IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 73-79
www.iosrjournals.org

Data Preprocessing, Sentiment Analysis & NER On Twitter Data.

Mr.SanketPatil, Prof.VarshaWangikar, Prof. K. Jayamalini
(Computer, SLRTCE, India)
(Information Technology, K.C.College of Engineering and Management Studies and Research College, India)
(Computer, SLRTCE, India)

Abstract: *The world today has become more advanced where people talk, chat and comment their viewswithout their physical appearance just by dealing with words and not expressions. For instance, Twitter is being used my many people to raise their voice against a particular issue, now the job of finding out and segregating the tone of the comments as positive, negative or neutral is a difficult but crucial task. This information can be used by someone who is a manufacturer of a particular product to know about the impact of his product or Quality of service in the market. It might also be used by a celebrity to how common people react to a particular issue raised by him. This Paper aims to contribute to the field of sentiment analysis, Data Preprocessing and NER. Preprocessing method plays a very important role in text mining techniques and applications. It is the first step in the text mining process. In this we mainly focus on Stop word removal and slang word to remove special characters. Sentiment analysis is a very effective tool used in social media today. It is extremely useful in twitter as it enables to understand the perspectives of the common people behind a topic .Labeling a text with entities is* ***done by using Stanford's*** *NER and CRF.*

Keywords –*CRF, NER, Sentiment Analysis, Social media, twitter*

I. Introduction

Twitter is a social networking service that allows members to register and then to broadcast short messages called tweets. A user on Twitter can broadcast tweets o**r follow other user's tweets. Mobile** phones, desktops clients can be used to send tweets and replies and it can also be done by posting on the website twitter.com website. Twitter has become one of the most important communication channels with its ability of providing the most up-to-date and newsworthy information. A treasure of information research lies for researchers considering millions of monthly active users and tweets sent per day and also it attracts attention of not only academics but also organizations to extract user interests.

Twitter messages are also called as Tweets. We will use these tweets as raw data. We will use a method that automatically extracts tweets into either negative, or positive or neutral sentiments. By using the sentiment analysis the customer can know the feedback about the product or services before making a purchase. The views of the customers can easily be identified by the use of sentiment analysis which helps in analyzing the requirements of the customers and improving the quality of the product. Due to the flooding of information obtained from social websites, sentimental Analysis [1] has become a popular area of research online forums, and blogs. Here, live tweets are picked up from the site and preprocessed to identify the name of organization (if any) in the tweet, the location and the person. Twitter Use in Organizations- Researchers in Public Relations have conducted several works regarding how organizations leverage Twitter for stakeholder communication. It investigate tweets in a group of 93 companies with an active Twitter account in Fortune 500 companies to understand the dialogic features of Twitter and the target public groups of those companies. They found the dialogic features when analyzing tweets, for exampl**e responses to users' posts (60.2%), posting newsworthy** information about the company (58.1%), and posing questions (30.1%). They also found that a many of the tweets are addressed to general audience, which is not explicitly identified (74.5%), while a very few tweets are sent to customers, which are specific users with @username in the tweets (0.9%).Thus knowing the location of organization and any comments or perspectives related to the same can easily be acknowledged. The tweet is then segregated to see **if it's positive, negative or neutral. This is done through sentimental analysis or opinion** mining. Opinion mining can be very useful in many ways. It provides a tool for the marketers to evaluate how a specific ad or the campaign of a new product is successful and identify which demographics like or dislike particular product features. For example, a review on a website might be broadly positive about a digital camera, but be specifically negative about how heavy it is. The manufacturer gets a much clear idea about his product and his services rather than an unsystematic approach of surveys or focus groups do, because the data is created by the customer. A very interesting feature of this is that it can identify the sarcasm used in the sentence and can **judge the comment on the basis of tone in addition to the words. An example of this is**

the statement "Congrats, you lost the match again". Most humans would be able to quickly interpret that the person was being sarcastic. It is known that for losing a match is not a great experience. The contextual understanding can lead to a negative judgment by any human being but having a look at the contextual side the sentence above might see the word

"Congrats " and categories it as positive. It is very difficult to teach a machine to analyze various cultural variations, grammatical nuances, the slang and misspellings .Teaching a machine to understand how context can affect tone is even more difficult. Here is where sentimental analysis plays its role. So, in spite, of having the word superb, the tweet will have a negative scoring because of the word lost. The machine intelligence will give its final verdict as negative. Social media monitoring tools like Brandwatch Analytics make that process quicker and easier than ever before, thanks to real-time monitoring capabilities.

The principal algorithms used are:

1. NER:

It is very important sometimes to identify a part of speech of words in areas such as speech recognition and natural language processing tasks. Conditional Random Field (CRF)[5] is alike to that approach. NER trains its classifier on a big or large set of data having a sequence of words, each word annotated by an entity(if any). Sliding window is a technique that takes into account the words that prefix and the words that suffix the tokens in question and it programs a label for each of the word which could be either the name of an organization, a person or a location. Since the entire theory is based on probability, there are a few chances of error which can be taken care of by using Sentimental Analysis [2].

2. Sentiment analysis:

A parameter is essential to check written or spoken language which does not support facial expressions. This parameter can be used to test the expression or the tone of the person as favorable, unfavorable or neutral. This parameter is defined as Sentiment analysis [4] or opinion mining. Sentiment Analysis can handle a large chunk of valuable feedback of customers accurately. If paired with text analytics, it reveals the perspective and the views of the customers about the products and services.

II. Proposed System

2.1 Data Pre-Processing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often not complete or consistent, and also lacks in certain trends or behaviors, and is bound to have some errors. It is a proven method of resolving such issues. Data preprocessing prepares raw data for further preprocessing [6].

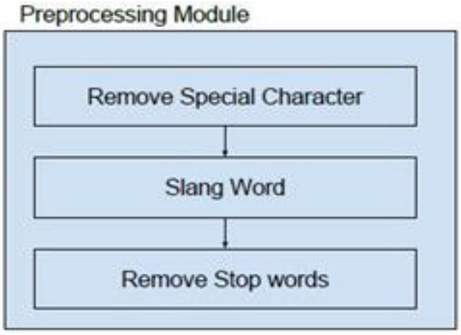


Fig. 2.1 Preprocessing Module

Data preprocessing is used in various real-time applications such as neural networks and is a database-driven application. Data goes through a series of steps during preprocessing:

a) Data Cleaning: Processes such as filling of missing values, resolving the inconsistencies in the information or smoothing the noisy data.
b) Data Integration: It is the assembling of various representations of data and problems within the data is taken care of.
c) Data Transformation: Data is aggregated, normalized, and generalized.
d) Data Reduction: This step aims to present a reduced representation of the data in a data warehouse.
e) Data Discretization: The range of attribute intervals is reduced by reducing the no of values.

Pre-processing of extracted data:

After retrieval of tweets Sentiment analysis [4] tool is applied on raw tweets but in most of cases results to very poor performance. Therefore, preprocessing techniques are necessary for obtaining better results. We extract tweets i.e. short Messages from twitter which are used as raw data. This raw data needs to be preprocessed. So, preprocessing involves

Following steps which construct in grams:

i) **Filtering:** Filtering is nothing but cleaning of raw data. In this step, URL links (E.g. http://twitter.com), special **words in twitter (e.g. "RT" which means ReTweet), user names in twitter (e.g. @Ron** -@ symbol indicating a user name), emoticons are removed.

ii) **Tokenization:** Tokenization is nothing but Segmentation of sentences. In this step, we will tokenize or segment text with the help of splitting text by spaces and punctuation marks to form container of words.

iii) **Removal of Stop words: Articles such as "a", "an", "the" and other stop words such as "to", "of", "is", "are", "this", "for "removed in this step.**

iv) **Construction of n -grams:** Set of n-**grams can make out of consecutive words. Negation words such as "no","not" is attached to a** word which follows or**Precedes it. For Instance: "I do not like remix music" has two bigrams: "I do not","do+not like", "not+like remix music". So the accuracy of the classification improves by** such procedure, because negation plays an important role in sentiment analysis. Negation needs to be taken into account, because it is a very common linguistic construction that affects polarity

2.2 Classification

When performing classification, every tweet is checked for the positive and negative words from a fixed set of list. Then the average is taken out for both positive and negative, depending on the higher score the tweet label is saved as positive /negative or if the score is 0 its neutral.

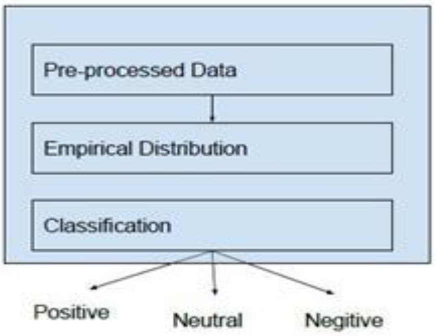


Fig. 2.2 Classification Methods

The tweet words are also checked from sentiword.net and the score is again calculated for every tweet and labeled as positive if greater than 0 and labeled as negative if less than 0.

Then the two labels are compared if both the tweets share the same sentiment the label does not change, but if the labels are opposite then the score from each method is checked and the higher score label is assigned. For e.g.

"I love India but India is not that clean**."** method 1: positive method 2: negativeMethod1: score= 0.9 Method2: score= -1.9
|method1:score|<|method2:score| Tweet is labeled as negative.

2.3 Ner [Named Entity Recognizer]

Named identity recognizer (NER) is an implementation of Java. It labels a group or sequence of words in a file which are the name of a person, a gene, a company name or things such as protein names. Stanford NER has a well-built feature extractor for Named Entity Recognition and various options for defining some feature extractor. The three classes i.e. ORGANIZATION, PERSON, LOCATION are good named entity recognizers for English.

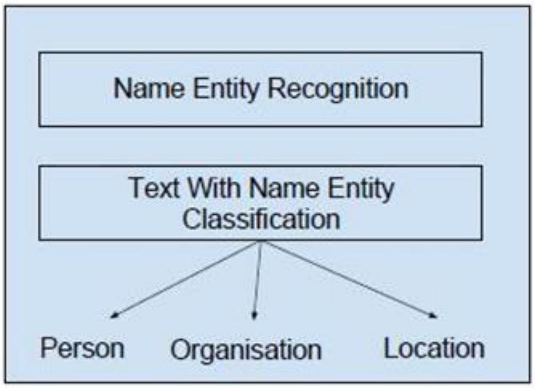


Figure 2.3 NER Representation

An NER labels sequences of words in which there are names of entities such as people, Organizations, locations etc. We utilize the Stanford NER to extract such named entities. Below is an example of a tweet and the entities extracted by the Stanford NER.@billgates<ORGANIZATION>Microsoft</ORGANIZATION> expects tough market for Windows 7: <ORGANIZATION>Microsoft</ORGANIZATION> executive <PERSON>Bill Koefoed</PERSON>

The Stanford NER can't automatically classify names tags (@person) and URLs as entities so we augment it byincluding these types of entities. It seems intuitive to collect person tags since this is how Twitter users express their opinion or communicate with the entity being tagged. Similarly, opinions about URLs that are shared on Twitter, since disseminating URLS is of the chief use cases for Twitter.
<NAME>@billgates</NAME><ORGANIZATION>Microsoft</ORGANIZATION>

Expects tough market for Windows 7: <ORGANIZATION>Microsoft </ORGANIZATION> executive <PERSON>Bill</PERSON> answered some questions about the..<URL>http://tinyurl.com/l4a4z9</URL>

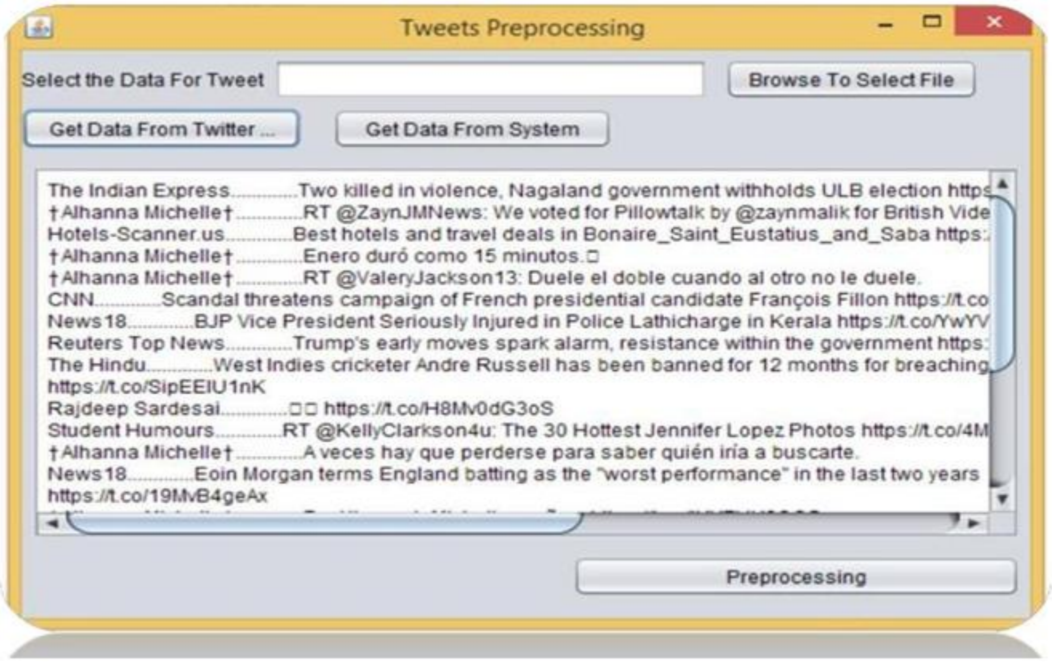
The NER implementation comes with four pre-trained classifiers. Each of these classifiers comes with a Second version that uses a distributional similarity lexicon to improve performance. We used the

Former three class version with the distributional similarity lexicon. It is interesting to note that running even the most trivial of operations on a dataset of 250 million lines (3-4 lines of meta data per tweet) has its own computational challenges. For running the NER on the Twitter corpus we used a home-grown Map-Reduce approach to utilize the four cores of the quad-core machine available to us.

III. Results And Discussion

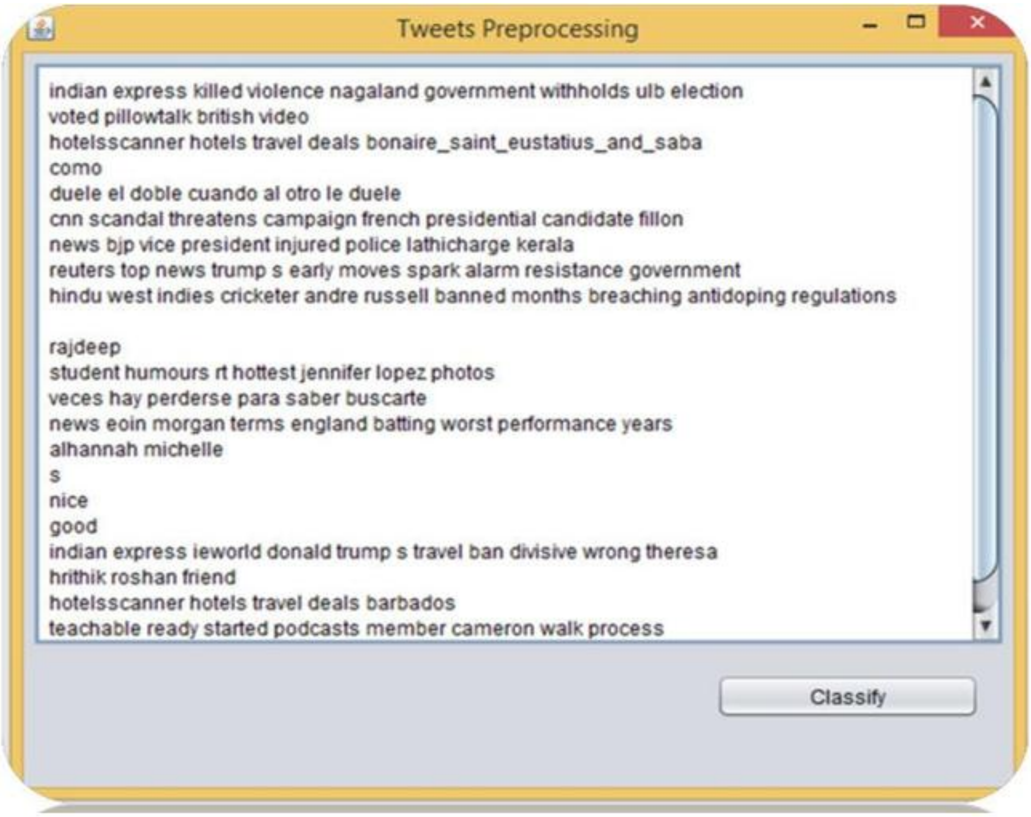
The main aim of the system is to analyze the twitter data. Analysis of twitter data is done using sentiment analysis and Name Entity Recognition (NER). Firstly, the live data is retrieved from the twitter which is shown is Fig. 3.1. Secondly, the raw data is pre-processed using stop-word removal method to remove the slang words & special characters. The processed data is shown in Fig. 3.2.

The processed data received after pre-processing is applied as input to NER and Sentiment analysis model. The NER model then segregates the data in respective entity category i.e. Name, Location and Organization which is shown in Fig. 3.3. Finally, the processed data received after pre-processing is also applied to the Classification model, so as to classify the opinions of the tweets as Positive, Negative and Neutral which is shown in Fig. 3.4.



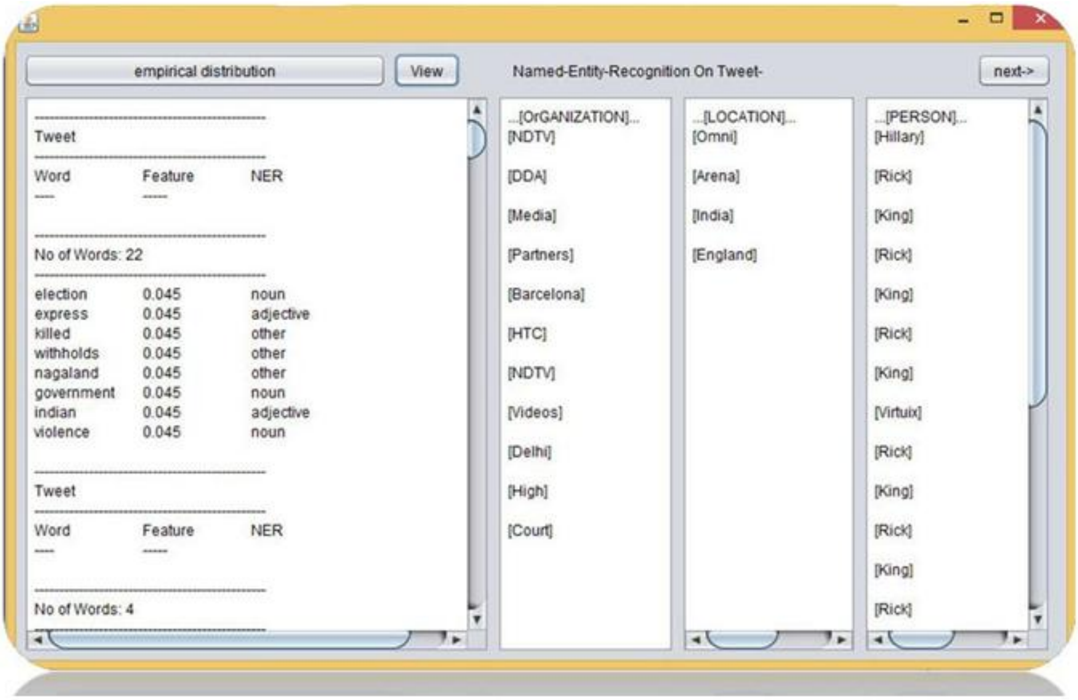
```
The Indian Express.............Two killed in violence, Nagaland government withholds ULB election https
†Alhanna Michelle†.............RT @ZaynJMNews: We voted for Pillowtalk by @zaynmalik for British Vide
Hotels-Scanner.us.............Best hotels and travel deals in Bonaire_Saint_Eustatius_and_Saba https:
†Alhanna Michelle†.............Enero duró como 15 minutos.□
†Alhanna Michelle†.............RT @ValeryJackson13: Duele el doble cuando al otro no le duele.
CNN.............Scandal threatens campaign of French presidential candidate François Fillon https://t.co
News18.............BJP Vice President Seriously Injured in Police Lathicharge in Kerala https://t.co/YwYV
Reuters Top News.............Trump's early moves spark alarm, resistance within the government https:
The Hindu.............West Indies cricketer Andre Russell has been banned for 12 months for breaching
https://t.co/SipEEIU1nK
Rajdeep Sardesai.............□□ https://t.co/H8Mv0dG3oS
Student Humours.............RT @KellyClarkson4u: The 30 Hottest Jennifer Lopez Photos https://t.co/4M
†Alhanna Michelle†.............A veces hay que perderse para saber quién iría a buscarte.
News18.............Eoin Morgan terms England batting as the "worst performance" in the last two years
https://t.co/19MvB4geAx
```

Fig.3.1 Tweet Data Retrieval



```
indian express killed violence nagaland government withholds ulb election
voted pillowtalk british video
hotelsscanner hotels travel deals bonaire_saint_eustatius_and_saba
como
duele el doble cuando al otro le duele
cnn scandal threatens campaign french presidential candidate fillon
news bjp vice president injured police lathicharge kerala
reuters top news trump s early moves spark alarm resistance government
hindu west indies cricketer andre russell banned months breaching antidoping regulations

rajdeep
student humours rt hottest jennifer lopez photos
veces hay perderse para saber buscarte
news eoin morgan terms england batting worst performance years
alhannah michelle
s
nice
good
indian express ieworld donald trump s travel ban divisive wrong theresa
hrithik roshan friend
hotelsscanner hotels travel deals barbados
teachable ready started podcasts member cameron walk process
```

Fig.3.2 Pre-processed data after removing special characters & slang words.



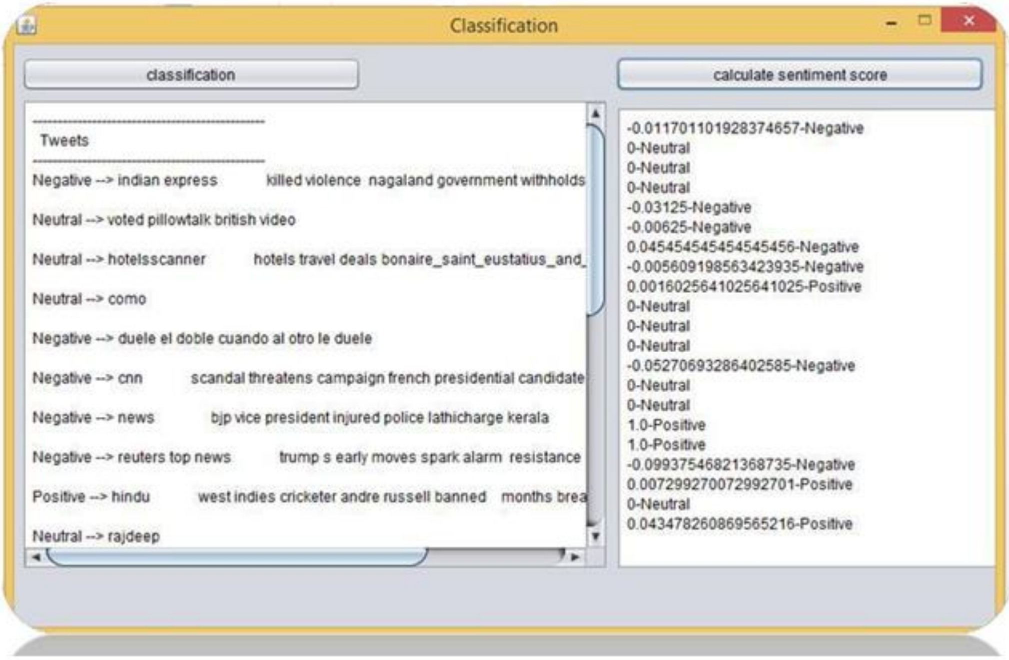
```
----------------------------------------
Tweet
----------------------------------------
Word            Feature         NER
----            -----

----------------------------------------
No of Words: 22
----------------------------------------
election        0.045           noun
express         0.045           adjective
killed          0.045           other
withholds       0.045           other
nagaland        0.045           other
government      0.045           noun
indian          0.045           adjective
violence        0.045           noun

----------------------------------------
Tweet
----------------------------------------
Word            Feature         NER
----            -----

----------------------------------------
No of Words: 4
----------------------------------------
```

```
...[OrGANIZATION]...
[NDTV]

[DDA]

[Media]

[Partners]

[Barcelona]

[HTC]

[NDTV]

[Videos]

[Delhi]

[High]

[Court]
```

```
...[LOCATION]...
[Omni]

[Arena]

[India]

[England]
```

```
...[PERSON]...
[Hillary]

[Rick]

[King]

[Rick]

[King]

[Rick]

[King]

[Virtuix]

[Rick]

[King]

[Rick]

[King]

[Rick]
```

Fig. 3.3 Categorized data after applying NER.



```
----------------------------------------
 Tweets
----------------------------------------
Negative --> indian express        killed violence  nagaland government withholds

Neutral --> voted pillowtalk british video

Neutral --> hotelsscanner        hotels travel deals bonaire_saint_eustatius_and_

Neutral --> como

Negative --> duele el doble cuando al otro le duele

Negative --> cnn        scandal threatens campaign french presidential candidate

Negative --> news         bjp vice president injured police lathicharge kerala

Negative --> reuters top news         trump s early moves spark alarm  resistance

Positive --> hindu        west indies cricketer andre russell banned   months brea

Neutral --> rajdeep
```

```
-0.011701101928374657-Negative
0-Neutral
0-Neutral
0-Neutral
-0.03125-Negative
-0.00625-Negative
0.045454545454545456-Negative
-0.005609198563423935-Negative
0.0016025641025641025-Positive
0-Neutral
0-Neutral
0-Neutral
-0.05270693286402585-Negative
0-Neutral
0-Neutral
1.0-Positive
1.0-Positive
-0.09937546821368735-Negative
0.007299270072992701-Positive
0-Neutral
0.043478260869565216-Positive
```

Fig. 3.4 Opinion of Tweets

IV. Conclusion

Twitter is a proficient social networking site which can be used as a great advantage by people. We have thus added to the importance of Twitter by enabling a person to know the name of the person, organiztion name and location by counting on the live tweet. The same feature is available offline as well. Also, the seperation of the comments as favourable, unfavourable or neutral helps positively.

References

[1]. VarshaSahayak,VijayaShete , ApashabiPathan,, *International Journal of Innovative Research in Advanced Engineering (IJIRAE)Issue 1, Volume 2 (January 2015)*

[2]. Apoorv Agarwal, BoyiXie ,Ilia Vovsha ,Owen Rambow, Rebecca Passonneau,, *Sentiment Analysis of Twitter Data*

[3]. MandefroLegesseKejela,DejeneEjiguDedefa,*NamedEntity Recognition for Afan Oromo Lap Lambert Academic PublishingGmbH KG, 2012*

[4]. Alec Go, et.al, *Twitter Sentiment Analysis, CS224N - Final Project Report, June 6, 2009.*

[5]. Charles A. Sutton ,*Efficient Training Methods for Conditional Random Fields 2008.*

[6]. Carl French, *Data Processing and Information Technology, 10th Edition.*