**Individual In-depth Report**

**Member name:** [Gautham Vijayaraj](mailto:gvijaya6@asu.edu)

**Evaluated by:** [Krupaben Kothadia](mailto:kkothadi@asu.edu)

**Date: 09/17/2023**

**Tasks Assigned:**

* Focus on my question to work upon as my area of research for this project.
* The question I chose was to do my research and find a system that maintains the balance between detecting suspicious activity in social media and safeguarding the user's privacy and maintaining data integrity.
* Review of [**Sentiments Analysis Of Twitter Data Using Data Mining**](https://drive.google.com/file/d/1SGNHuQyCJ--drDbYgOgs12MNIApcoHFT/view?usp=drive_link)

**Summary:**

* The main focus area of this paper is to find an approach to analyze the sentiments of users using data mining classifiers.
* Experimental results obtained demonstrate that k-nearest neighbor classifier gives very high predictive accuracy.
* According to the survey there are 304 million monthly active users of Twitter. Approximately 500 million tweets get tweeted per day on twitter.
* Existing systems propose a hybrid approach of extracting opinion using direct and indirect features of Twitter data based on Support Vector Machines (SVM), Naive Bayes, Maximum Entropy and Artificial Neural Networks based supervised classifiers.
* This paper presents a mechanism to predict the overall sentiments inclination of Indian people towards political situations and issues.
* After collecting raw tweets various preprocessing methods get applied to clean the data.
* The sentiment analysis is performed is using the following steps:
  + **Data Collection:** Training and testing tweets collected from twitter by using twitter searched API v 1.1 for various political leaders and parties in india.
  + **Preprocessing:** All userID,twitterId, userinfo from the tweets is removed. Duplicate tweets and retweets also get removed from the training dataset. After applying all cleaning methods text that is only with the tweeted text remains.
  + **Training DataSet and Testing Dataset:** For classifying a dataset, three classes, Positive, Negative, Neutral are used. To classify tweets into these categories, “SentiWordNet 3.0.0.” dictionary is used.
* Basically a single tweet is split into words, after splitting, the polarity of words is calculated from SentiWordNet 3.0.0. After deciding the polarity of each word all positive and negative words get added separately. And then comparison between positive words polarity with negative words polarity gets done.
* If sentences have more positive polarity then classify sentence(tweet) as positive polarity. Same for negative and neutral polarity. Duplicate tweets were not considered in a training dataset.

**Outcome:**

* The polarity value of the tweets can be further taken into consideration for finding suspicious activities in Twitter.
* 2,102,52 tweets were collected about various political leaders and parties. k-nearest neighbor gives the highest accuracy of all classifiers with an **accuracy of 99.6456%.**

**References** *(with citation)*  
  
[9] Anurag P. Jain, Vijay D. Katkar, “Sentiments analysis of Twitter data using data mining,” in *“International Conference on Information Processing (ICIP)”*, Pune, India, 2015, pp.807-810

**Evaluation of Report**

**Evaluation summary with justification.**

This paper explores sentiment analysis of Indian political tweets using data mining classifiers. It achieves high predictive accuracy with the k-nearest neighbor classifier. Given Twitter's extensive user base and tweet volume, this research is crucial. It introduces a hybrid approach with supervised classifiers, offering insights into Indian users' political sentiments.

**The quality of the major result(s) with justification.**  
Result Quality: Rigorous experimentation and data mining techniques establish the k-nearest neighbor classifier's robust and reliable high predictive accuracy.

**The usefulness of the paper to the overall project.**   
 It discusses the data mining technique to process the twitter dataset and also it explores one of the varied dataset: an Indian twitter dataset.

**Other comments**

None

**Evaluation Approval  
  
Evaluation by:** [Krupaben Kothadia](mailto:kkothadi@asu.edu) **Date: 09/18/2023**

**Is the written report of the in-depth study complete with all the major result(s) of the paper(s)? If not, provide as many examples of the major result(s) missing in the written report as possible. (in bullet form). [Normally within 100 words]**Yes, because the report discusses the data mining pipeline as a whole by providing a proper explanation of each step, along with the results.

**Is each section of the guidelines sufficiently completed? If not, point out what is missing. [Normally within 40 words].**

Yes

**Is the quality of this version of the written report satisfactory? If not, then why not? [Normally within 40 words]**

Yes, the quality of this version of the written report is satisfactory.

**Approval.  
  
Approved by:** [Krupaben Kothadia](mailto:kkothadi@asu.edu) **Date: 09/18/2023  
  
Is the quality of this written in-depth study report and Evaluation report satisfactory? If not, then why not? (limit: 40 words)**

Yes, the quality of the in-depth study report is satisfactory, as the approach is discussed clearly. Also, the evaluation report is satisfactory as the quality of content mentioned, denotes that the study report has been thoroughly understood by the evaluator.