**TWITTER SENTIMENTAL ANALYSIS USING HYBRID APPROACH**

**ABSTRACT**

Sentimental Analysis is the process of determining the opinion of the writer based on their messages on various social networking sites. Twitter is one of the famous social networking sites where user can read and post messages about a person, an event, a product and the current happenings all over the world. These are normally 140-280 characters in length. In this system, the tweets are used as the raw data. The tweets are collected through Twitter API using the secret tokens. Then they are preprocessed using text mining package to reduce the noise in words. The score is computed for each pre-processed tweet using Dictionary-Based Approach. For positive tweet the score is 1, for negative tweet the score is -1 and 0 for neutral tweet. The pre-processed tweets along with the scores are stored in CSV format for further process. The train data and test data is provided in the ratio 60:40 to construct the classification model. After classification, Convolutional Neural Network is implemented to compute the probability of the tweets. The system uses K-fold cross validation method to improve over the holdout method. Finally, as the result the opinion of the sentiment related to the given tweets is predicted using probability of the positive tweets by hybrid approach. This system produces a better performance measure when compared to other method.

|  |  |  |
| --- | --- | --- |
|  |  |  |

**CHAPTER 1**

**INTRODUCTION**

* 1. **SENTIMENTAL ANALYSIS**

Sentiment Analysis is the computational investigation of individuals' suppositions, attitudes and feelings toward an element that utilize data mining procedures and methods to extricate information for investigation to perceive the subjective sentiment of a record, similar to blog entries, audits, news articles and online networking nourishes like tweets and notices. The element may be people, occasions or subjects. Sentiment Analysis is not only applied on item surveys however can likewise be connected on stock markets [12],news articles [20], or political debates [9]. It is a progressing field of research in text mining.

There are three fundamental order levels in Sentiment Analysis: document-level, sentence-level, and aspect-level Sentiment Analysis. Document-level Sentiment Analysis aims to group a conclusion from an opinion document as communicating a positive or negative sentiment. It considers the entire document as a fundamental data unit. Sentence-level Sentiment Analysis expects to order conclusion communicated in each sentence. The initial step is to recognize whether the sentence is subjective or objective. It recognizes the sentiment communicated in a content at that point and then examines it. Aspect-level Sentiment Analysis aims to classify the sentiment with respect to the specific aspects of entities. In this manner, the sentiment analysis techniques automatically detect the polarity of text.

* + 1. **Dictionary-Based Approach**

Using a dictionary approach to compile sentiment words is an obvious approach because most dictionaries list synonyms and antonyms for each word. Thus, a simple technique in this approach is to use a few seed sentiment words to bootstrap based on the synonym and antonym structure of a dictionary. Specifically, this method works as follows: A small set of sentiment words with known positive or negative orientations is first collected manually, which is very easy. The algorithm then grows this set by searching in the online dictionary for their synonyms and antonyms. The latterly found words are added to the seed list. The next iteration begins. The iterative process ends when no more new words can be found. After the process completes, a manual inspection step was used to clean up the list.

**1.1.2 Deep Learning**

Deep learning has emerged as a powerful machine learning technique that learns multiple layers of representations or features of the data and produces state-of-the-art prediction results [17].The main concept in deep leaning algorithms is automating the extraction of representations from the data. Deep learning algorithms use a huge amount of unsupervised data to automatically extract complex representation. These algorithms are largely motivated by the field of artificial intelligence, which has the general goal of emulating the human brain’s ability to observe, analyze, learn, and make decisions, especially for extremely complex problems. Work pertaining to these complex challenges has been a key motivation behind Deep Learning algorithms, which strive to emulate the hierarchical learning approach of the human brain.

In contrast, Deep Learning architectures have the capability to generalize in non-local and global ways, generating learning patterns and relationships beyond immediate neighbors in the data. Deep learning is in fact an important step toward artificial intelligence. It extracts representations directly from unsupervised data without human interference. A key concept underlying Deep Learning methods is distributed representations of the data, in which a large number of possible configurations of the abstract features of the input data are feasible, allowing for a compact representation of each sample and leading to a richer generalization. The number of possible configurations is exponentially related to the number of extracted abstract features.

Deep learning algorithms lead to abstract representations because more abstract representations are often constructed based on less abstract ones. An important advantage of more abstract representations is that they can be invariant to the local changes in the input data. Learning such invariant features is an ongoing major goal in pattern recognition. Deep learning algorithms are actually Deep architectures of consecutive layers. Each layer applies a nonlinear transformation on its input and provides a representation in its output.

**CHAPTER 2**

**LITERATURE SURVEY**

1. **Sentiment Analysis on Social Media**

Carlo Aliprandi, Federico Neri, Federico Capeci ,et.al.

The web acts as a big platform where the users can share their opinions. Their reviews are very significant in providing good product, efficient plans. This paper proposes a system that uses facebook as a social media site to gather the opinion of the public through facebook posts [8]. 1000 facebook posts of La7 and Rail news programs of Italy are collected. This used a knowledge mining system used by Italy security agencies for government. K-Means clustering algorithm is applied for the clustering of news posts of La7 and Rail. To enhance the results, semantic and linguistic approaches are included. Each sentence is analyzed to assign the context with the appropriate meaning to provide accurate result. Then the posts are preprocessed by removing different Part Of Speech i.e. noun, verb, adverb, adjective to decrease the complexity. Sentimental analysis is performed through polarity mining and syntactic tree. The Bayesian learning method is used for classification process. Recall and precision is adopted as the evaluation measures for computing the performance of the proposed system.

**Disadvantage:**

* Only news data sets can be analyzed.

1. **Comparative analysis of Twitter data using supervised classifiers**

Joshi, Rohit, and Rajkumar Tekchandani.

Twitter achieves greatest measure of consideration in the field of research areas related to product, movie reviews, stock exchange, government policies etc. Twitter users vary from person to person, as user can be politician, film stars, celebrities, sportsman and many leaders across the country containing data of different languages [10]. The proposed system performs sentiment prediction on movie reviews using machine-learning algorithms. This uses supervised machine-learning algorithms like SVM, Max Ent and NB. The unigram, bigram and hybrid i.e. unigram + bigram features are used for classifying data. The system uses 15000 movie reviews, 5000 tweets for training set and 2000 tweets for test set. The stop words, repeated words and links are preprocessed to avoid unwanted process. Finally, the tweets are categorized into positive, negative and neutral reviews. Accuracy is used as the performance measure in this approach.SVM using hybrid feature outperforms all other classifiers and selection feature with accuracy of 84% for movie reviews.

**Disadvantage:**

* Difficult for large data sets.

1. **Sentiment Analysis Based on Dictionary Approach**

Apeksha Pande , Reshma Bhonde , Binita Bhagwat ,et.al.

Sentiment analysis has become upcoming topic in recent years, and sentiment dictionaries are essential for research in this field [3].Consumers regularly face the trade-off in purchase decisions so nowadays if one wants to buy a consumer product one prefer user reviews and discussion in public forums on web about the product. Many consumers use reviews posted by other consumers before making their purchase decisions. The tweets retrived are pre-processed into a more structured information. Tokenization is used to identify all words in a given text. After pre-processing is done the analysis is done. The analysis step is usually considered the core of text mining, because this is when some type of useful, nontrivial knowledge is extracted from the text. The performance of sentiment classification is evaluated by using accuracy.

**Disadvantage:**

* Requires more pre-processing steps.

1. **Using Machine Learning Techniques for Sentiment Analysis**

Oscar Romero Llombart.

The Sentiment Analysis is the name of the problem that with a sentence or text, the machine gets capable to analyze and predict with the maximum precision possible the sentiment that will be obtained by a person when reads it or the contextual opinion related to something [15].In this paper, five different machine learning methods has been used with a good performance in the sentiment analysis problem. The effect of apply transformations on the data can improve the performance of the classification methods but the type of transformations depends on the dataset and the type of the language it has. The data is processed by making a feature selection, applying transformations and filtering the data that have less importance and the information can make the machine learning method learn more efficiently and generalize better, because these days the machines have limitations and can’t handle all the data without any kind of prior process. In general, the machine learning methods tends to give similar results and, again, the results depend on the type of the data.

**Disadvantage:**

* More investigation must be done on learn with less transformations.

1. **Twitter Sentiment Analysis using Deep Learning Methods**

Adyan Marendra Ramadhani and Hong Soon Goo

Deep learning is the neural network that consists of many hidden layers and reflects the process of neurons in human brain to provide accurate result. The contribution of this system is to increase the accuracy of the predictions using neural network [1].This proposed system uses tweets from Twitter for sentimental analysis. The data set is downloaded from Twitter API. To clean the words preprocessing steps are performed which includes stemming, removal of numbers, punctuations, stop words, white spaces and converting all the tweets into same case letters. Then the sentence is tokenized. The data set consists of 2000 training data set and 2000 testing data set in both Korean and English languages. Then FNN is applied on the data set. This system used Tensorflow platform for creating neural network. Accuracy is used as the evaluation measure. This inputs 100 neurons for processing. This neural network uses three hidden layers for sentiment prediction. Multilayer perceptron produces accuracy of 52.60% as a result.

**Disadvantage:**

* Accuracy of the training data is low.

**CHAPTER 3**

**SYSTEM ANALYSIS**

**3.1 EXISTING SYSTEM**

Textual information has become one of the most important sources of data to extract useful and heterogeneous knowledge. Texts can provide factual information, such as descriptions or even instructions, and opinion-based information, which would include reviews, emotions, or feelings. Subjective information can be expressed through different textual genres, such as blogs, forums, and reviews, but also through social networks and microblogs. These sites involve a large amount of subjective information, due to millions of users share opinions on different aspects of their everyday life. Extracting this subjective information has a great value for both general and expert users.

**3.1.1 Dictionary-Based Approach**

Sentiment analysis involves categorizing viewpoints in text within categories such as positive, negative and neutral [8]. The input feed into the sentiment analysis system is a corpus of documents. The chief component or part of the system is the document analysis module, which utilizes the linguistic tools to gloss the pre-processed records with sentiment annotations. The observations may be attached to whole documents, to individual sentences or to specific features of entities. These observations are the outcome of the system and they may be represented to the user utilizing different varieties of visualization tools.

**Disadvantages:**

* Less accuracy rate

**3.1.2 CNN**

It is possible to predict user satisfaction of a product, happiness with some particular environment or destroy situation after disasters. Recently, deep learning is able to solve problems in computer vision or voice recognition, and CNN works good for image analysis and image classification. The biggest reason to adopt CNN in image analysis and classification is due to CNN can extract an area of features from global information, and it is able to consider the relationship among these features. For natural language processing, texts data features also can be extracted piece by piece and to consider the relationship among these features. And currently, CNN has a convolutional layer to extract information by a larger piece of text, so we work for sentiment analysis with CNN, and we design a simple CNN model and test it on benchmark.

**Disadvantages:**

* Difficult to handle large number of datasets
* Number of incorrectly classified tweets are high
* False positive rate is high

**3.2 PROPOSED SYSTEM:**

Sentiment analysis has been a popular topic in the field of machine learning. It is largely applied to data that comes with self-labeled information reviews from Twitter API. To access Twitter API, an app is developed to join as developer member of Twitter. After creating the app, Twitter provides 4 types of tokens to provide authentication to retrieve tweets. The tweets collected are pre-processed. The pre-processing steps includes removal of links, numbers, punctuations, retweets, stop words, prepositions, references to other screen names, hashtags, tabs, new line, spaces at the beginning and end , emoticons and convert the tweet to lower case. The ability to identify the positive or negative sentiment behind a piece of text is even more interesting when it comes to social data. Each review is computed score using dictionary-based approach. A collection of positive and negative word list is read and it is compared with the tweets to find score of the tweet. If the tweet is positive then the score is 1, for negative tweet it is -1 and 0 for neutral review. After that, the tweets are stored with score in CSV file. Convolutional Neural Network is made up of neurons that have learnable weights and biases. The train set and test set is provided. Document-term matrix and vocabulary is constructed for model creation. Then the sentiment is analyzed using deep learning process. Finally, a probability graph depicting the probability of positive tweets is presented.

**Advantages:**

* Improved accuracy rate.
* Higher precision.
* Higher recall.

**CHAPTER 4**

**SYSTEM SPECIFICATION**

**4.1 HARDWARE REQUIREMENTS:**

|  |  |
| --- | --- |
| Processor | Dual core processor 2.6.0 GHZ |
| RAM | 4 GB |
| Hard disk | 500 GB |
| Compact Disk | 650 Mb |
| Keyboard | Standard keyboard |
| Monitor | 15 inch color monitor |

**4.2 SOFTWARE REQUIREMENTS**

|  |  |
| --- | --- |
| Operating system | Windows OS |
| Tool | RStudio-1.1.423 |
| Language | R x32 3.4.4 |

**CHAPTER 5**

**SOFTWARE ENVIRONMENT**

**5.1 R LANGUAGE:**

R is an [open source](https://en.wikipedia.org/wiki/Open_source) [programming language](https://en.wikipedia.org/wiki/Programming_language) and software environment for [statistical computing](https://en.wikipedia.org/wiki/Statistical_computing) and graphics that is supported by the R Foundation for Statistical Computing. The R language is widely used among [statisticians](https://en.wikipedia.org/wiki/Statistician) and [data miners](https://en.wikipedia.org/wiki/Data_mining) for developing [statistical software](https://en.wikipedia.org/wiki/Statistical_software)and [data analysis](https://en.wikipedia.org/wiki/Data_analysis). Polls, [surveys of data miners](https://en.wikipedia.org/wiki/Rexer%27s_Annual_Data_Miner_Survey), and studies of scholarly literature databases show that R's popularity has increased substantially in recent years. R is a [GNU package](https://en.wikipedia.org/wiki/List_of_GNU_packages). The [source code](https://en.wikipedia.org/wiki/Source_code) for the R software environment is written primarily in [C](https://en.wikipedia.org/wiki/C_(programming_language)), [Fortran](https://en.wikipedia.org/wiki/Fortran), and R. R is freely available under the [GNU General Public License](https://en.wikipedia.org/wiki/GNU_General_Public_License), and pre-compiled binary versions are provided for various [operating systems](https://en.wikipedia.org/wiki/Operating_system). R is a programming language and software environment for statistical analysis, graphics representation and reporting. R was created by Ross Ihaka and Robert Gentleman at the University Of Auckland, New Zealand, and is currently developed by the R Development Core Team.

**5.2 R STUDIO:**

RStudio is a  [free and open-source](https://en.wikipedia.org/wiki/Free_and_open-source) [integrated development environment](https://en.wikipedia.org/wiki/Integrated_development_environment) for [R](https://en.wikipedia.org/wiki/R_(programming_language)), a [programming language](https://en.wikipedia.org/wiki/Programming_language) for [statistical computing](https://en.wikipedia.org/wiki/Statistical_computing) and graphics. [Hadley Wickham](https://en.wikipedia.org/wiki/Hadley_Wickham) is the Chief Scientist at RStudio. RStudio is available in two editions: RStudio Desktop, where the program is run locally as a regular [desktop application](https://en.wikipedia.org/wiki/Desktop_application); and RStudio Server, which allows accessing RStudio using a web browser while it is running on a remote [Linux](https://en.wikipedia.org/wiki/Linux) server. Prepackaged distributions of RStudio Desktop are available for [Windows](https://en.wikipedia.org/wiki/Windows), [macOS](https://en.wikipedia.org/wiki/MacOS), and [Linux](https://en.wikipedia.org/wiki/Linux). RStudio is available in open source and commercial editions and runs on the desktop or in a browser connected to RStudio Server.

**5. 3 FEATURES:**

As stated earlier, R is a programming language and software environment for statistical analysis, graphics representation and reporting. The following are the important features of R −

* R is a well-developed, simple and effective programming language which includes conditionals, loops, user defined recursive functions and input and output facilities.
* R has an effective data handling and storage facility,
* R provides a suite of operators for calculations on arrays, lists, vectors and matrices.
* R provides a large, coherent and integrated collection of tools for data analysis.
* R provides graphical facilities for data analysis and display either directly at the computer or printing at the papers.

**CHAPTER 6**

**SYSTEM DESIGN**

**6.1 ARCHITECTURE DIAGRAM**

**Tweets Retrieval**

Search Tweets for specified term

Load R packages and functions

**Environment Setup**

Twitter API Authentication

**Tweets**

**Pre-processing**

**Polarity Mining using Dictionary-Based Approach**

Read positive and negative data set

Positive score = 1

**Store as tweets.csv with scores**

Negative score = -1

Neutral score = 0

**CNN**

Provide train and test data

**Performance Measure**

Create Vocabulary & Document-Term Matrix

Precision

Accuracy

Classification Model

Recall

Probability Graph

**Fig: 6.1 Architecture diagram**

**6.2 USE-CASE DIAGRAM**

A use case diagram in the Unified Modeling Language is a type of behavioral diagram defined by and created from a Use-case analysis. It’s purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals and dependencies between those use case. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actor in the system can be depicted.

**Twitter User**

**Admin**

**Fig: 6.2 Use Case diagram**

**6.3 CLASS DIAGRAM**

A class diagram provides an overview of system by showing its classes and the relationships among them . UML class notation is rectangle divided into three parts: Class Name, Fields, Methods, Names of abstract class and Interfaces. Relationship between classes is connecting links.

|  |
| --- |
| Upload data |
| +Twitter API |
| +Read tweets() |

|  |
| --- |
| Tweets Pre-processing |
|  |
| +Stop words removal()  +Prepositions removal()  +Lowercase()  +Punctuations removal()  +Expressions removal()  +Digits removal()  +Links removal()  +Spaces removal()  +Symbols removal()  +Retweet removal()  +Hashtags removal  +Tab removal  +Reference to other screen names removal() |

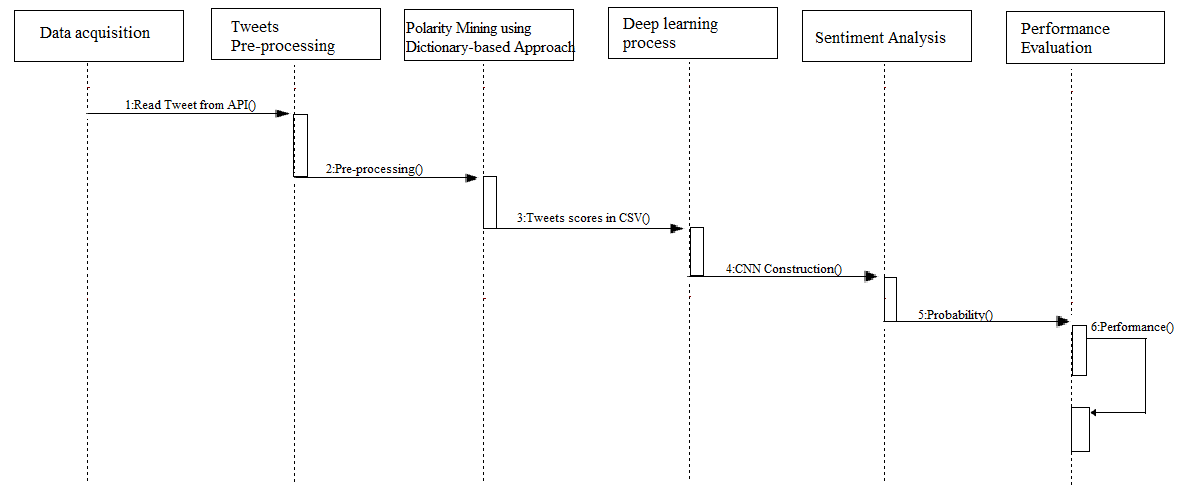
|  |
| --- |
| Feature Vector Construction |
| +CNN construction |
| +Sentiment Analysis()  +Performance() |

|  |
| --- |
| Polarity Mining using Dictionary-based Approach |
| +Positivewords.CSV  +Negativewords.CSV |
| +positive()  +negative()  +neutral() |

**Fig: 6.3 Class diagram**

**6.4 SEQUENCE DIAGRAM**

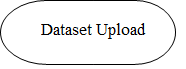
Sequence diagram is an interaction diagram that shows how processes operate with one another and what order. It’s a construct of a Message Sequence Chart. A sequence diagram shows object interactions arranged in time sequence. It depicts the object and classes involved in the scenario and sequence of the message exchanged between the objects needed to carry out the functionalities of scenario. Sequence diagram are typically with associate with the use case realization in the logical view of the system under development. Sequence diagrams are sometimes called Event diagrams or Event Scenarios.

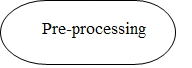
****

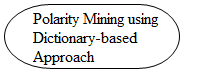
**Fig: 6.4 Sequence diagram**

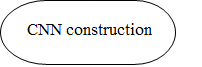
**6.5 ACTIVITY DIAGRAM**

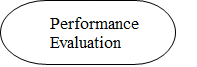
Activity diagram is a type of diagram used in computer science and related fields to describe the behavior of the systems. Activity diagram require that the system described composed of a finite number of states. Activity diagrams are used to give an abstract description of the behavior of the system. This behavior is analyzed and represented in series of events that could occur in one or more possible state.

****



****

****



**Fig: 6.5 Activity diagram**

**CHAPTER 7**

**SYSTEM IMPLEMENTATION**

**7.1 MODULES DESCRIPTION**

* Data Collection
* Tweet Preprocessing
* Polarity Mining using Dictionary-Based Approach
* Sentiment Analysis using Hybrid Approach
* Performance Evaluation

**7.1.1 Data Collection**

Social media is generating a large volume of sentiment rich data in the form of tweets, status updates, blog posts, comments, reviews, etc. For analyzing the tweets, an app is created to retrieve tweets from Twitter API. After authentication by the user the tweets are retrieved. To perform polarity mining using dictionary-based approach positive words and negative words are stored in text document format. Once the tweets are pre-processed the data is stored as tweetfile.CSV.

**7.1.2 Tweet Preprocessing**

Data cleaning or pre-processing is the process of detecting and correcting the data and then replacing, modifying, or deleting the dirty or coarse data [18]*.* Once the tweets are retrieved, the cleaning process is done.

The basic preprocessing steps involved are:

* Removal of Digits
* Removal of Expressions
* Removal of links
* Removal of symbols
* Removal of Stop-words
* Removal of Prepositions
* Removal of Punctuations
* Removal of retweet
* Clean hashtags
* Remove tab
* Removal of reference to other screen names
* Convert tweets to lower case
* Removal of Spaces

**7.1.3 Polarity Mining using Dictionary-Based Approach**

After the pre-processing is done, the positive and negative words are read and saved for comparison [8]. The pre-processed tweets is compared with the positive and negative words and classified into positive, negative and neutral. The score for positive, negative and neutral tweet is 1, -1 and 0 is allotted respectively[11] . Then the tweets are stored with scores in CSV file. The graph is plotted to represent the sentimental analysis using Dictionary-Based Approach.

**7.1.4 Sentiment Analysis using Hybrid Approach**

Once polarity mining is done using Dictionary-Based Approach, Sentimental Analysis is performed using CNN. The tweets stored in the CSV file is retrieved[16] . Then the train and test data is provided in the ratio 60:40. The vocabulary and document-term matrix is formed. The CNN model is trained, using the train and test dataset. This uses K-fold cross validation to estimate how [accurately](https://en.wikipedia.org/wiki/Accuracy) a [predictive model](https://en.wikipedia.org/wiki/Predictive_modelling) will perform in practice. In a prediction problem, a model is usually given a dataset of known data on which training is run, and a dataset of unknown data against which the model is tested. The goal of cross validation is to define a dataset to "test" the model in the training phase , in order to limit problems like [overfitting](https://en.wikipedia.org/wiki/Overfitting), give an insight on how the model will generalize to an independent dataset. Here k is assumed to be 5. The positive probability of each tweet is determined and stored as matrix. Then the total positive, negative and neutral tweets are known. Finally, the probability graph is plotted. This graph shows the sentiment probability from 0 to 1 in y-scale and created time in x-scale. The probability of each tweet is plotted on the graph.

**Construction of CNN model**

CNN has archived impressive result in image recognition several years ago and has archived remarkable results in NLP recently, there is a convolutional layer to make a piece of words can be considered together. We adopt convolutional neural network as our sentiment analysis model because in image analysis and classification field, CNN can extract an area of features from global information, with the convolution operation, a piece of data information can be extract together as the features, and it is able to consider the relationship among these features[5]. For computer vision, such as image analysis, it is able to extract a part of pixel data information, not only extract the pixels one by one, the features information can be extracted piece by piece, the piece contains multi pixels data information; when we transfer the text into matrix, it can also be considered as same as an image pixels’ matrix, so we can do the same operation to the text data to make the input features to the model can be trained in another effective way.

CNN has 3 layers namely,

* **Convolutional Layer,**
* **Pooling Layer**
* **Fully-Connected Layer**

**Fig: 7.1.4 CNN Construction**

**7.1.4 Performance Evaluation**

We can evaluate the performance using various metrics namely:

* Precision
* Recall
* Accuracy

**7.1.4.4 Precision**

[Precision](http://en.wikipedia.org/wiki/Accuracy_and_precision#Accuracy_and_precision_in_binary_classification) measures the exactness of a classifier [19]. A higher precision means less [false positives](http://en.wikipedia.org/wiki/Type_I_and_type_II_errors#Type_I_error), while a lower precision means more false positive rate.

Precision=TP/ (TP + FP) 1

**7.1.4.5 Recall**

Recall measures the completeness, or [sensitivity](http://en.wikipedia.org/wiki/Sensitivity_and_specificity), of a classifier. It is also known as Sensitivity [14]. Higher recall means less [false negatives](http://en.wikipedia.org/wiki/Type_I_and_type_II_errors#Type_II_error), while lower recall means more false negative rate.

Recall=TP/ (TP + FN) 2

**7.1.4.6 Accuracy**

A measure of how often a sentiment rating was correct. It tracks how many of the tweets were rated correctly.

Accuracy = (TP+TN) / (TP+TN+FP+FN) 3

**CHAPTER 8**

**EXPERIMENTAL RESULT**

We choose the tweets from Twitter API as dataset for sentimental analysis.

|  |  |
| --- | --- |
| **DATASET** | Twitter API |
| **NO. OF TWEETS** | 2500 |
| **LANGUAGE** | English |

**Table 8.1: Detail of Dataset**

The tweets are pre-processed and score is calculated using Dictionary-based approach.

|  |  |
| --- | --- |
| **NO.OF NEGATVIE WORDS** | 2195 |
| **NO. OF POSTIVE WORDS** | 4096 |
| **LANGUAGE** | English |

**Table 8.2: Train data for Dictionary-based Approach**

We split the pre-processed dataset with scores into 2 groups, train set which to be input as training samples and the development set to verify the accuracy of the checkpoint of the CNN. For each of the dataset, we set the train set around 60% of the whole data amount and the development set is around 40%. For the testing, we train the CNN model.

|  |  |  |
| --- | --- | --- |
| **NO. OF**  **DATASET** | **NO. OF**  **TRAIN DATASET** | **NO. OF**  **TEST DATASET** |
| 2500 | 1500 | 1000 |

**Table 8.3: Detail of Train and Test dataset**

Confusion matrix for Dictionary-based approach and CNN is calculated and with the help of confusion matrix precision, recall and accuracy is calculated.

|  |  |  |
| --- | --- | --- |
| **N=2500** | **Reference** | |
| **Prediction** | 0 | 1 |
| 0 | TN | FP |
| 1 | FN | TP |

**Table 8.4: Confusion Matrix for Dictionary-Based Approach + CNN**

**Fig.: 8.1: Comparison between Dictionary-Based Approach, CNN and Hybrid Approach**

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Recall** | **Precision** | **Accuracy** |
| Dictionary-based approach | 45% | 52% | 36% |
| CNN | 51% | 53% | 45% |
| Hybrid Approach | 53% | 58% | 55% |

**Table 8.5: Comparison of accuracy**

Compared to the existing system i.e., Dictionary –Based Approach and CNN the hybrid approach produces high rates in terms of precision, recall and accuracy.

**CHAPTER 9**

**SYSTEM TESTING AND MAINTENANCE**

**9.1 TEST CASE SCENARIOS**

Software testing is a method of assessing the functionality of a software program. There are many different types of software testing but the two main categories are dynamic testing and static testing. Dynamic testing is an assessment that is conducted while the program is executed; static testing, on the other hand, is an examination of the program's code and associated documentation. Dynamic and static methods are often used together.

**9.2 SYSTEM TESTING**

Testing is a set activity that can be planned and conducted systematically. Testing begins at the module level and work towards the integration of entire computers based system. Nothing is complete without testing, as it is vital success of the system.

**Testing Objectives:**

1. Testing is a process of executing a program with the intent of finding an error
2. A good test case is one that has high probability of finding an undiscovered error.
3. A successful test is one that uncovers an undiscovered error.

**9.3 TYPES OF TESTING**

**9.3.1 Unit Testing**

The first test in the development process is the unit test. The source code is normally divided into modules, which in turn are divided into smaller units called units. These units have specific behavior. The test done on these units of code is called unit test. Unit tests ensure that each unique path of the project performs accurately to the documented specifications and contains clearly defined inputs and expected results.

**9.3.2 Functional Testing**

Functional test can be defined as testing two or more modules together with the intent of finding defects, demonstrating that defects are not present, verifying that the module performs its intended functions as stated in the specification and establishing confidence that a program does what it is supposed to do.

**9.3.3 Integration Testing**

Testing in which modules are combined and tested as a group. Modules are typically code modules, individual applications, source and destination applications on a network, etc. Integration Testing follows unit testing and precedes system testing. Betas are often widely distributed or even distributed to the public at large in hopes that they will buy the final product when it is released.

**9.3.4 White box Testing**

Testing based on an analysis of internal workings and structure of a piece of software. This testing can be done sing the percentage value of load and energy. The tester should know what exactly is done in the internal program. Includes techniques such as Branch Testing and Path Testing. Also known as Structural Testing and Glass Box Testing.

**CHAPTER 10**

**CONCLUSION AND FUTURE ENHANCEMENT**

**10.1 CONCLUSION**

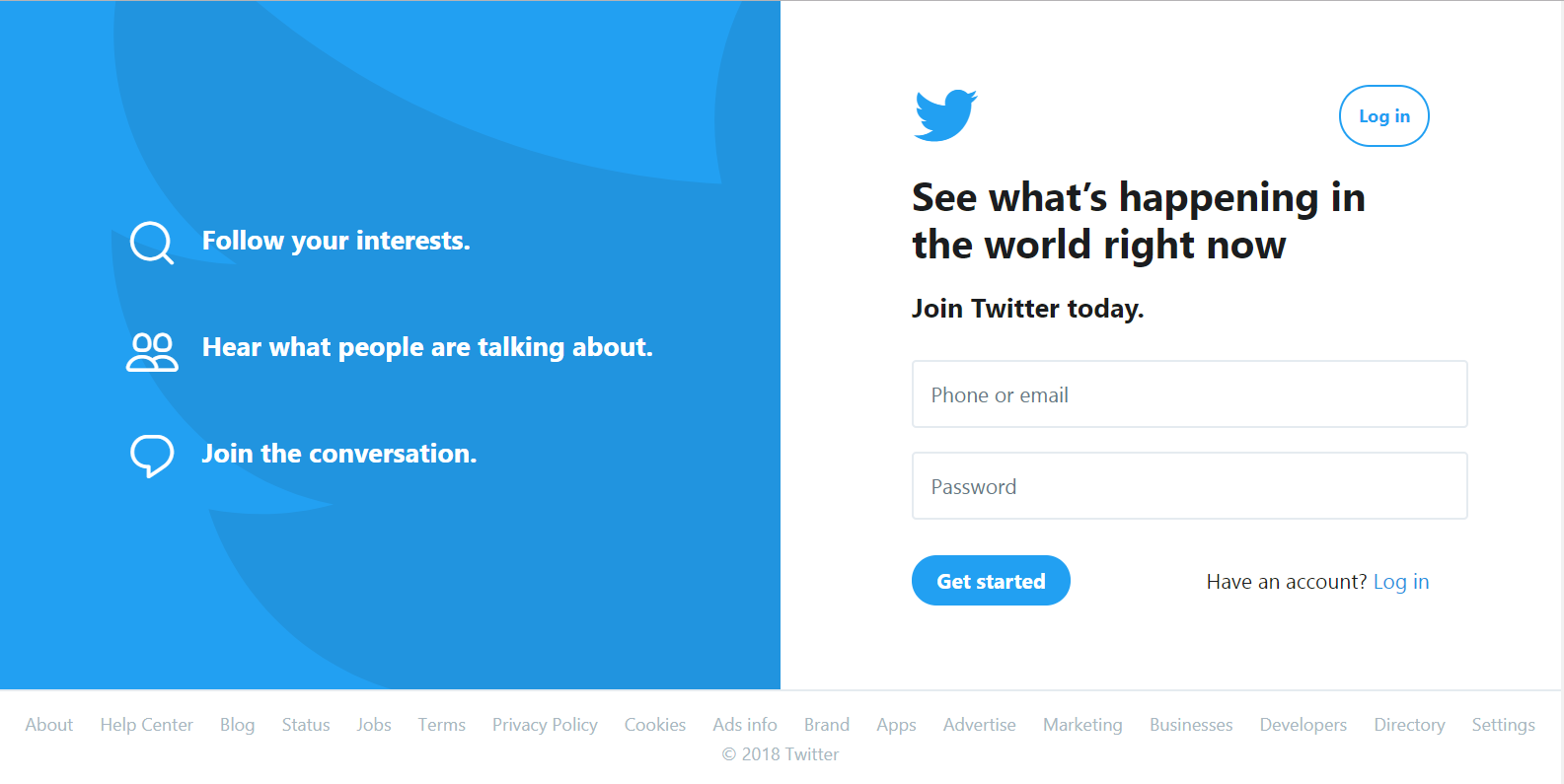
Twitter sentiment analysis is difficult because it is very tough to identify emotional words form tweets and also due to the presence of the repeated characters, slang words, white spaces, misspellings etc. To handle these problems the feature vector is created. Before creating feature vector pre-processing is done on each tweet. The n it is stored with scores in CSV format. Then CNN model is constructed. Finally the probability of the positive tweets is calculated and the performance is evaluated. The proposed method classified the tweets in positive, negative and neutral sentiments. The work of proposed model has gone through preprocessing stage, features generation stage and classifiers learning stage. The analytical evaluation of proposed model is done in terms of precision, recall and accuracy. The comparative observations are taken against the Dictionary-based approach and CNN methods. The comparative results shows that the proposed model has improved the precision, recall and accuracy of tweet class prediction.

**10.2 FUTURE ENHANCEMENTS**

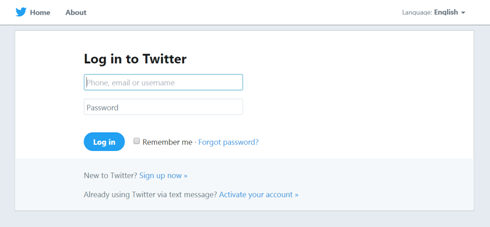
A method to predict or deduct the location of a tweet based on the tweet’s information and the user’s information should be identified in the future. It is proposed to stream real time live tweets from twitter using Twitter API, and to perform sentimental analysis on different languages.

**APPENDIX 2**

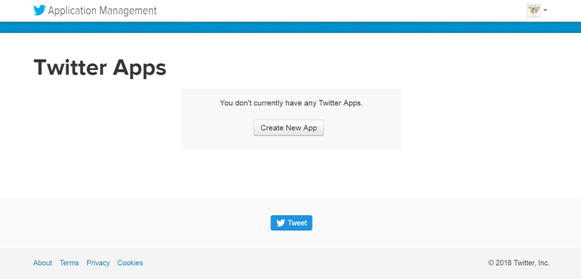
**SCREEN SHOTS**

****

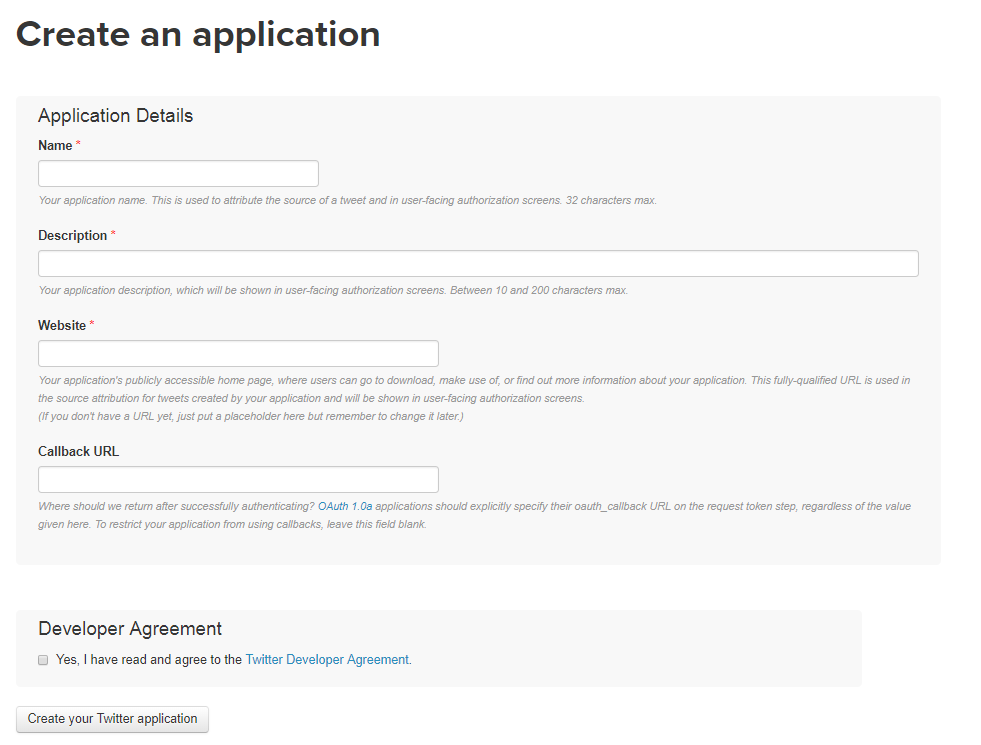
**Fig.: A 2.1 Create Twitter Account**

****

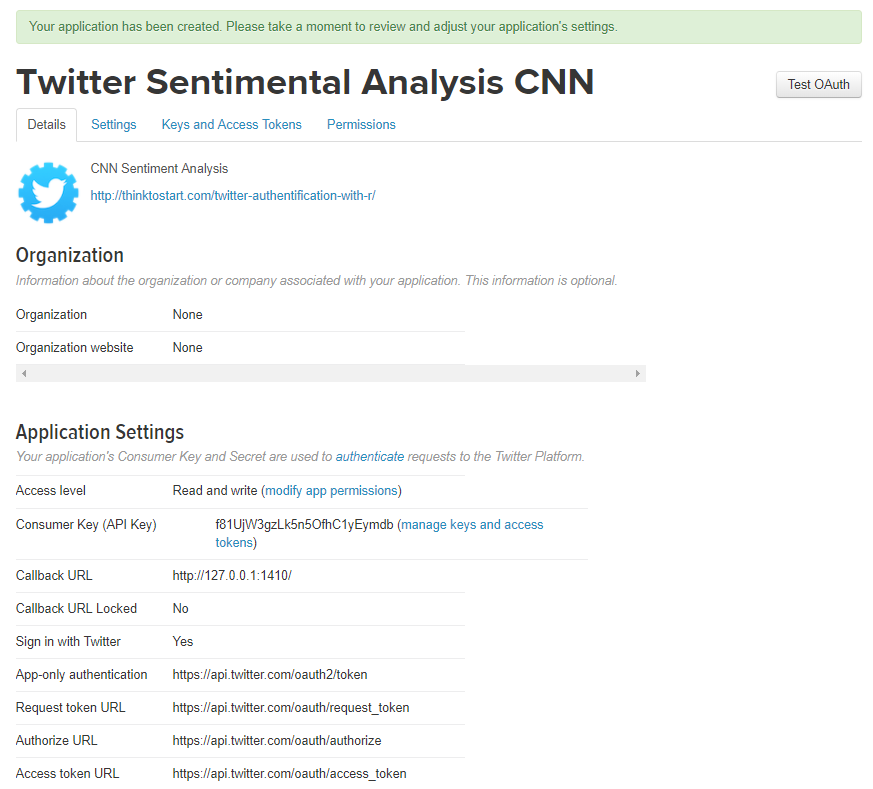
**Fig.: A 2.2 Join on developer side**

****

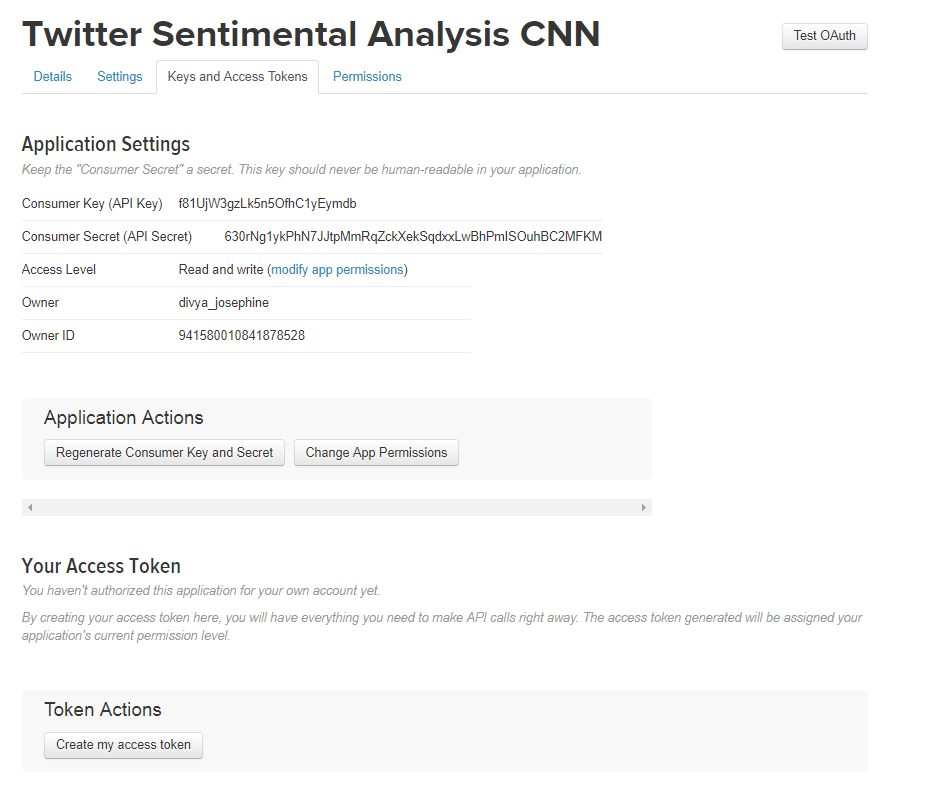
**Fig.: A 2.3 Create app to access Tweets**



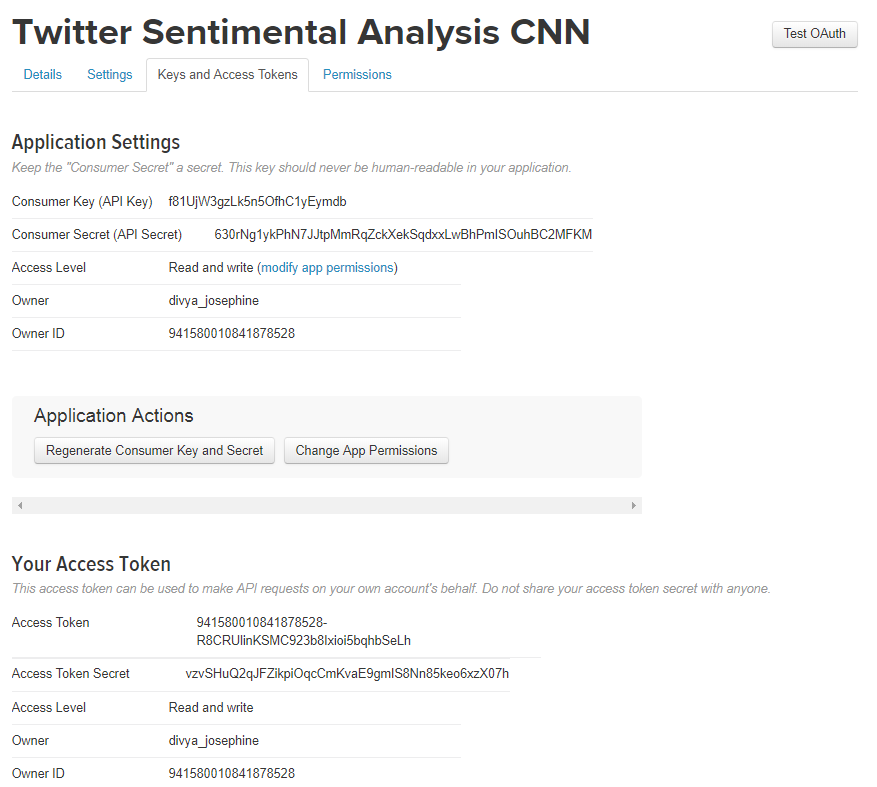
**Fig.: A 2.4 Specify Application Details**



**Fig.: A 2.5 API Key**



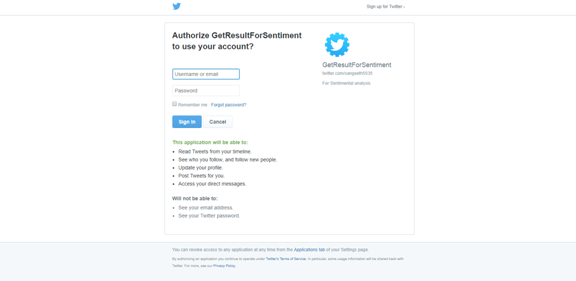
**Fig.: A 2.6 API Key and API Secret Key**



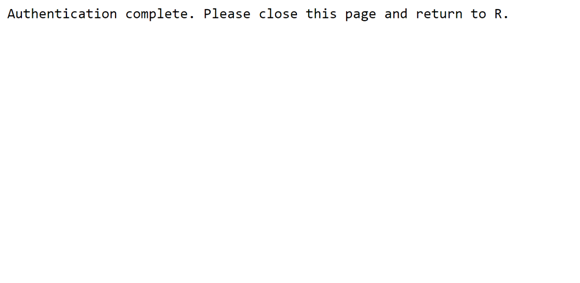
**Fig.: A 2.7 Access Token**



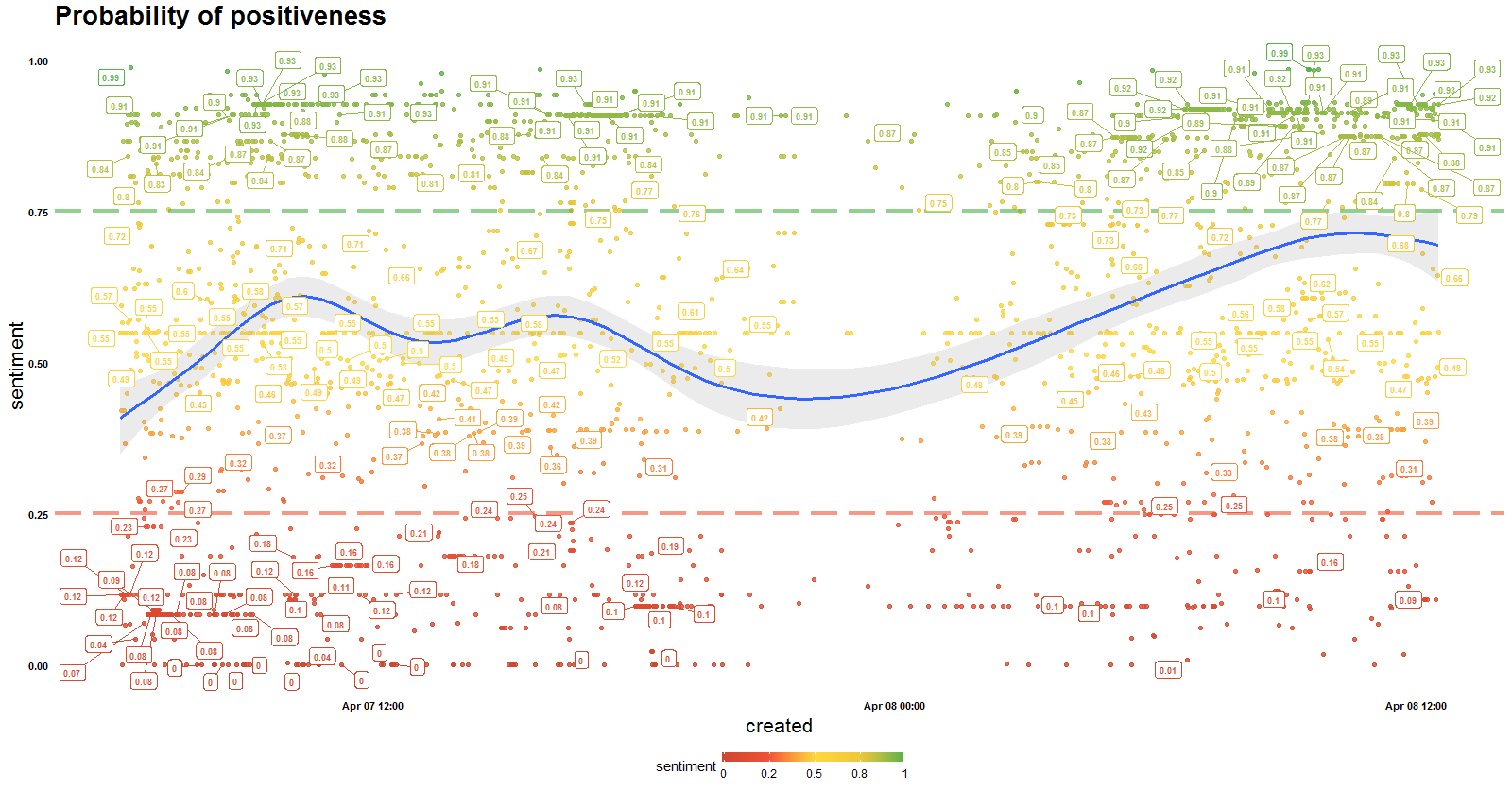
**Fig.: A 2.8 Authorize App**

****

**Fig.: A 2.9 Twitter API Login**

****

**Fig.: A 2.10 Twitter Authentication**

****

**Fig.: A 2.12 Probability Graph**

**REFERENCES**

1. Adyan Marendra Ramadhani and Hong Soon Goo, "Twitter Sentiment Analysis using Deep Learning Methods." Engineering Seminar (InAES), 2017 7th International Annual.
2. Ali Selamat and Nurulhuda Zainuddin , “Sentiment Analysis Using Support Vector Machine” 2014 IEEE 2014 International Conference on Computer, Communication, and Control Technology, 2014 .
3. Apeksha Pande , Reshma Bhonde , Binita Bhagwat ,et.al, “Sentiment Analysis Based on Dictionary Approach”, International Journal of Emerging Engineering Research and Technology Volume 3, Issue 1, January 2015, PP 51-55.
4. Bhumika , Jadav and Vimalkumar Vaghela, “Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis.” International Journal of Computer Applications Volume 146 – No.13, July 2016.
5. Bowen Baker, Otkrist Gupta, Nikhil Naik, et.al,“Designing Neural Network Architectures using Reinforcement learning” Published as a conference paper at ICLR 2017.
6. Carlo Aliprandi, Federico Neri, Federico Capeci ,et.al, "Sentiment Analysis on Social Media." IEEE Computer Society Washington, 2012
7. Daniel Jurafsky and James H. Martin “Naive Bayes and Sentiment Classification” Published on August 7,2017.
8. Fatehjeet Kaur Chopra and Rekha Bhatia, “Sentiment Analyzing by Dictionary based Approach” International Journal of Computer Applications(0975-8887).
9. Isa Maks and Piek Vossen, “A lexicon model for deep sentiment analysis and opinion mining applications.” Elsevier Science Publishers B. V. Amsterdam, 2012-11-01.
10. Joshi, Rohit, and Rajkumar Tekchandani, "Comparative analysis of Twitter data using supervised classifiers." Inventive Computation Technologies International Conference on. Vol. 3. IEEE, 2016.
11. Laura Cruz , Jose Ochoa, Mathieu Roche, et.al, “Dictionary-Based Sentiment Analysis applied to specific domain using a Web Mining approach.”Published 2016 in SIMBIG.
12. Liang-Chih Yu, Jheng-Long Wu, Chang, et.al, “Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news.”  Knowledge-Based Systems 41:89-97 · March 2013.
13. Lopamudra Dey, Sanjay Chakraborty, et.al , “Sentiment Analysis of Review Datasets using Naïve Bayes’ and K-NN Classifier”. I.J. Information Engineering and Electronic Business, 2016, 4, 54-62 Published Online July 2016 in MECS.
14. Marina Sokolova and Guy Lapalme, “A systematic analysis of performance measures for classification tasks” Information Processing and Management: an International Journal archive Volume 45 Issue 4, July, 2009.
15. Oscar Romero Llombart, “Using Machine Learning Techniques for Sentiment Analysis.” Conference on School of Engineering (UAB) June of 2017.
16. Shiyang Liaoa, Junbo Wangb, et.al ,“CNN for situations understanding based on sentiment analysis of twitter data” 8th International Conference on Advances in Information Techonology, IAIT2016, 19-22 December 2016, Macau,China.
17. Shuai Wang and Bing Liu, “Deep Learning for Sentiment Analysis: A Survey Lei Zhang, LinkedIn Corporation.”National Science Foundation (NSF) under grant no IIS1407927 and IIS-1650900, 2018.
18. Tajinder Singh and Madhu Kumari, “Role of Text Pre-Processing in Twitter Sentiment Analysis” Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016).
19. Vinothini and Chandrasekaaran, “A comparative performance evaluation of neural network based approach for sentiment classification of online reviews” Journal of King Saud University - Computer and Information Sciences archive Volume 28 Issue 1, January 2016.
20. Xu Tao, Qinke Peng and Yinzhao Cheng, “Identifying the semantic orientation of terms using S-HAL for sentiment analysis.” Knowl-Based Syst, 35 ,2012, pp. 279-289.”,2017.