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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Opinion of Tweets Using Sentimental Analysis

Sailaja.Gullipalli[1], Sravani.Grandhi[2]

Asst. Professor, Dept. of CSE, Thandra Paparaya Institute of Sciences & Technology College, JNTU, Kakinada, A.P, India [1]

B.Tech. Student, Dept. of CSE, Thandra Paparaya Institute of Sciences & Technology College, JNTU, Kakinada, A.P, India. [2]

ABSTRACT: Sentimental Analysis (SA) deals with the classifying and identifying opinions or sentiments expressed in the form of source text. SA of the user generated data is very useful in knowing opinion of the clan/grass roots. In this paper, we investigate the utility of polyglot features for detecting the sentiment of tweets in twitter. Twitter SA is difficult when compared to general SA due to the presence of buzzword/dialect and misspellings. The maximum limit of characters that are allowed in Twitter is 140. We use a supervised approach from micro-blogging to evaluate the usefulness of existing lexical resources as well as features that capture information about the informal and creative language.

KEYWORDS: Sentimental Analysis (SA), micro-blogging, twitter.

I. INTRODUCTION

There has been a colossal growth in the use of micro-blogging platforms such as Twitter, compared to the last few years. Spurred by that growth, many companies and media organizations are increasingly seeking their ways to mine Twitter with their feelings and opinions on products and services. Millions of messages are appearing daily in popular websites that provide services for micro-blogging such as Twitter, Face-book.

Internet users tend to shift from traditional communication tools (such as mails) to micro-blogging services as it provides a free format of messages and an easy accessibility. As more and more users express their views on the products and services they use, micro-blogging web-sites become valuable source's opinions and sentiments. Such data can be efficiently used for marketing or social studies.

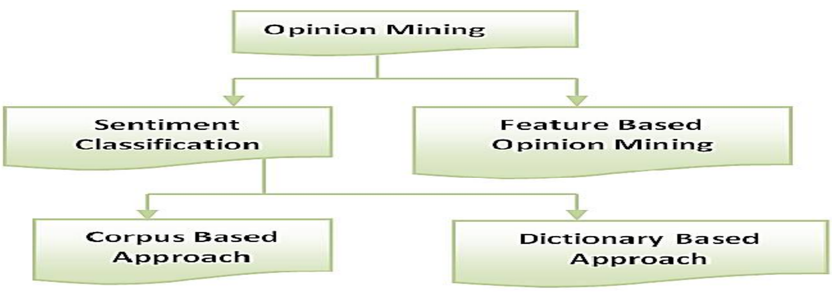


Fig 1: Classification of approaches in Sentimental Analysis

Twitter contains a very large number of very short messages created by the users of this micro-blogging platform which vary from personal thoughts to public statements. Features such as automatic parts-of-speech tags and resources such as sentiment lexicons have proved useful for SA in other domains, but will they also prove useful for SA in Twitter? In this paper, we begin to investigate this question.

Another challenge of micro-blogging is the incredible breadth of topic that is covered. It is not an aggravation to say that people tweet about anything and everything. In this paper, we explore one method for building such data:

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

using Twitter hash tags (e.g., #job, #news, #quote) to identify positive, negative, and neutral tweets to use for training three-way sentiment classifiers.

II. RELATED WORK

SA is a growing area of Natural Language Processing with research ranging from document level classification (Pang and Lee 2008) to learning the polarity of words and phrases (e.g., (Hatzivassiloglou and McKeown 1997; Esuli and Sebastiani 2006)). Given the character limitations on tweets, classifying the sentiment of Twitter messages is most similar to sentence-level sentiment analysis (e.g., (Yu and Hatzivassiloglou 2003; Kim and Hovy 2004)); however, the informal and specialized language used in tweets, as well as the very nature of the micro-blogging domain make Twitter SA a very different task. It's an open question how well the features and techniques used on more well-formed data will transfer to the micro-blogging domain. Just in the past year there have been a number of papers looking at Twitter sentiment buzz (Jansen et al. 2009; Pak and Paroubek 2010; O'Connor et al. 2010; Tumasjan et al. 2010; Bifet and Frank 2010; Barbosa and Feng 2010; Davidov, Tsur, and Rappoport 2010).Other researchers have begun to explore the use of part-of-speech features but results remain mixed. Features common to micro-blogging (e.g., emotions) are also common, but there has been little investigation into the usefulness of existing sentiment resources developed on non-microblogging data.

Researchers have also begun to investigate various ways of automatically collecting training data. Several researchers rely on emoticons for defining their training data (Pak and Paroubek 2010; Bifet and Frank 2010). (Barbosa and Feng 2010) exploit existing Twitter sentiment sites for collecting training data. (Davidov, Tsur, and Rappoport 2010) also use hashtags for creating training data, but they limit their experiments to sentiment/non-sentiment classification, rather than 3-way polarity classification, as we do.

III.DATA

We use three different corpora of Twitter messages in our experiments. For development and training, we use the hashtagged data set (HASH), which we compile from the Edinburgh Twitter corpus, and the emotion data set (EMOT) from http://twittersentiment.appspot.com. For evaluation we use a manually annotated data set produced by the iSieve Corporation (ISIEVE). The number of Twitter messages and the distribution across classes is given in Table 1.

	Positive	**Negative**	**Neutral**	**Total**
HASH	31,871(14%)	84,750(35%)	125,859(52%)	2,42,480
EMOT	230,811(57%)	170,970(43%)	-	4,01,781
ISIEVE	1,770(41%)	300(7%)	2,253(52%)	4323

Table I: *Corpus Statistics*

Hashtag	**Frequency**	**Synonyms**
#nowplaying	255,715	
#followfriday	227,530	#ff
#jobs	181,205	#tweetajob
#fb	144,835	#facebook
#39;s	110,150	
#formspringtome	85,775	
#musicmonday	78,585	#mm
#tcot	77,294	
#tinychat	58,310	
#quote	33,554	
#letsbehonest	32,732	#tobehonest
#omgfacts	30,042	
#fail	23,007	#epicfail

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

#factsaboutme	20,143	
#news	19,179	
#random	17,180	
#shoutout	17,445	

Table II: *Most frequent hashtags in the Edinburgh corpus*

HASHTAGGED DATA SET

The Edinburgh Twitter corpus is a superset of the hashtagged data set. The Edinburgh corpus contains 97 million tweets collected over a period of two months. To create the hashtagged data set, we first filter out duplicate tweets, non-English tweets, and tweets that do not contain hashtags. From the remaining set (about 5 million), we investigate the distribution of hashtags and identify what we hope will be sets of frequent hashtags that are indicative of positive, negative, and neutral messages. These hashtags are used to select the tweets that will be used for development and training.

Table 2 lists the 17 most-used hashtags in the Edinburgh corpus. In additional to the very common hashtags that are part of the Twitter folksonomy (e.g., #followfriday, #musicmonday), we find hashtags that would seem to indicate message polarity: #fail, #omgthatsotrue, #iloveitwhen, etc.

To select the final set of messages to be included in the HASH data set, we identify all hashtags that appear at least 1,000 times in the corpus. From these, we selected the top hashtags that we felt would be most useful for identifying positive, negative, and neutral tweets. These hashtags are given in Table 3. Messages with these hashtags were included in the final dataset, and the polarity of each message is determined by its hash tag.

Positive	#iloveitwhen, #thingslike, #bestfeeling, #bestfeelingever, #omgthatssotrue, #imthankfulfor, #thingsilove, #success
Negative	#fail, #epicfail, #nevertrust, #worst, #worse, #worstlies, #imtiredof, #somethingaintright, #somethingsnotright, #ihate
Neutral	#job, #tweetajob, #omgfacts, #news, #listeningto, #lastfm, #hiring, #cnn

Table III: *Top positive, negative, and neutral hashtags used to create the HASH data set*

EMOTION DATA SET

The Emotion data set was created by Go, Bhayani, and Huang for a project at Stanford University by collecting tweets with positive '+' and negative '-'emotions. Messages containing both positive and negative emotions were omitted. They also hand-tagged a number of tweets to use for evaluation, but for our experiments**,** we only use their training data. This set contains millions of tweets. For experiment, the set contains 401,781 tweets, 230,811 positive and 170,970 negative and the majority of these messages doesn't contain any hashtags.

ISIEVE DATA SET

We use ISieve data set exclusively for evaluation. It contains approximately 4350 tweets, which was collected and hand-annotated by the ISieve Corporation. The data in this collection was selected to be on certain topics, and the label of each tweet reflects its sentiment (positive, negative, or neutral) towards the tweet's topic.

PREPROCESSING

The data preprocessing consists of three steps. They are 1) Tokenization, 2) Normalization, 3) Parts-of-speech (POS) tagging. These are described below.

1) *Tokenization*: Here, emotions and abbreviations (e.g., OMG, BRB, GM) are identified as part of the tokenization process and treated as individual tokens.
2) *Normalization*: In this process, the presence of abbreviations within a tweet is noted and then abbreviations are replaced by their actual meaning (e.g., GM-> *good morning*). We also identify informal intensifiers such as

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

all-caps (e.g., I LOVE this place!!) and character repetitions (e.g., "you are so luckyyyyyyy"), note their presence in the tweet. All-caps words are made into lower single character. Finally, the presence of any special Twitter token is noted (e.g., #hashtags, user tags, and URLs) and placeholders indicating the token type are substituted.

3) *POS*: This is the last preprocessing step. The performance of this POS tagger is improved by the normalization process.

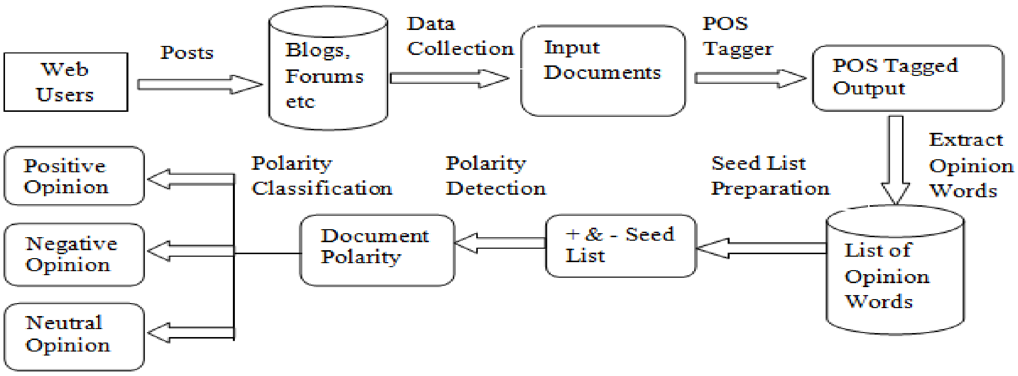


Fig 2: Document polarity of Sentimental Analysis

IV.FEATURES

We use a variety of features for our classification experiments. For the baseline, we use unigrams and bigrams. We also use features like sentiment lexicon, which represents information and POS features from sentiment analysis. Finally, we include features to capture some of the more domain-specific language of micro-blogging.

n- gram FEATURES

To identify a set of useful *n*-grams, we first remove stop-words. We then perform rudimentary negation detection by attaching the word 'not' to a word that precedes or follows a negation term**.** This has proved useful in previous work (Pak and Paroubek, 2010). Finally, all unigrams and bigrams are identified in the training data and ranked according to their information gain, measured using Chi-squared. For our experiments, we use the top 1,000 *n*-grams in a bag-of-words fashion.

LEXICON FEATURES

Words listed the MPQA subjectivity lexicon (Wilson, Wiebe, and Hoffmann 2009) are tagged with their prior polarity: positive, negative, or neutral. We create three features based on the presence of any words from the lexicon.

PART-OF-SPEECH FEATURES

For each and every tweet, we have features for counts of the number of verbs, adverbs, adjectives, nouns, and any other parts of speech.

MICRO-BLOGGING FEATURES

We create binary features that capture the presence of positive, negative, and neutral emotions and abbreviations and the presence of intensifiers (e.g., all-caps and character repetitions). For the emotions and abbreviations, we use the Internet Lingo Dictionary (Wasden 200) and various internet slang dictionaries available online.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

EXPERIMENTS AND RESULTS

Our goal for these experiments is two-fold. 1) We want to evaluate whether our training data with labels from hashtags and emotions is useful for training sentiment classifiers for Twitter. 2) We want to evaluate the effectiveness of the features from section for sentiment analysis in Twitter data.

There are some questions like, How useful is the sentiment lexicon developed for formal text on the short and informal tweets? How much gain do we get from the domain-specific features?

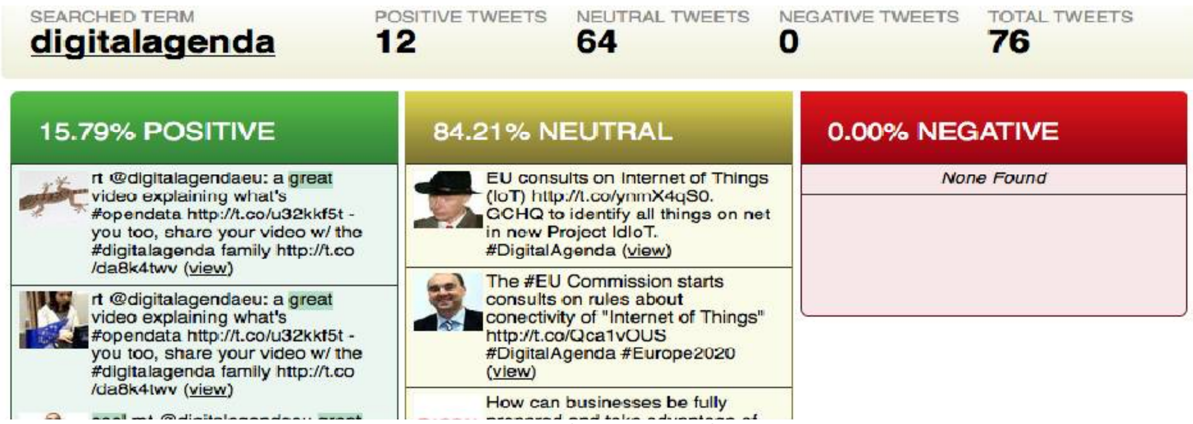
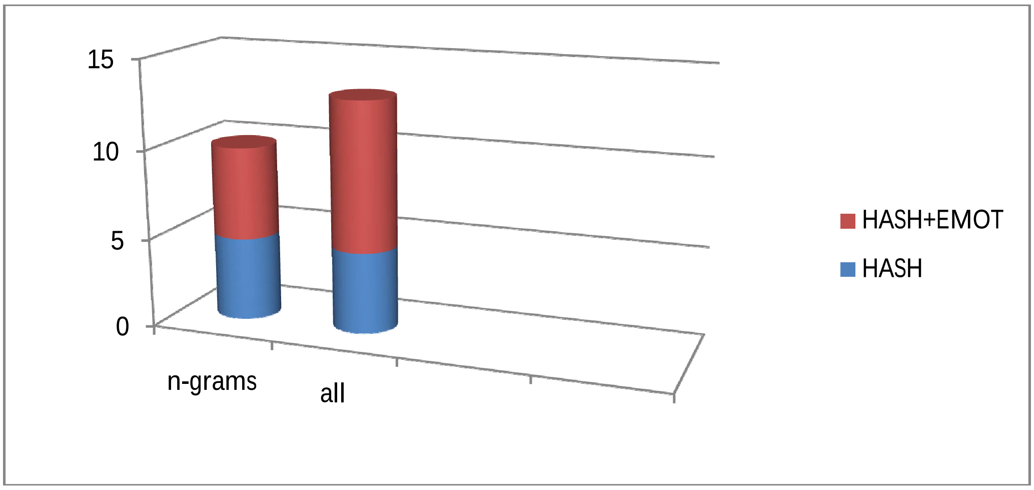


Fig 3: Example for reviews of opinions

For our first set of experiments we use the HASH and EMOT data sets. The process is started by randomly sampling 10% of the HASH data to use as a validation set. This validation set is used for *n*-gram feature selection and for parameter tuning. The remainder of the HASH data is used for training. To train a classifier, we sample 22,247 tweets from the training data and use this data to train AdaBoost.MH (Schapire and Singer 2000) models with 500 rounds of boosting. We repeat this process ten times and average the performance of the models.



Graph 1: Average F-measure on the validation set over models trained on the HASH and HASH+EMOT data.

Our experiments involve 3-way classification. As EMOT data set doesn't contain neutral data, so it is not included in the initial experiments. Instead, we explore whether it is useful to use the EMOT data to expand the HASH data and improve sentiment classification. 21,000 messages from the EMOT data set, divided equally between positive and negative, are randomly selected and added to the HASH data and the experiments are repeated.

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

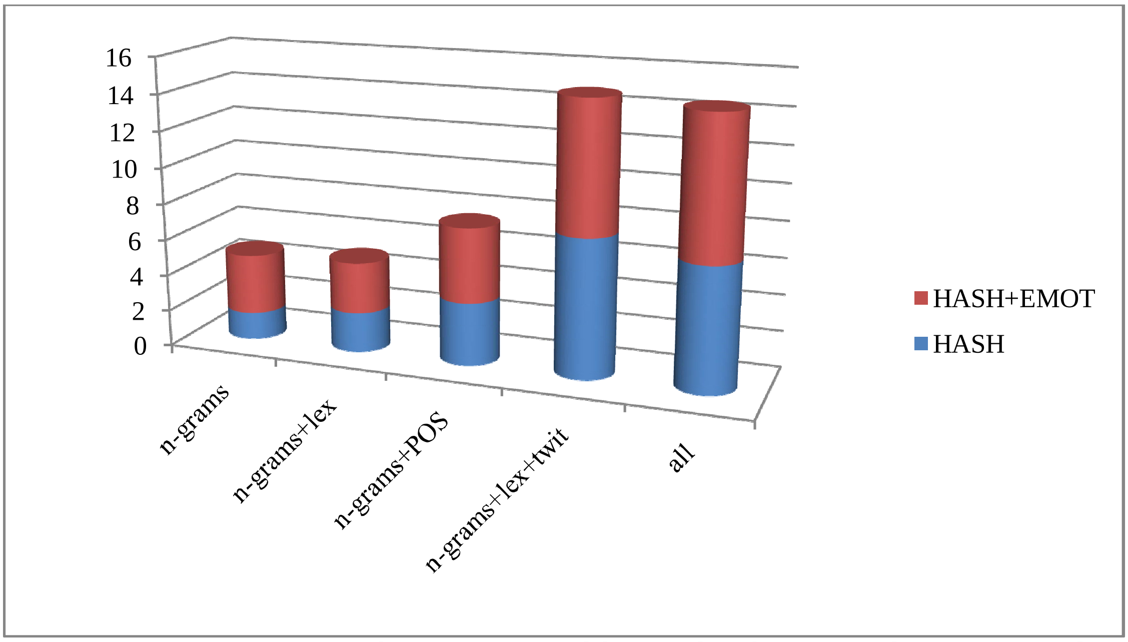
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

To get a sense for an upper-bound on the performance we can expect for the HASH-trained models and whether including the EMOT data may yield improvements, we first check the results of the models on the validation set. Fig 1 shows the average F-measure for the n-gram baseline and all the features on the HASH and the HASH+EMOT data. On this data, adding the EMOT data to the training does lead to improvements, particularly when all the features are used.



Graph 2: *Average F-measure on the test set over models trained on the HASH and HASH+EMOT data.*

Turning to the test data, we evaluate the models trained on the HASH and the HASH+EMOT data on the ISIEVE data set. Figure 2 shows the average F-measure for the baseline and four combinations of features: n-grams and lexicon features (n-gram +lex), n-grams and part-of-speech features (n-gram +POS), n-grams, lexicon features and micro-blogging features (n-grams + lex +twit), and finally all the features combined. Figure 3 shows the accuracy for these same experiments.

Interestingly, the best performance on the evaluation data comes from using the n-grams together with the lexicon features and the micro-blogging features. Including the part-of- speech features actually gives a drop in performance. Whether this is due to the accuracy of the POS tagger on the tweets or whether POS tags are less useful on micro-blogging data will require further investigation.

Also, while including the EMOT data for training gives a nice improvement in performance in the absence of micro-blogging features, once the micro-blogging features are included, the improvements drop or disappear. The best results on the evaluation data comes from the n-grams, lexical and Twitter features trained on the hash-tagged data alone.

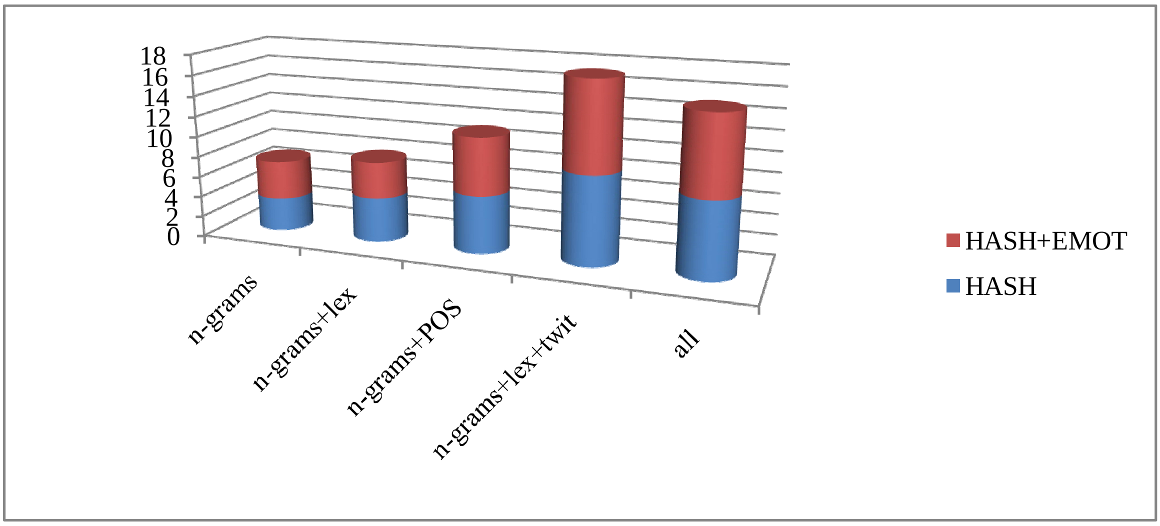
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017



Graph 3: *Average accuracy on the test set over models trained on the HASH and HASH+EMOT data.*

V. CONCLUSIONS

In our experiments on twitter sentiment analysis show that part-of-speech features may not be useful for sentiment analysis in the micro-blogging domain. To determine whether the POS features are just of poor equality due to the results of the tagger or whether POS features are just less useful for sentiment analysis in this domain, more research is needed. Features from existing sentiment lexicon were somewhat useful in conjunction with micro-blogging features, but the micro-blogging features (i.e., the presence of intensifiers and positive/negative/neutral emotions and abbreviations) were clearly the most useful.

In this paper, we use hashtags to collect training data based on positive and negative emotions. However, which method produces the better training data and whether the two sources of training data are complementary may depend on the type of features used. Therefore, our experiments show that when micro-blogging features are included, the benefit of emotion training data is lessened.

REFERENCES

1. Yu, H., and Hatzivassiloglou.V, 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In Proc. of EMNLP.
2. Barbosa.L, and Feng.J, 2010. Robust sentiment detection on twitter from biased and noisy data. In Proc. of Coling.
3. Kim.S. –M., and Hovy.E, 2004. Determining the sentiment of opinions. In *Proceeding of Coling.*
4. Tumasjan.A, Sprenger, T.O.; Sandner, P.; and Welpe,I. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of ICWSM.*
5. Bifet, A., and Frank.E. 2010. Sentiment Knowledge discovery in twitter streaming data. In *Proc. of 13th International Conference on Discovery Science*.
6. Pang.B, and Lee.L, 2008. Opinion mining and sentiment analysis. *Foundations and trends in Information Retrieval* 2(1-2):1-135.
7. Esuli.A, and Sebastiani.F, 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC.*
8. Wilson.T, Wiebe.j, and Hoffmann.P, 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics* 35(3):399-433.
9. Davidov.D; Tsur.O; and Rappoport.A, 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of Coling.*
10. Schapire.R.E; and Singer.Y, 2000. BoosTexter: A boosting-based system for text categorization. *Machine Learning* 39(2/3):135-168.
11. O'Connor.B; Balasubramanyan.R; Routledge.B; and Smith.N, 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of ICWSM.*
12. Jansen.B.J; Zhang.M; Sobel.K; and Chowdury.A 2009. Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology* 60(11):21 9-2188.
13. Pak.A; and Paroubek.P, 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proc. of LREC.*
14. Wasden.L, 2006. Internet Lingo Dictionary: A Parent's Guide to Codes Used in Chat Rooms. Instant Messaging, Text Messaging, and Blogs. Technical report, Idaho Office of the Attorney General.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

BIOGRAPHY

G.Sailaja, M.Tech Received in Computer Science and Engineering with Specialization in Computer Science and Technology from Andhra University, Visakhapatnam. Now she is working as Assistant professor in Thandra Paparaya Institute of Science &Technology (TPIST) Komatapalli Bobbili, Vizianagram Dist, Andhra Pradesh from Dec 2014 onwards. Her Area Of interest incuding C, C++ and JAVA are web technologies, Computer networks, compiler design, flat, hci, mc and cns.

G.SRAVANI Receiving her B.Tech degree in Computer Science and Engineering from TPIST, Komatapalli JNTU Kakinada in 2017. Her area of Interest is java, C Programming, Hadoop and cns.