



POS TAGGING AND SENTIMENT ANALYSIS OF TWEETS USING STATE MACHINE

Vishnu Priya R, Sparsha Saha, Sridhar Swain

School of SCSE, VIT Chennai, Tamil Nadu, India 600127

{vishnupriya.r@vit.ac.in, sparsha.saha2015@vit.ac.in, sridhar.swain2015@vit.ac.in}

Abstract – With the rapid growth of social networking sites such as Facebook, Instagram, Twitter etc. which are very popular among people as they provide a platform to share and express their views on the current trending topics and have discussion with different people, post messages across the world, comment and share their opinions. So it's important to recognize the sentiment of each and every person in their post and classify whether it is a negative or a positive feedback. This can be done by using sentiment analysis which involves classifying of emotions, attitude, opinions, feedbacks, reviews of the post provided by the users. The sentiments of the words in a sentence can be vague, manipulating, contradicting, provoking or mixed emotions which can be easily understood by the humans but it is difficult for a machine to get the exact sentiment of the users. The dataset consist of tweets by the 'President of America, Donald Trump'. Assuming that the tweets are in English language with no slang words used. Sentiment analysis can be done using opinion mining, word embedding, Part of speech tagging and microblogging. So, it is not easy to determine the sentiment of the sentences just from the single approach. Hence the main focus is to use part of speech for tagging each word in a sentence and then passing to a finite state machine which propose a state machine which would analyze the tweets and categorize them into positive, negative or neutral sentiments using POS tags.

Nowadays, the way people express their opinions and views is highly changed by the internet. Online forums, Product reviews website, social media, blog post are the interactive platforms where the user inform, express, influence others. Social Media websites like Facebook, Instagram, Twitter and YouTube produce a huge amount of the data which can be in the form of posts, comments, images, tweets and videos respectively. This huge data involves different semantic dimensions for a given dataset resulting in multiple views, contradicting opinions, rational opinion and manipulating views. So, to manage this vagueness of the opinions of people from different backgrounds on the same topic. Hence, the core for this research paper is to come up with the powerful state machine which can deal with the multiple views of people and estimate the morphological syntax for a given tweet. The sentiments are handled by the state machine which helps to find a generalized sequence in sentences from tweets. By tagging each word of it by parsing the sentence through the nodes of the state machine, i.e. we are going to create a graph of nodes with 36 tags (<https://cs.nyu.edu/grishman/jet/guide/PennPOS.html>) and connect them as if they occur in sequence in a sentence. Furthermore, it will find how a word with a particular tag in a sentence can influence its sentiment. In addition to exhaustive search of sentiment can be performed which can be further classify into the following:

Keywords–Part-of-speechTagging, TextBlob, Tweets, Polarity, Tags, tweepy library ,python

1. INTRODUCTION

1. The words that can convert a sentence with positive or 0 sentiment to negative.

2. The words that can convert a sentence with negative or 0 sentiment to positive.

After segregating the words, values will be assigned through the state machine to find the tag of the word and then check if the word exists in either of the above mentioned classification. Assuming the initial sentiment of the sentence is zero. If the classification matches then the alteration of the current sentiment depending on the following:

1. Current Sentiment
2. Word
3. Tag of the word

For a given input sentence the sentiment value depends on the pos tag of the initial word and the value keep on changes as we traverse the whole sentence and the final sentiment of the sentence will be the value of the last word of input sentence.

2. LITERATURE SURVEY –

The paper ‘Sentiment analysis of Twitter Data’ published under ‘International Journal of Innovations in Engineering and Technologies’ uses the approach of opinion mining and classifying the sentiment into positive, negative or neutral. The four name task in their paper is entity extraction and categorization, aspect extraction and categorization, aspect sentiment classification and opinion generation to analyze the sentiment of the tweets. The paper extracted the twitter data from Twitter API. The data preprocessing simplifies the data by handling the punctuations, spelling correction, stemming etc. to remove the noise and then applying the classification algorithm and result of the sentiment analysis is provided using pie chart.

The paper ‘Sentiment analysis of Twitter Data : A Survey of Techniques’ published under

‘International Journal of Computer Applications’ uses machine learning techniques with the feature extracted as vectors. The most commonly used

feature in sentiment classification are term presence and its term frequency, part of speech information, negations and opinion words. Level of sentiment analysis taken into consideration are document level, sentence or phrase level, feature level and word level. The classification used in this paper are Naïve Bayes and Support Vector Machine (SVM).

The paper ‘Sentiment Analysis : Determining People’s emotions in Facebook’ published by N.Azmina M.Zamani, SITI Z Z. Abidin, M.Z.Z Abidin, Nasiroh Omar. This paper uses a computerized tool to extract the words in a Facebook post and comments so that people emotion towards the issue that can be determine if the overall feedback of the people is happy, unhappy or emotionless.

The paper ‘Sentiment analysis of Twitter Data’ published under ‘International Journal of Innovative Research in Advanced Engineering’. This paper uses two resource a hand notated dictionary for emoticons and acronym dictionary gathered from web. The used different machine learning and feature extractor. The machine learning classifier are Naïve Bayes, Maximum Entropy and Support Vector Machine whereas the feature extractor consists of unigrams, bigrams, unigrams and bigrams and unigram with part of speech tags. The steps involved in the propose system of this paper are retrieval of tweets, preprocessing of the extracted data, parallel processing, sentiment scoring module and output sentiments.

The paper ‘Sentiment analysis of Twitter Data’ published by Apoorv Agarwal, Boyi Xie, Illia Vovsha, Owen Rambow, Rebecca Passonneau. The contribution of this paper are POS specific prior polarity features and use of a tree kernel to obviate the need for tedious feature engineering. The compilation of an acronym dictionary used in this paper helps to translate abbreviated words and emoticons like lol is ‘Laughing out Load’ and so on. Furthermore the design of tree kernel for calculating the similarity of two trees is done by partial tree kernel whereas this partial tree which

includes the sub tree in which non-adjacent branches become adjacent by excising other branches.

The paper ‘Sentiment analysis on Twitter’ published by Akshi Kumar and Teeja Mary Sebastian from International Journal of Computer Science .In this paper , the microblogging sites used is twitter which allows corpus to have different dimensions and diverse application. The main motivation of this paper is to propose a model which retrieve tweets on a topic through the use of twitter API and calculates the sentiment based on score. The categorization of tweets are based on message length, writing techniques, availability ,topics and real time with addition of basic terminology related to twitter like emoticons, targets, hashtags and special symbols. To uncover the sentiment firstly extraction of opinion words in tweets as they are combination of adjective along with verbs and adverbs .The overall tweet sentiment is then calculated by use of linear equation with incorporation of emotion intensifiers.

The paper “Sentiment Analysis on feedback data from global support service ” by Gamon published in 2004. The main aim of their paper is to analyze the role of linguistic features like POS tags in message which extensive feature analysis and feature selection to illustrate abstract linguistic features for their classifier accuracy .The dataset used consist of 100 abstract linguistic features as well as hard unigram baseline.

3. METHODOLOGY

3.1 Dataset Description-

Tweets from twitter carry different sentiments which challenges the sentiment analysis while shaping the hidden sentiments in it. The dataset consist of the latest trending hashtags on twitter

Dataset used here is 800(for the sake of experiment) from the recent tweets of President Donald Trump. The tweets are extracted using

tweepy library in python which enable us to fetch tweets of Donald Trump who goes by username ‘realdonaldtrump’.

Received message last night from XI JINPING of China that his meeting with KIM JONG UN went very well and that KI
M
For years and through many administrations everyone said that peace and the denuclearization of the Korean Penins
u
THE SECOND AMENDMENT WILL NEVER BE REPEALED As much as Democrats would like to see this happen and despite the wo

I am very pleased to welcome the opioid memorial to the President s Park in April I encourage all to visit and re
m
RT realDonaldTrump Because of the 700 amp 716 Billion Dollars gotten to rebuild our Military many jobs are create
d and our Military is
Trade talks going on with numerous countries that for many years have not treated the United States fairly In the
end all will be
Great news MAGA
So much Fake News Never been more voluminous or more inaccurate But through it all our country is doing
The economy is looking really good It has been many years that we have seen these kind of numbers The underlying
HowieCarrShow just wrote a book which everyone is talking about He was a great help He is a veteran journalist
President Donald J Trump Proclaims March 25 2018 as Greek Independence Day A National Day of Celebration of Gr

Figure1. Dataset

3.2 TAGS - Taking the tags stated in the page of Pennsylvania institute of tech. And considering these are the only tags possible, which has been mapped in reverse index fashion :

```
{'CC':0,'CD':1,'DT':2,'EX':3,'FW':4,'IN':5,'JJ':6,'JJR':7,'JJS':8,'LS':9,'MD':10,'NN':11,'NNP':13,'NNPS':14,'NNS':12,'PDT':15,'POS':16,'PRP':17,'PRP$':18,'RB':19,'RBR':20,'RBS':21,'RP':22,'SYM':23,'TO':24,'UH':25,'VB':26,'VBD':27,'VBG':28,'VBN':29,'VBP':30,'VBZ':31,'WDT':32,'WP':33,'WP$':34,'WRB':35,'end':36}
```

Tags	POS Description
JJ	Adjective
JJR	Adjective Comparative
JJS	Adjective superlative
RB	Adverb
RBR	Adverb comparative
RBS	Adverb superlative
VB	Verb
VBD	Verb Past Tense
VBG	Verb present participle

VBN	Verb past participle
VBP	Verb, known third person singular present
VBZ	Verb third person singular Present.

TABLE2.1

Sentiment of a sentence is determined by Verb, Adjective and Adverbs, Hence 'JJ','JJR','JJS', 'RB', 'RBR', 'RBS', 'VB', 'VBD', 'VBG', 'VBN', 'VBP' and 'VBZ' are the tags which are considered to change the sentiment of a sentence.

3.3 Experimental Set Up - After creating a twitter developer account and getting the ACCESS_TOKEN and ACCESS_SECRET, the tweepy library fetches tweets of a particular user. As per preprocessing of text, the case of words has to be changed and removal of all special characters but space for delimiting the words.

Next step is to initialize the reverse indices of tags and initiate a 36X36 matrix all 0 initially. TextBlob is used as reference and training API. TextBlob uses NLP techniques to tag and find sentiment of a given sentence.

3.4 Sentiment Conversion –

The sentence consist of verb ,adjective and adverb which mostly affect the sentiment in a sentence. Checking how removing these words affect the polarity value . Below are few sentiment converting words from postoneg and negtopos respectively.

posToNeg words:

Words like terrible, hard, insecure, sadly, weak, etc. are the words affecting the sentiment in a negative way.

```
{('stupid', 'JJ'): 1, ('hard', 'RB'): 1, ('insecure', 'JJ'): 1, ('half', 'JJ'): 1, ('very', 'RB'): 1, ('sadly', 'RB'): 1, ('not', 'RB'): 1, ('weak', 'JJ'): 1, ('terrible', 'JJ'): 1, ('unnecessary', 'JJ'): 1, ('tough', 'JJ'): 1, ('simple', 'JJ'): 1, ('past', 'JJ'): 1, ('disappointed', 'VBD'): 1, ('dishonest', 'JJ'): 1, ('corrupt', 'JJ'): 1, ('small', 'JJ'): 1, ('heavy', 'JJ'): 1, ('difficult', 'JJ'): 1, ('secret', 'JJ'): 1, ('unable', 'JJ'): 1, ('long', 'JJ'): 1, ('approximately', 'RB'): 1, ('desperate', 'JJ'): 1, ('unfair', 'JJ'): 1, ('boring', 'JJ'): 1, ('illegal', 'JJ'): 1, ('weakly', 'JJ'): 1, ('outdated', 'VBN'): 1, ('horrible', 'JJ'): 1, ('tragic', 'JJ'): 1, ('long', 'RB'): 1, ('worse', 'JJR'): 1, ('foreign', 'JJ'): 1, ('little', 'JJ'): 1, ('come', 'VB'): 1, ('failed', 'JJ'): 1, ('roughly', 'RB'): 1, ('previous', 'JJ'): 1, ('everyday', 'JJ'): 1, ('other', 'JJ'): 1, ('bad', 'JJ'): 1, ('hard', 'JJ'): 1, ('fake', 'JJ'): 1, ('desperately', 'RB'): 1, ('stupidly', 'RB'): 1, ('expected', 'VBN'): 1, ('sad', 'JJ'): 1, ('guilty', 'JJ'): 1, ('down', 'RB'): 1, ('cutting', 'VBG'): 1, ('political', 'JJ'): 1, ('hate d', 'JJ'): 1, ('sloppy', 'JJ'): 1}
```

Figure 2. Neg word

negToPos words: Words like happy, successful, brave, important, innocent can convert sentiment of negative sentence to positive.

```
{('successful', 'JJ'): 1, ('happy', 'JJ'): 1, ('brave', 'VBP'): 1, ('important', 'JJ'): 1, ('most', 'RBS'): 1, ('innocent', 'JJ'): 1, ('more', 'RBR'): 1, ('including', 'VBG'): 1, ('free', 'JJ'): 1, ('best', 'JJS'): 1, ('written', 'VBN'): 1, ('easy', 'JJ'): 1, ('incredible', 'JJ'): 1, ('extraordinary', 'JJ'): 1, ('far', 'RB'): 1, ('right', 'JJ'): 1, ('live', 'JJ'): 1, ('unpopular', 'JJ'): 1, ('able', 'JJ'): 1, ('filled', 'VBN'): 1, ('really', 'RB'): 1, ('dishonest', 'JJ'): 1, ('fantastic', 'JJ'): 1, ('meaningful', 'JJ'): 1, ('fast', 'RB'): 1, ('true', 'JJ'): 1, ('secure', 'VB'): 1, ('strong', 'JJ'): 1, ('first', 'JJ'): 1, ('right', 'VBD'): 1, ('large', 'JJ'): 1, ('legal', 'JJ'): 1, ('win', 'VB'): 1, ('proud', 'JJ'): 1, ('much', 'RB'): 1, ('greatest', 'JJS'): 1, ('new', 'JJ'): 1, ('nearly', 'RB'): 1, ('winning', 'VBG'): 1, ('first', 'RB'): 1, ('brave', 'JJ'): 1, ('welcome', 'VB'): 1, ('ready', 'JJ'): 1, ('proudly', 'RB'): 1, ('been', 'VBN'): 1, ('good', 'JJ'): 1, ('powerful', 'JJ'): 1, ('highly', 'RB'): 1, ('many', 'JJ'): 1, ('major', 'JJ'): 1, ('very', 'RB'): 1, ('nice', 'JJ'): 1, ('tremendous', 'JJ'): 1, ('full', 'JJ'): 1, ('loved', 'VBN'): 1, ('real', 'JJ'): 1, ('higher', 'JJR'): 1, ('beloved', 'JJ'): 1, ('great', 'JJ'): 1, ('better', 'JJR'): 1, ('intense', 'JJ'): 1, ('serious', 'JJ'): 1, ('amazing', 'JJ'): 1, ('popular', 'JJ'): 1, ('strongly', 'RB'): 1, ('most', 'JJS'): 1, ('interested', 'VBN'): 1, ('interested', 'JJ'): 1, ('quickly', 'RB'): 1, ('special', 'JJ'): 1, ('wonderful', 'JJ'): 1, ('smart', 'VB'): 1, ('more', 'JJR'): 1, ('beautiful', 'JJ'): 1, ('love', 'VBP'): 1, ('appreciated', 'VBN'): 1, ('be', 'VB'): 1, ('complete', 'JJ'): 1, ('wants', 'VBZ'): 1, ('live', 'VB'): 1, ('positive', 'JJ'): 1, ('looking', 'VBG'): 1, ('much', 'JJ'): 1, ('top', 'JJ'): 1}
```

Figure 3. Pos words

3.5 ALGORITHM –

Initialize posToNeg = [] //will contain words that converts pos to neg

Initialize negToPos = [] //will contain words that converts neg to pos

For all tweet in tweets{

blob = TextBlob(tweet) //process the sentence

tagged = blob.tags

for i in [1 until tags.length]{

matrix[x][y] +=1 //Create and edge between the nodes if it occurs sequentially in sentence

```

        polarity = blob.getPolarity() // Get
        polarity of the sentence
        for tag in tags{
            if(polarity(tag)>=0    and
            polarity<1) //If the word can convert -ve
            sentence to +ve
                negToPos.push(tag)
            if(polarity(tag)<=0    and
            polarity>1 )//If the word can convert +ve
            sentence to -ve
                }
        }
    }
}

```

3.6 Scoring Module

The sentiment property returns a namedtuple in the form of S[Polarity, subjectivity] and Polarity score is floating number ranging from -1.0 to 1.0. Subjectivity range from 0-1 where 1 is subjective and 0 is very objective.

To determine the polarity:

A determiner is attached to the every node in the graph except Subject(S), Conjunction(C) and Sentence Conjunction(SC). This determiner will determine the polarity of the given the Object, Adjective, Noun or Adverb. A determiner can be words like “not”, “been”, etc. Negative determiner with negative node gives positive local polarity.

POLARITY TABLE 2.2

Polarity	Node	Result	Points
-ve	-ve	+ve	1
-ve	+ve	-ve	-1
	+ve	-ve	-1
	+ve	+ve	1
	-ve/+ve	N	0

4. Observation

The adjacency matrix formed by passing each word to a state machine and forming the links between adjective, verb and adverb. The positive and negative tags are given according to the polarity score calculated from scoring module.

The values below are provided by TextBlob for finding the sentiment affecting words and tagging is done by state machine.

Example-1

Sentence : Wishing all of those celebrating Hanukkah around the world a happy and healthy eight nights in the company

TAGS:

```

[('Wishing', 'VBG'), ('all', 'DT'), ('of', 'IN'), ('thos
e', 'DT'), ('celebrating', 'VBG'), ('Hanukkah', 'NNP'),
('around', 'IN'), ('the', 'DT'), ('world', 'NN'), ('a', 'D
T'), ('happy', 'JJ'), ('and', 'CC'), ('healthy', 'JJ'),
('eight', 'CD'), ('nights', 'NNS'), ('in', 'IN'), ('the',
'DT'), ('company', 'NN')]

```

Polarity = 0.65 i.e +1

Removing ‘**Wishing**’ from the sentence does no change

Polarity(“all of those celebrating Hanukkah around the world a happy and healthy eight nights in the company”) = **0.65**

Removing “**celebrating**” from the sentence also does no change Polarity(“Wishing all of those Hanukkah around the world a happy and healthy eight nights in the company”) = **0.65**

Removing “**happy**” from the sentence ,Polarity(Wishing all of those celebrating Hanukkah around the world a and healthy eight nights in the company) = **0.5**

That means happy is one of the word which gives positive sentimental affect to the sentence.

However it cannot be considered to change the polarity of the sentence as it is still positive.

Further removing “healthy” Polarity(“Wishing all of those celebrating Hanukkah around the world a and eight nights in the company”) = 0.0 that means combination of happy and health supported the +ve sentiment of the sentence

Example 2

Sentence : The Tax Cut Reform Bill including Massive Alaska Drilling and the Repeal of the highly unpopular Individual.

```
[('The', 'DT'), ('Tax', 'NNP'), ('Cut', 'NNP'), ('Reform', 'NNP'), ('Bill', 'NNP'), ('including', 'VBG'), ('Alaska', 'NNP'), ('Drilling', 'NNP'), ('and', 'CC'), ('the', 'DT'), ('Repeal', 'NNP'), ('of', 'IN'), ('the', 'DT'), ('highly', 'RB'), ('unpopular', 'JJ'), ('Individual', 'NNP')]
```

Polarity of while sentence = 0.53333

Removing “massive” Polarity(“The Tax Cut Reform Bill including Alaska Drilling and the Repeal of the highly unpopular Individual”) = 0.08 the sentiment did decrease but did not change.

Removing “highly” Polarity(“The Tax Cut Reform Bill including Massive Alaska Drilling and the Repeal of the unpopular Individual”) = 0.3 the polarity is not changed yet.

Removing(“unpopular”) Polarity(“The Tax Cut Reform Bill including Massive Alaska Drilling and the Repeal of the Individual”) = 0.0
Hence combination of highly and unpopular can affect the polarity of the sentence.

5. RESULT

By the help of the designed state machine, the result for sentiment analysis on twitter concludes that the accuracy of the testing data set is 81.8% .

The test experiment is done by getting all such words and/or combinations of them. After running

the algorithm for the tweets by Donanld Trump the following state machine is obtained -

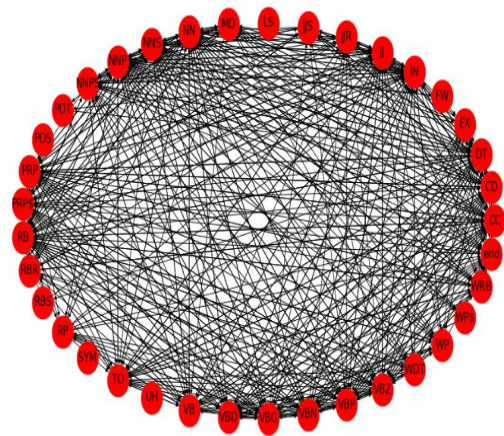


Figure 4

6. CONCLUSION

Testing sentences:

- "I ate a delicious hamburger.",
- "Zack joins the Sprockets Academy Explorers Club at school",
- "Zack gets paired up with Seth, the class bully, and that's dreadful but Zack is excited when he finds a huge galaxy gemmite",
- "A gemmite that large was found in last 100 years",
- "My favorite part is that the bird tried and learned that she could fly",
- "The bird has to decide if it will try to fly",
- "One fateful day they make the acquaintance of the Cotton family",
- "I wish I could fly."
- "He is good"
- "Abhimanyu is not as good as Anirudha."

The above sentences were tested on the proposed algorithm and the following difference were obtained on the predicted and targeted score. Hence ,the accuracy came out to be 90%.

Index	Predicted	Target	
1	1.0	1.0	[('delicious', 'JJ')]
2	0.0	0.0	[]
3	-1.0	-0.07	[('s', 'VBZ')]
4	1.0	0.11	[('large', 'JJ')]
5	1.0	0.65	[('fly', 'VB')]
6	1.0	0.8	[('fly', 'VB')]
7	0.0	0.0	[]
8	1.0	0.8	[('fly', 'VB')]
9	1.0	0.7	[('good', 'JJ')]
10	-1.0	0.7	[('not', 'RB'), ('as', 'RB'), ('good', 'JJ')]
Accuracy : 0.9			

Hence, the selection features are POS tagging and passing it to the proposed state machine. The state machine form links as the verb ,adjective and adverb occurs in a sentence and the polarity value is given by removing and checking respective word whether it is bringing the difference in the polarity value of the given sentence. This way we can figure out the vital word which actually affects the sentiment of the sentences. The polarity score achieved by analyzing the tweets is not different from the sentiment analysis from other genres.

7. REFERENCES

- [1] <http://www.cs.columbia.edu/~julia/papers/Agarwaletal11.pdf>
- [2] <https://www.ijcsi.org/papers/IJCSI-9-4-3-372-378.pdf>
- [3] <https://arxiv.org/ftp/arxiv/papers/1601/1601.06971.pdf>
- [4] <http://ijiet.com/wp-content/uploads/2016/04/37.pdf>
- [5] <https://www.ijcsi.org/papers/IJCSI-9-4-3-372-378.pdf>
- [6] <https://pdfs.semanticscholar.org/abbc/1e4150e8fc40300d3b5d1ac05153583f4f95.pdf>
- [7] <http://www.ijcsmc.com/docs/papers/April2017/V6I4201712.pdf>
- [8] <http://sentic.net/wisdom2017fuentes.pdf>
- [9] <https://www.hindawi.com/journals/cmmm/2017/5140631/>
- [10] <http://ieeexplore.ieee.org/document/7346686/>
- [11] <http://ieeexplore.ieee.org/document/7976920/>
- [12] <https://pdfs.semanticscholar.org/c114/7f3d9b46ff0a0c7c43b668123cb15a26120d.pdf>
- [13] <http://www.cs.columbia.edu/~julia/papers/Agarwaletal11.pdf>
- [14] <https://arxiv.org/ftp/arxiv/papers/1601/1601.06971.pdf>
- [15] <https://pdfs.semanticscholar.org/abbc/1e4150e8fc40300d3b5d1ac05153583f4f95.pdf>
- [16] <http://www.ijirst.org/articles/IJIRSTV3I10001.pdf>
- [17] <http://www.hindawi.com/journals/cmmm/2017/5140631/>
- [18] <https://pdfs.semanticscholar.org/d5b7/264fc852e65bfaecdde1d42e7af42d9deb06.pdf>
- [19] <http://www.cs.columbia.edu/~julia/papers/Agarwaletal11.pdf>
- [20] <https://pdfs.semanticscholar.org/c114/7f3d9b46ff0a0c7c43b668123cb15a26120d.pdf>
- [21] http://ceur-ws.org/Vol-917/SDAD2012_6_Spencer.pdf
- [22] <http://www.aclweb.org/anthology/S16-1001>
- [23] Sayyad Samee , K.Vengatesan ,(2017) ,Smart City Automatic Garbage Collecting System for a better tomorrow , International Journal of Pure and Applied Mathematics , Volume 114,No- 9, pp:455-463

