system is an active field of research and we can still further in working more on the algorithms, tying out different things in pre-processing and checking which ones gets the best precision metrics.

In conclusion, while sentiment analysis techniques can provide valuable insights, they should be seen as part of a broader toolkit for understanding Twitter data radicalization. By combining sentiment analysis with other analytical approaches and human expertise, we can gain a more comprehensive understanding of the issue and develop more effective strategies to address it.

5. FUTURE WORK

Sentiment Analysis Techniques: Experiment with different sentiment analysis techniques to accurately classify tweets as positive, negative, or neutral. Explore both rule-based approaches (e.g., using predefined sentiment lexicons) and machine learning-based methods (e.g., utilizing deep learning models) to identify the most effective approach for your research.

Radicalization Indicators: Develop a set of indicators or features that can help identify tweets associated with radicalization. These may include specific keywords, mentions of extremist ideologies or figures, aggressive language, or other contextual cues. Combine sentiment analysis with these indicators to detect potentially radicalized content.

Model Training and Evaluation: Train and fine-tune your sentiment analysis models using annotated data that reflects the sentiment and radicalization labels of the tweets. Evaluate the performance of your models using appropriate metrics such as accuracy, precision, recall, and F1-score.

Understanding Radicalization Dynamics: Analyze the relationship between sentiment and radicalization. Examine whether there is a correlation between negative sentiment and the likelihood of radicalization. Explore how sentiment evolves over time within specific radicalization pathways or extremist communities.

Network Analysis: Incorporate network analysis techniques to study the propagation of radicalized content on Twitter. Analyze the connections between users, identify influential accounts, and explore the role of sentiment in spreading radical ideas within social networks.

Mitigation Strategies: Investigate potential strategies for mitigating the spread of radicalized content on Twitter. For example, you could explore the use of sentiment analysis to flag or prioritize content for human moderation, or develop algorithms that can detect and suppress radicalized tweets from appearing in users' feeds.

Remember that studying radicalization and its relationship with sentiment analysis is a complex and sensitive topic. It's important to approach it with caution, adhering to ethical guidelines and considering potential biases in the data and models used.

References

Mohammad A.Hassonah, RizikAl-Sayyed, AliRodan, Ala' M.Al-Zoubi, IbrahimAljarah, and HossamFaris, "An efficient hybrid filter and evolutionary wrapper approach for sentiment analysis of various topics on Twitter", Knowledge- Based Systems, vol. 192, 15 March 2020.

FengXu, ZhenchunPan, and RuiXia, "E-commerce product review sentiment classification based on a naïve Bayes continuous learning framework", Information Processing & Management, Available online 13 February 2020**.**

Deepa, K. & Sangita, H. & Shruthi, H.. (2022). Sentiment Analysis of Twitter Data Using Machine Learning. 10.1007/978-981-19-2177-3_26.

Gupta, Itisha & Joshi, Nisheeth. (2021). A Review on Negation Role in Twitter Sentiment Analysis. International Journal of Healthcare Information Systems and Informatics. 16. 1-19. 10.4018/IJHISI.20211001.oa14.

Arun, K. & Srinagesh, Ayyagari. (2020). Multi-lingual Twitter sentiment analysis using machine learning. International Journal of Electrical and Computer Engineering (IJECE). 10. 5992. 10.11591/ijece.v10i6.pp5992-6000.

Mollah, Md. (2022). An LSTM model for Twitter Sentiment Analysis. 10.48550/arXiv.2212.01791.

Samih, Amina & Ghadi, Abderrahim & Fennan, Abdelhadi. (2022). IWVTSA: IMPROVED WORDS VECTORS FOR TWITTER SENTIMENTS ANALYSIS. Journal of Theoretical and Applied Information Technology.

Singh, Aaryan & Srivastava, Harsh & Aman, Mohd & Dubey, Gaurav. (2023). Sentiment Analysis on User Feedback of a Social Media Platform. 826-832. 10.1109/ICSCDS56580.2023.10105082.

Parameswarappa, Priya & Saideep, Sunkari & Bejgam, Rahul. (2022). Twitter Sentimental Analysis for Businesses Using Python Web Services in Salesforce Cloud. 1-7. 10.1109/INCOFT55651.2022.10094352