ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

Analysing the Impact of Government Programmes

Akash Mahajan[1], Rushikesh Divyavir [2], Nishant Kumar[3], Chetan Gade[4], L.A.Deshpande[5]

B.E Student, Department of Computer Engineering, Vishwakarma Institute of Information Technology Pune, India[1,2,3,4]

Professor, Department of Computer Engineering, Vishwakarma Institute of Information Technology Pune, India[5]

ABSTRACT: Sentiment Mining is technique which is used to determine whether given piece of sentence of a speaker is positive negative or neutral. Nowadays sentiment analysis is hot topic of research, many prior researchers focused on commercial approach like analyzing product reviews, movie reviews, and analysis of twitter data. The purpose of this paper is to analyze the opinions of peoples about the different government schemes and predicting either scheme is successful at people's level or not and also predicting the scope of the scheme.

KEYWORDS: Data Mining, Data Extraction, Data Pre-processing, Data Classification, Stemming.

I. INTRODUCTION

Opinion mining and sentiment analysis is a fast growing topic with various world applications, from polls to advertisement placement. Traditionally individuals gather feedback from their friends or relatives before purchasing an item, but today the trend is to identify the opinions of a variety of individuals around the globe using micro Blogging data. This paper discusses an approach where a publicized stream of tweets from the micro Blogging site are preprocessed and classified based on their emotional content as positive, negative and irrelevant; and analyzes the performance of various classifying algorithms based on their precision and recall in such cases.

In the past few years, there has been a huge growth in the use of micro blogging platforms. While there has been a fair amount of research on how sentiments are expressed in genres such as on-line reviews and news articles, how sentiments are expressed given the informal language and message-length constraints of micro blogging has been much less studied[3] In this paper we are developing a system which can determine the response of the individual. We are extracting the data from any social media sites like my gov. in, after extracting the data we are Pre-processing the data in which related and unrelated posts are found out. All unrelated posts are considered to be outliers. We are focusing only on related posts.
Another challenge of micro blogging is the incredible breadth of topic that is covered. It is not an exaggeration to say that people tweet about anything and everything. Therefore, to be able to build systems to mine sentiment about any given topic, we need a method for quickly identifying data that can be used for training [3].

II. RELATED WORK

This section includes the work done on related topics by various researchers. Following is the brief description of some of them:
G.Angulakshmi et al. [1], proposed a paper which discusses about an overview of Opinion Mining in detail with the techniques and tools.
Hansi Senaratne et al. [2], proposed a work in which they have analyzed spatial-temporal data also called VGI (volunteered geographical information) through which they have characterizes the trajectories or paths by using the concept of drift analysis.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

Efthymios Kouloumpis et al. [3], proposed a paper in which they have analyzed the micro blogging data in which they used features from existing sentiment lexicon in conjunction with micro blogging features where they found out that this micro blogging features(i.e., positive/negative/neutral emoticons ,presence of intensifiers and abbreviations) play vital role in analysis.
Sudipta Roy el al. [4], proposed a paper which represents the importance of sentiment or opinion in social networking Sites, they also reflect the overview of sentiment mining in social networks.
Apoorv Agarwal et al. [5] proposed a paper in which they have examine the twitter data using POS-specific prior polarity features and tree kernel to avoid tedious features.
Fuchs.G.Andrienko et al. [6] proposed a paper in which it we review these two key assumptions based on the results of applying a visual analytics approach to a dataset of georeferenced Tweets from Germany over eight months witnessing several large-scale flooding situations throughout the country. Our results con ram the potential of Twitter as a distributed 'social sensor' but at the same time highlight some caveats in interpreting immediate results. To overcome these limits we explore incorporating evidence from other data sources including further social media and mobile phone network metrics to detect, confirm and refine events with respect to location and time. We summarize the lessons learned from our initial analysis by proposing recommendations and outline possible future work directions.

Akshi Kumar et al. [7], in this paper they expound a hybrid approach using both corpus based and dictionary based methods to determine the semantic orientation of the opinion words in tweets. They propose and investigate a paradigm to mine the sentiment from a popular real-time micro blogging service, Twitter, where users post real time reactions to and opinions about everything.

Alexander Pak et al. [8], focuses on using Twitter, the most popular micro blogging platform, for the task of sentiment analysis. We show how to automatically collect a corpus for sentiment analysis and opinion mining purposes. They perform linguistic analysis of the collected corpus and explain discovered phenomena. Using the corpus, we build a sentiment classifier, which is able to determine positive, negative and neutral sentiments for a document.

Hassan Saif et al. [9], proposed a paper which introduce a novel approach of adding semantics as additional features into the training set for sentiment analysis. For each extracted entity (e.g. iPhone) from tweets, we add its semantic concept (e.g. Apple product) as an additional feature, and measure the correlation of the representative concept with negative/positive sentiment. They apply this approach to predict sentiment for three different Twitter datasets.

Varsha Sahayak et al. [10], proposed a paper we will discuss the existing analysis of twitter dataset with data mining approach such as use of Sentiment analysis algorithm using machine learning algorithms. An approach is introduced that automatically classifies the sentiments of Tweets taken from Twitter dataset as in. These messages or tweets are classified as positive, negative or neutral with respect to a query term. This is very useful for the companies who want to know the feedback about their product brands or the customers who want to search the opinion from others about product before purchase. They use machine learning algorithms for classifying the sentiment of Twitter messages using distant supervision which is discussed in. The training data consists of Twitter messages with emoticons, acronyms which are used as noisy labels discussed in. They also examine sentiment analysis on Twitter data.

III. ALGORITHM

A. *Stemming Algorithm*

Stemming algorithm is a process of normalization of language, in which various forms of the words are reduced to its common form [13] for example, the words connection, connections, connective, connected, connecting gets converted to their common form i.e. "connect".
In these languages words tend to be constant at the front, and to vary at the end [13]:
-ion
-ions
Connect-ive
-ed
-ing

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

The variable part is the 'ending', or 'suffix'. Taking these endings off is called 'suffix stripping' or 'stemming', and the residual part is called the stem [13]. Another way of looking at endings and suffixes is to think of the suffix as being made up of a number of endings. For example, the French word confirmative can be thought of as 'confirm' with a chain of endings,
-atif (adjectival ending - morphological)
Plus -e (feminine ending - grammatical)
Plus -s (plural ending - grammatical)
Endings fall into two classes, grammatical and morphological. The addition of -s in English to make a plural is an example of a grammatical ending. The word remains of the same type. There is usually only one dictionary entry for a word with all its various grammatical endings. Morphological endings create new types of word. In English -ise or -ize makes verbs from nouns ('demon', 'demonize'), -ly makes adverbs from adjectives ('foolish', 'foolishly'), and so on. Usually there are separate dictionary endings for these creations [13].

IV. PROPOSED SYSTEM

Sentiments are the words or sentences that represent view or opinion that is held or expressed that can be positive, negative or neutral [10]. The field of Sentiment Analysis expresses the feelings of an individual about some particular topic. We are going to propose a system which extracts the data from my-gov.in site, and predicts the future scope of the scheme weather the scheme will be successful or not. The proposed Sentiment Analysis on data is based on two important parts via Data Extraction, pre-processing of extracted data and classification.

I. Data Sources

User opinion is major criteria for the improvement of quality of service. Blogs, review sites, data and micro blogs provide a good understanding for the deliverable level of the products and services provided to customers [1].
There has been a lot of prior research done on review sites, blogs, and micro blogging site which improves the quality of service provided to the customer. In the present day, the challenge for governments is how to move on from focusing on service delivery to providing people-centered applications. In other words, government's success relies on effectively communicate their messages to citizens and build strong alliances with them by empowering their participation in the decision-making process. For our idea we have taken the official government site data i.e. my-gov.in, which brings the government closer to the common.
Citizens by the use of on-line platform creating an interface for exchanging the ideas and views between citizen and experts to contribute in social and economic transformation of the country.
It contributes various attributes like:

a. Discussion
b. Tasks
c. Talk
d. Polls
e. Blogs

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

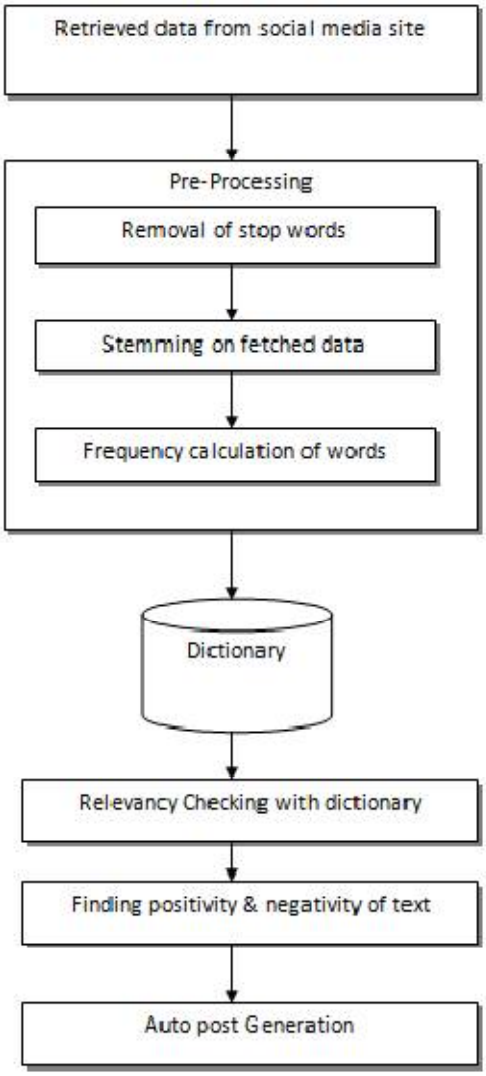

Fig.1. System Architecture

Around 1.78 million users are part of this platform which contributes their ideas on different issues and about 10,000 Posts were generated per week on different issues. It keeps the citizens stick to the important policy issues and governance like Clean Ganga, Skill Development, Digital India and many more. The site contains around 45 different schemes of different issues.

For our project we took 3 different schemes:

a. Swachh Bharat (Clean India)
b. Beti Bachao Beti Padhao
c. Digital India

II. Background of system

The system is divided into different modules.

1. Data Extraction Module: This module performs the operations of data fetching. There are various tools are available for extracting the data from heterogeneous sources and also various API's are also available.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

2. Data Pre-processing Module: The data which is extracted is completely raw in nature. The real time data contains lot of errors and inconsistencies. Before storing it into system database it should go through cleaning operations. It performs removal of stop words, Stop words are usually refer to the most common words in a language like a, an, for, the, I, will, when etc. Next phase in data pre-processing is applying Stemmer algorithm which will make the source date more compact and versatile.

Fig.2. Stop words list used in our project

Tasks in Data Pre-processing:

a. Data cleaning: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies [12].
b. Data integration: using multiple databases, data cubes, or files [12].
c. Data transformation: normalization and aggregation [12].
d. Data reduction: reducing the volume but producing the same or similar analytical results [12].
e. Data discretization: part of data reduction, replacing numerical attributes with nominal ones [12].

3. Dictionary: This is the database that will store the predefined keywords.
4. Data Classification Module: Now the posts/comments are classified into relevant tweet or irrelevant tweet depending upon the dictionary. When any new data is extracted first the relevancy of the data is checked. The data is relevant if and only if it matches with the keyword which is mention in dictionary. If the data is relevant then it is further analyzed for looking into response whether it is positive opinion or negative opinion, simultaneously positive and negative counters are declared which will count the number of positive and negative responses.
5. Auto post Module: This module takes the input from its predecessor i.e. Data classification Module. For generation of automated post the module depends upon the two factors: Number of posts available with us, number of positive and negative value.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

Condition 1: No. of posts/comments.

Condition 2: No. of positive and negative counts.

Condition 1	Condition 2	Statement
TRUE	TRUE	Scheme is showing good response at people's level, it can be used in upcoming years.
TRUE	FALSE	People are much aware about the scheme but it does not show good response from people's side.
FALSE	TRUE	People are not much aware about the scheme, scheme needs awareness.
FALSE	FALSE	Government should improve the scope of the scheme and also it should think over the scheme that it should be continued in future or not.

Fig. 3. Following table showing the possible outcomes.

Input1: - Number of posts/comments.
Input2: - values from positive and negative counter.
Output: - Depending upon the Input1 and Input 2 an automated post is generated.

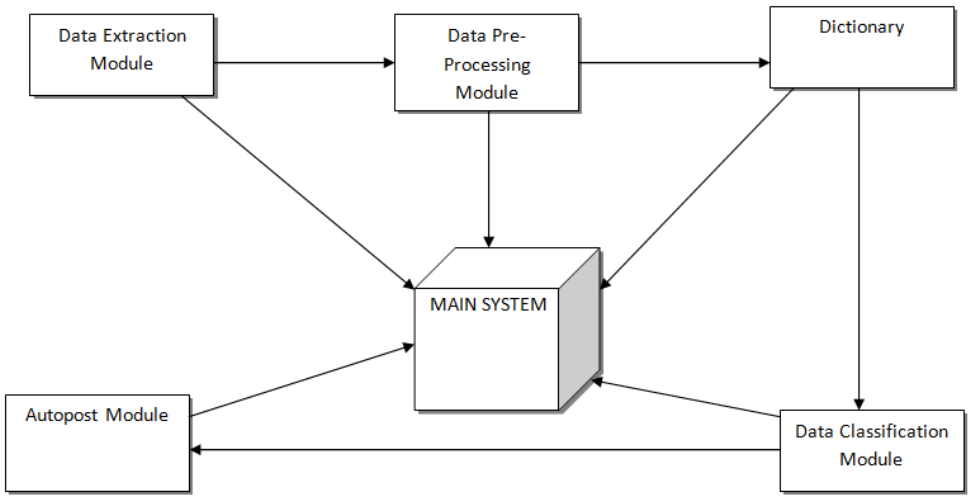

Fig. 4. System Modules

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

V. RESULTS

For our case study we have extracted the data from government site WWW.my gov. In. MYGOV is a citizen engagement platform formed by government of India to promote the participation of citizens in their countries development [11].
It consists of various government schemes onto which various users put forward their views on particular schemes. It also allows users to upload documents in various formats like PDF, image file, audio etc. The website is hosted and managed by National Informatics Center (NIC) [11].
For our project we take different schemes like Swachh Bharat Abhiyan, Beti Bachao Andolan, Digital India Etc. which are the Indian government programs. These are the programs which gets good response from the Indian peoples. From these 3 different types of schemes we extracted comments of peoples from these 3 different schemes and we stored the extracted comments in a file.

Fig. 5. Extracted Data

Now the data that we have is raw data. Raw data is called as primary data it is the data collected from multiple sources, here the sources are different schemes from which we are extracting the data, but this extracted data is not subjected to processing. This raw data contains many errors, inconsistencies etc. so for that purpose we need to pre process the data. The stop words and stemmer algorithm is applied on it which will make the source date more compact and versatile.
The next phase is the frequency count where a upper limit is set for the keywords A counter is initialized which counts the frequency of each word and then only those word get considered into context which will having the occurrence over the limit which has been set. All these words will be stored in a dynamic dictionary which increments as the newly source file will be preprocessed and the frequently occurring words will be keep added to it. The dictionary contains some predefined keywords according to each scheme which is used for matching purpose initially.

Here we are giving some keywords like
For:-

1. Swachh Bharat Abhiyan: clean, dirty, toilets, society, NGO, human health organization etc.

2. Beti Bachao Beti Padhao: school infrastructure, urban areas, Woman, Victims etc.

3. Green India: Natural resources, Green cooking fuels, Woody Biomass, Solar panels etc.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

4. Rural development: Pension, card holders, BPL, Aadhar card, agriculture, MNREGA, Panchayat etc.

5. Clean Ganga: Wastage, Tourism, Recycle Water, fertilization, Clean Ganga, Dead Bodies etc.

The words in the dictionary will be simultaneously classified in the positive, negative or average words irrespective of any other context. When the source file gets fetched then the whole process being applied over it and matching of the words is done of file with the dictionary. According to that the measure of positiveness (relevancy), negativity (irrelevancy) and neutral will be taken out. From the desired data we can also generate bar graphs and pie charts so that the data can be further classified according to the time stamps.

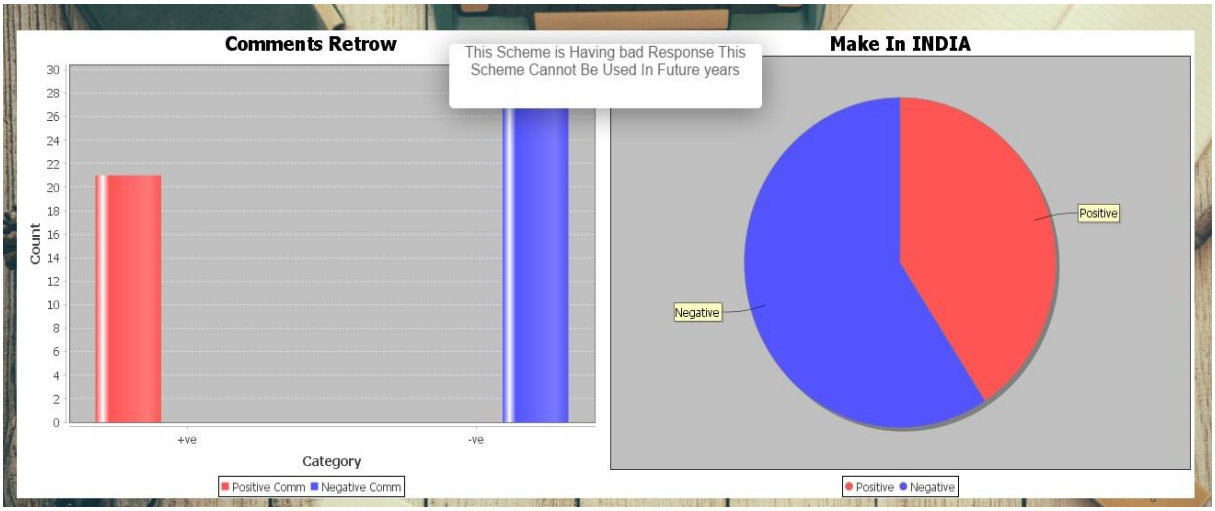

Fig.6. Scheme 1 Analysis

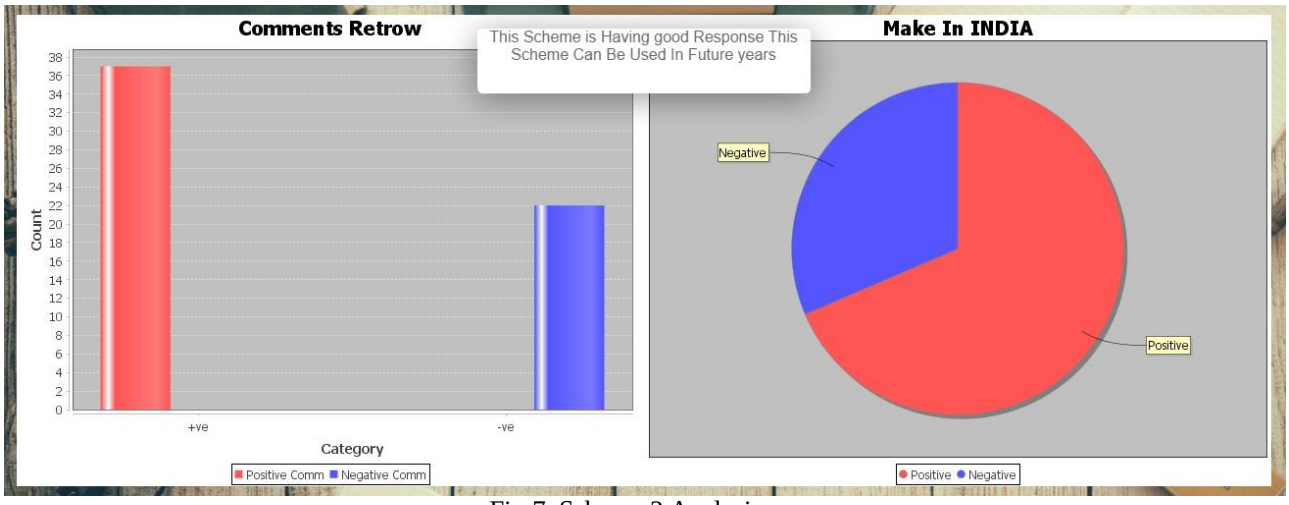

Fig.7. Scheme 2 Analysis

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

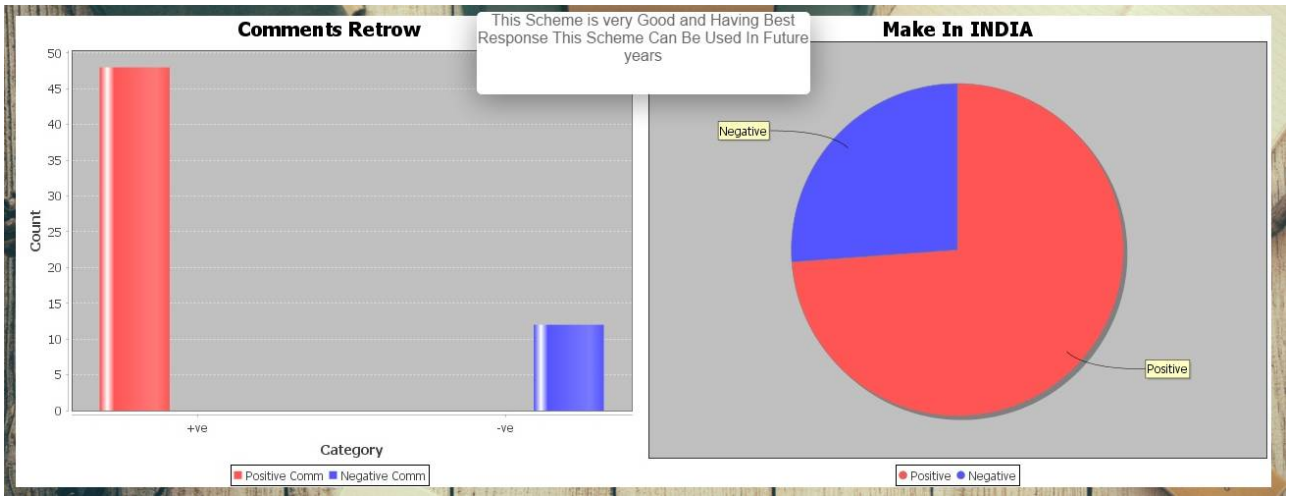

Fig.8. Scheme 3 Analysis

VI. CONCLUSION

Our experiments show that how we are analyzing the mood depending upon the user's response. In our experiment we have taken three different schemes via Bharat Swachh Abhiyan, Digital India, and Beti Bachao Andolan. We have extracted the data from these three schemes, we analyzed the data after analyzing we are predicting the future scopes of these schemes.

REFERENCES

1. G.Angulakshmi, and Dr.R.ManickaChezian, "An Analysis on Opinion Mining: Techniques and Tools" International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE) Vol. 3, Issue 7, July 2014.
2. Hansi Senaratne, Arne Bröring, Tobias Schreck, and Dominic Lehle, "Moving on Twitter: Using Episodic Hot spot and Drift Analysis to Detect and Characterize Spatial Trajectories".
3. Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore, "Twitter Sentiment Analysis: The Good the Bad and OMG!"
4. Sudipta Roy, Sourish Dhar, Arnab Paul, Saprativa Bhattacharjee, Anirban Das, and Deepjyoti Choudhury, "Current Trends Of opinion mining and sentiment analysis in social network" International Journal of Research in Engineering and Technology (IJRET) Vol. 2, Issue 2, Dec 2013.
5. Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau, "Sentiment Analysis of Twitter Data".
6. Fuchs, G., Andrienko, N., Andrienko, G., Bothe, S. Stange, H. (2013). "Tracing the German Centennial Flood in the Stream of Tweets: First Lessons Learned", Paper presented at the 2nd ACM SIGSPATIAL International Workshop on Crowd sourced and Volunteered Geographic Information (GEOCROWD) 2013, 5 Nov 2013, Orlando, FL, US.
7. Akshi Kumar, and Teja Sebastian, "Sentiment Analysis on Twitter", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012.
8. Alexander Pak and Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining".
9. Hassan Saif, Yulan He, and Harith Alani, "Semantic Sentiment Analysis of Twitter".
10. Varsha Sahayak, Vijaya Shete, and Apashabi Pathan, "Sentiment Analysis on Twitter Data", International Journal of Innovative Research in Advanced Engineering (IJIRAE) Vol. 2, Issue 1, Jan 2015.
11. https://en.wikipedia.org/wiki/MyGov.in
12. https://www.cs.ccsu.edu/markov/ccsu_courses/datamining.html
13. https://xapian.org/docs/stemming.html