**REPORT ON**

**“Sentiment Analysis”**

A report Submitted in

Partial Fulfillment of requirements for the Degree

*Of*

**Bachelor of Technology**

*In*

**Computer Science and Engineering**

Under the Supervision of

Dr. Harsh Dev (Professor)

*By*

**Kushagra Manoj Gupta(1716410129)**

**Kushagra Sharma(1716410130)**

**Manmohan Krishna(1716410136)**

**Kushal(1716410131)**

**Batch (2017-2021)**



**Pranveer Singh Institute of Technology**

Kanpur-Agra- Delhi National Highway-2

Bhauti – Kanpur – 209305

Dr. A.P.J. Abdul Kalam Technical University

Lucknow

June 2021

***DECLARATION***

*We hereby declare that this submission is our own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.*

|  |  |
| --- | --- |
| *Signature*  *Name: Kushagra Manoj Gupta*  *Roll No.: 1716410129* | *Signature*  *Name: Kushagra Sharma*  *Roll No.: 1716410130* |
| *Signature*  *Name: Kushal Omar*  *Roll No.: 1716410131*  *Date: June 8th 2021* | *Signature*  *Name: Manmohan Krishna*  *Roll No.: 1716410136* |

**Certificate**

This is to certify that Project Report entitled **“Sentiment Analysis”** which is submitted by **Kushagra Manoj Gupta**(CS) – 1716410129, **Kushagra Sharma**(CS) – 1716410130, **Kushal Omar**(CS) – 1716410131, **Manmohan Krishna**(CS) – 1716410136 in partial fulfillment of the requirement for the award of degree **B.Tech.** in Department of **Computer Science and Engineering** of **Pranveer Singh Institute of Technology, affiliated to Dr. A.P.J. Abdul Kalam Technical University, Lucknow** is a record of the candidates own work carried out by them under **Dr. Harsh Dev**. The project embodies result of original work and studies carried out by the students themselves and the contents of the project do not form the basis for the award of any other degree to the candidate or to anybody else.

|  |  |  |
| --- | --- | --- |
| Signature:  Dr. Vishal Nagar  Head of Department  CSE Department,  PSIT, Kanpur |  | Signature:  Dr. Harsh Dev  Professor  CSE Department,  PSIT, Kanpur |

Date:

**Acknowledgment**

Every project big or small is successful largely due to the effort of a number of people who have always given their valuable advice or a helping hand to resolve problems. I sincerely appreciate the inspiration; support and guidance of all those people who have been taking efforts in making this project a success.

We have to express our appreciation to the **Prof. Dr. Harsh Dev** for sharing her pearls of wisdom and precious time with me during the course of this research. I would also like to thank all the faculty members of **Pranveer Singh Institute of Technology** for their critical advice and guidance without which this project would not have been possible.

Last but not the least I place a deep sense of gratitude to my family members and my friends who have been constant source of inspiration during the preparation of this project work.

**Abstract**

Sentiment analysis or opinion mining is used to automate the detection of subjective information such as opinions, attitudes, emotions, and feelings. Sentiment analysis becomes an important source in people trust. As search engine land\* statistics and bright local survey\*, 92% of consumers trust online reviews as much as personal recommendations. Many researchers spend long time searching for useful papers. Online reviews on papers are the essential source to help them. There are several challenges that are obstacles in sentiment evaluation. The difficulty of understanding computer human sentences or linguistics. Thus, online sentiments can save the researcher's time, it provides effort and paper cost. Analyzing scientific papers domain is hard. Evaluating sentiments with respect to several properties is hard. This domain requires scientific lexicons for parameters or features.

In this thesis, we propose a new technique to analyze online reviews in the scientific research domain called: "Sentiment Analysis Of Online Papers" (SAOOP). SAOOP aims to support researchers and save their time and effort by enabling them to report the total evaluation for the papers. SAOOP introduces a hybrid model and creates a new criteria for evaluating scientific papers. This hybrid model of an enhanced Bag-of-Words and Part-of-Speech models. SAOOP improves accuracy with solving several sentiment challenges. SAOOP consists of two evaluations for each research paper: Sentiment score and System score. Sentiment score is an evaluation of online sentiments. System score is a new criteria of evaluation topic parameters.

SAOOP employs several techniques including natural language processing, text analysis and opinion mining in the sentiment analysis evaluation process. SAOOP is a new technique that introduces an enhancement for the bag-of-words (BOW) model in sentiment analysis. It introduces several sentiment evaluation challenges based on a review structure. This review structure. The review structure is structured, short, and formal. It also solves two essential bag-of-words model weaknesses that are low accuracy and manual working approach. The enhanced bag-of-words model is an automated model for analyzing sentiment reviews and evaluating them. This enhancement aims at improving accuracy and analyzing sentiments. This thesis presents a comparison between standard and proposed enhanced Bag-of-Words models to measure the effect on accuracy results. Although SAOOP works on the word level sentiment analysis, it evaluates words with respect to the perspective of sentence level and review level of sentiment analysis. In order, to deal with order of words and grammar in review sentences with logical meaning. It also generates a new miniature lexicon to avoid the problems in the

standard bag-of-words model. This newly generated lexicon can help, avoid repetition and duplication and enables searching in it easily and fast. In addition, the proposed lexicon can deal with adjectives, nouns, verbs, adverbs, adjectives, prefixes, suffixes and other grammatical classes into synonym. This lexicon is based on a word level sentiment analysis by using similarity and difference algorithms. The lexicon includes 4500 words with 1 <tag> called<Noun>. SAOOP also categorizes reviews based on the topic essential features and keywords. This categorization has a big effect on the meaning of sentiment polarity and score of the words. This classification is summarized into five classes: (place of publication, publishing date, authors, citation number and the topic name). SAOOP provides the sentiment classification polarity levels with divides it into five classes. These classes are (Very Negative, Negative, neutral, positive, Very Positive). They uses to estimate the sentiment polarity strength.

The proposed technique presents solutions for some of the sentiment challenges to improve accuracy. With scanning the online scientific research reviews, we determine these challenges. These challenges are bi-polar words, negation, world Knowledge, topic domain features, and create huge lexicon. The proposed technique evaluates the sentiment polarity of words and sentences with respect to the implicit and explicit negative. It also extracts the topic features, keywords and characteristics to support the evaluation, and understands the fuzzy or bi-polar words in sentiments. There is a challenge in understanding the world knowledge such as recognize famous scientists names. So the proposed technique produces a solution with similarity and differences algorithms and the hierarchal database model in nouns for deal with this problem. We propose a new algorithm of evaluating negative challenge based on negative strength.

The second part of evaluating a significant paper is system score. This score is based on three significant parameters in scientific papers domain, these parameters are: 1) the place of publication, 2) number of citations and 3) the publishing date. The total evaluation of a research papers can support researchers to retrieve the useful papers for their research in a short time while saving their effort. Finally, the efficiency evaluation of SAOOP applies on making a comparison with other two techniques. The comparison terms are based on measuring accuracy, performance. The comparison depends on a model of data sets and subsets.

**Keywords:** Sentiment Analysis; Opinion Mining; Reviews; Text Analysis; Bag-Of-Words; Sentiment Analysis Challenges; Accuracy.

**Table of Contents**

**Page**

DECLARATION……………………………………………………

CERTIFICATE……………………………………………………

ACKNOWLEDGEMENTS………………………………………

ABSTRACT…………………………………………………………

LIST OF TABLE…………………………………………………..

LIST OF FIGURES…………………………………………………

LIST OF ABBREVIATIONS………………………………………

[CHAPTER 1: INTRODUCTION………………………………… 1](#_TOC_250039)

[1.1 Sentiment Analysis- An Introduction....................................................... 2](#_TOC_250038)

[1.2 Problem Definition.......................................................................................... 3](#_TOC_250037)

[1.3 Thesis Contribution......................................................................................... 4](#_TOC_250036)

[1.4 Thesis Organization........................................................................................ 6](#_TOC_250035)

[Chapter 2 : Related work 7](#_TOC_250034)

[2.1 Related work in Sentiment analysis................................................................. 8](#_TOC_250033)

2.2 What is Lexicon-drive Methodology? ............................................................ 11

[2.3 Related work in Sentiment analysis Challenges.............................................. 13](#_TOC_250032)

[2.4 Related work in web Scrapping/extracting...................................................... 20](#_TOC_250031)

2.5 Chapter Summary............................................................................................ 21

Chapter 3: Background of Sentiment Analysis 23

[3.1 Introduction..................................................................................................... 24](#_TOC_250030)

[3.2 What is sentiment analysis? ........................................................................... 24](#_TOC_250029)

3.3 How Does Sentiment Analysis Works? ......................................................... 25

[3.4 Why Sentiment analysis is important? ........................................................... 26](#_TOC_250028)

[3.5 Why Sentiment analysis is hard? ................................................................... 26](#_TOC_250027)

3.6 Sentiment Analysis (SA) & NLP.................................................................. 27

[3.7 Sentiment Analysis Models............................................................................. 32](#_TOC_250026)

3.7.1 Bag-of-words (BOW)................................................................................ 32

[3.7.2 Lexicon....................................................................................................... 32](#_TOC_250025)

3.7.3 Part of Speech (POS) Tagging................................................................... 33

[3.7.4 Semantic relationships............................................................................... 34](#_TOC_250024)

3.8 The Sentiment Analysis problems.................................................................... 35

3.8.1 Sentiment problems Defintions.................................................................. 35

[3.8.2 Sentiment Analysis Evaluation……………................................................ 36](#_TOC_250023)

3.8.3 Sentiment classification Challenges............................................................ 36

[3.9 Sentiment Analysis Lexicons & Tools.............................................................. 38](#_TOC_250022)

[3.9.1 Sentiment lexicons.……………………………………………………..... 40](#_TOC_250021)

[3.9.2 Sentiment Tools & Applications………………………………………… 40](#_TOC_250020)

[3.10 Sentiment Analysis Research......................................................................... 41](#_TOC_250019)

[3.11 Chapter Summary........................................................................................... 44](#_TOC_250018)

[Chapter 4 : Proposed Technique 45](#_TOC_250017)

4.1 The proposed Technique Overview................................................................ 46

4.2 The proposed Technique Methodology.......................................................... 50

[4.2.1 Evaluate System Score........................................................................ 50](#_TOC_250016)

[4.2.2 Evaluate Sentiment Score................................................................... 52](#_TOC_250015)

[4.2.3 SAOOP Technique............................................................................ 52](#_TOC_250014)

[4.3 Enhancement Bag-Of-Words (BOW) Model…............................................ 53](#_TOC_250013)

4.3.1 Enhancement Bag-Of Words (BOW) Model................................... 53

[4.3.2 The Comparison............................................................................... 55](#_TOC_250012)

[4.4 Proposed Solutions for Sentiment Analysis Challenges............................... 85](#_TOC_250011)

4.4.1 Topic Domain Indecence.................................................................. 59

4.4.2 Negation......................................................................................... 61

4.4.3 Create Lexicon................................................................................. 62

[4.4.4 World Knowledge Requirement...................................................... 64](#_TOC_250010)

[4.4.5 Bi-Polar Sentiments......................................................................... 64](#_TOC_250009)

[4.5 Design and Implementation.......................................................................... 65](#_TOC_250008)

[4.5.1 Design................................................................................................ 65](#_TOC_250007)

4.5.2 Implementation.................................................................................. 66

[4.6 Chapter Summary......................................................................................... 66](#_TOC_250006)

[Chapter 5 : Experimental Results 68](#_TOC_250005)

[5.1 Experiment1: SAOOP Case Study.................................................................. 69](#_TOC_250004)

* 1. Experiment 2: Comparison between Standard BOW and Enhancement BOW 78
  2. Experiment 3: Comparison among Proposed SAOOP Technique and Two 81

Sentiment Techniques............................................................................................

5.3.1 Datasets................................................................................................ 81

5.3.2 Comparison Measures............................................................................ 82

5.3.3 Comparison Results............................................................................... 83

[5.4 Chapter Summary..................................................................................... 91](#_TOC_250003)

Chapter 6 : Conclusions and Future Work 92

* 1. [Conclusion. 93](#_TOC_250002)
  2. [Future Work. 94](#_TOC_250001)

[References 95](#_TOC_250000)

**List of Tables**

**Page**

**Table 3.1:** Sentiment Analysis Tools and Applications..41

**Table 4.1:**The Pseudo code of The Enhancement BOW Model…………...53

**Table 4.2:**Comparison between Standard BOW and the Enhancement BOW...56

**Table 4.3:** Comparison between standard BOW and Enhancement Bow in

SWOT Analysis 57

**Table 4.4** Review Classification in Scientific Papers Domain 60

**Table 5.1** Extracting Data for a Case Study 70

**Table 5.2:** Steps of Proposed Technique 71

**Table 5.3:**Results of Case study 72

**Table 5.4:**Sentiment Score Evaluation 74

**Table 5.5:**System Score Evaluation 76

**Table 5.6:**Average Results For All Datasets 79

**Table 5.7:**Difference Between Standard BOW and Enhancement BOW

Models................................................................................. 80

**Table 5.8:**Measurement Metrics Table ..83

**Table 5.9:**Final Results for the Three Datasets .87

**Table 5.10:**Percentage of Accuracy between Techniques 88

**List of Figures**

**Page**

**Figure 3.1:**The Sentiment Analysis Challenges…………………………37

**Figure 4.1:**SAOOP Evaluation Components.............................................47

**Figure 4.2:**Researcher Guide based on ACM library................................49

**Figure 4.3:**SAOOP Technique Overview..................................................51

**Figure 4.4:**Boundary of Sentiment Analysis Challenges Solutions..........59

**Figure 4.5:**SAOOP Classifications Levels Perspective………………….65

**Figure 4.6:**Proposed Design of Database..................................................66

**Figure 5.1:**Comparison between Accuracy of BOW and Enhanced BOW…………………………………………………………80

**Figure 5.2:**Correctness Part of Speech Comparison among Sentiment Techniques……………85

**Figure 5.3:**Percentage of Handling Sentiment Analysis Challenges for The

Three Techniques.....................................................................86

**Figure 5.4:**Differences among Accuracy of Three Techniques................87

**Figure 5.5:**Differences among F-measure Performance of three techniques ………………………89

**Figure 5.6:**Differences among runtime Performance of Three

Techniques...............................................................................91

**List of Abbreviations**

|  |  |
| --- | --- |
| **WWW** | : World Wide Web |
| **NLP** | : Natural Language Processing |
| **ML** | : Machine Learning |
| **SA** | : Sentiment Analysis |
| **BOW** | : Bag-Of-Words |
| **SO** | : Sentiment Orientation |
| **POS** | : Part-Of-Speech |
| **NLTK** | : Natural Language Toolkit |
| **NLPS** | : Natural Language Processing Stanford |
| **EasywebExtract** | : Easy Web Extract Tool |
| **SAOOP** | : Sentiment Analysis Of Online Papers |
| **SWOT** | : SWOT Analysis (Strength , Weaknesses, Opportunities, and Threats |
|  |  |

**CHAPTER 1:**

**INTRODUCTION**

The World Wide Web has become the most popular communication platforms to the public reviews, opinions, comments and sentiments. These sentiments refers to opinions aboutproducts, places, books or re-search papers become daily text reviews. The number of active user bases and the size of their reviews created daily on online websites are massive. There are 2.4 billion active online users, who write and read online around the world . Although the scientific domain is huge as a big world of journals and conferences, there are more than 4000 rated conferences and 5000 ranked journals. According to a new survey conducted by dimensional research, April 2013: 90% of customer’s decisions depends on online reviews. As the result, a large number of studies and research have monitored the trending increase of online research resources year by year. In this thesis, we try to achieve trusted scientific reviews evaluation to be useful for researchers and facilitate the selection of papers that match their research direction. We create a new proposed technique for evaluating scientific papers based on sentiment analysis and domain parameters. We present a new proposed miniature lexicon.

##### **Sentiment Analysis: An Introduction**

Recently, several websites encourage researchers to express and exchange their views, suggestions and opinions related to scientific papers. Sentiment analysis aims at determining the attitude of a writer with respect to some topics or the overall sentiment polarity of a text, such as positive or negative. Sentiment analysis depends on two issues sentiment polarity and sentiment score. Sentiment polarity is a binary value either positive or negative. On the other hand, sentiment score relies on one of three models. Those models are Bag-of-words model (BOW), part of speech (POS), and semantic relationships. BOW model is the most popular for researchers and based on the representation of terms. The term refers to words in Bag-of-Words model. It neglects language grammar and words ordering. POS tagging is a grammatically tagging model especially verbs, adjectives and adverbs. **For example**; (The bag is not good.) declaring in (The/DT bag/NN is/VBZ not/RB good/JJ. /.). In the example DT refers to "Determiner", NN refers to "Noun", singular or mass, VBZ refers to "Verb", RB refers to "Adverb", and JJ refers to "Adjective". But it neglects logical meaning. The last model called a semantic relationship that is the most complex method. It is based on the relationship between concepts or meanings **for example**; antonym, synonym, homonym etc.

Sentiment analysis refers to the use of natural language processing, text analysis, and computational linguistics to identify and extract subjective information in source materials. Sentiment analysis is widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service. The objective of Sentiment Analysis is evaluating the sentiments and opinions of a writer respectively, one topic domain or multi-topic domain. It calculates the aggregate sentiment polarity of online real reviews for one topic based on sentiment classification levels, such as positive or negative.

##### **Problem Definition**

In this thesis, there are three problems in this thesis. First: sentiment analysis problem declares in sense of the word a difference at the level of word and sentence level or review level. Second: sentiment evaluation problem: there are several challenges that are obstacles in sentiment evaluation and the difficulty of understanding computer human sentences or linguistics. The negative is one of the most challenges faced in the evaluation process with respect to the similar structures for several expressions. The world knowledge is an obstacle in analyzing sentiments, it requires to identify famous knowledge and information. Other challenges appear in the bipolar sentiments and short sentence like abbreviations. These challenges have a bad effect on the understanding of reviews and the sentiment evaluation. There is a research gap between the sentiment challenges and sentiment evaluation. There are hundreds of thousands of re-searcher, who write and read online papers daily. In research domain to save a long time in searching for the suitable paper and save efforts for researchers. Third problem in hardiness of scientific papers domain analysis is sentiments evaluation with respect to several properties is hard and requires scientific lexicons for parameters or features.

**1.3 Thesis Contribution**

This thesis aims to fill this research gap between the sentiment challenges and evaluation. It presents a new technique to analyze online sentiment reviews. It is called sentiment analysis of online papers (SAOOP). The target of the new technique is analyzing the scientific papers domain and evaluate them. That depends on the evaluation of sentiment reviews entitled sentiment score and the topic domain parameters score which is called system score. The proposed technique introduces an enhancement for the bag-of-words (BoW) model to improve accuracy and understand the meaning better. The proposed technique works on the word level sentiment analysis. It uses the bag-of-words model, but it solves the major weakness of it. This technique classifies reviews based on extracting the keywords and features for the topic domain. Our contributions are as follows:

###### **An evaluation approach for online research papers that is based on:**

* **The Sentiment analysis evaluation**

This evaluation relies on word level sentiment analysis (SA). The reviews structures are short and formal so the most efficient model to analyze it is a bag-of- words (BOW) model. We present an enhancement Bag-of-Words model to analyze and evaluate sentiment reviews. The proposed enhanced BOW also introduces solutions for sentiment analysis challenges to improve accuracy. It classifies the topic features or keywords into five classes (Topic, author, publishing date and place of publication). Each class of reviews has some features and keywords and has a different effect on the evaluation score. We also classify the sentiment polarity levels into five classes (Very negative, negative, neutral, positive and very negative).

* **The System Score evaluation**

The system evaluation score is a new proposed criteria for domain parameters evaluation. These parameters are place of publication whether journal or conference, number of citations for each paper, and the publishing paper date of the research paper.

**2.Comparative Evaluation:**

For measuring the efficiency evaluation, makes a comparison between the proposed technique and other two techniques. The comparison terms are based on the following three issues:

* **Accuracy:** The accuracy represents the rate at which the method predicts results correctly (Acc.). The precision also called the positive predictive rate, calculates how close the measured values are to each other (P). Accuracy is also used as a statistical measure of how well a [binary classification](https://en.wikipedia.org/wiki/Binary_classification) test correctly identifies or excludes a condition.
* **Performance:** there are two types of performance. First**: F-Measure** is a measure of a test’s accuracy and relies on both the precision and recall. It is one of algorithmic efficiency to analyze and measure resources usage. Second: **Runtime** of an algorithm quantifies the amount of time taken by an algorithm to run as a function of the length of the string representing the input.
* **And the correctness precinct in understanding (part-of-speech) in reviews:** It is a percentage of understanding wholesale whether nominal or actual according to the review classification for each sentence in the review.

**1.4 Thesis Organization**

The thesis is organized as follows:

* **Chapter 2: Related work:** gives an overview of the sentiment analysis and the research concluded in this field.
* **Chapter 3: Background of Sentiment Analysis:** overview for sentiment analysis and the important definitions in this domain. It examines the differentiate sentiment analysis techniques architecture, the importance of sentiment analysis and the sentiment analysis challenges.
* **Chapter 4: Proposed Technique:** presents the proposed technique and describe a hybrid model and how to evaluate sentiment and system score.
* **Chapter 5: Experimental Results:** presents there experimental results we achieved through comparing our proposed approach to existing relevant techniques.
* **Chapter 6: Conclusion and Future Work:** It concludes the thesis and mentions the possible direction for future work.

**CHAPTER 2**

**RELATED WORK**

Although sentiment analysis and opinion mining became one of the most important sources in decision making in business still several challenges need further attention. In the following, we discuss the related work with sentiment analysis and challenges:

##### **Related Work in Sentiment Analysis**

Shukla presented a tool which judges the quality of text based on annotations on scientific papers. Its methodology collects sentiments of annotations in two approaches. It counts all the annotation produces the documents and calculates total sentiment scores. Its problem declares in a relationship between annotations that is complex. The technique needs to have a big query knowledge base containing metadata. Kasper &Vela proposed a “Web Based Opinion Mining system” for hotel reviews. The paper introduced an evaluation system for online user’s reviews and comments to support quality controls in hotel management system. It is capable of detecting and retrieving reviews on the web and deals with German reviews. It has multi-topic domain and is based on multi-polarity classification; the system could recognize the neutral e.g., “don’t know” to “classify sentiment polarity that as neutral” and the multi-topic cases identified in their corpus.

Mobile devices products reviews were analyzed in. This research can help in evaluate accuracy. It is useful in a judgment of the product quality and status in the market. This research used three machine learning algorithms (Naïve Base Classifier, K-nearest neighbor, and random forest) to calculate the sentiments accuracy. The random forest improves the performance of the classifier. There are some ways in analyzing sentiments and opinions. (Godbole, et-al) analyzed news sentiments and blogs. It splits prior work in the context of their specific task (sentiment analysis for news and blogs) into two categories. First category which - regards with techniques for automatically creating sentiment lexicon and the second one which relates to systems that analyze sentiment for entire documents.

Further, Esuli & Srinivasi's research splits related work in two other classification: the first one works with detecting the term orientation and the other works with detecting the term subjectivity. These divisions only refer to research study on the term level classification, and not document-level classification. The purpose of this research is the sentiment evaluation which refers to get the sentiment polarity (positive, negative, or neutral) of a text reviews data and evaluate the sentiment score of the text review.

The previous research with Hearst on sentiment-based categorization of the input documents has implicated either the using models inspired mostly by cognitive linguistics or the manual or semi-manual construction of discriminant-word lexicons introduced by Das & Chen. For instance, in Turney's paper introduced a new method for sentiment extraction in real time in the domain of finance; which is working based on messages from web-based stock message boards, attempt to automatically label each such message as a \buy", \sell" or \neutral" recommendation. It presented classifier reaches to the accuracy of 62% (the upper bound, human agreement rate, was 72%).

The Corpus-based approaches inspect the incorporation with seed words based on large groups of text or search for the context-dependent labels by considering the local constraints depending on (et-al) in. Alternatively, people have searched into investigating knowledge encoded in Word Net as the relations (synonymy, antonymy, and hyponymy) and glosses. Subjectivity detection research for a sentiment polarity classification mostly suppose the input documents to be opinionated. According with many tools and applications, there is a need to make a decision about the given document includes subjective information or not, or recognize which portions of the document are subjective. A subjective sentence expresses some feelings, views, or beliefs. With sentence- level subjectivity, rather than individual words, each sentence in a given document is analyzed and checked to be subjective. When necessary, the subjective sentence can be further classified as being of positive or negative semantic orientation. Then, using minimum cuts formulation, they integrated inter-sentence level contextual information with traditional bag-of-words features. They report considerable improvements over a baseline word vector classifier. The researchers introduced a recursive neural models have in common: word vector representations and classification. Their technique provides multi-topic domain. But the sentiment requires wider supervised training and evaluation resources.

The Research on opinion mining on YouTube performed with Jin & Ho for discussing how social media can be utilized to observe a person. The research idea which illustrates in Crawling, a global social networking platform, such as YouTube, has the potential to unearth content and interaction aimed at radicalization of those with little or no apparent prior interest in violent Jihadism. They got together a large dataset from a collection within YouTube that was recognized as potentially having a radicalizing agenda. The data is analyzed using social network analysis and sentiment analysis tools. It also examines the topics discussed and what the sentiment polarity (positive or negative) is towards these topics. Particularly, they focused on gender differences in this group of users, suggesting most extreme and less tolerant views among female users.

With labeled data collected from the online websites, the researchers approached the related task of detecting a sentiment polarity in reviews via supervised learning approaches. While they did experiment with a set of different features in the previous research by Brody & Elhadad, their essential focus was not on feature engineering.

The research in sentiment analysis trend is not limited yet. In order to improve accuracy and performance of the proposed techniques, applications, or algorithms. It enables them to more compatible with understanding meaning and features. But still there are some problems and challenges in text analysis of reviews and evaluate sentiment scores.

**What is Lexicon-Driven Methodology?**

(Andranik, et-al) in and (Wiebe, et-al) in utilized the MPQA sentiment lexicon to recognize the tweets sentiments about the president Barack Obama. It can enable categorizing tweets and count them if it includes more positive polarities or negative polarities of words whereby the sentiment lexicon. Although this approach is simple, they who decided an important co-relation between the aggregate sentiment in tweets and the Gallups opinion surveys

The bootstrapping approach is presented in Carmen's research for constructing subjective lexicon for under-resourced languages. This approach constructed a subjective lexicon by using a small seed list (60 words), an online lexicon (Romanian Dictionary) and a small annotated corpora. They utilized word level similarity (LSA and PMI) for filtering words. In this approach the initial seed list was

manually selected and including 60 words, which were evenly distributed among adjectives, adverbs, nouns, and verbs.

**Related Work in Sentiment Analysis Challenges**

For the purpose of this thesis, recognize the “sentiment challenges” means to find the sentiment challenges in evaluation and detection polarities for reviews and find the affects solutions for improving accuracy for text. We can minimize the key of sentiment challenges in ten sentiment challenges that face the evaluation process of sentiment reviews. Spam & fake detection, Implicit & Explicit Negation, Bipolar sentiments, world knowledge, domain dependence, huge lexicon, Natural language processing overheads, Pragmatics, Thwarted Expectations, Anaphora/co-reference Resolution, and Ambiguity.

**Spam and Fake Detection:** The World Wide Web (WWW) contains both authentic and spam contents. For effective Sentiment classification, this spam content should be eliminated before processing. Reviews face this challenge from reviewer's generated contents to express personal reviews about objects. So it becomes the existence of inserting dishonest ratings or inserting unreliable comments is an obstacle in evaluate the sentiment score accurately. There are three levels of the challenge spam:

**The duplicate reviews:** The duplicate reviews are assumed to be fake reviews. Sometime the website has duplicate sentiments reviews by considering the same reviewer, this causes the problem occurs in evaluate the real number of review and the evaluation will be repeated although the reviewer is the same in many review.

**The empty reviews:** This problem clears if we count the number of reviews or evaluate the number factor for each idea/product.

**The reviews have some words holding polarity but they don't refer to the scope or the topic:** This problem has a significant impact in sentiment evaluation score. This problem appears in Emails such as the improved e-mail classification techniques based on Artificial Immune System for reducing the false positive and create spam detectors, but it becomes effects in sentiment reviews issues. So we need to detect them though classification as example in two classes (spam and non-spam).

There are some approaches to solve this problem as an approach which used for evaluation the sentiment score from the natural language text based on a shallow dependency parser with (Zhang, et-al) in [38]. A set of discriminative rules are presented through intuitive observation. The discriminative rules are combined with the time series method to find out suspicious stores. Other challenge in social spam detection level declares in the distinct characteristics and properties of social media services. The experimental results demonstrate the effectiveness of the proposed framework as well as the roles of different types of information.

**For example:** *Review 1:"The cartoon is good" & Review2: "the horror cartoon are unrealistic*".-

**Observation:** the first review has a word has sentiment score [good] with positive polarity and the second review has a word of [unrealistic] with the negative polarity. But if we evaluate a sentiment score for determine film, the first review will be useful but the second will be spam because it not talk about the intended film or his entities.

**Implicit and Explicit Negation:** Negation is one of the biggest challenges in sentiment analysis. This challenge splits into two types: explicitly and implicitly negative.

**Explicitly** is deliberately formed and are easy to self-report and by keywords. **For example**: “I do [not like+] – this cartoon”, is to detect the negative sentiment polarity because the word (not) and convert the sentence operator to negative.

**Implicitly** is the unconscious level, are involuntarily formed and are typically unknown to us without any keywords of negative. **For example**: “I [hope to [improve] +] **-** your research work”, although the word [improve] has a positive polarity, but the [hope] word refers it is not good enough so we need to improve it, so the sentiment polarity will be negative based on implicitly negative.

Other challenge in negation indicates in some **expressions** have bi-polar values in the most cases have positive polarity although they declare with the negative words such as "Not only", "No one", and "Nobody". A method often followed in handling negation explicitly in sentences like: “[Not only+] I [like+] this algorithm, but also [easy+] to understand and apply.” the polarity is not reversed after “not” due to the presence of “only”. So this type of combinations of “not” with other words like “only” has to be kept in mind while designing the algorithm.

**For example:** *“I do [not like+] – this cartoon”*

**Observation:** the review is to determine a negative sentiment polarity because the word (not) and convert the sentence operator to negative.

**For example:** *“The viewers of the cartoon [wish] from artists to [enhance] +]*

*their performance”,*

**Observation:** Although the word [*enhance*] has a positive polarity, but the [*wish*] word refers it is not good enough so we need to improve it, so the sentiment polarity will be negative based on implicitly negative.

There are some meanings in the reviews refer to negative polarity implicitly such as *“your duties are good but you [need to achieve to degree of success] in the work”*. In the last example we observe need to also refers to the negative polarity in the review.

**Domain-independence:** One of the essential problem faced by opinion mining. This challenge is an influential relevant to the negative challenge as a survey study. There is a difference effect of the topic domain and multi-topic domain models in evaluation of sentiment analysis.

**Topic domain:** Adsod & Chopde [38] investigated the sentiment dependency in joint sentiment and topic analysis. A novel model, called Dependency-Sentiment- LDA, is proposed with extension of their joint sentiment and topic model, Sentiment- LDA. They also can employ the local dependency among sentiments by their model. They used the domain-specific and general information about sentiment expressions and combine them in the co-training setting. The experiments indicate that co-training is an effective way to combine the two information with respect to sentiment classification performance.

**Multi-topic domain:** According to the multi-topic domain, the researchers summarized an approach to explain the sentiment classification by using a cross- domain sentiment sensitive lexicon. The problem of the classifying sentiments with cross-domain consideration has to face two major challenges. First is to train classifier with the help of one or even multiple domains that we call the source domains. So the researchers suggest that this problem needs a framework that learns using the information that shows how the features of source domain are related to the features of the target domain. It is very essential to look into these to challenges and overcome them to achieve a proper cross-domain sentiment classifier.

**For example:** *Review 1:* [*This is bad*] *& Review2: [This is a bad room service in this hotel].*

**Observation:** Review 1 refers to multi-topic domain and we evaluate a word [bad] with negative polarity. But review 2 refers to one topic domain is [hotel domain] with feature of this domain [service] so [bad] - has a negative polarity with specific domain.

**Moreover the multilingual problem:** A general purpose sentiment/emotion analysis tool is also required to be working in other languages than English. Most of the work related to multilingual is tied to subjectivity analysis, a simpler sentiment-like analysis which consists in determining whether a text convey an objective or subjective assessment. Several solutions are possible, for instance training classifiers on translated corpora, using translated lexicons, building lexicons for targeted language. Recent experiments with

**Related work in web Scrapping/extracting**

Jan & Widom presented that the data mining web term [44] which is a technique utilize to crawl through different web resources to get required information, which enables an individual or a company to promote business, understanding marketing dynamics, new promotions floating on the Internet, etc. There is a growing trend among companies, organizations and individuals alike to gather information through web data mining to utilize that information in their best interest. The data mining technology is going through a huge evolution and new and better techniques are made available all the time to gather whatever information is required. There is an amount of personal information available on the internet and web data mining had helped to keep the idea of the need to secure that information at the forefront [44]. There are many tools for web scraping and data extracting online, there are very beneficial. But the big challenge faced users with them that they are very expensive. In the following, we discuss some beneficial tools in web scraping and data extracting such as:

**CHAPTER 3**

**BACKGROUND ON SENTIMENT ANALYSIS**

##### **3.1 INTRODUCTION**

Sentiment analysis is called: also opinion mining which is a computational study of reviews, sentiments, opinions, evaluations, attitudes, subjective, views, emotions, etc., expressed in the text.

**Sentiments** can be recognized as emotions, or as judgments, opinions or ideas prompted or colored by emotions or susceptibility or feelings . In Computational Linguistics, the focus is on **opinions and sentiments** rather than on feelings or emotions, and the words ‘sentiment’ and ‘opinion’ are often used alternately, also in this thesis. There are two types of textual information: facts and opinions information. While the facts are objective expressions about objects, features, entities, events and their characteristics, opinions are ordinarily subjective expressions that identify people’s sentiments, views or feelings toward objects, entities, events and their characteristics .

**Sentiment analysis** (also called: sentiment mining, sentiment classification, opinion mining, subjectivity analysis, review mining or appraisal extraction, and in some cases polarity classification) can deal with the computational handling of subjective, sentiment, and opinion in the text . It plans to realize the attitude or opinion of a writer with respect to a certain topic or goal. The attitude could reflect his/her opinion and evaluation, his/her effective situation (what are the feelings of the writer at the time of recording the opinion) or the purpose emotional communication (What is the effect which is situated on the reader when reading the opinion of the writer). Moreover, it should be noted that in this context ‘subjective’ does not mean that something is not true . In sentiment analysis, studying the subjective language: language used to rapid a private situation in the context of a text or conversation. The

research identified a private situation as a general encasement term for opinions, evaluations, emotions, and assessment. The subjective expressions has three essential types: references to private situations (e.g. “He was boiling with anger”), references to speech (or writing) events expressing private situations (e.g. “The editors of the left- leaning paper attacked the new House Speaker”) and expressive subjective entities (see below, e.g. “That doctor is a quack").

##### **3.2 HOW DOES SENTIMENT ANALYSIS WORK**

There is an approach to use sentiment analysis is with constructing a lexicon with information about which **words and phrases** are positive and which are negative. **For example**, SentiWordNet is an overtly obtainable lexical resource in which each WordNet. Synset is ascribed three numerical scores describing how objective, positive, and negative the terms in the synset are . This lexicon can either compile manually or be acquired automatically. The annotation of lexical or corpora is usually done by hand, and classifiers are then trained with large sets of features to classify a new batch of words or phrases. There are other approaches to analyze sentiments focus on the mining of **sentences or entire documents,** rather than to depend on the parity of words. This approach usually works with corpora of text documents. The essential problem with document classification (polarity classification) which is that it has to determine the overall sentiment characteristics of an entire document, while the expressed sentiment can be included in just one sentence or word. In other cases, the sentiment can be expressed implicitly, which makes it even more difficult to detect and classify. However, the context surrounding these ‘hidden’ sentiments can provide very beneficial information for classifying it. Based on this division of the field of sentiment analysis, we often speak of **word-level**, **sentence-level** and **document-level** sentiment classification.

On other hand, we find another approach in the mining of sentiment is **on the web**. Web opinion mining aims at extracting summarize, and track various aspects of subjective information on the Web . This can prove helpful for advertising companies or trend watchers. By a synopsis of **Sentiment analysis** defection (also called as opinion mining) that refers to the use of [natural language processing](https://en.wikipedia.org/wiki/Natural_language_processing) (NLP), [text analysis](https://en.wikipedia.org/wiki/Text_analytics) (TA) and [computational linguistics](https://en.wikipedia.org/wiki/Computational_linguistics) (CL) to identify and extract subjective information in source materials. Sentiment analysis is widely utilized for online reviews and social media for a variety of applications, ranging fr[om marketing](https://en.wikipedia.org/wiki/Marketing) to [customer service](https://en.wikipedia.org/wiki/Customer_relationship_management).

##### **Why sentiment analysis is Important?**

There are millions online users, who write and read online and Internet usage around the world. Online daily sentiments becomes the most significant issue in making a decision. According to a new survey conducted by Dimensional Research, the survey discuss the perecntage of trust online customer reviews as much as personal recommendations. According to 2011 Study: 74% of customer’s confidence is based on online personal recommendation reviews, 60% in 2012 study, and 57% in 2013 Study. But this percentage increases with respect to 2014 Study: 94% of customer’s trust on online sentiment reviews .

##### **Why Sentiment Analysis is hard?**

Sentiment analysis is one of the toughest problems to be addressed by Computer. Identifying some entities, features or patterns is hard for machines or even impossible while it is easy for human beings. Below you can find some intractable situations for computers:

* + 1. Dealing with ironies or sarcasm, it is difficult to understand that the opposite meaning of a sentence is required. Sometimes ironies can be recognized through special punctuation marks such as (!!!) but it is not that common to be a rule or sign for these types of expressions.
       1. Pronoun resolution is another daunting task. Although there are some techniques and algorithms that can solve, it is still demanding task in sentiment analysis. For instance; there are opinion words in a sentence but because the corresponding feature is a pronoun, it is not easy to find which feature is expressed by those sentiment words.
       2. Defining the Strength of an opinion also should be recognized as a demanding task in this area. Opinions have different strengths. Some of them are very strong:

“This camera is a piece of junk” and some of them are weak: “I think this camera is fine”. Although it is possible to create a seed list of weak or strong opinion words depending on the application need, it is still not doable for computers when the strength of opinion mixes with the position of that opinion and changes the polarity of the document completely.

**For example:**

“*The scenario was superficial, Artists played awful, sound quality was terrific, but I liked it*” has positive polarity in the context of movie review although it has lots of negative opinion words. Another difficulty is extracting the implicit keywords or features, in order to identify the expressed sentiment. Despite there are some research which have investigated on identifying the Implicit Sentiment of texts, most of the time, explicit features are taken into account to perform opinion mining tasks.

So we have discussed that sentiment analysis in abstract till now. At this time, we want to focus on the classification task from a practical point of view. First, we are going to look at document level sentiment analysis: approaches, classifiers, features and other related sub-tasks and then focus on Sentence-Level sentiment classification. But before going further we have to know about opinion words and their generation in detail as these is an important component in sentiment analysis .

##### **Sentiment Analysis (SA) & Natural Langauge processing (NLP):**

In the following discussion, we present some of the main definitions for sentiment analysis:

* **Natural language processing (NLP):**

It is an area of the [computer science](https://en.wikipedia.org/wiki/Computer_science), [artificial intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence), and [computational](https://en.wikipedia.org/wiki/Computational_linguistics) [linguistics](https://en.wikipedia.org/wiki/Computational_linguistics) interested in the interactions between [computers](https://en.wikipedia.org/wiki/Computer) and [human (natural)](https://en.wikipedia.org/wiki/Natural_language) [languages.](https://en.wikipedia.org/wiki/Natural_language) Intrinsically, NLP is related to the field of [human–computer interaction](https://en.wikipedia.org/wiki/Human%E2%80%93computer_interaction) (HCI). Many challenges in NLP include [natural language understanding](https://en.wikipedia.org/wiki/Natural_language_understanding), that is, enabling computers to deduce meaning from human or natural language input, and

others involve [natural language generation](https://en.wikipedia.org/wiki/Natural_language_generation).The term of Natural Language Processing involves a wide set of techniques for automated generation, manipulation and analysis of natural or human languages. Despite most NLP techniques inherit largely from Linguistics and Artificial Intelligence, they are also affected by relatively newer domains such as Machine Learning, Computational statistics and Cognitive Science.

* + **Token**: Before any real processing can be done on the input text, it needs to be segmented into linguistic units such as words, punctuation, and numbers or alphanumeric. These units are recognized as tokens.
  + **Sentence**: This refers to an ordered sequence of tokens.
  + **Tokenization**: The operation of splitting a sentence into its constitutive tokens.
  + **Corpus**: This means a body of text, usually including a large number of sentences.
  + **Part-of-speech (POS) Tag**: A word can be categorized into one or more of a set of lexical or part-of-speech classes such as Nouns, Verbs, Adjectives and Articles, to name a few. A POS tag is a symbol representing such a lexical category – NN (Noun), VB (Verb), JJ (Adjective), AT (Article). One of the oldest and most commonly used tag sets is the Brown Corpus tag set.
* **Parse Tree**: It represents a tree defined over a given sentence that interprets the syntactic structure of the sentence as identified with a formal grammar.
  + **Part-Of-Speech (POS) Tagging**: Given a sentence and a set of POS tags, a mutual language processing task is to automatically specify POS tags to each word in the sentences. **For example**, given the sentence "The ball is red", the output of a POS tagger would be the /AT ball/NN is/VB red/JJ. State-of-the-art POS taggers can reach to higher accuracy as 96%. Tagging text with part-of-speech turns out to be much beneficial for more complicated NLP tasks such as parsing and machine translation.
  + **Computational Morphology**: Natural languages consist of a very large number of words that are constructed upon basic building blocks known as morphemes (or stems), the smallest linguistic units possessing meaning. Computational morphology which is interested in the discovery and analysis of the internal structure of words using computers.
  + **Parsing**: In the parsing task, a parser builds the parse tree given a sentence. There are some parsers assume the existence of a set of grammar rules to parse but recent parsers are smart enough to deduce the parse trees directly from the given data using complex statistical models . Most parsers also operate in a supervised setting and require the sentence to be POS-tagged before it can be parsed. Statistical parsing is an area of active research in NLP.
* **Subjective Sentence:** It is a sentence in which the writer expresses his or her feelings or sentiments toward entities, events and their properties. **For example:** “I like swimming”.
* **Objective Sentence:** It is a factual sentence about entities, events, and their properties, **For example:** “The schedule contains swimming, diving and …”
* **Opinion:** It is a belief or judgment based on special knowledge towards a topic. Opinions are sometimes expressed explicitly like: “The voice quality of this phone is amazing.” But sometimes they are hidden in the sentiment of a sentence, for instance; “The earphone broke in two days”. Since the concept of opinion is very wide, sentiment classification mostly concentrates on the general feeling expressed by opinions (Positive / Negative). In fact, positivity or negativity is determining the Polarity of an opinion. In other words, one of the main subtasks of sentiment analysis is determining the polarity of documents or in more details, determining the polarity of each subjective sentence in a document.
* **Opinion words:** They are words that are commonly used to express positive or negative sentiment. **For example:**

{Beautiful, pretty, love}  Positive sentiment

{Ugly, awful, hate}  Negative sentiment

* **Sentiment Orientation-SO (Polarity):** It indicates whether the expressed opinion by opinion words is positive, negative or neutral. **For example:**

"The camera takes wonderful pictures"  Positive.

* **Opinion Sentence:** It is a sentence which contains one or more opinion words.

**For example:** "The story was amazing as was the playing of actors".

* **Object / Features:** So far, we have used “topic” to refer the main subject in reviews which is going to be discussed. Hence, we call it “Object”. In opinionated documents, objects and their components or attributes are going to be reviewed and sentiments toward them are expressed in terms of “opinion words”; these components or attributes are called: “object-features” . **For Example:**

"The voice quality of the phone is good".

**Object:** phone

Explicit object- feature: voice quality

**Opinion word**: good

In this example: the explicit feature is voice quality, but sometimes object features should be inferred from the sentence. This kind of feature is called: "implicit feature".

**For example:**

"The phone is too large"

**Object** phone Implicit feature: size **Opinion word:** large

It is a function to classify different objects and label them as an output. In sentiment analysis, classifiers are used to determine the polarity of a subjective sentence with respect to the topic. There are two types of classification: Supervised Classification, such that a classifier is inferred from the training set. The classifier should predict the correct label (positive or negative) for any valid input object. In contrast unsupervised classification infers the hidden structure of raw data. In sentiment analysis, both classification types are widely used. The main task of Sentiment Analysis is extracting suitable features and constructing an engineered feature vector as an input for classifier

**Text mining**  also referred to as text [data mining,](https://en.wikipedia.org/wiki/Data_mining) almost synonymous to [text](https://en.wikipedia.org/wiki/Text_mining#Text_mining_and_text_analytics) [analytics,](https://en.wikipedia.org/wiki/Text_mining#Text_mining_and_text_analytics) refers to the process of deriving high-quality [information](https://en.wikipedia.org/wiki/Information) from [text](https://en.wikipedia.org/wiki/Plain_text). High- quality information is typically derived through the devising of patterns and trends through means such as statistical. Text mining ordinarily involves the operation of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a [database](https://en.wikipedia.org/wiki/Database)), deriving patterns within the structured, and lately evaluation and interpretation of the output. 'High quality' in the text mining usually refers to some combination of

[relevance,](https://en.wikipedia.org/wiki/Relevance_(information_retrieval)) [novelty,](https://en.wikipedia.org/wiki/Novelty_(patent)) and interestingness. Typical text mining tasks include [text](https://en.wikipedia.org/wiki/Text_categorization) [categorization,](https://en.wikipedia.org/wiki/Text_categorization) [text clustering,](https://en.wikipedia.org/wiki/Text_clustering) [concept/entity extraction](https://en.wikipedia.org/wiki/Concept_mining), production of granular taxonomies, [sentiment analysis,](https://en.wikipedia.org/wiki/Sentiment_analysis) [document summarization](https://en.wikipedia.org/wiki/Document_summarization), and entity relation modeling (i.e., learning relations between [named entities](https://en.wikipedia.org/wiki/Named_entity_recognition)).

**Text mining** is a new field of computer science which takes care of strong connections with natural language processing, data mining, machine learning, information retrieval and knowledge management. Text mining looks for extracting helpful information from unstructured textual data through the identification and exploration of interesting patterns . **Text analysis** involves [information](https://en.wikipedia.org/wiki/Information_retrieval) [retrieval,](https://en.wikipedia.org/wiki/Information_retrieval) [lexical analysis](https://en.wikipedia.org/wiki/Lexical_analysis) to study word frequency distributions, [pattern](https://en.wikipedia.org/wiki/Pattern_recognition) [recognition,](https://en.wikipedia.org/wiki/Pattern_recognition) [tagging](https://en.wikipedia.org/wiki/Tag_(metadata))/[annotation](https://en.wikipedia.org/wiki/Annotation), [information extraction,](https://en.wikipedia.org/wiki/Information_extraction) [data mining](https://en.wikipedia.org/wiki/Data_mining) techniques including link and association analysis, [visualization](https://en.wikipedia.org/wiki/Information_visualization), and [predictive analytics](https://en.wikipedia.org/wiki/Predictive_analytics). The comprehensive goal is, essentially, to turn text into data for analysis, via application of [natural language processing](https://en.wikipedia.org/wiki/Natural_language_processing) (NLP) [49, 50] and analytical methods. Textual analysis is the approach communication researchers use to describe and interpret the characteristics of a recorded or visual message. The purpose of textual analysis is to describe the content, structure, and functions of the messages contained in the texts. The significant considerations in textual analysis contain selecting the types of texts to be studied, acquiring appropriate texts, and determining which particular approach to employ in analyzing them. There are two general categories of texts:

* + Transcripts of communication (verbatim recordings)
  + Outputs of communication (messages produced by communicators) In terms of acquiring.

The last definition is **Computational linguistics** which is an [interdisciplinary](https://en.wikipedia.org/wiki/Interdisciplinary) field concerned with the statistical or rule-based modeling of [natural language](https://en.wikipedia.org/wiki/Natural_language) from a computational perspective. Computational linguistics which has the theoretical and applied components, where theoretical computational linguistics takes up issues in [theoretical linguistics](https://en.wikipedia.org/wiki/Theoretical_linguistics) and cognitive science, and applied computational linguistics focuses on the practical outcome of modeling human language use.

##### **3.7 Sentiment Analysis Models**

Earlier, the sentiment text analysis or more exactly positive/negative classification relies on using a dataset and a classifier. The documents apply the classifier into two sets: positive and negative. The increased documents will be exhibited informatively in test phase, the more accurate the result of classification will be. Since, finding the best document representatives that can describe it (document) better is of a vital importance in sentiment analysis.

###### 

###### **Part-of- Speech Tags**

Part-of-speech tagging (POS tagging or POST), also known [grammatical](https://en.wikipedia.org/wiki/Grammar) tagging or [word-category](https://en.wikipedia.org/wiki/Lexical_category) disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular [part of speech,](https://en.wikipedia.org/wiki/Parts_of_speech) based on both its definition, as well as its context—i.e. [relationship with adjacent and](https://en.wikipedia.org/wiki/Lexicography) [related words](https://en.wikipedia.org/wiki/Lexicography) in a [phrase,](https://en.wikipedia.org/wiki/Phrase) [sentence,](https://en.wikipedia.org/wiki/Sentence_(linguistics)) or [paragraph.](https://en.wikipedia.org/wiki/Paragraph) Tagging the grammatical features of each word is also a very good strategy to improve accuracy ratios and detect useful patterns for classifications. Subjective texts tend to use simple past tense instead of the past participle. As opposite to the positive set, the negative set contains more often verbs in the past tense, because many authors express their negative sentiments about their loss or disappointment.

###### **Semantic relationships**

Semantic relationships are the associations that there exist between the meanings of words (semantic relationships at word level), between the meanings of phrases, or between the meanings of sentences (semantic relationships at phrase or sentence level). Following is a description of such relationships. The Semantic approach gives sentiment values directly and relies on different principles for computing the similarity between words. This principle gives similar sentiment values to semantically close words. WordNet for example; provides different kinds of semantic relationships between words used to calculate sentiment polarities. WordNet could be used too for obtaining a list of sentiment words by iteratively expanding the initial set with synonyms and antonyms and then determining the sentiment polarity for an unknown word by the relative count of positive and negative synonyms of this word.

Semantic methods can be mixed with the statistical methods to perform SA task as the work who used both methods to find product weakness from online reviews. Their weakness finder extracted the features and group explicit features by using morpheme-based method to identify feature words from the reviews. They used Hownet-based similarity measure to find the frequent and infrequent explicit features which describe the same aspect . They identified the implicit features with collocation statistics-based selection method PMI. They have grouped products feature words into corresponding aspects by applying semantic methods. They have utilized sentence-based SA method to determine the polarity of each aspect in sentences taking into consideration the impact of adverbs of degree. They could find the weaknesses of the product, as it was probably the most unsatisfied aspect in customers’ reviews, or the aspect which is more unsatisfied when compared with their competitor’s product reviews. Their results expressed the good performance of the weakness finder .

##### **Sentiment analysis problems**

In this section, we will discuss the problems faced sentiment analysis shortly. The research point of an abstraction of the problems enables us to see a rich set of inter-related sub-problems which make up the sentiment analysis problem. It is often said that if we cannot structure a problem, we probably do not understand the problem. The objective of the definitions is thus to abstract a structure from the complex and intimidating unstructured natural language text. From a practical application point of view, the definitions let practitioners see what sub-problems need to be solved in a practical system, how they are related, and what output should be produced.

###### **Sentiment Problem Definitions**

Sentiment analysis mainly studies opinions and sentiments which express or imply positive or negative sentiments. There are main types of sentiment problems [9]: text analysis, understanding meaning of the text, and in the evaluation of sentiment analysis. In this section, we discuss the problem in this context. In the following, we discuss the types of sentiment problems [60].

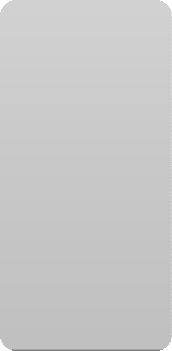
###### **Sentiment Analysis Evaluation**

**Sentiment classification polarity:** the sentiment classifies the polarities to understand the meaning nearly and be more accurate. So there are some ways to classify polarities such as two level (positive, and negative) .

* **For example:** *"The book is great work. And having an efficient algorithm."*
* **Observation:** With applying text analysis on the previous review with splitting the reviews into sentences and tokenizing sentences into some words. Remove stop list and punctuation list. Then detect the sentiment polarity for the words [*great*] + with positive polarity, [*efficient*] + with positive polarity and evaluate score for each word, but we have [*algorithm*] word does not have polarity. We can detect a polarity of a word [*algorithm*] is neutral in other way is hierarchal classification with (negative, neutral, and positive). Or the more specified classification assuming four with (positive, fair, mixed, negative). Moreover a classified by five sentiment classes (Very positive, Positive, Fair, Negative, Very Negative) for being more accurate and dependence.
* **For example:** *"The book is a great work. And having an efficient algorithm."*
* **Observation:** The [*great*] and [efficient] words have in very positive class. An [algorithm] word is a fair class.

###### **Sentiment Analysis Challenges:**

For the purpose to recognize the “sentiment analysis challenges” means to find the sentiment challenges in evaluation and detection polarities for reviews and find the most effected solutions for the highest accuracy for text. There are a lot of research in the evaluation challenges [10, 11], we divide challenges into Theoretical and Technical challenges as our survey study for forty-seven papers, as in the following:



Negation

Extracting

Topic Features

World Knowledge

**Theoretical**

Spam or fake

reviews

Domain Dependance

**Sentiment**

**Analysis Challenges**

NLP

Overheads

Thwarted Expections

Huge lexicon

**Technical**

Temporal

Relations

Bi-polar words

Figure 3.1: The Sentiment Analysis Challenges [10]

Figure 3.1 which illustrates the main sentiment challenges. These challenges become obstacles in sentiment evaluation process and in understanding the meaning of the sentiments reviews.

***Four technical Challenges:***

1. **Temporal Relations:** The time of reviews may be important for sentiment analysis.
   * **For example:** The reviewer may think that Windows Vista is good in 2008, but now he may have negative opinion in 2009 because of new Windows7.
   * **Observation:** So assessing this kind of sentiments that are changed with time may enhance the sentiment analysis performance. This supports us to notice if a certain product has improved with time, or people change their sentiment about a product.
2. **Thwarted Expectations:** Sometimes the writer intentionally sets up situation only to disprove it at the end. English text to illustrate the concept of thwarted expectations. There are several fuzzy words that we can't recognize it or determine the sentiment polarity. Although we may identify the topic domain, we also can't the polarity of them.
3. **Bi-polar sentiments:** There are some words and phrases have bi-polar meaning that depend on the topic and features or domain keywords implicit meaning.

* **For example:** Review 1: [*Old conference in the data mining field.*] & Review 2: [*Old author in this field*] & Review 3: [*Old topic*].
* **Observation:** almost cases the word [Old] has a negative sentiment score but in the previous reviews logically has a positive polarity is review 1 and review 2 but has a negative polarity is review 3. So we need to recognize a features or keywords or the topic domain to know how to identify the polarity.

1. **Generate huge lexicons:** this has an obstacle in creating huge lexicons to cover the data evaluation.

The previous types of sentiment problems have a big effect on the understanding sentence. There are some sentiments and words have more than meaning and other some sentiments relies on world knowledge. There are some challenges contain that asymmetry in the availability of opinion mining software and Term Position.

##### **Sentiment Analysis lexicons & Tools:**

Sentiment analysis is receiving an increasingly growing interest from the natural language processing (NLP) community, which is particularly motivated by the wide- spread need for sentiment/ opinion based applications, such as product reviews, entity tracking and analysis and sentiment summarization. The existence of the World Wide Web has changed the way that people express their views and opinions, and has provided researchers with a huge source of user-generated content. Wanting to buy a product no longer involves questioning friends or family; wanting consumer opinions about your own product no longer needs to rely on focus groups or external consultants. However, this huge resource of valuable information, the Web, is unstructured, and sentiment analysis is able to automatically discover opinions and present them in a structured manner.

Sentiment mining has become a useful tool for the commercial activities of both companies and individual consumers. They want to sort out opinions about products, services, or brands that are scattered in online texts such as product review articles or forums [65]. In the following, paragraphs we sum up a few important (future) applications of sentiment analysis. Sentiment analysis can be used for determining critics’ opinions about a given product (e.g. a digital camera, movie, etc.) by classifying online product reviews from websites such as Amazon and C|Net, RottenTomatoes.com and IMDb, and can also prove very helpful for opinion- oriented questions in question answering. Tracking the shifting attitudes of the general public toward a political candidate by mining online forums is also a useful application. It can furthermore be used to alert customer services of dissatisfied customers that utter their frustrations on forums or discussion boards. Tracking (mood) trends of bloggers is also becoming a valued research field since it can be used for research in trends or consumer preferences.

In other words, applications resulting from sentiment analysis research can help a great deal in marketing research (i.e. quality control, automatic information gathering from the Internet instead of bothering customers with surveys, etc.), and can consequently be of great service for advertising and market intelligence companies and trend watchers. In this respect, sentiment analysis can contribute to collective intelligence research: the study of the combination of behavior, preferences, or ideas of a group of people to create novel insight.

##### **Sentiment Analysis Research**

###### **Introduction**

As discussed above, pervasive real-life applications are only part of the reason why sentiment analysis is a popular research problem. It is also highly challenging as a NLP research topic and covers many novel sub-problems as we will see later.

Additionally, there was little research before the year 2000 in either NLP or in linguistics . Part of the reason is that before then there was little opinion text

available in digital forms. Since the year 2000, the field has grown rapidly to become one of the most active research areas in NLP. It is also widely researched in data mining, Web mining, and information retrieval. In fact, it has spread from computer science to management sciences.

**2.Different Levels of Analysis**

We now give a brief introduction to the main research problems based on the level of granularities of the existing research. In general, sentiment analysis has been investigated mainly at four levels:

**3.Document-level sentiment classification:** The task at this level is to classify whether a whole opinion document expresses a positive or negative sentiment . For example, given a product review, the system determines whether the review expresses an overall positive or negative opinion about the product. This task is commonly known as document-level sentiment classification. This level of analysis assumes that each document expresses opinions on a single entity (e.g., a single product). Thus, it is not applicable to documents which evaluate or compare multiple entities.

**4. Sentence level sentiment classification:** The task at this level goes to the sentences and determines whether each sentence expressed a positive, negative, or neutral opinion. Neutral usually means no opinion. This level of analysis is closely related to subjectivity classification , which distinguishes sentences (called: objective sentences) that express factual information from sentences (called: subjective sentences) that express subjective views and opinions. However, we should note that subjectivity is not equivalent to sentiment as many objective sentences can imply opinions, e.g., “We bought the car last month and the windshield wiper has fallen off.” Researchers have also analyzed clauses, but the clause level is still not enough, e.g., “Apple is doing very well in this lousy economy”.

**The differences between document level and sentence level**: is classifying an opinionated document as expressing a positive or negative opinion [68] and Sentence Level Sentiment Classification: classifying a sentence as subjective or objective and for a subjective sentence, classifying it as expressing positive, negative or natural opinion.

Document level sentiment classification is often used to give a dominant sentiment towards a topic as a whole. It does not take into account the different angles of the object. For example in the context of movie reviews, when the document is positive it means that the writer likes the movie in general, but if he or she concentrates on different aspects such as the story or the playing of actors, maybe the opinions are negative. In this case, if each feature and related opinion is extracted from each subjective sentence, the result is going to be a Sentence-Level sentiment classification. Sentence-level sentiment classification is required in order to obtain an accurate analysis of an object; for example in terms of products, such a precise analysis is necessary in order to make product improvements by distinguishing between what features (components or attributes) of a product are liked and disliked by consumers. Such information is not obtained by Document-Level sentiment classification.

**5. Entity and Aspect level sentiment classification:** Both the document level and the sentence level analyzes do not discover what exactly people liked and did not like. Aspect level performs finer-grained analysis. Aspect level was earlier called: feature level (feature-based opinion mining and summarization). Instead of looking at language constructs (documents, paragraphs, sentences, clauses or phrases), aspect level directly looks at the opinion itself. It is based on the idea that an opinion consists of a sentiment (positive or negative) and a target (of opinion).

An opinion without its target being identified is of limited use. Realizing the importance of opinion targets also helps us understand the sentiment analysis problem better. For example, although the sentence “although the service is not that great, I still love this restaurant” clearly has a positive tone, we cannot say that this sentence is entirely positive. In fact, the sentence is positive about the restaurant (emphasized), but negative about its service (not emphasized). In many applications, opinion targets are described by entities and/or their different aspects. Thus, the goal of this level of analysis is to discover sentiments on entities and/or their aspects. For example, the sentence “The iPhone’s call quality is good, but its battery life is short” evaluates two aspects, call quality and battery life, of iPhone (entity). The sentiment on iPhone’s call quality is positive, but the sentiment on its battery life is negative. The call quality and battery life of iPhone are the opinion targets. Based on this level of analysis, a structured summary of opinions about entities and their aspects can be produced, which turns unstructured text to structured data and can be used for all kinds of qualitative and quantitative analyzes.

**Word level sentiment classification**: In more recent works, text analysis granularity has been taken down to sentence level, e.g., by using presence of opinion- bearing lexical items (single words or n-grams) to detect subjective sentences, or by exploiting association rule mining for a feature based analysis of product reviews. So word level is working on the evaluation of each word. Sometimes this classification deals with ordering of part-of-speech to achieve the evaluation score. For example, "The book is good", this sentence has a 'good' word has score which refers to positive polarity.

**The difference between sentence level and word level**: declares in the total evaluation score. They are similar in analysis of word based. But the differences appears in the mechanism of evaluation this words. In sentence level, the sentence polarity relies on the majority of polarity (positive or negative). That becomes not accurate enough, **for example:** "The story of movie is good but it is worst film" in this case in sentiment sentence level, the polarity is neutral not positive and not negative. But the word level is based on term frequency and total weight not polarity only.

**3.11 Chapter Summary:**

The objective of Sentiment Analysis is evaluating the sentiments and opinions of a writer respectively, on topic domain or multi-topic domain. It calculates the aggregated sentiment polarity of a text or online reviews for one topic based on sentiment classification levels, such as positive or negative. This sentiment can be his or her judgments or evaluations or any other emotional reviews. With the growing availability of opinion resources such as social networks or review websites, the challenge for seeking out the opinions of others has increased as well. The computational study for opinions, sentiments, and emotions expressed in texts or reviews are the challenges of natural language processing. It has been presented that an integration of Information Retrieval and Natural Language Processing approaches can be useful in terms of opinion mining and sentiment analysis.

**CHAPTER 4**

**PROPOSED TECHNIQUE:**

**SENTIMENT ANALYSIS OF SENTENCES**

**4.1 SOFTWARE DEVELOPMENT MODEL**

**4.1.1 Software development Lifecycle:**

Software Development Life Cycle (SDLC) is a process used by the software industry to design, develop and test high quality softwares. The SDLC aims to produce a high-quality software that meets or exceeds customer expectations, reaches completion within times and cost estimates.

* SDLC is the acronym of Software Development Life Cycle.
* It is also called as Software Development Process.
* SDLC is a framework defining tasks performed at each step in the software development process.
* ISO/IEC 12207 is an international standard for software life-cycle processes. It aims to be the standard that defines all the tasks required for developing and maintaining software.

## 4.1.2 What is SDLC?

SDLC is a process followed for a software project, within a software organization. It consists of a detailed plan describing how to develop, maintain, replace and alter or enhance specific software. The life cycle defines a methodology for improving the quality of software and the overall development process.

The following figure is a graphical representation of the various stages of a typical SDLC.

**Diagram

Description automatically generated**

A typical Software Development Life Cycle consists of the following stages −

### Stage 1: Planning and Requirement Analysis

Requirement analysis is the most important and fundamental stage in SDLC. It is performed by the senior members of the team with inputs from the customer, the sales department, market surveys and domain experts in the industry. This information is then used to plan the basic project approach and to conduct product feasibility study in the economical, operational and technical areas.

Planning for the quality assurance requirements and identification of the risks associated with the project is also done in the planning stage. The outcome of the technical feasibility study is to define the various technical approaches that can be followed to implement the project successfully with minimum risks.

### Stage 2: Defining Requirements

Once the requirement analysis is done the next step is to clearly define and document the product requirements and get them approved from the customer or the market analysts. This is done through an **SRS (Software Requirement Specification)** document which consists of all the product requirements to be designed and developed during the project life cycle.

### Stage 3: Designing the Product Architecture

SRS is the reference for product architects to come out with the best architecture for the product to be developed. Based on the requirements specified in SRS, usually more than one design approach for the product architecture is proposed and documented in a DDS - Design Document Specification.

This DDS is reviewed by all the important stakeholders and based on various parameters as risk assessment, product robustness, design modularity, budget and time constraints, the best design approach is selected for the product.

A design approach clearly defines all the architectural modules of the product along with its communication and data flow representation with the external and third party modules (if any). The internal design of all the modules of the proposed architecture should be clearly defined with the minutest of the details in DDS.

### Stage 4: Building or Developing the Product

In this stage of SDLC the actual development starts and the product is built. The programming code is generated as per DDS during this stage. If the design is performed in a detailed and organized manner, code generation can be accomplished without much hassle.

Developers must follow the coding guidelines defined by their organization and programming tools like compilers, interpreters, debuggers, etc. are used to generate the code. Different high level programming languages such as C, C++, Pascal, Java and PHP are used for coding. The programming language is chosen with respect to the type of software being developed.

### Stage 5: Testing the Product

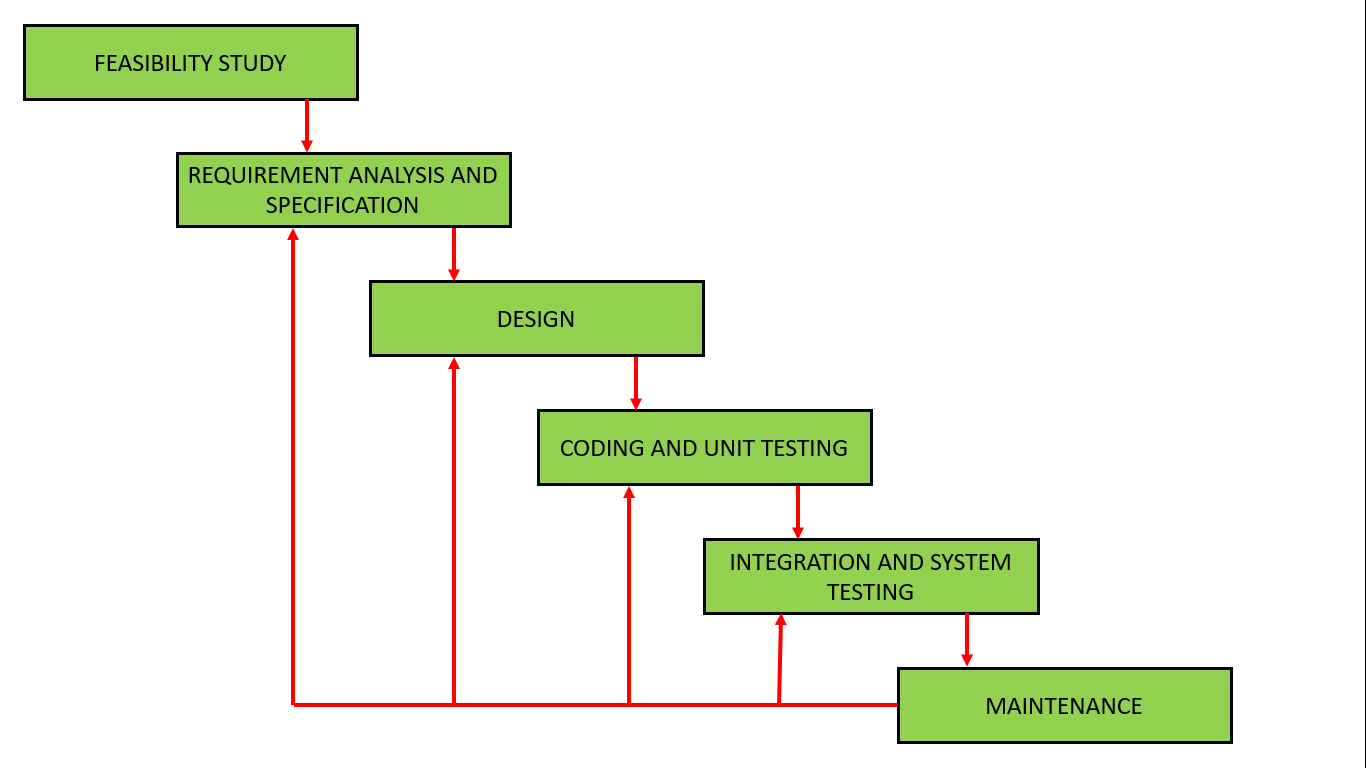
This stage is usually a subset of all the stages as in the modern SDLC models, the testing activities are mostly involved in all the stages of SDLC. However, this stage refers to the testing only stage of the product where product defects are reported, tracked, fixed and retested, until the product reaches the quality standards defined in the SRS.

### Stage 6: Deployment in the Market and Maintenance

Once the product is tested and ready to be deployed it is released formally in the appropriate market. Sometimes product deployment happens in stages as per the business strategy of that organization. The product may first be released in a limited segment and tested in the real business environment (UAT- User acceptance testing).

Then based on the feedback, the product may be released as it is or with suggested enhancements in the targeting market segment. After the product is released in the market, its maintenance is done for the existing customer base.

|  |  |
| --- | --- |
| **Sr.No.** | **Operator & Description** |
| 1 | [**Button**](https://www.tutorialspoint.com/python/tk_button.htm)  The Button widget is used to display buttons in your application. |
| 2 | [**Canvas**](https://www.tutorialspoint.com/python/tk_canvas.htm)  The Canvas widget is used to draw shapes, such as lines, ovals, polygons and rectangles, in your application. |
| 3 | [**Checkbutton**](https://www.tutorialspoint.com/python/tk_checkbutton.htm)  The Checkbutton widget is used to display a number of options as checkboxes. The user can select multiple options at a time. |
| 4 | [**Entry**](https://www.tutorialspoint.com/python/tk_entry.htm)  The Entry widget is used to display a single-line text field for accepting values from a user. |
| 5 | [**Frame**](https://www.tutorialspoint.com/python/tk_frame.htm)  The Frame widget is used as a container widget to organize other widgets. |
| 6 | [**Label**](https://www.tutorialspoint.com/python/tk_label.htm)  The Label widget is used to provide a single-line caption for other widgets. It can also contain images. |
| 7 | [**Listbox**](https://www.tutorialspoint.com/python/tk_listbox.htm)  The Listbox widget is used to provide a list of options to a user. |
| 8 | [**Menubutton**](https://www.tutorialspoint.com/python/tk_menubutton.htm)  The Menubutton widget is used to display menus in your application. |
| 9 | [**Menu**](https://www.tutorialspoint.com/python/tk_menu.htm)  The Menu widget is used to provide various commands to a user. These commands are contained inside Menubutton. |
| 10 | [**Message**](https://www.tutorialspoint.com/python/tk_message.htm)  The Message widget is used to display multiline text fields for accepting values from a user. |
| 11 | [**Radiobutton**](https://www.tutorialspoint.com/python/tk_radiobutton.htm)  The Radiobutton widget is used to display a number of options as radio buttons. The user can select only one option at a time. |
| 12 | [**Scale**](https://www.tutorialspoint.com/python/tk_scale.htm)  The Scale widget is used to provide a slider widget. |
| 13 | [**Scrollbar**](https://www.tutorialspoint.com/python/tk_scrollbar.htm)  The Scrollbar widget is used to add scrolling capability to various widgets, such as list boxes. |
| 14 | [**Text**](https://www.tutorialspoint.com/python/tk_text.htm)  The Text widget is used to display text in multiple lines. |
| 15 | [**Toplevel**](https://www.tutorialspoint.com/python/tk_toplevel.htm)  The Toplevel widget is used to provide a separate window container. |
| 16 | [**Spinbox**](https://www.tutorialspoint.com/python/tk_spinbox.htm)  The Spinbox widget is a variant of the standard Tkinter Entry widget, which can be used to select from a fixed number of values. |
| 17 | [**PanedWindow**](https://www.tutorialspoint.com/python/tk_panedwindow.htm)  A PanedWindow is a container widget that may contain any number of panes, arranged horizontally or vertically. |
| 18 | [**LabelFrame**](https://www.tutorialspoint.com/python/tk_labelframe.htm)  A labelframe is a simple container widget. Its primary purpose is to act as a spacer or container for complex window layouts. |
| 19 | [**tkMessageBox**](https://www.tutorialspoint.com/python/tk_messagebox.htm)  This module is used to display message boxes in your applications. |



**4.2 Front-End Technology Used:**

**4.2.1 Python Tkinter**

Tkinter is the standard GUI library for Python. Python when combined with Tkinter provides a fast and easy way to create GUI applications. Tkinter provides a powerful object-oriented interface to the Tk GUI toolkit.

## 4.2.2 Tkinter Widgets

Tkinter provides various controls, such as buttons, labels and text boxes used in a GUI application. These controls are commonly called widgets.

There are currently 15 types of widgets in Tkinter. We present these widgets as well as a brief description in the following table –

**4.3 Backend Technology Used:**

**4.3.1 nltk**

**4.3.2 Stemming:**

Stemming is a technique used to extract the base form of the words by removing affixes from them. It is just like cutting down the branches of a tree to its stems. For example, the stem of the words ***eating, eats, eaten*** is ***eat***.

Search engines use stemming for indexing the words. That’s why rather than storing all forms of a word, a search engine can store only the stems. In this way, stemming reduces the size of the index and increases retrieval accuracy.

## Various Stemming algorithms

In NLTK, **stemmerI**, which have **stem()** method, interface has all the stemmers which we are going to cover next. Let us understand it with the following diagram

**Diagram

Description automatically generated**

## 4.3.2.1 Porter stemming algorithm

It is one of the most common stemming algorithms which is basically designed to remove and replace well-known suffixes of English words.

### PorterStemmer class

NLTK has **PorterStemmer** class with the help of which we can easily implement Porter Stemmer algorithms for the word we want to stem. This class knows several regular word forms and suffixes with the help of which it can transform the input word to a final stem. The resulting stem is often a shorter word having the same root meaning. Let us see an example −

First, we need to import the natural language toolkit(nltk).

import nltk

Now, import the **PorterStemmer** class to implement the Porter Stemmer algorithm.

from nltk.stem import PorterStemmer

Next, create an instance of Porter Stemmer class as follows −

word\_stemmer = PorterStemmer()

Now, input the word you want to stem.

word\_stemmer.stem('writing')

### Output

'write'

word\_stemmer.stem('eating')

### Output

'eat'

### Complete implementation example

import nltk

from nltk.stem import PorterStemmer

word\_stemmer = PorterStemmer()

word\_stemmer.stem('writing')

### Output

'write'

## 4.3.2.2 Lancaster stemming algorithm

It was developed at Lancaster University and it is another very common stemming algorithms.

### Lancaster Stemmer class

NLTK has **LancasterStemmer** class with the help of which we can easily implement Lancaster Stemmer algorithms for the word we want to stem. Let us see an example −

First, we need to import the natural language toolkit(nltk).

import nltk

Now, import the **LancasterStemmer** class to implement Lancaster Stemmer algorithm

from nltk.stem import LancasterStemmer

Next, create an instance of **LancasterStemmer** class as follows −

Lanc\_stemmer = LancasterStemmer()

Now, input the word you want to stem.

Lanc\_stemmer.stem('eats')

### Output

'eat'

### Complete implementation example

import nltk

from nltk.stem import LancatserStemmer

Lanc\_stemmer = LancasterStemmer()

Lanc\_stemmer.stem('eats')

### Output

'eat'

## 4.3.2.3 Regular Expression stemming algorithm

With the help of this stemming algorithm, we can construct our own stemmer.

### RegexpStemmer class

NLTK has **RegexpStemmer** class with the help of which we can easily implement Regular Expression Stemmer algorithms. It basically takes a single regular expression and removes any prefix or suffix that matches the expression. Let us see an example −

First, we need to import the natural language toolkit(nltk).

import nltk

Now, import the **RegexpStemmer** class to implement the Regular Expression Stemmer algorithm.

from nltk.stem import RegexpStemmer

Next, create an instance of **RegexpStemmer** class and provides the suffix or prefix you want to remove from the word as follows −

Reg\_stemmer = RegexpStemmer(‘ing’)

Now, input the word you want to stem.

Reg\_stemmer.stem('eating')

### Output

'eat'

Reg\_stemmer.stem('ingeat')

### Output

'eat'

Reg\_stemmer.stem('eats')

### Output

'eat'

### Complete implementation example

import nltk

from nltk.stem import RegexpStemmer

Reg\_stemmer = RegexpStemmer()

Reg\_stemmer.stem('ingeat')

### Output

'eat'

## 4.3.2.4 Snowball stemming algorithm

It is another very useful stemming algorithm.

### SnowballStemmer class

NLTK has **SnowballStemmer** class with the help of which we can easily implement Snowball Stemmer algorithms. It supports 15 non-English languages. In order to use this steaming class, we need to create an instance with the name of the language we are using and then call the stem() method. Let us see an example −

First, we need to import the natural language toolkit(nltk).

import nltk

Now, import the **SnowballStemmer** class to implement Snowball Stemmer algorithm

from nltk.stem import SnowballStemmer

Let us see the languages it supports −

SnowballStemmer.languages

### Output

(

'arabic',

'danish',

'dutch',

'english',

'finnish',

'french',

'german',

'hungarian',

'italian',

'norwegian',

'porter',

'portuguese',

'romanian',

'russian',

'spanish',

'swedish'

)

Next, create an instance of SnowballStemmer class with the language you want to use. Here, we are creating the stemmer for ‘French’ language.

French\_stemmer = SnowballStemmer(‘french’)

Now, call the stem() method and input the word you want to stem.

French\_stemmer.stem (‘Bonjoura’)

### Output

'bonjour'

### Complete implementation example

import nltk

from nltk.stem import SnowballStemmer

French\_stemmer = SnowballStemmer(‘french’)

French\_stemmer.stem (‘Bonjoura’)

### Output

'bonjour'

## 4.3.2.5 What is Lemmatization?

Lemmatization technique is like stemming. The output we will get after lemmatization is called ‘lemma’, which is a root word rather than root stem, the output of stemming. After lemmatization, we will be getting a valid word that means the same thing.

NLTK provides **WordNetLemmatizer** class which is a thin wrapper around the **wordnet** corpus. This class uses **morphy()** function to the **WordNet CorpusReader** class to find a lemma. Let us understand it with an example −

### Example

First, we need to import the natural language toolkit(nltk).

import nltk

Now, import the **WordNetLemmatizer** class to implement the lemmatization technique.

from nltk.stem import WordNetLemmatizer

Next, create an instance of **WordNetLemmatizer** class.

lemmatizer = WordNetLemmatizer()

Now, call the lemmatize() method and input the word of which you want to find lemma.

lemmatizer.lemmatize('eating')

### Output

'eating'

lemmatizer.lemmatize('books')

### Output

'book'

### Complete implementation example

import nltk

from nltk.stem import WordNetLemmatizer

lemmatizer = WordNetLemmatizer()

lemmatizer.lemmatize('books')

### Output

'book'

**4.4 Vader Sentiment Analysis:**

**VADER (Valence Aware Dictionary and sentiment Reasoner)** is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. **VADER** uses a combination of A sentiment lexicon is a list of lexical features (e.g., words) which are generally labeled according to their semantic orientation as either positive or negative. **VADER** not only tells about the Positivity and Negativity score but also tells us about how positive or negative a sentiment is.

# import SentimentIntensityAnalyzer class

# from vaderSentiment.vaderSentiment module.

from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer

# function to print sentiments

# of the sentence.

def sentiment\_scores(sentence):

    # Create a SentimentIntensityAnalyzer object.

    sid\_obj = SentimentIntensityAnalyzer()

    # polarity\_scores method of SentimentIntensityAnalyzer

    # oject gives a sentiment dictionary.

    # which contains pos, neg, neu, and compound scores.

    sentiment\_dict = sid\_obj.polarity\_scores(sentence)

    print("Overall sentiment dictionary is : ", sentiment\_dict)

    print("sentence was rated as ", sentiment\_dict['neg']\*100, "% Negative")

    print("sentence was rated as ", sentiment\_dict['neu']\*100, "% Neutral")

    print("sentence was rated as ", sentiment\_dict['pos']\*100, "% Positive")

    print("Sentence Overall Rated As", end = " ")

    # decide sentiment as positive, negative and neutral

    if sentiment\_dict['compound'] >= 0.05 :

        print("Positive")

    elif sentiment\_dict['compound'] <= - 0.05 :

        print("Negative")

    else :

        print("Neutral")

# Driver code

if \_\_name\_\_ == "\_\_main\_\_" :

    print("\n1st statement :")

    sentence = "Geeks For Geeks is the best portal for \

                the computer science engineering students."

    # function calling

    sentiment\_scores(sentence)

    print("\n2nd Statement :")

    sentence = "study is going on as usual"

    sentiment\_scores(sentence)

    print("\n3rd Statement :")

    sentence = "I am vey sad today."

    sentiment\_scores(sentence)

**Output :**

**Graphical user interface, text, email

Description automatically generated**

The Compound score is a metric that calculates the sum of all the lexicon ratings which have been normalized between -1(most extreme negative) and +1 (most extreme positive).

positive sentiment : (compound score >= 0.05)  
neutral sentiment : (compound score > -0.05) and (compound score < 0.05)  
negative sentiment : (compound score <= -0.05)

**4.4 Library Used for Data Extraction:**

**4.4.1 ImageTk Module**

The [**ImageTk**](https://pillow.readthedocs.io/en/stable/reference/ImageTk.html#module-PIL.ImageTk) module contains support to create and modify Tkinter BitmapImage and PhotoImage objects from PIL images.

For examples, see the demo programs in the Scripts directory.

***class*PIL.ImageTk.BitmapImage(*image=None*, *\*\*kw*)**[**[source]**](https://pillow.readthedocs.io/en/stable/_modules/PIL/ImageTk.html#BitmapImage)

A Tkinter-compatible bitmap image. This can be used everywhere Tkinter expects an image object.

The given image must have mode “1”. Pixels having value 0 are treated as transparent. Options, if any, are passed on to Tkinter. The most commonly used option is **foreground**, which is used to specify the color for the non-transparent parts. See the Tkinter documentation for information on how to specify colours.

**Parameters**

**image** – A PIL image.

**height()**[**[source]**](https://pillow.readthedocs.io/en/stable/_modules/PIL/ImageTk.html#BitmapImage.height)

Get the height of the image.

**Returns**

The height, in pixels.

**width()**[**[source]**](https://pillow.readthedocs.io/en/stable/_modules/PIL/ImageTk.html#BitmapImage.width)

Get the width of the image.

**Returns**

The width, in pixels.

***class*PIL.ImageTk.PhotoImage(*image=None*, *size=None*, *\*\*kw*)**[**[source]**](https://pillow.readthedocs.io/en/stable/_modules/PIL/ImageTk.html#PhotoImage)

A Tkinter-compatible photo image. This can be used everywhere Tkinter expects an image object. If the image is an RGBA image, pixels having alpha 0 are treated as transparent.

The constructor takes either a PIL image, or a mode and a size. Alternatively, you can use the **file** or **data** options to initialize the photo image object.

**Parameters**

* **image** – Either a PIL image, or a mode string. If a mode string is used, a size must also be given.
* **size** – If the first argument is a mode string, this defines the size of the image.
* **file** – A filename to load the image from (using **Image.open(file)**).
* **data** – An 8-bit string containing image data (as loaded from an image file).

**height()**[**[source]**](https://pillow.readthedocs.io/en/stable/_modules/PIL/ImageTk.html#PhotoImage.height)

Get the height of the image.

**Returns**

The height, in pixels.

**paste(*im*, *box=None*)**[**[source]**](https://pillow.readthedocs.io/en/stable/_modules/PIL/ImageTk.html#PhotoImage.paste)

Paste a PIL image into the photo image. Note that this can be very slow if the photo image is displayed.

**Parameters**

* **im** – A PIL image. The size must match the target region. If the mode does not match, the image is converted to the mode of the bitmap image.
* **box** – A 4-tuple defining the left, upper, right, and lower pixel coordinate. See [Coordinate System](https://pillow.readthedocs.io/en/stable/handbook/concepts.html#coordinate-system). If None is given instead of a tuple, all of the image is assumed.

**width()**[**[source]**](https://pillow.readthedocs.io/en/stable/_modules/PIL/ImageTk.html#PhotoImage.width)

Get the width of the image.

**Returns**

The width, in pixels.

**4.4.2 TextBlob:**

The approach that the TextBlob package applies to sentiment analysis differs in that it’s rule-based and therefore requires a pre-defined set of categorized words. These words can, for example, be uploaded from the NLTK database. Moreover, sentiments are defined based on semantic relations and the frequency of each word in an input sentence that allows getting a more precise output as a result.

Once the first step is accomplished and a Python model is fed by the necessary input data, a user can obtain the sentiment scores in the form of polarity and subjectivity that were discussed in the previous section. We can see how this process works in this paper by [Forum Kapadia](https://www.cs.rit.edu/usr/local/pub/GraduateProjects/2165/fjk9481/Report.pdf):

TextBlob’s output for a **polarity** task is a float within the range [-1.0, 1.0] where -1.0 is a negative polarity and 1.0 is positive. This score can also be equal to 0, which stands for a neutral evaluation of a statement as it doesn’t contain any words from the training set.

Whereas, a **subjectivity/objectivity** identification task reports a float within the range [0.0, 1.0] where 0.0 is a very objective sentence and 1.0 is very subjective.

There are various examples of Python interaction with TextBlob sentiment analyzer: starting from a model based on different [Kaggle](https://www.kaggle.com/) datasets (e.g. movie reviews) to calculating tweet sentiments through the Twitter API.

But, let’s look at a simple analyzer that we could apply to a particular sentence or a short text. We first start with importing the TextBlob library

*# Importing TextBlob*

from textblob import TextBlob

Once imported, we'll load in a sentence for analysis and instantiate a TextBlob object, as well as assigning the sentiment property to our own analysis:

*# Preparing an input sentence*

sentence = '''The platform provides universal access to the world's best education, partnering with top universities and organizations to offer courses online.'''

*# Creating a textblob object and assigning the sentiment property*

analysis = TextBlob(sentence).sentiment

print(analysis)

The sentiment property is a namedtuple of the form Sentiment(polarity, subjectivity).

Where the expected output of the analysis is:

Sentiment(polarity=0.5, subjectivity=0.26666666666666666)

Moreover, it’s also possible to go for polarity or subjectivity results separately by simply running the following:

from textblob import TextBlob

*# Preparing an input sentence*

sentence = '''The platform provides universal access to the world's best education, partnering with top universities and organizations to offer courses online.'''

analysisPol = TextBlob(sentence).polarity

analysisSub = TextBlob(sentence).subjectivity

print(analysisPol)

print(analysisSub)

Which would give us the output:

0.5

0.26666666666666666

One of the great things about TextBlob is that it allows the user to choose an algorithm for implementation of the high-level NLP tasks:

* PatternAnalyzer - a default classifier that is built on the pattern library
* NaiveBayesAnalyzer - an NLTK model trained on a movie reviews corpus

To change the default settings, we'll simply specify a NaiveBayes analyzer in the code. Let’s run sentiment analysis on tweets directly from [Twitter](https://twitter.com/):

from textblob import TextBlob

The last step in this example is switching the default model to the NLTK analyzer that returns its results as a namedtuple of the form: Sentiment(classification, p\_pos, p\_neg):

*# Applying the NaiveBayesAnalyzer*

blob\_object = TextBlob(tweet.text, analyzer=NaiveBayesAnalyzer())

*# Running sentiment analysis*

analysis = blob\_object.sentiment

print(analysis)

Finally, our Python model will get us the following sentiment evaluation:

Sentiment(classification='pos', p\_pos=0.5057908299783777, p\_neg=0.49420917002162196)

Here, it's classified it as a *positive* sentiment, with the p\_pos and p\_neg values being ~0.5 each

**4.4.3 Newspaper: Article:**

Newspaper is a Python module used for extracting and parsing newspaper articles. Newspaper use advance algorithms with web scrapping to extract all the useful text from a website. It works amazingly well on online newspapers websites. Since it use web scrapping too many request to a newspaper website may lead to blocking, so use it accordingly.

Installation:

pip install newspaper3k

Newspaper supports following languages:

**input code** **full name**

ar Arabic

da Danish

de German

el Greek

en English

it Italian

zh Chinese

......... and many more

Some Useful functions

To create an instance of article

article\_name = Article(url, language="language code according to newspaper")

To download an article

article\_name.download()

To parse an article

article\_name.parse()

To apply nlp(natural language procesing) on article

article\_name.nlp()

To extract article’s text

article\_name.text

To extract article’s title

article\_name.title

To extract article’s summary

article\_name.summary

To extract article’s keywords

article\_name.keywords

**Code-**

from newspaper import Article

#A new article from TOI

url = "http:// timesofindia.indiatimes.com/world/china/chinese-expert-warns-of-troops-entering-kashmir/articleshow/59516912.cms"

#For different language newspaper refer above table

toi\_article = Article(url, language="en") # en for English

#To download the article

toi\_article.download()

#To parse the article

toi\_article.parse()

#To perform natural language processing ie..nlp

toi\_article.nlp()

#To extract title

print("Article's Title:")

print(toi\_article.title)

print("n")

#To extract text

print("Article's Text:")

print(toi\_article.text)

print("n")

#To extract summary

print("Article's Summary:")

print(toi\_article.summary)

print("n")

#To extract keywords

print("Article's Keywords:")

print(toi\_article.keywords)

Output:

Article's Title:

India China News: Chinese expert warns of troops entering Kashmir

**4.4.4 Tkinter filedialog**

Python Tkinter (and TK) offer a set of dialogs that you can use when working with files. By using these you don’t have to design standard dialogs your self. Example dialogs include an open file dialog, a save file dialog and many others. Besides file dialogs there are other standard dialogs, but in this article we will focus on file dialogs.

File dialogs help you open, save files or directories. This is the type of dialog you get when you click file,open. This dialog comes out of the module, there’s no need to write all the code manually.

Native Load/Save Dialogs

The following classes and functions provide file dialog windows that combine a native look-and-feel with configuration options to customize behaviour. The following keyword arguments are applicable to the classes and functions listed below:

*parent* - the window to place the dialog on top of

*title* - the title of the window

*initialdir* - the directory that the dialog starts in

*initialfile* - the file selected upon opening of the dialog

*filetypes* - a sequence of (label, pattern) tuples, ‘\*’ wildcard is allowed

*defaultextension* - default extension to append to file (save dialogs)

*multiple* - when true, selection of multiple items is allowed

**Static factory functions**

The below functions when called create a modal, native look-and-feel dialog, wait for the user’s selection, then return the selected value(s) or None to the caller.

tkinter.filedialog.**askopenfile**(*mode="r"*, *\*\*options*)

tkinter.filedialog.**askopenfiles**(*mode="r"*, *\*\*options*)

The above two functions create an [Open](https://docs.python.org/3/library/dialog.html#tkinter.filedialog.Open) dialog and return the opened file object(s) in read-only mode.

tkinter.filedialog.**asksaveasfile**(*mode="w"*, *\*\*options*)

Create a [SaveAs](https://docs.python.org/3/library/dialog.html#tkinter.filedialog.SaveAs) dialog and return a file object opened in write-only mode.

tkinter.filedialog.**askopenfilename**(*\*\*options*)

tkinter.filedialog.**askopenfilenames**(*\*\*options*)

The above two functions create an [Open](https://docs.python.org/3/library/dialog.html#tkinter.filedialog.Open) dialog and return the selected filename(s) that correspond to existing file(s).

tkinter.filedialog.**asksaveasfilename**(*\*\*options*)

Create a [SaveAs](https://docs.python.org/3/library/dialog.html#tkinter.filedialog.SaveAs) dialog and return the selected filename.

tkinter.filedialog.**askdirectory**(*\*\*options*)

Prompt user to select a directory.

Additional keyword option:

*mustexist* - determines if selection must be an existing directory.

*class*tkinter.filedialog.**Open**(*master=None*, *\*\*options*)

*class*tkinter.filedialog.**SaveAs**(*master=None*, *\*\*options*)

The above two classes provide native dialog windows for saving and loading files.

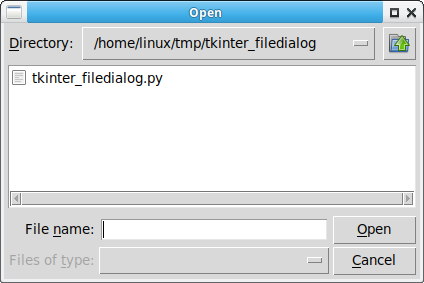
Tkinter does not have a native looking file dialog, instead it has the customer tk style. You can see these below.

The file dialog will work on all desktop platforms.

The tkinter filedialog comes in several types. Which type you need really depends on your applications needs. All of them are methods calls.

You can open a single file, a directory, save as file and much more. Each dialog made with the example below is a different type of dialog.

|  |
| --- |
| import tkinter.filedialog   tkinter.filedialog.asksaveasfilename() tkinter.filedialog.asksaveasfile() tkinter.filedialog.askopenfilename() tkinter.filedialog.askopenfile() tkinter.filedialog.askdirectory() tkinter.filedialog.askopenfilenames() tkinter.filedialog.askopenfiles() |



You can create an open file dialog which asks for a filename, and then returns the name of the selected dialog.

|  |
| --- |
| import tkinter as tk from tkinter import filedialog as fd   def callback():  name= fd.askopenfilename()   print(name)   errmsg = 'Error!' tk.Button(text='Click to Open File',   command=callback).pack(fill=tk.X) tk.mainloop() |

The appearance of the dialog is different on every operating system. It will look different on Windows, Mac and Linux (gnome).  
Other file dialogs work similar to the example shown above.

**CHAPTER 5:**

**EXPERIMENTAL RESULTS**

**CHAPTER 6:**

**CONCLUSION AND FUTURE WORK**

In this thesis, we present a new technique for analyzing online sentiment sentiments. A proposed technique targets performing statistical and numerical analysis on online sentiments for scientific papers domain. It is a hybrid model (the enhancement Bag-of-words (BOW) model combing with Part-of-Speech POS model) for English sentiments. The proposed technique can improve accuracy and support understanding implicit and explicit meaning. The evaluation of these papers based on online researcher's sentiments and parameters and features of the scientific domain. Then we measure the newly proposed technique efficiency by making a comparison among it and two techniques based on the accuracy and performance.

##### **6.1 Conclusion**

Sentiment analysis becomes the most important source in decision making. Almost people depend on it to achieve the efficient product. Although, there are hundreds of thousands of researcher, who write and read online papers daily, the research in this field finds not enough till now. Because analyzing scientific papers domain is hard. It has special features and characteristics effects on the sentiment polarity evaluation.

In this thesis, we introduced the new technique for analyzing scientific papers domain based on sentiment analysis. This technique aims at supporting researchers in selecting the useful papers for their research. This technique includes two evaluation parts on the scientific paper: sentiment score and system score. First: sentiment score is based on the online reviews evaluation. Second: system score is a new criteria for essential topic parameters evaluation. This technique is called sentiment analysis of online papers “SAOOP”. It improves accuracy and understanding of the online sentiment reviews.

The approach of evaluating the sentiments consists of the enhancement of Bag- of-words model and solutions for the essential sentiment challenges in this domain with respect to sentence level sentiment analysis and review level. The enhancement Bag-of-words model solves the two major weaknesses of the standard one: low

accuracy and manual evaluation approach. It is an automated model for evaluating sentiments and depends on each word weight replacing term frequency of each word. It also classifies a sentiment strength into five sentiment polarity classification levels. It also introduces solutions for the most significance sentiment challenges for improving accuracy. These challenges are negative, bi-polar reviews, extracting and evaluate topic features or keywords, world knowledge, and create a huge lexicon. The system score evaluates the most significance parameters in the scientific research domain. These parameters are place of publication, publishing date, and a number of citations for paper. SAOOP creates a new proposed miniature lexicon for sentiment evaluation.

For evaluating the proposed technique SAOOP efficiency, we make a comparison between it and two famous techniques. The results have a comparison between the accuracy and performance between the three techniques when the researchers apply the techniques on three data sets (training, test and verified). The comparison results illustrate how the proposed technique can increase accuracy and performance with facing many language coverage cases and solving some sentiment analysis challenges. The accuracy results show in NLTK (62%), NLPS (72%) and SAOOP (83%) approximately. The technique with the highest F-measure was faced sentiment analysis challenges is SAOOP.

##### **Future work:**

The future works that can be related to the thesis are mention as follows:

* + 1. We have some problems in handling phrases and we try to improve it in future work.
    2. We think that more works need to be done with many online websites.
    3. Improving proposed technique to work on multi-languages in the scientific domain.
    4. Analyzing online sentiments on the researchgate website.
    5. Proposing an additional evaluation criteria for sentiment properties.

**References**

[1] Louis, F.& Wojciech, C., Book "Advances in The Human Side of Service Engineering", 5th International Conference on Applied Human Factors and Ergonomics, Volume Set, Proceedings of the 5th AHEE Conference 19-23 July 2014.

[2] Larsen, Peder Olesen, and Markus von Ins. “The Rate of Growth in Scientific Publication and the Decline in Coverage Provided by Science Citation Index.”,Scientometrics 84.3 (2010): 575–603. PMC. Web. 25 Sept. 2015.

[3] Peng, L., Cui, G., Zhuang, M., and Li, C., “What do seller manipulations of online product reviews mean to consumers? ” (HKIBS Working Paper Series 070-1314). Hong Kong: Hong Kong Institute of Business Studies, Lingnan University, 2014.

[4] Thomas, B., Keep Social Honest , “What Consumers Think about brands on social media, and what businesses need to do about it” Report, 2013.

[5] Liu, B., Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, 2012.

[6] Yin, Z., Rong, J., and Zhi-Hua, Z., “Understanding Bag-of-Words Model: A Statistical Framework”, International Journal of Machine Learning and Cybernetics, 2010.

[7] Nisha, J., & Dr.E. Kirubakaran, “M-Learning sentiment analysis with Data Mining Techniques”, International Journal of Computer Science and Telecommunications, Volume 3, Issue 8, 2012.

[8] Richard, S., Alex, P., Jean, Y.W., Jason, C., Christopher, D.M., Andrew, Y.N. and Christopher, P., “Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank”, Conference on Empirical Methods in Natural Language Processing, 2013.

[9] Farah, B., Carmine, C., & Diego.R., “Sentiment analysis: Adjectives and Adverbs are better than Adjectives Alone”, [International Conference on Weblogs and Social](http://libra.msra.cn/Conference/2773/icwsm-international-conference-on-weblogs-and-social-media) [Media](http://libra.msra.cn/Conference/2773/icwsm-international-conference-on-weblogs-and-social-media) -ICWSM , Boulder, CO USA, 2007.

[10] Saifee, V., & Jay, T., “Applications and Challenges for Sentiment Analysis: A Survey”, International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue2, 2013.

[11] Hassena, R.P., “Challenges and Applications”, International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 3, Issue 5, 2014.

[12] Apoorv, A., Boyi, X., Ilia, V., Owen, R., and Rebecca, P.,"sentiment Analysis of Twitter Data", proceeding LSM '11 Proceedings of the Workshop on Languages in Social Media, Pages 30-38, USA, 2011.

[13] Anuj, S. & Shubhamoy, D.,"Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis",Special Issue of International Journal of Computer Applications (0975 – 8887) on Advanced Computing and Communication Technologies for HPC Applications -

ACCTHPCA, June 2012

[14] Andranik, T., Timm, O., Philip, G.S., and Isabel, M.W., "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment", Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment,

Association for the advacement of artifical intelligence 2010

[15] Mitchell, P.M., Mary A.M., and Beatrice, S., “Building a Large Annotated Corpus of English: The Penn Treebank”, Computational Linguistics Journal , Vol. 19, Number 21993.

[16] Schukla, A., “Sentiment analysis of document based on annotation”, CORR Journal, Vol. abs/1111.1648, 2011.

[17] Kasper, W. & Vela, M., “Sentiment analysis for hotel reviews”, proceedings of the computational linguistics-applications, Jacharanka Conference, 2011.

[18] Zhang, L., Hua, K., Wang, H., and Qian, G.,“Sentiments reviews for mobile devices products”, The 11th International Conference on Mobile Systems and Pervasive Computing (MobiSPC-2014) ,procedia computer scinence, Volume 34, 2014.

[19] Godbole, N., Srinivasaiah, M., and Skiena, S., “Large-Scale Sentiment Analysis for News and Blogs”, ICWSM’2007 Boulder, Colorado, USA, 2007.

[20] Esuli A., & Sebastiani, F.,“SentiWordNet: A High-Coverage Lexical Resource for Opinion Mining”, Kluwer Academic Publishers. Printed in the Netherlands, 2006.

[21] Hearst, M., “Direction-based text interpretation as an information access refinement”, In Paul Jacobs, editor, Text-Based Intelligent Systems. Lawrence Erlbaum Associates, 1992.

[22] Das, S., and Chen, M., “Yahoo! for Amazon: Extracting market sentiment from stock message boards”, In Proc. of the 8th Asia Pacific Finance Association Annual Conference (APFA 2001), 2001.

[23] Turney, P., “Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews”. In Proc. of the ACL, 2002.

[24] Argamon-Engelson, S., Koppel, M., and Avneri, G., “Style-based text categorization: What newspaper am I reading? ”, In Proc. of the AAAI Workshop on Text Categorization, pages 1–4, 1998.

[25] Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A., Shaked, T., Soderland, S., Weld, D. and Yates, A., “Unsupervised named-entity extraction from the web: An experimental study”, Artificial Intelligence, 165(1):91–134, 2005.

[26] Pang, B. & Lee, L., “A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts”, Association of Computational Linguistics (ACL), 2004.

[27] Jin, W., & HO H. H., “A novel lexicalized HMM-based learning framework for web opinion mining”, Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Quebec, Canada, ACM: 465-472, 2009.

[28] Brody, S., & Elhadad, N., “An unsupervised aspect-sentiment model for online reviews”, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, California, Association for Computational Linguistics: 804-812, 2010.

[29] Wiebe, J., Wilson, T., and Cardie, C., “Annotating expressions of opinions and emotions in language”. Language Resources and Evaluation, 39(2-3):165–210, 2005.

[30] Cui, A., Zhang, M., Liu, Y., and Ma, S., “Emotion tokens: Bridging the gap among multilingual twitter sentiment analysis”, In Information Retrieval Technology, pages 238–249. Springer, 2011.

[31] Kumar, A., & Sebastian, T. M., “Sentiment Analysis On Twitter”. IJCSI International Journal of Computer Science, Issues 9(3):372–378, 2012.

[32] Nielsen, F. A. “ANEW: Evaluation of a word list for sentiment analysis in microblogs”. In Proceedings of the ESWC2011 Workshop on ‘Making Sense of Microposts’: Big things come in small packages, volume 718 of CEUR Workshop Proceedings, pages 93–98, 2011.

[33] Minocha, A. & Singh, N., “Generating domain specific sentiment lexicons using the web directory”. Advanced Computing: an International Journal, 3, 2012.

[34] Kamps, M., Marx, R.J. Mokken, and Rijke, M.D., “Using wordnet to measure semantic orientation of adjectives”, procedding in National Institute for (2004) 1115– 1118, 2004.

[35] Kim, S., & Hovy, E., “Identifying and analyzing judgment opinions”. In: Proceedings of HLT/NAACL-2006. (2006) 200–207.

[36] Rao, D., & Ravichandran, D., “Semi-supervised polarity lexicon induction”. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. EACL ’09, Stroudsburg, PA, USA, Association for Computational Linguistics (2009) 675–682.

[37] Carmen Banea, R.M., & Wiebe, J., “A bootstrapping method for building subjectivity lexicons for languages with scarce resources”, In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Tapias, D., eds.: Proceedings of the Sixth International Language Resources and Evaluation (LREC’08), Marrakech, Morocco, European Language Resources Association (ELRA) , May 2008.

[38] Radulescu, C., Dinsoreanu, M., and Potolea, R., “Identification of spam comments using natural language processing techniques”, [Intelligent Computer](http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6926641) [Communication and Processing (ICCP), 2014 IEEE International Conference on,](http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6926641) Sepbtemebr 2014.

[39] Zhang, Q., Wu, Y., Li, T., Ogihara, M., Johnson, J., and Huang, X., “Mining Product Reviews Based on Shallow Dependency Parsing”, SIGIR’09, July 19–23, Boston, Massachusetts, USA. ACM 978-1-60558-483-6/09/07, 2009.

[40] Emus*,* R., “Modeling and Representing Negation in Data-driven Machine Learning-based Sentiment Analysis”, In Proceedings of the 1st International

Workshop on Emotion and Sentiment in Social and Expressive Media (ESSEM)*,* 2013.

[41] Adsod, A. A., & Chopde, N. R., “A review on web mining”, International Journal of Engineering Trends and Technology(IJETT), Volume 10 Number 3, 2014.

[42] Balahur, A. & Turchi, M., “Multilingual Sentiment Analysis using Machine Translation? ”, In Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis. Korea, 2012.

[43] Banea, C., Mihalcea, R., and Wiebe, J., “Multilingual Sentiment and Subjectivity Analysis”, In Multilingual Natural Language Processing”, editors Imed Zitouni and Dan Bikel, Prentice Hall, 2011.

[44] Baccianella, S., Esuli, A. and Sebastiani, F., “SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining”, Proceedings of the Seventh conference on International Language Resources and Evaluation, 2010, pp. 2200-2204).

[45] Jeh, G., & Widom, J., “Mining the Space of Graph Properties”, Proceedings of the Tenth {ACM} {SIGKDD} International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August, 2004.

[46] Liu, B., “Sentiment Analysis and Subjectivity”. In Nitin, I. & Fred, J. (eds). Handbook of Natural Language Processing. 2nd Ed, Machine Learning & pattern recognition series, Chapman& Hall/CRC, 2010.

[47] Bo, P., & Lillian, L., “Opinion mining and sentiment analysis”, Journal Foundations and Trends in Information Retrieval, Vol. 2, 2008.

[48] Wilson, T., Wiebe, J., and Hoffmann, P., “Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis”. In Computational Linguistics 35 (3), p. 1-34, 2009. (95 in chapter 2)

[49] Halleh, K., Comparing Feature Sets and Classifiers for Sentiment Analysis of Opinionated Free Text, MSc. Thesis, July 2012, page 13.

[50] Bird, S., Klein, E., and Loper, E., Natural Language Processing with Python, 1st Edition, O’Reilly Media publisher, 2009.

[51] Choudhari, and Rajankar, Introduction to Natural Language Processing With Python, International Journal of Modern Trends in Engineering and Research (IJMTER)Volume 02, Issue 04, [April – 2015].

[52] Bespalov, D., Bai, B., Qi, Y., and Shokoufandeh, A., “Sentiment classification based on supervised latent n-gram analysis”, In Proceeding of the ACM conference on Information and knowledge management (CIKM- 2011), 2011.

[53] Partha, C., and Amit, N., Software Innovations in Clinical Drug Development and Safety book, page 56.

[54] Ian, H.W., Eibe, F., and Mark A. H., Data Mining: Practical machine learning tools and techniques. The Morgan Kaufmann series in data management systems, 3rd Edition, Morgan Kaufmann Publishers is an imprint of Elsevier, 2011.

[55] Liu, B., “Sentiment Analysis: A Multi-Faceted Problem”, IEEE Intelligent Systems, 2010.

[56] Yumusak, S., Dogdu, E., and Kodaz, H., “Tagging Accuracy Analysis on Part-of- Speech Taggers”, Journal of Computer and Communications, 2014.

[57] Presutti, V., Stankovic, M., Cambria, E., Cantador, I., Di Iorio, A., Di Noia, T., Lange, C., Reforgiato Recupero, D., Tordai, A., Semantic web evaluation challenge. 1st Edition, Series title: Communications in Computer and Information Science, Springer International Publishing, 2014.

[58] Andreas, H., Gerhard, P., and Andreas, N., “A Brief Survey of Text Mining”, LDV Forum - GLDV Journal for Computational Linguistics and Language Technology, 2005.

[59] Dushyant, B.R., & Samrat, K., “A Review on Emerging Trends of Web Mining and IT’s Application”, International Journal of Engineering Development and Research | IJEDR, 2013.

[60] Tamilsevi, A. & M.Parveen, T., “Sentiment analysis of micro blogs using opinion mining classification algorithm”, International Journal of Science and Research (IJSR), Vol. 2 Issue 10, 2013.

[61] Chenghua, L., & Yulan, H., “Joint Sentiment/Topic Model for Sentiment Analysis”, [CIKM '09](http://www.ischool.drexel.edu/cikm2009/about/index.asp) Proceedings of the 18th ACM conference on Information and knowledge management, Hong Kong, China, ACM 978-1-60558-512-3/09/11, 2009.

[62] Hang, C., Vibhu, M., & Mayur, D., “Comparative experiments on sentiment classification for online product reviews”, American Association for Artificial Intelligence (AAAI Conference), 2006.

[63] Bas, H., Paul, v.I., Alexander, H., Flavius, F., and Uzay, K., “Accounting for Negation in sentiment analysis”, DIR, Amsterdam, the Netherlands, 2011.

[64] Namita, M., Basant, A., Garvit, C., Nitin, B., & Prateek, P., “Sentiment analysis of Hindi Review based on Negation and Discourse Relation”, International Joint Conference on Natural Language Processing, Nagoya, Japan, 2013.

[65] Mark, C., Oliver, D., and Faith, U., “Potential and Limitations of Commercial Sentiment Detection Tools”, proceedings of the first international workshop on emotion and sentiment in social and expressive mead: approaches and perspective from AI, ESSEM, 2013.

[66] Segran, T., Programming Collective Intelligence, Sebastopol: O’Reilly Media publisher, Inc, August 2007.

[67] Diana, M., & John.C.,“Disambiguating Nouns, verbs, and adjectives using automatically acquired Selectional preferences”, Association for Computational Linguistics , Vol. 29, 2003.

[68] Hiroshi, K., Tetsuya, N., and Hideo, W., “Deeper Sentiment Analysis Using Machine Translation Technology”. In Proceedings of the 20th International Conference on Computational Linguistics, 2004.

[69] Alberto, H.F.L., Berthier, A.R-N, Altigran, S. d. S., and Juliana, S. T., “A brief survey of web data extraction tools”, ACM SIGMOD, Vol.31, Issue 2,, New York

, USA, June 2002.

[70] Seyed, M.M, Mustafa, E., Dincturk, Salman, H., Gregor, V. B., Guy-Vincent, J., and Iosif, V.O., “A brief history of web crawlers", CASCON'13 proceedings of the 2013 Conference of the center for advanced studies on collaborative research, pages 40-54, USA, 2013.

[71] Tamlin M. P., “Recent Progress in Development of SWOT river Discharge Algorithms”, Department of Geological Sciences, University of North Carolina and participants in the 2012 SWOT Discharge Algorithms Workshop, June 18-20, Chapel Hill, NC, 2012.

[72] Goranczewski, B., & Puciato, D., “SWOT Analysis in the formaulation of Tourism Development Strategies for Destinations”, Tourism 2010, 20/2, 2011.

[73] Gelman, A. & Hill, J., “Data analysis using regression and multilevel/ hierarchical models”, Cambridge university 2007.

[74] Yang, L., & Boley, H., “A weighted-Tree Simplicity Algorithm for SIMILARITY matching of partial product Descriptions”, in Proc. of ISCA 14 th International Conference on Intelligent and Adaptive Systems and Software Engineering”, Toronto, 2005.

[75] Quoc, L., & Tomas, M., “Distributed Representations of Sentences and Documents”, Proceedings of the 31 st International Conference on Machine Learning, Beijing, China. JMLR: W&CP, volume 32, 201