Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 6, June 2015, pg.768 – 777

RESEARCH ARTICLE

Opinion Mining on Social Media: Based on Unstructured Data

Poornima Singh[1], Gayatri S Pandi (Jain)[2]
[1,2]Department of Computer Engineering, L. J. Institute of Engineering and Technology, Ahmedabad, Gujarat, India
[1] poornimasingh0806@gmail.com

Abstract— Opinion Mining, also known as sentiment analysis, refers to the identification and extraction of subjective information from the source materials. Recently, microblogging has been a very popular tool for communication among Internet users. It is well known that there are plenty of raw data being posted by people as real time messages about their views on a variety of products and topics in daily life. Therefore, it is imperative to develop a method which will collect and analyze these data, which could be used by users or managers and enable them to make informed decisions. The scope of the research work, described in this thesis, is to enhance the efficiency of existing Opinion Mining methods and also give a novel approach or method to handle complex negative sentences. The main aim of this research is to handle both the subjective and objective sentences, which are very often, exist on the web in form of Tweets, News Articles, Blogs, Reviews of Products and Movie etc.

Keywords— Opinion mining, Social network analysis, Twitter, Data mining, Sentiment analysis

I. INTRODUCTION

With the growing popularity of the Web 2.0, we are increasingly provided with documents expressing opinions on different topics. Micro-blogging is a very popular communication tool among Internet users. Recently, new research approaches were defined in order to automatically extract such opinions on the Internet [1].

There are various Micro-blogging sites are present now days, in which Facebook and Twitter is widely used. Now days, information is generated and managed through either computer or mobile devices by one person and used by many other persons. In social media, people posting real time messages about their opinions with lots of raw data contained on that information. To collect and analyse these data is a worthwhile research endeavour.

Opinion mining is the field of study that analyzes people’s sentiments, attitudes, opinions, appraisals, evaluations, and emotions towards entities such as products, services, events, organizations, individuals, issues, topics and their attributes [2]. Opinion mining determines whether the comments are positive, negative or neutral. It also aims to extract attributes and components of the object that have been commented on the documents. Individuals as well as organizations are using this approach too. The main task of opinion mining is to analyze the sentiments of people, accessing their emotions, attitudes and views.

Opinions can be of either Regular and Comparative Opinions or Explicit and Implicit Opinions. Opinion in the literatures is often referred to as regular opinion. Regular opinions consist of two main sub-types: direct opinions and Indirect Opinions. A relation of similarities or differences between two or more entities is expressed by comparative opinion. An explicit opinion is a subjective statement that gives a regular or comparative opinion whereas an implicit opinion is an objective statement that implies a regular as well as comparative opinion. An implicit opinion is also called implied opinion.

With the rapid growth of social media that is Twitter, forum discussions, reviews, blogs, comments and postings in social network site on the Web, persons and organizations are more and more using the content of these media for decision making. Whenever one need to make a decision one wants to hear other's opinions that is why opinions of people's always play an important role in decision making process. Individuals as well as organizations are using this approach too.

The main task of opinion mining is the analysis of sentiments of people and also accessing their emotions, attitudes and opinions. Similarly, because of the viral nature of social media some issues rapidly and unpredictably become important through rumour; the necessity to gain a real-time understanding of people's concerns has grown. Due to this inconsistency, the public dispute in social media is characterized by thoughtlessness and auto-referentiality. Similarly, the total amount of unprocessed data is also an opportunity to better make sense of opinions.

The origin of Web 2.0 and social media content has encouraged much enthusiasm and created abundant opportunities for understanding the opinions of the general public and consumers toward company strategies, social events, product fondness, political movements, and marketing campaigns [3].

A. Applications of Opinion Mining

There are various applications that use the techniques developed in this field to know the customers interested in brand tracking and market perception. There are variety of activities that may be involved are as follows:

- Automated content analysis helps dealing out large amount of qualitative data.
- Tracking combined user views for rating of products and services.
- Opinions matter a great deal in politics. Some work has focused on understanding what voters are thinking.
- Evaluating feedback in multiple languages.
- To understand the voice of the customer as expressed in everyday communications.
- For organizing the policy statements in a logical way, argument mapping software is used.
- Measuring response to company-related events and incidents.
- Analyzing consumer trends, competitors and market buzz.

B. Key Challenges in Opinion Mining

Currently, the solutions for opinion mining which is also known as sentiment analysis are evolving rapidly, with reduced human interventions needed to classify the comments. Following are the key challenges in opinion mining:

- Grouping of synonyms is also a difficult task. The different words could be used to refer to the same feature of the object that creates problem. Therefore, such words must be recognized and grouped together.
- Another key challenge is orientation classification of opinion where the opinion's orientation is identified.
- At the document level, classification of evaluating texts does not give knowledge about likes and dislikes of the opinion proprietor.
- The main challenge in opinion mining is that there is a difference between comments entered by different peoples. The use of language, knowledge and abbreviations varies from person to person.
- Due to various reasons the task of extracting the opinions expressed in texts is another challenge. For example, the same word could have dissimilar polarities based on the context of matter.
- Identify comparison words totally depends on their context. So, this is another issue in opinion mining.
- Monitoring opinions changing with the course of time is another challenge.
- Some peoples are expressing positive as well as negative comments in the same sentence which is another very big challenge for opinion mining. When people communicate with each other via informal mediums, this issue is very common.
- There are misleading opinions due to spam opinion which are:
    - ➢ Important terms repetition
    - ➢ Removal of many unrelated terms

II. Related Work

From machine learning, bag-of-words approaches to rule-based techniques, there has been a wide range of research done on Opinion mining. The three main research directions of opinion mining operate on the document level [4,5], the sentence level and the phrase level, Where the first two classification methods are usually based on the identification of opinion words or phrases using basically two types of approaches. The first approach is lexicon-based approach whereas the second one is the rule-based approaches. In the lexicon based approach, a lexicon table with each word in this table belonging to positive or negative evaluations will be built first then echo count of positive words and negative words will be calculated. Whereas the parts-of-speech (POS) taggers will first be used to tag each word in rule-based approaches with co-occurrence patterns of words and tags will be found to determine the sentiments.

In this research work, we are more focusing on real time data which is found over microblogging sites like Facebook and Twitter on which users post real time reactions. We are considering twitter's data for performing opinion mining. The tweets are much unstructured in nature because they are short, complex and filled with slangs. Habitually people do not care about the grammar of their Tweets. A number of current approaches on Opinion Mining take this into account and perform opinion classification that classifies opinion texts or sentences as positive, neutral or negative [6].

Sentence-Based Sentiment Analysis introduced by Alexandre and Francesc [7] uses the Semeval 2007 dataset and a Twitter corpus to evaluate the efficiency of Sentiment Analysis. These dataset are used for due to their affective nature and their granularity at the sentence level that is appropriate for an expressive TTS scenario. They were developed an EmoLib framework which is used for structuring prototypes that allows studying the appropriateness of different strategies.

A model called Opinion Miner was introduced by Liang and Dai [8] that can automatically analyze the sentiments of micro-blog messages. In this research work, the system is combined with manually annotated data from Twitter which is a most popular microblogging platform. By using this approach, machines can learn how to automatically extract the set of messages which contain opinions and determine their sentiment directions i.e. positive or negative by using this system and sort out non-opinion messages.

A novel method to identify opinion features from online reviews by exploiting the difference in opinion feature statistics across two corpora, one domain-independent corpus and one domain-specific corpus was introduced by Zhen et al. [9]. They built a model which Identifies Features in Opinion Mining via Intrinsic and Extrinsic Domain Relevance.

There are numerous methods of Opinion Mining were developed by various researchers. The comparison among them is shown in Table I, which shows the data source taken by the researchers with their accuracies.

TABLE I
COMPARISON OF VARIOUS EXISTING METHODS FOR OPINION MINING

ML Algorithm	Data Source	Year	Accuracy	References
Naïve Bayes	Microblogs	2013	70.39%	[8]
	Cantonese reviews	2011	93%	[10]
	Movie reviews	2011	85.8%	[11]
	Microblogs	2010	82.5%	[12]
	Microblogs	2009	82.7%	[6]
	Reviews	2008	77.5%	[13]
	Chinese sentiment corpus	2008	85.58%	[14]
	Reviews	2006	77.5%	[15]
	Car reviews	2005	86%	[16]
	Reviews	2003	81.9-87.0%	[4]
	Reviews	2002	81.5%	[5]
Maximum Entropy	Movie review	2011	85.4%	[11]
	Microblogs	2009	83%	[6]
	Reviews	2008	77.1%	[17]
	Reviews	2002	81.0%	[5]
	Tweets	2013	73.0%	[7]
	Movie review	2011	86.4%	[11]
	Microblogs	2009	71.56%	[6]

Support Vector Machines	Reviews	2008	77.4%	[13]
	Chinese sentiment corpus	2008	86.64%	[14]
	Reviews	2006	84.6%	[15]
	Movie reviews	2006	86.2%	[18]
	Reviews	2002	82.9%	[5]
K-nearest Neighbor	Chinese sentiment corpus	2008	86.64%	[14]
	Tweets	2010	66.0-87.0 %	[19]

III. SYSTEM ARCHITECTURE

An opinion is a quintuple, (e_i, a_{ij}, s_{ijkl}, h_k, t_l), where e_i is the name of an entity, a_{ij} is an aspect of e_i, s_{ijkl} is the sentiment on aspect a_{ij} of entity e_i, h_k is the opinion holder, and t_l is the time when the opinion is expressed by h_k. The sentiment s_{ijkl} is positive, negative, or neutral, or expressed with different strength levels [2]. Based on this theory, an empirical approach is proposed in this research work.

In this proposed approach, tweets from Twitter are crawled first then due to unstructured format of tweets, some pre-processing steps will be performed. After performing pre-processing on tweets, the tweets which contain opinion will be extracted and non-opinion tweets will be discarded. After extracting the opinionated tweets, the classification over them will be performed. Then training data will be made based on three categories i.e. positive neutral and negative and combined together to build the classifier. In this proposed model SVM classifier is used to classify the opinions into positive, neutral and negative categories.

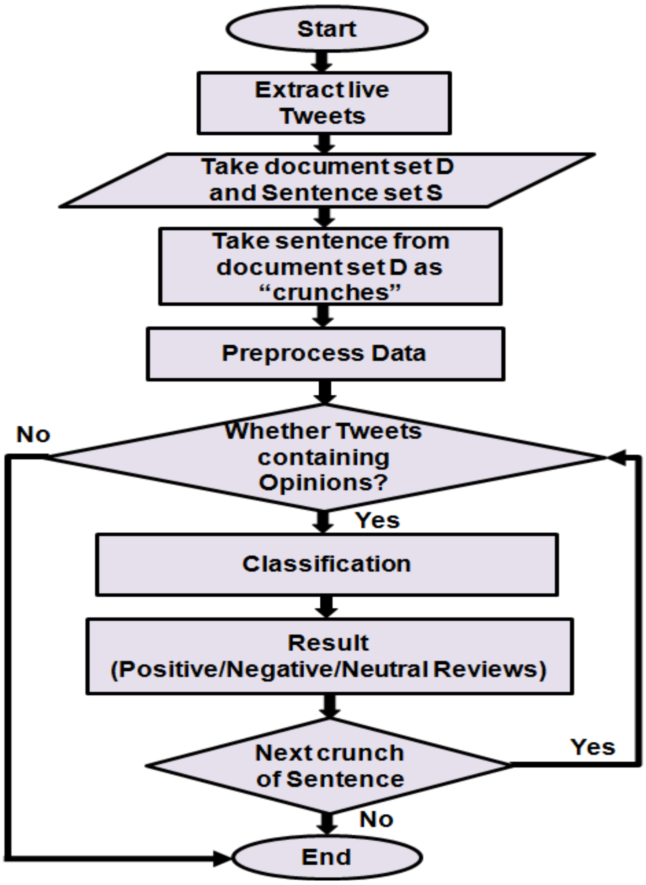

Start
Extract live Tweets
Take document set D and Sentence set S
Take sentence from document set D as "crunches"
Preprocess Data
Whether Tweets containing Opinions?
No
Yes
Classification
Result (Positive/Negative/Neutral Reviews)
Next crunch of Sentence
Yes
No
End

Fig. 1 Flowchart of proposed work

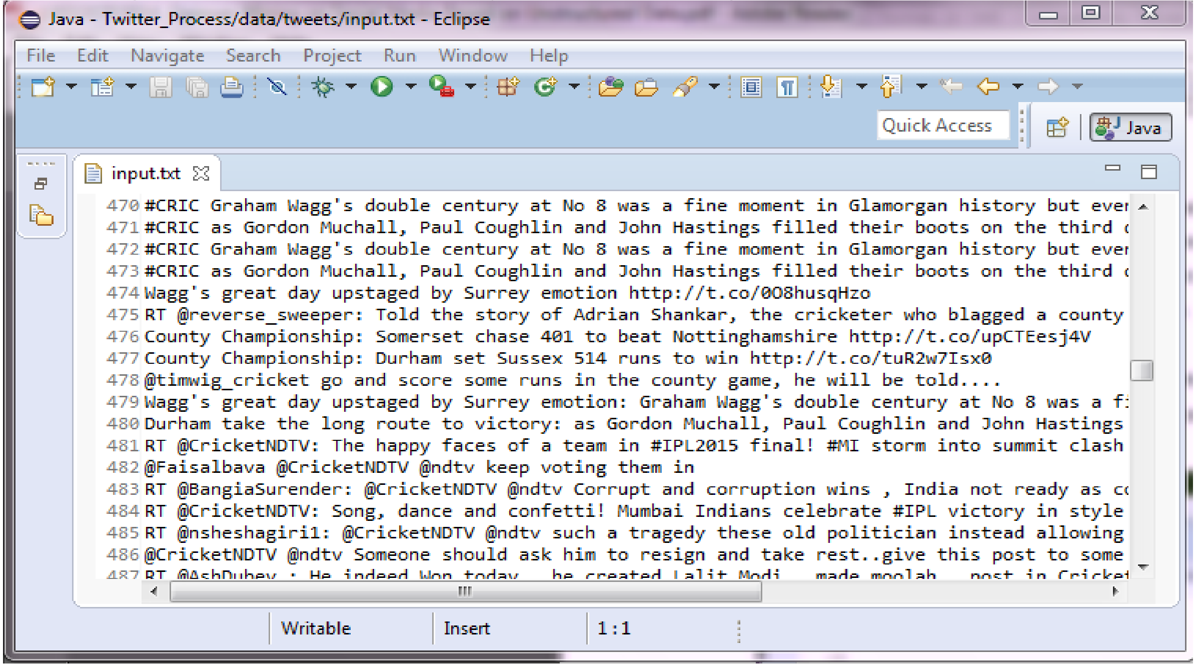
IV. Experimental Details

In this section, evaluation of whole system is done and a result for predicting the semantic orientations on Twitter is presented. The main task of the system is to classify tweets to positive versus neutral versus negative categories.

Part A describes the datasets used in the experiment and pre-processing of data is shown in part B. Result of each step and the final result of this research are shown in part C whereas Part D defines the comparison of proposed system with other existing models.

A. Data Sets

In this research work, Twitter API is used to collect data by extracting tweets from twitter which is a social media website. Tweets are crawled based on 10 distinct categories of cricket (i.e. world cup, Indian premier league, BPL T20, T20Blast, Natwest T20 Blast, County Cricket etc.) because Twitter API has a limit of 100 tweets in a response for any request. Language parameter is set to English in twitter API as to test the system on tweets in English.


Java - Twitter_Process/data/tweets/input.txt - Eclipse
File Edit Navigate Search Project Run Window Help
Quick Access
Java
input.txt
Writable
Insert
1 : 1

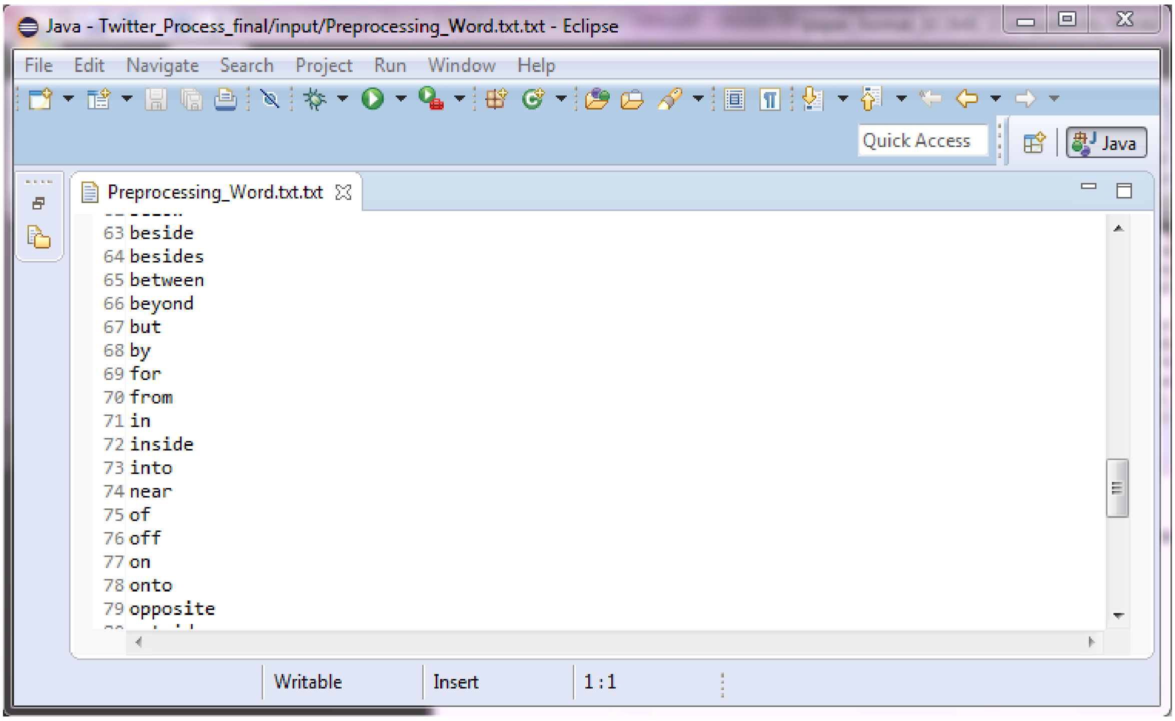
```
#CRIC Graham Wagg's double century at No 8 was a fine moment in Glamorgan history but ever
#CRIC as Gordon Muchall, Paul Coughlin and John Hastings filled their boots on the third c
#CRIC Graham Wagg's double century at No 8 was a fine moment in Glamorgan history but ever
#CRIC as Gordon Muchall, Paul Coughlin and John Hastings filled their boots on the third c
Wagg's great day upstaged by Surrey emotion http://t.co/0O8husqHzo
RT @reverse_sweeper: Told the story of Adrian Shankar, the cricketer who blagged a county
County Championship: Somerset chase 401 to beat Nottinghamshire http://t.co/upCTEesj4V
County Championship: Durham set Sussex 514 runs to win http://t.co/tuR2w7Isx0
@timwig_cricket go and score some runs in the county game, he will be told....
Wagg's great day upstaged by Surrey emotion: Graham Wagg's double century at No 8 was a fi
Durham take the long route to victory: as Gordon Muchall, Paul Coughlin and John Hastings
RT @CricketNDTV: The happy faces of a team in #IPL2015 final! #MI storm into summit clash
@Faisalbava @CricketNDTV @ndtv keep voting them in
RT @BangiaSurender: @CricketNDTV @ndtv Corrupt and corruption wins , India not ready as co
RT @CricketNDTV: Song, dance and confetti! Mumbai Indians celebrate #IPL victory in style
RT @nsheshagiri1: @CricketNDTV @ndtv such a tragedy these old politician instead allowing
@CricketNDTV @ndtv Someone should ask him to resign and take rest..give this post to some
RT @AshDubey : He indeed Won today , he created Lalit Modi , made moolah , post in Cricket
```

Fig. 2 Extraction of live tweets from Twitter

B. Pre-Processing

Based on extracted tweets, one dataset is generated which is unstructured in nature so it's pre-processing is necessary. Tweets are pre-processed as follows:

- All the tweets are eliminated which are not in English.
- All words are transformed into lower case.
- All the special characters are removed.
- All the stop words are removed from tweets.
- All the tweets are eliminated which have just a URL.


Java - Twitter_Process_final/input/Preprocessing_Word.txt.txt - Eclipse
File Edit Navigate Search Project Run Window Help
Quick Access
Java
Preprocessing_Word.txt.txt
Writable
Insert
1 : 1

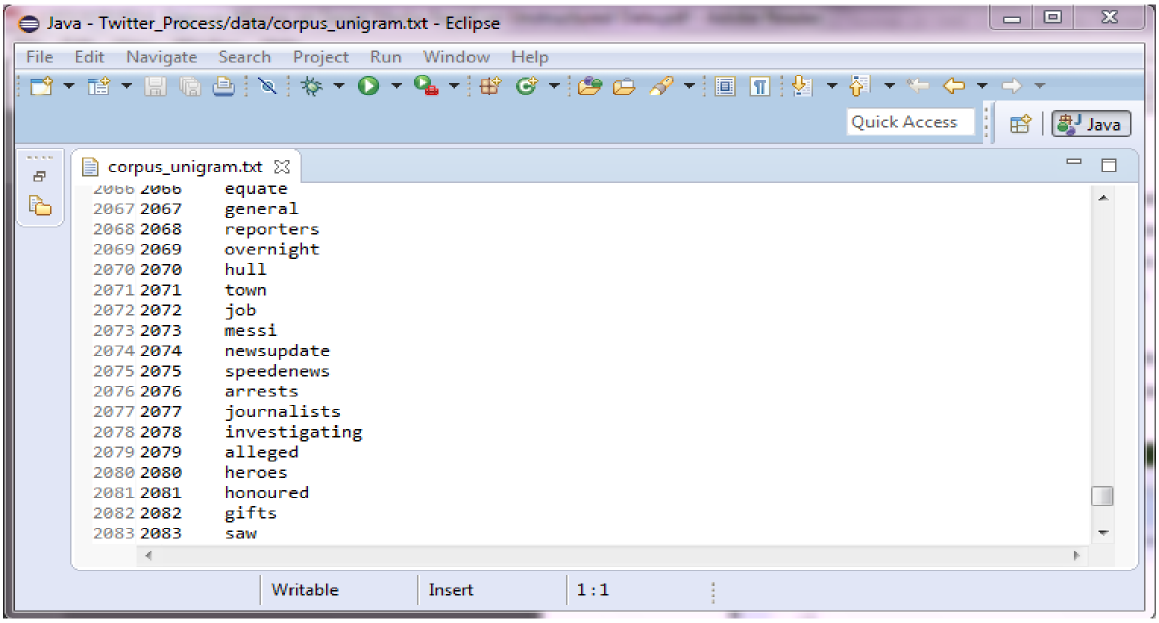
```
beside
besides
between
beyond
but
by
for
from
in
inside
into
near
of
off
on
onto
opposite
```

Fig. 3 Stop Word Removal

In computing, stop words are words which are filtered out before or after processing of natural language data. Fig. 3 shows the list of stop words which are removed from the tweets to make analysis process more efficient.

All the unique words are assigned by a token value which is unique for each word. Fig. 4 shows the token generation process where each word from tweets is assigned by a unique token value.

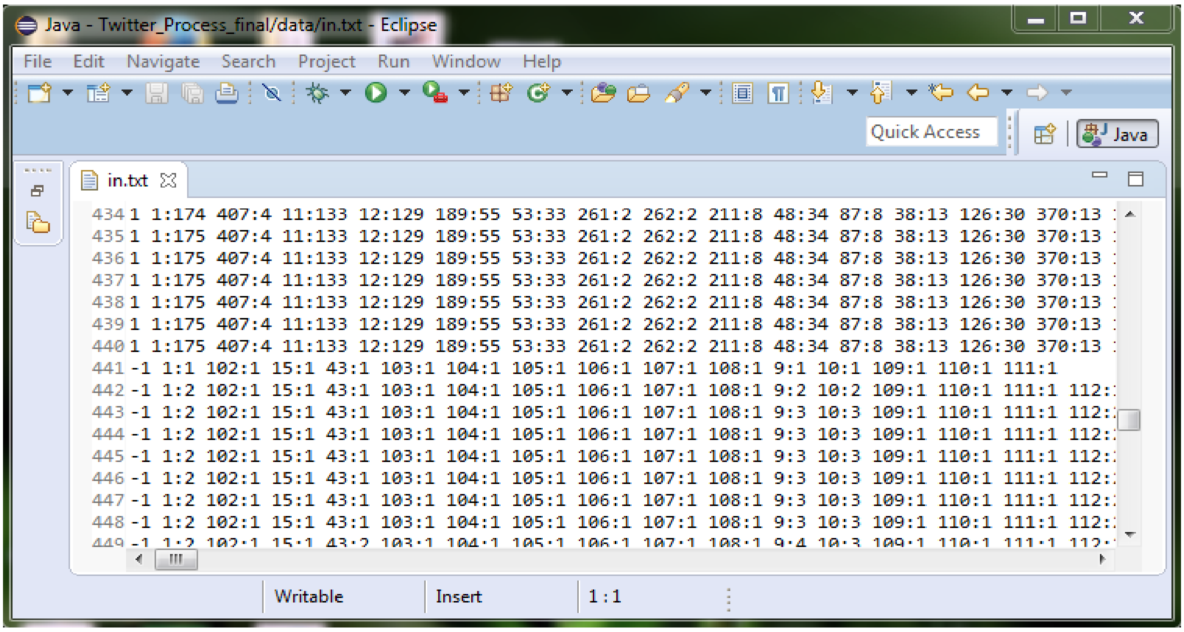
After the token generation for each world, the next step is to prepare training dataset for three distinct categories which is positive, negative and neutral. For each category, a unique label is assigned which differentiates them from each other. The label for positive category is 1, label for neutral category is 0 and for negative category, the label is -1.


Java - Twitter_Process/data/corpus_unigram.txt - Eclipse
File Edit Navigate Search Project Run Window Help
Quick Access
Java
corpus_unigram.txt
Writable
Insert
1 : 1

```
2066    equate
2067    general
2068    reporters
2069    overnight
2070    hull
2071    town
2072    job
2073    messi
2074    newsupdate
2075    speedenews
2076    arrests
2077    journalists
2078    investigating
2079    alleged
2080    heroes
2081    honoured
2082    gifts
2083    saw
```

Fig. 4 Token Generation

As shown in Fig. 5, the training dataset for all three categories (i.e. positive/ neutral/ negative) are generated. Each tweet with their token value and word count is mentioned in the training dataset.


Java - Twitter_Process_final/data/in.txt - Eclipse
File Edit Navigate Search Project Run Window Help
Quick Access
Java
in.txt
Writable
Insert
1 : 1

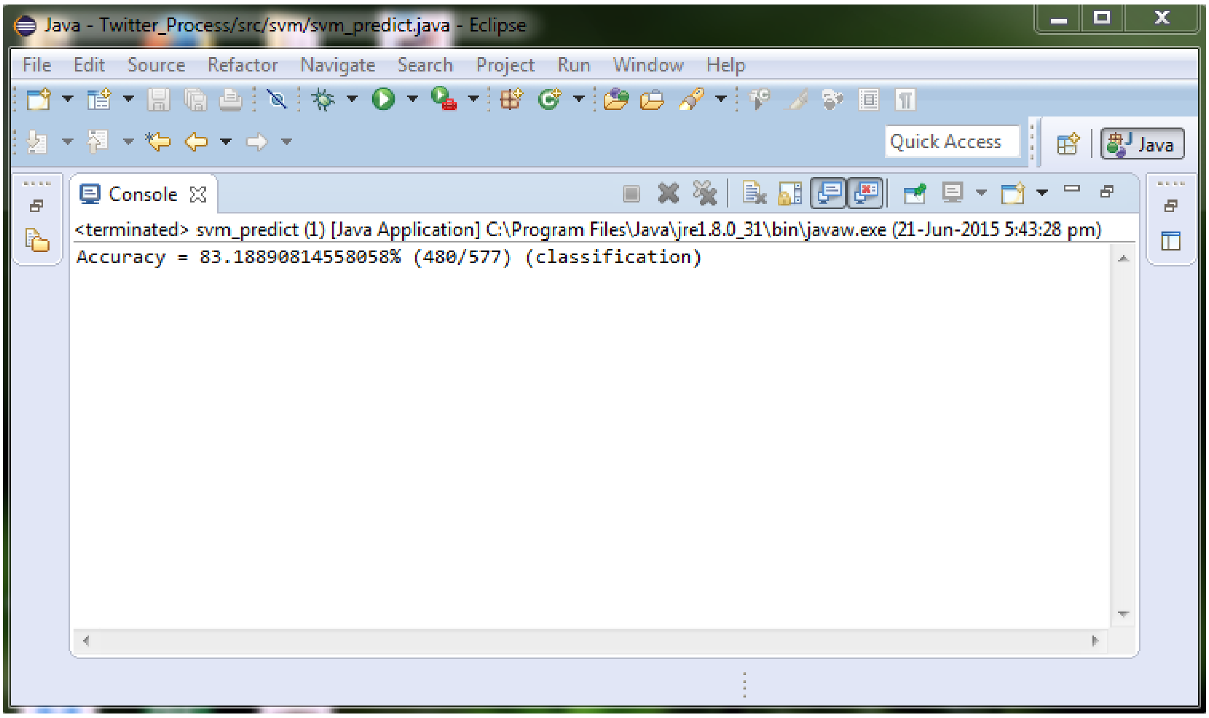
```
1 1:174 407:4 11:133 12:129 189:55 53:33 261:2 262:2 211:8 48:34 87:8 38:13 126:30 370:13 :
1 1:175 407:4 11:133 12:129 189:55 53:33 261:2 262:2 211:8 48:34 87:8 38:13 126:30 370:13 :
1 1:175 407:4 11:133 12:129 189:55 53:33 261:2 262:2 211:8 48:34 87:8 38:13 126:30 370:13 :
1 1:175 407:4 11:133 12:129 189:55 53:33 261:2 262:2 211:8 48:34 87:8 38:13 126:30 370:13 :
1 1:175 407:4 11:133 12:129 189:55 53:33 261:2 262:2 211:8 48:34 87:8 38:13 126:30 370:13 :
1 1:175 407:4 11:133 12:129 189:55 53:33 261:2 262:2 211:8 48:34 87:8 38:13 126:30 370:13 :
1 1:175 407:4 11:133 12:129 189:55 53:33 261:2 262:2 211:8 48:34 87:8 38:13 126:30 370:13 :
-1 1:1 102:1 15:1 43:1 103:1 104:1 105:1 106:1 107:1 108:1 9:1 10:1 109:1 110:1 111:1
-1 1:2 102:1 15:1 43:1 103:1 104:1 105:1 106:1 107:1 108:1 9:2 10:2 109:1 110:1 111:1 112::
-1 1:2 102:1 15:1 43:1 103:1 104:1 105:1 106:1 107:1 108:1 9:3 10:3 109:1 110:1 111:1 112::
-1 1:2 102:1 15:1 43:1 103:1 104:1 105:1 106:1 107:1 108:1 9:3 10:3 109:1 110:1 111:1 112::
-1 1:2 102:1 15:1 43:1 103:1 104:1 105:1 106:1 107:1 108:1 9:3 10:3 109:1 110:1 111:1 112::
-1 1:2 102:1 15:1 43:1 103:1 104:1 105:1 106:1 107:1 108:1 9:3 10:3 109:1 110:1 111:1 112::
-1 1:2 102:1 15:1 43:1 103:1 104:1 105:1 106:1 107:1 108:1 9:3 10:3 109:1 110:1 111:1 112::
-1 1:2 102:1 15:1 43:1 103:1 104:1 105:1 106:1 107:1 108:1 9:3 10:3 109:1 110:1 111:1 112::
-1 1:2 102:1 15:1 43:2 103:1 104:1 105:1 106:1 107:1 108:1 9:4 10:3 109:1 110:1 111:1 112::
```

Fig. 5 Final Training Dataset

C. Result Analysis

After generating training dataset for each category, all the dataset are combined together to form the final training dataset to test the classifier. In this research work SVM classifier is used to test the dataset.

Fig. 6 shows the accuracy of the overall opinion mining system generated in this approach.

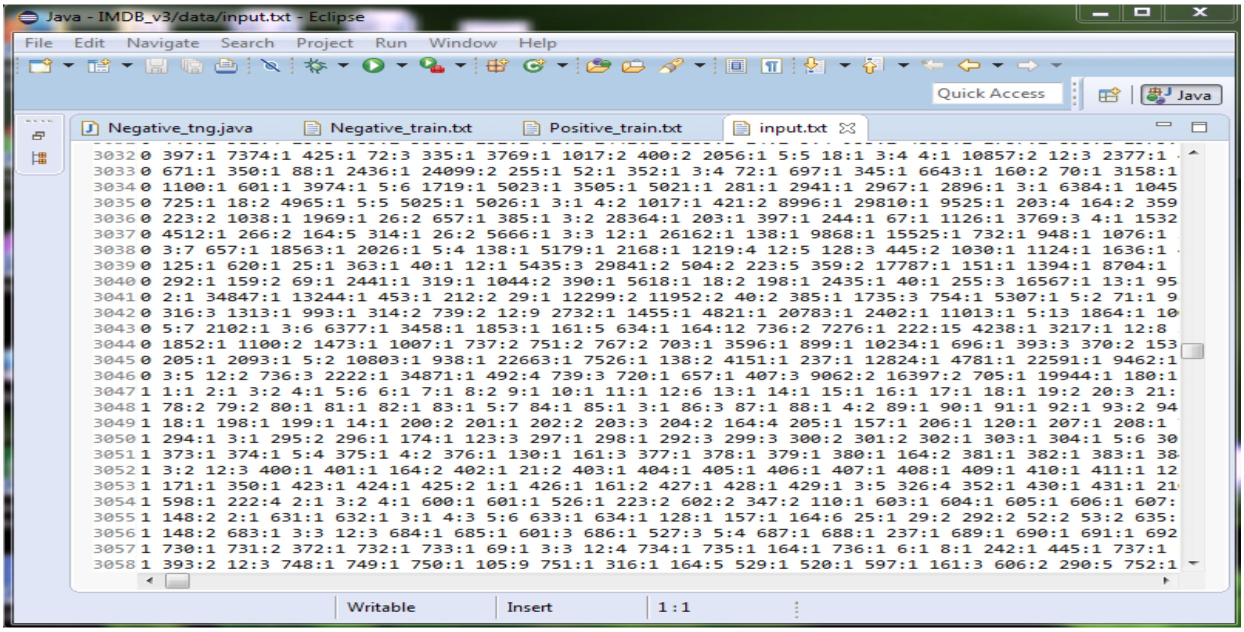

Java - Twitter_Process/src/svm/svm_predict.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
Quick Access
Java
Console

```
<terminated> svm_predict (1) [Java Application] C:\Program Files\Java\jre1.8.0_31\bin\javaw.exe (21-Jun-2015 5:43:28 pm)
Accuracy = 83.18890814558058% (480/577) (classification)
```

Fig. 6 Classifier Accuracy

The overall accuracy of this system is 83.19 % which is shown in fig. 6.

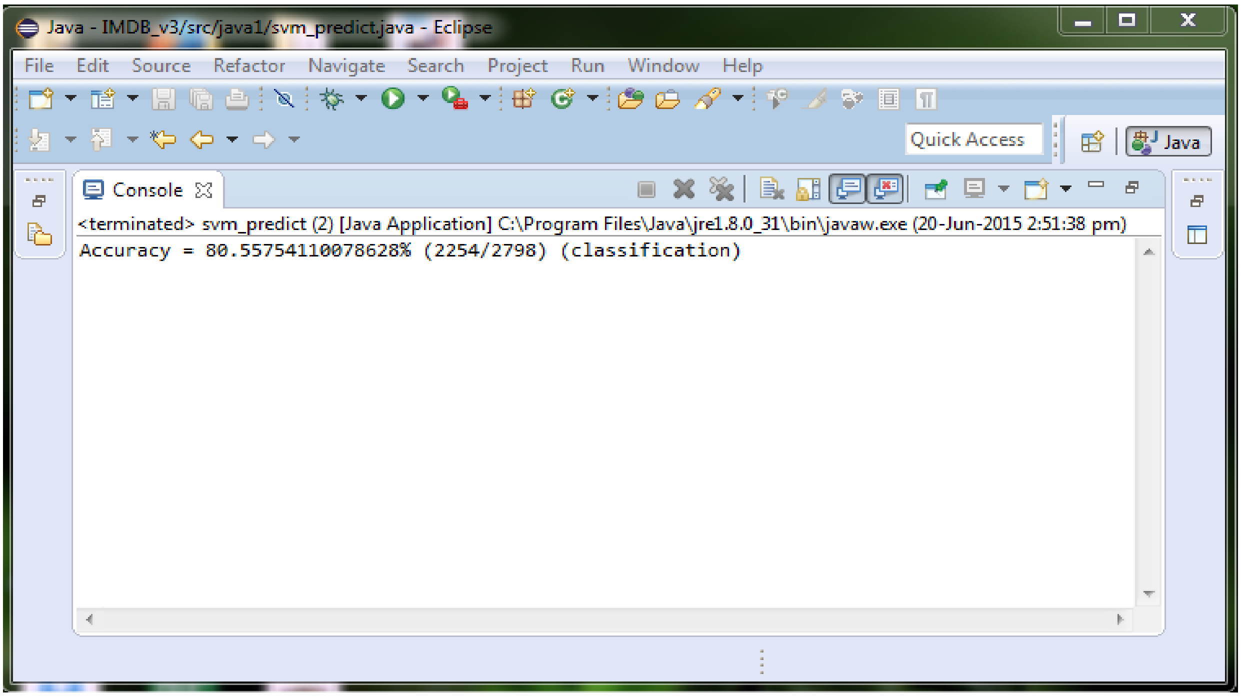
We have also built a system for opinion mining based on a standard IMDB dataset available on web. All the implementation steps using this dataset are same as discussed earlier with twitter's manually annotated dataset.



```
0 397:1 7374:1 425:1 72:3 335:1 3769:1 1017:2 400:2 2056:1 5:5 18:1 3:4 4:1 10857:2 12:3 2377:1
0 671:1 350:1 88:1 2436:1 24099:2 255:1 52:1 352:1 3:4 72:1 697:1 345:1 6643:1 160:2 70:1 3158:1
0 1100:1 601:1 3974:1 5:6 1719:1 5023:1 3505:1 5021:1 281:1 2941:1 2967:1 2896:1 3:1 6384:1 1045
0 725:1 18:2 4965:1 5:5 5025:1 5026:1 3:1 4:2 1017:1 421:2 8996:1 29810:1 9525:1 203:4 164:2 359
0 223:2 1038:1 1969:1 26:2 657:1 385:1 3:2 28364:1 203:1 397:1 244:1 67:1 1126:1 3769:3 4:1 1532
0 4512:1 266:2 164:5 314:1 26:2 5666:1 3:3 12:1 26162:1 138:1 9868:1 15525:1 732:1 948:1 1076:1
0 3:7 657:1 18563:1 2026:1 5:4 138:1 5179:1 2168:1 1219:4 12:5 128:3 445:2 1030:1 1124:1 1636:1
0 125:1 620:1 25:1 363:1 40:1 12:1 5435:3 29841:2 504:2 223:5 359:2 17787:1 151:1 1394:1 8704:1
0 292:1 159:2 69:1 2441:1 319:1 1044:2 390:1 5618:1 18:2 198:1 2435:1 40:1 255:3 16567:1 13:1 95
0 2:1 34847:1 13244:1 453:1 212:2 29:1 12299:2 11952:2 40:2 385:1 1735:3 754:1 5307:1 5:2 71:1 9
0 316:3 1313:1 993:1 314:2 739:2 12:9 2732:1 1455:1 4821:1 20783:1 2402:1 11013:1 5:13 1864:1 10
0 5:7 2102:1 3:6 6377:1 3458:1 1853:1 161:5 634:1 164:12 736:2 7276:1 222:15 4238:1 3217:1 12:8
0 1852:1 1100:2 1473:1 1007:1 737:2 751:2 767:2 703:1 3596:1 899:1 10234:1 696:1 393:3 370:2 153
0 205:1 2093:1 5:2 10803:1 938:1 22663:1 7526:1 138:2 4151:1 237:1 12824:1 4781:1 22591:1 9462:1
0 3:5 12:2 736:3 2222:1 34871:1 492:4 739:3 720:1 657:1 407:3 9062:2 16397:2 705:1 19944:1 180:1
1 1:1 2:1 3:2 4:1 5:6 6:1 7:1 8:2 9:1 10:1 11:1 12:6 13:1 14:1 15:1 16:1 17:1 18:1 19:2 20:3 21:
1 78:2 79:2 80:1 81:1 82:1 83:1 5:7 84:1 85:1 3:1 86:3 87:1 88:1 4:2 89:1 90:1 91:1 92:1 93:2 94
1 18:1 198:1 199:1 14:1 200:2 201:1 202:2 203:3 204:2 164:4 205:1 157:1 206:1 120:1 207:1 208:1
1 294:1 3:1 295:2 296:1 174:1 123:3 297:1 298:1 292:3 299:3 300:2 301:2 302:1 303:1 304:1 5:6 30
1 373:1 374:1 5:4 375:1 4:2 376:1 130:1 161:3 377:1 378:1 379:1 380:1 164:2 381:1 382:1 383:1 38
1 3:2 12:3 400:1 401:1 164:2 402:1 21:2 403:1 404:1 405:1 406:1 407:1 408:1 409:1 410:1 411:1 12
1 171:1 350:1 423:1 424:1 425:2 1:1 426:1 161:2 427:1 428:1 429:1 3:5 326:4 352:1 430:1 431:1 21
1 598:1 222:4 2:1 3:2 4:1 600:1 601:1 526:1 223:2 602:2 347:2 110:1 603:1 604:1 605:1 606:1 607:
1 148:2 2:1 631:1 632:1 3:1 4:3 5:6 633:1 634:1 128:1 157:1 164:6 25:1 29:2 292:2 52:2 53:2 635:
1 148:2 683:1 3:3 12:3 684:1 685:1 601:3 686:1 527:3 5:4 687:1 688:1 237:1 689:1 690:1 691:1 692
1 730:1 731:2 372:1 732:1 733:1 69:1 3:3 12:4 734:1 735:1 164:1 736:1 6:1 8:1 242:1 445:1 737:1
1 393:2 12:3 748:1 749:1 750:1 105:9 751:1 316:1 164:5 529:1 520:1 597:1 161:3 606:2 290:5 752:1
```

Fig. 7 Training Data for Standard Dataset

Fig. 7 shows the combined training dataset which is given to the SVM for training purpose.


Java - IMDB_v3/src/java1/svm_predict.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
Quick Access
Java
Console

```
<terminated> svm_predict (2) [Java Application] C:\Program Files\Java\jre1.8.0_31\bin\javaw.exe (20-Jun-2015 2:51:38 pm)
Accuracy = 80.55754110078628% (2254/2798) (classification)
```

Fig. 8 Classifier Accuracy for standard dataset

Fig. 8 shows the overall accuracy of this approach in which standard dataset is used.

D. Comparison of models

The proposed method was compared with two existing models namely Expressive TTS system and Opinion miner which showed the accuracy of 73.0% and 70.39%, respectively (Table II). However, the proposed method resulted into 83.19% accuracy which was better than the aforementioned models.

TABLE II
ACCURACY COMPARISON OF DIFFERENT MODELS

Model	**Accuracy (%)**	**References**
Expressive TTS system	73.0	[10]
Opinion miner	70.39	[11]
Proposed method	83.19	- -

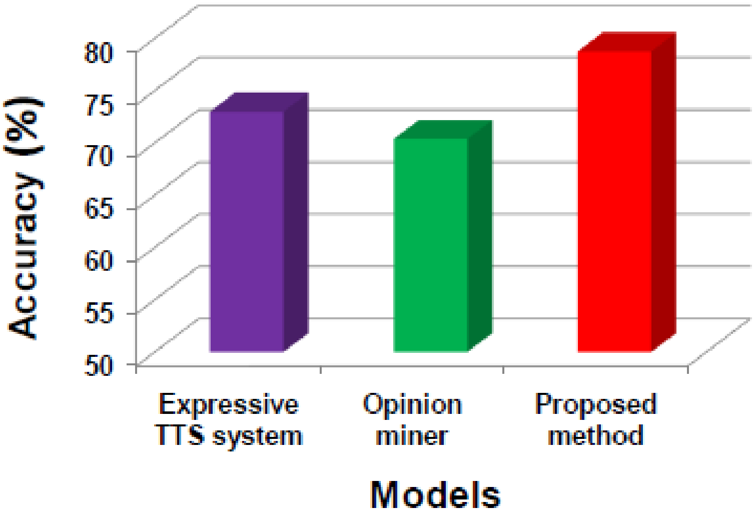


Fig. 9 Comparison of proposed model with existing model

Fig. 9 shows the comparison of accuracies of different models with our proposed system.

V. CONCLUSION & FUTURE WORK

In this research work, a system is generated which integrated machine learning techniques and domain-specific data. The experimental results demonstrated the effectiveness of the whole system. Machine learning performed well in the classification of sentiments in tweets. In proposed work, support vector machine is used to classify the sentiments in tweets. In this work, the use of domain-specific training data to build the model was demonstrated, and obtained a very positive performance.

In future, further improvement and refinement in this technique could be made in order to enhance the accuracy of the method. With this in mind, the following is a list of possible research directions. Distinct machine learning techniques can be strategically deployed in different parts, to analyze which method is more suitable. Finally, rule-based models or methods of natural language processing can be incorporated into our system.

REFERENCES

[1] P. Singh and G. S. Pandi, "Opinion Mining Techniques for Social Network Analysis: A Survey," *International Journal for Scientific Research & Development*, vol. 2, pp. 350-354, 2015.

[2] B. Liu, "Sentiment Analysis and Opinion Mining," *Synthesis Lectures on Human Language Technologies*, Ch. 1, ISBN: 9781608458851, pp. 1-167, 2012.

[3] M. S. Chen, J. Han, and P. S. Yu, "Data mining: an overview from a database perspective," *Knowledge and Data Engineering, IEEE Transaction*, ISSN: 1041-4347, pp. 866-883, 1996.

[4] K. Dave, S. Lawrence, and D. M. Pennock, *"Mining the peanut gallery: opinion extraction and semantic classification of product reviews," presented at the Proceedings of the 12th international conference on World Wide Web*, Budapest, Hungary, ISBN: 1-58113-680-3, pp. 34-44, 2003.

[5] Bo Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," *presented at the Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, DOI: 10.3115/1118693.1118704, pp. 79-86, 2002.

[6] A. Go, R. Bhayani, and H. Lei, "Twitter sentiment classification using distant supervision," *CS224N Project Report*, Stanford, pp. 1-12, 2009.

[7] A. Trilla and F. Alias, "Sentence-Based Sentiment Analysis for Expressive Text-to-Speech," *Audio, Speech, and Language Processing, IEEE Transaction* , ISSN: 1558-7916, pp. 223-233, 2013.

[8] PW. Liang and BR. Dai, "Opinion Mining on Social Media Data," *in Mobile Data Management (MDM), IEEE 14th International Conference*, ISBN: 978-1-4673-6068-5, pp. 91-96, 2013.

[9] Z. Hai, K. Chang, JJ. Kim, and CC. Yang, "Identifying Features in Opinion Mining via Intrinsic and Extrinsic Domain Relevance," *Knowledge and Data Engineering, IEEE Transaction*, ISSN: 1041-4347, pp. 623-634, 2014.

[10] Z. Zhang, Q. Ye, Z. Zhang, and Y. Li, "Sentiment classification of Internet restaurant reviews written in Cantonese," *Expert Systems with Applications*, ISSN: 0957-4174, pp. 7674-7682, 2011.

[11] R. Xia, C. Zong and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," *Information Sciences*, ISSN: 0020-0255, pp. 1138-1152, 2011.

[12] B. Albert and F. Eibe, "Sentiment knowledge discovery in twitter streaming data," *presented at the Proceedings of the 13th international conference on Discovery science,* Canberra, Australia, ISSN: 0302-9743, pp. 1-15, 2010.

[13] M. Annet And G. Kondrak, "A Comparison of Sentiment Analysis Techniques: Polarizing Movie Blogs," *in Advances in Artificial Intelligence, Springer Berlin Heidelberg*, ISSN: 0302-9743, pp. 25-35, 2008.

[14] S. Tan and J. Zhang, "An empirical study of sentiment analysis for chinese documents," *Expert Systems with Applications*, ISSN: 0957-4174, pp. 2622-2629, 2008.

[15] C. Chaomei, IS. Fidelia, Sanjuan, Eric and C. Weaver, "Visual Analysis of Conflicting Opinions," *in Visual Analytics Science And Technology*, IEEE Symposium, ISBN: 1-4244-0591-2, pp. 59-66, 2006.

[16] M. Gamon and A. Aue, "Automatic identification of sentiment vocabulary: exploiting low association with known sentiment terms," *presented at the Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, Michigan, pp. 57-64, 2005

[17] K. Shimada and T. Endo, "Seeing Several Stars: A Rating Inference Task for a Document Containing Several Evaluation Criteria," *in Advances in Knowledge Discovery and Data Mining, Springer Berlin Heidelberg*, ISSN: 0302-9743, pp. 1006-1014, 2008.

[18] A. Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," *Computational Intelligence*, ISSN: 0824-7935, pp. 110-125, 2006.

[19] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using Twitter hashtags and smileys," *presented at the Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, Beijing, China, pp. 241-249, 2010.