ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

Document Level Opinion Mining of Reviews of Mobile Phone Companies

Sunidhi Dwivedi [1], Shama Parveen [2]

M.Tech Scholar, Department of Computer Science & Engineering, Amity University, Lucknow, U.P., India[1]

Assistant Professor, Department of Computer Science & Engineering, Amity University, Lucknow, U.P., India[2]

ABSTRACT: Human beings express themselves by giving opinions, feedback, suggestions or ideas about any entity. Opinions can be expressed in many ways such as it can be expressed on facebook, twitter, blogs etc. World and technology is growing at faster rate so people are using other's opinion present on blogs, social networking sites or shopping websites for taking decisions such as buying a product, voting for a politician etc. Nowadays, if someone has to buy a product he/she is no longer limited to asking friends and relatives. They can easily view other buyer's reviews on the internet and take the decision accordingly. Now organizations do not need to conduct surveys to gather public opinions as it was a time taking task. They can easily find it on internet and can improve their product quality accordingly. There are thousands of opinions on the web about an entity so taking decisions might be difficult. Therefore, opinion mining is used for classifying the polarity of the reviews as positive, negative or neutral. Opinion mining or sentiment analysis is a natural language processing and information extraction task that aims for distinguishing the emotions expressed within the reviews and classifying its polarity as positive, negative or neutral. In this paper, we will discuss the various document level opinion mining techniques and perform opinion mining on movie reviews and classify them on the basis of their sentimental scores**.**

KEYWORDS: Opinions, Human Factor, Opinion Mining, Sentimental Score

I. INTRODUCTION

World and technology is growing at a faster rate, thousands of people are using internet for posting their views, ideas, comments or opinions about any product, person, movies on blogs, review sites, micro-blogging sites such as facebook or twitter and various shopping websites. Opinions are the key influencers in decision making so it is important to analyze them. As there are thousands of reviews present about a product so it is difficult to know that what the overall polarity about the product is. To simplify this task, sentiment analysis technique is used to determine the attitude of speaker and classify them as positive, negative or neutral. Opinion mining or sentiment analysis is a natural language processing and information extraction task that aims for distinguishing the emotions expressed within the reviews and classifying its polarity as positive, negative or neutral. Opinion mining task can be performed at three levels-

- Document level- In this overall sentiment of the whole document is extracted and classified as positive, negative and neutral.
- Sentence level- It is a fine-grained level than document level sentiment analysis.
- Feature level- In this level, feature words are extracted and classified as positive, negative or neutral.

In document level opinion mining, overall sentiment classification of the whole document is done. Assumption in this approach is that the whole document expresses opinions about a single entity. Most of the reviews on various sites are present in form of documents so we are studying this technique for classification of opinions. In document level opinion mining, each sentence is individually classified, all the results are collected and whole document's polarity is summarized which is the required polarity.

Natural language processing and information retrieval are important fields of opinion mining task. Natural language processing [12] is a subarea of artificial intelligence that deals with human languages. It aims for providing flexible interfaces. The NLP techniques that can be used in opinion mining task include pattern matching, parsing, semantic

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

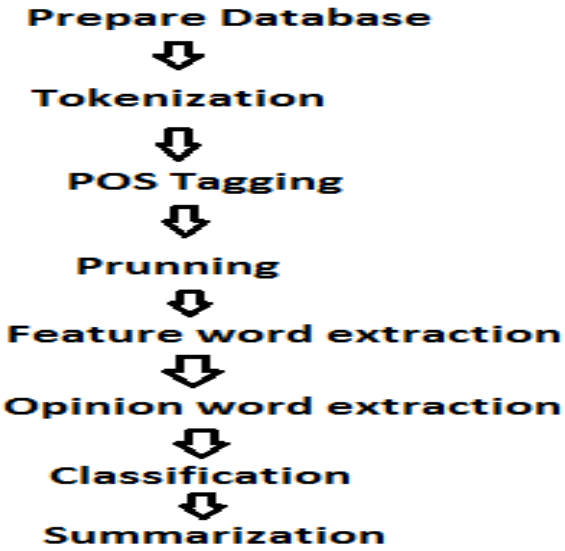
(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

grammars, tokenization etc. Information retrieval [13] is an activity of extracting relevant information from various sources. Performance and correctness measures include precision and recall that is to be calculated for showing the results while text mining.

Basic tasks that are involved in document level opinion mining include-

- Database collection- Data is collected from various websites.
- Pre-processing- This includes stop word removal, stemming, tokenization
- Pos-tag- Each word in document is tagged with their corresponding part-of-speech.
- Classification- In this individual sentences polarity is classified as positive, negative or neutral.
- Summarization- In this whole document polarity is identified.



In this paper, we study the various techniques of document level opinion mining and perform document-level opinion mining on mobile phone reviews.

II. RELATED WORK

Document level opinion mining classifies the overall polarity of the document as positive, negative or neutral with the assumption that the whole document describes about single entity. It helps in decision making process and makes it easier. It can be done by using supervised as well as unsupervised learning algorithm.

- In supervised learning technique, training and test corpus is created and then we need to train the training corpus with the help of test corpus. This will make the architecture more flexible.
- In unsupervised learning, classification is done on the basis of current dictionary without providing training.

Richa Sharma [1] proposed Document based Sentiment Orientation system. This system is based on unsupervised dictionary based technique. Dataset consists of the movie reviews which include user and critic reviews was collected from various websites. They had a seed list that consists of opinion word along with their polarity. Dataset was tagged and opinion words were extracted from them. If an opinion word gets matched with the word in the seed list, it is saved with the same polarity as present in the list. If the word is not matched, its synonyms are found using WordNet, if it gets matched then it is saved with the corresponding polarity. If its synonym does not match, antonym is found, if it gets matched it is saved with corresponding polarity. In this manner seed list keeps on increasing. Now, if there are more positive opinion words, document is classified as positive. If there are more negative opinion words, document is classified as negative and if number of positive and negative opinion words are same, document is classified as neutral.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

Ali Harb [2] presented the AMOD (Automatic Mining of Opinion Dictionaries) approach. It consists of three phases. Phase 1 is Corpora Acquisition Learning Phase. In this phase, they identified and extracted the relevant adjectives by automatic extraction of a training set. They prepared two seed lists, one of positive words and other of negative words. Phase 2 is Adjective extraction phase. In this phase, they extracted the adjectives that are highly correlated with seed words by using association rule algorithm. Phase 3 is classification in which documents are classified according to positive and negative opinions.

Shishir & Suresh [3] proposed the system of sentiment analysis that aimed to reduce computational complexity without losing structural properties of the language. They collected the dataset of around 3000 sentences. The dataset was filtered to make it standardized. The classifiers were trained to categorize the data in hierarchical manner. Further processing is done on the data that is domain specific. With the help of lingual language subroutine, system was able to work on language rules. Finally clustering is done and data is classified as positive, negative and neutral.

Noura[4] worked on Arabic documents and classified the polarity of the documents. They classified each sentence polarity in the document and then combined the results to find overall polarity of the document. They used known sentence classes as an input for their document classifier. Each document when passed through the classifier was divided into number of chunks. The number of chunks can be explicitly specified by the user. Beginning and end chunks contained opinionated sentences and middle chunks contained neutral sentences. Percentage of positive, negative and neutral for each sentence in the chunk is calculated and then the overall polarity is found out. They used SVM (Support Vector Machine) as their classifier.

Turney[5] presented the unsupervised learning algorithm. He classified the reviews as recommended or not recommended. Algorithm consists of three steps. It takes review as input and present classification as output. In first step, part-of-speech tagging of the document is done. In second step, semantic orientation of each phrase is estimated with help of PMI-IR algorithm. In this algorithm Pointwise Mutual Information (PMI) and information retrieval is used to measure similarity between words. In third step, document is classified as recommended i.e., positive or not recommended i.e., negative. The average accuracy achieved by algorithm is 74%.

Kalpana [6] divided the implementation structure into three layers. First is the presentation layer, it consists of user interface on which user can see the results. Second is the application layer, it consists of all the pre-processing steps such as stop word removal, stemming etc. Third is the database layer, it consists of dataset and the results having sentiment words along with the sentiment score.

Tanvir Ahmad [7] proposed the opinion mining system that consists of following modules. First, Document Parser is used for pre-processing of documents. It consists of Mark-up Language Tag filter that divides documents into record-size chunks. It cleans these chunks by removing ML tags and converts them into numeric vectors. Second, Subjective/Objective Analyzer classifies the sentences as subjective or objective with the help of decision free classifier. Subjective sentences are retained and rest are filtered out. Third, Document parser parses the subjective sentences and extracts product features from it. Fourth, Feature and Opinion Learner takes dependency tree as input and generates feasible information component as output. Last, Candidate pattern identification eliminates unfeasible feature with the help of FP-growth algorithm. This algorithm is based on divide and conquer approach but has a drawback that tree is too big to fit in main memory. So, we need to partition the database before processing and classify its polarity.

Zhaopeng [8] prepared the database on movie reviews. They used SVM-Light-TK toolkit for carrying out their experiments. They used Stanford parser to obtain constituency trees and then passed the resultant tree through Stanford constituency-to-dependency converter. Subset Tree (SST) and Partial Tree (PT) kernels were exploited for the above two parse trees. They manually determined degree of subjectivity as strong or weak and classified it as positive, negative or neutral.

Ainur [9] used the supervised learning technique and proposed the two-level model to address the aforementioned concerns. They tried to optimize document-level accuracy and improve performance. They presented supervised learning algorithm that was based on structural support vector machine (SVM).

III. PROPOSED WORK

The proposed system is based on unsupervised dictionary based approach. The system was implemented using JAVA on EclipseIDE. It is an unsupervised approach as we have not provided any prior training and did mining on the basis of current dictionary. Since it is a dictionary based technique so WordNet is used as a dictionary to find synonyms and

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

antonyms of unmatched words. Customer reviews of various mobile companies were collected from Flipkart (www.flipkart.com) and Amazon (www.amazon.in) websites. The polarity of the opinion words were detected on the basis of prepared seed list. After polarity detection, documents are classified as positive, negative or neutral by calculating sentimental score.

The document-level opinion mining system performs following task in various steps:-

1) **Data Pre-processing**
   The data pre-processing is done so that we can remove unwanted data and make the data more efficient. Following pre-processing steps were performed: sentence split, tokenization and pos tagging was done using Stanford CoreNLP[14] and stemming was performed using Porter Stemmer algorithm. A text file was taken as an input in which documents were present and pre-processed output was taken out in a text file.
2) **Extraction of Opinion words**
   Adjectives are considered to be the opinion words through which feelings of the opinion holder are expressed and nouns and noun phrases are considered to be features on which the opinions are expressed. Adjectives or opinion words are extracted using Apache OpenNLP[15] tool. Adjectives are extracted and saved in text file for further processing.
3) **Seed List Preparation**
   Seed list was manually prepared that contained some opinion words along with their polarity. We have prepared a seed list of 256 words that contained 156 positive words, 82 negative words and 18 neutral words. The extracted opinion words are matched with the words in the seed list to find their polarity.
4) **Polarity Detection**
   The polarity was detected using the prepared seed list and WordNet[16] dictionary. After the adjectives are extracted, they are matched with the words present in the seed list. If the word is present it is saved in a file along with its polarity. If the word is not present in seed list, its synonyms and antonyms are determined using WordNet. Its synonyms are matched in the seed list; if any synonym gets matched then it is saved in file as well as seed list gets updated. If synonyms are not matched then its antonyms are matched in seed list. If antonym is present then it is saved in file with the opposite polarity and seed list gets updated. In this way our seed list keeps on increasing with addition of new opinion words. After completion of following task, number of positive, negative and neutral words is calculated.
5) **Calculation of Sentimental Score**
   Sentimental score is calculated using frequency count of positive and negative words. It ranges between -1 to +1. If sentimental score is greater than zero, document is classified as positive. If it is less than zero, document is classified as negative and if sentimental score is equal to zero, document is classified as neutral. The formula for calculating sentimental score is:

$$\text{Sentimental Score} = \frac{(\text{number of positive words}) - (\text{number of negative words})}{(\text{number of positive words}) + (\text{number of negative words})}$$

IV. EXPERIMENTAL RESULTS

The experiment was conducted using customer reviews of mobile phones of various companies. All reviews were collected from amazon and flipkart websites. Our system performs mining in five steps:

- **Data Pre-processing**
  Input file was selected and taken as input for further processing. Data was pre-processed using Stanford CoreNLP and Poter Stemmer:

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

Fig 1: Perform sentence split, tokenization and pos-tagging

Fig 1 shows the result of sentence split, tokenization and POS tagging. Here each sentence of the document is individually split and indicated through "Sentence#" along with indicating the number of tokens in each line by assigning each token a token number. Each word is labelled with its corresponding part of speech along with indicating the lemma of each word. Lemma indicates dictionary form of the word.

Fig 2: Perform Stemming

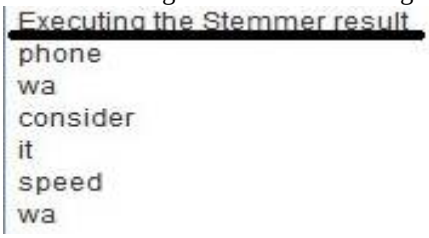


Fig 2 shows the result of stemming. Stemming is the process that chops off the end of words in order to get the root word.

- **Extraction of Opinion words and Seed List Preparation**
  Opinion words were extracted using Apache OpenNLP tool and the extracted words were manually tagged with the polarity as positive, negative or neutral to prepare seed list.

Fig 3: Prepared Seed List

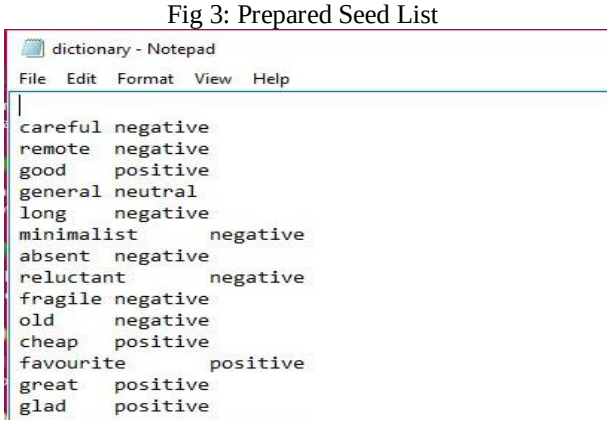


Fig 3 shows the manually prepared seed list which contains opinions words i.e., adjectives along with their polarity as positive, negative and neutral.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

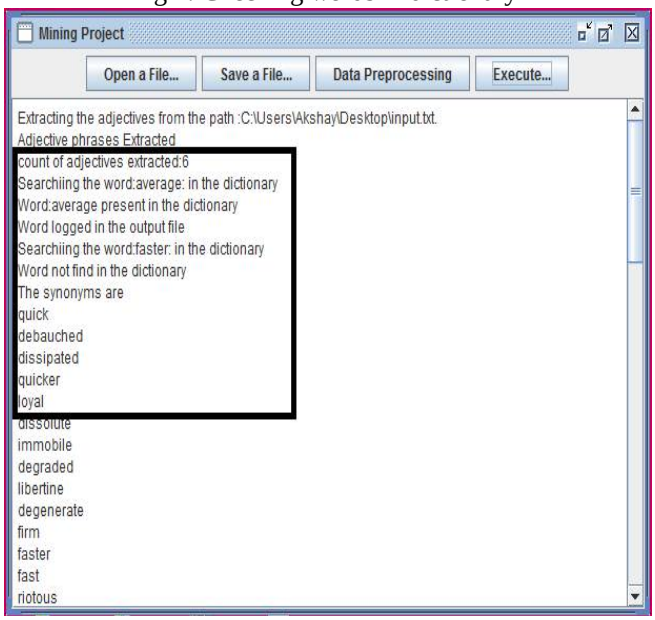
International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

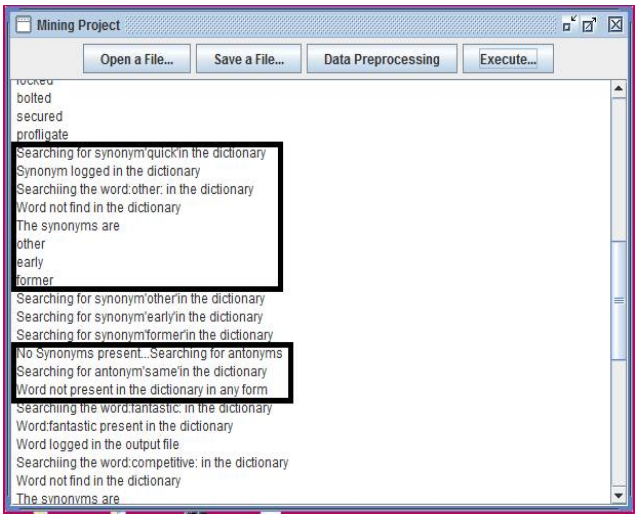
- **Polarity Detection**
  Input was taken from the input file and adjectives were extracted from them. Now those adjectives were matched with the words present in the seed list.

Fig 4: Checking words in dictionary



In Fig4 we can see six adjectives were extracted. It had searched for the word “average” in the dictionary. The word was found and saved in output file. Next it searched for the word “faster” in dictionary. It was not found so its synonyms were determined.

Fig 5: Checking for synonyms and antonyms



In Fig 5, “quick” which is the synonym of word “faster” that is determined using WordNet was present in the seed list. So the word “faster” was added in seed list and saved in output file along with its polarity. Next it searched for the

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

word “other” but it was not present in dictionary. So it searched for its synonyms, when no synonym was found then it determined its antonyms and checked its presence in dictionary.

- **Calculation of Sentimental Score**
  After checking all the adjectives extracted. It will calculate the number of positive, negative and neutral words present in output file and calculate the sentimental score to classify the documents. Sentimental score was calculated using the formula:

$$\text{Sentimental Score} = \frac{(\text{number of positive words}) - (\text{number of negative words})}{(\text{number of positive words}) + (\text{number of negative words})}$$

Fig 6: Calculating Sentimental Score

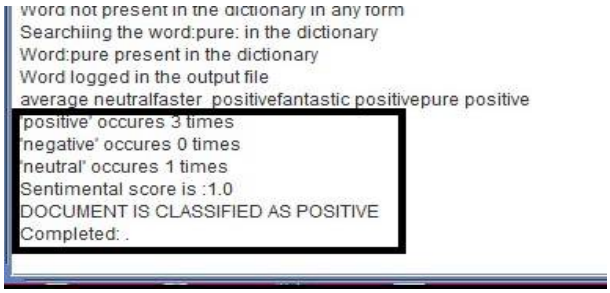


Fig 6 shows the number of positive, negative and neutral words and the calculated sentimental score. It has classified document on the basis of score. Since the score is “1.0” which is greater than 1 so document is classified as positive.
We have collected customer reviews of Apple mobiles, Samsung mobiles, Motorola mobiles and Lenovo mobiles and calculated sentimental scores of each of them. Table below shows their sentimental scores.

Table 1: Sentimental scores

MOBILE COMPANIES	SENTIMENTAL SCORE
Apple	0.55
Motorola	0.28
Samsung	0.32
Lenovo	0.17

Table 1 shows the sentimental scores of companies. It is clearly visible that Apple has more number of positive reviews than any other company.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

Fig 7: Comparing Sentimental Scores

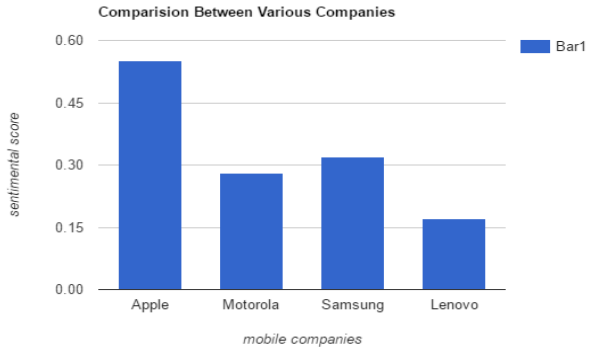


Fig 7 shows the bar chart that shows comparison between sentimental scores of 'Apple', 'Motorola', 'Samsung' and 'Lenovo'.

V. CONCLUSION

This paper gives brief about various document-level opinion mining techniques. It explains how data mining is useful in the real world and how people are benefited from it. It also describes how NLP and information extraction is useful in the process of opinion mining. A figure gives overview of the basic steps involved in text mining. In another section it describes various supervised and unsupervised approaches proposed by various authors. In the next section, I have explained my proposed work and then shown the comparison between the reviews of various mobile phone companies. In future, work can be done by developing various algorithms and system architecture to improve the accuracy of results and make system more feasible and reliable. Work can also be done on emoticons that are common for expressing feelings by human beings.

REFERENCES

[1] Rekha Jain, Richa Sharma and Shweta Nigam, "Opinion Mining of Movie Reviews at Document Level" , International Journal on Information Theory (IJIT), Vol.3, No.3, 2014
[2] Ali Harb, Gerard Dray, Pascal Poncelet, Mathieu Roche Francois Trousset and Michel Plantie, "Web Opinion Mining: How to extract opinions from blogs?", CSTST'08: International Conference on Soft Computing as Transdisciplinary Science and Technology, pp.7, 2008
[3] Shishir K. Shandilya and Dr. Suresh Jain, "Opinion Extraction & Classification of Reviews from Web Documents", IEEE International Advance Computing Conference (IACC 2009), 2009
[4] Elie Challita, Hazem Hajj, Noura Farra and Rawad Abou Assi, "Sentence-level and Document-level Sentiment Mining for Arabic Texts", IEEE International Conference on Data Mining Workshops, 2010
[5] Peter D. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 417-424, 2002.
[6] Abhinav Raj, Kalpana Razdan, Parth Srivastava, Mrunal Shinde, Vaidehi Dastapure, and Uma Nagraj, "Multi Aspect Based Document Level Sentiment Analysis for Educational Institute Analysis", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 5, May 2015
[7] Mohammad Najmud Doja and Tanvir Ahmad, "Opinion Mining using Frequent Pattern Growth Method from Unstructured Text", International Symposium on Computational and Business Intelligence, 2013
[8] Jennifer Foster, Josef van Genabith, Qun Liu, Shouxun Lin, Zhaopeng Tu and Yifan He "Identifuing High-Impact Sub-Structures for Convolution Kernels in Document-level Sentiment Classification", Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, July 2012
[9] Ainur Yessenalina, Claire Cardie and Yisong Yue, "Multi-level Structured Moels for Document –level Sentiment Classification", Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, October 2010
[10] Bing Liu, Minqing Hu, "Mining and Summarizing Customer Reviews", Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004
[11] Claire Cardie, Ellen Riloff , Jason Kessler, Janyce Wiebe, Paul Hoffmann, Siddharth Patwardhan, Swapna Somasundaran, Theresa Wilson and Yejin Choi, "OpinionFinder: A system for subjectivity analysis", Proceedings of HLT-Demo '05 Proceedings of HLT/EMNLP on Interactive Demonstrations, 2005
[12] Bing Liu, "Sentiment Analysis and Opinion Mining", Morgan & Claypool Publishers, May 2012.
[13] http://stanfordnlp.github.io/CoreNLP/
[14] https://opennlp.apache.org/cgi-bin/download.cgi
[15] https://wordnet.princeton.edu/wordnet/download