Analyzing Scientific Papers Based on Sentiment Analysis (First Draft)

Thesis · January 2016
DOI: 10.13140/RG.2.1.2222.6328

1 author:

Doaa Mohey El-Din
Cairo University

Information System Department
Faculty of Computers and Information
Cairo University

Analyzing Scientific Papers Based on Sentiment Analysis

By

Doaa Mohey El-Din Mohamed Hussein

A Thesis Submitted to the

Faculty of Computers and Information

Cairo University

In Partial Fulfillment of the

Requirements for the Degree of

Master of Science

In

INFORMATION SYSTEMS

Under the supervision of

Dr. Hoda M.O. Mokhtar
Information System Department
Faculty of Computers and Information
Cairo University

FACULTY OF COMPUTERS AND INFORMATION
CAIRO UNIVERSITY- EGYPT
February 2016

Declaration

I certify that this work has not been accepted in substance for any academic degree and is not being concurrently submitted in candidature for any other degree. Any portions of this thesis for which I am indebted to other sources are mentioned and explicit references are given.

Student Name: Doaa Mohey El-Din Mohamed Hussein

Signature:

Table of Contents

List of Tables

List of Figures

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)

Acknowledgments

First of all, I thank God for giving the support and providing me with the patience and guidance which I used in making this work.

I would like to express great gratitude to my supervisor **Assoc. Prof. Dr. Hoda M.O. Mokhtar** for providing her valuable time to express his professional thoughts. I would like to thank her for her help, continuous encouragement, fertile discussion and valuable suggestions and comments throughout the research and the thesis work. I am also grateful to **Dr. Osama Ismael** for spending time to support me and providing useful suggestions about this thesis. Thank you for pushing me to be a better researcher.

Last but not the least important, I owe more than thanks to my family members especially, my mother who is the queen of our life and thanks in everything that I have reached. Very thanks for my lovely sister for her support, love, and encouragement throughout my life. Many thanks for my father who supports me a lot in my life. I would like to thank my lovely brother for his support. Thanks a lot for my husband family for their encouragements especially, my second mother for her support and pray.

Special thanks are given to my husband for his care and his advices, but most of all for his patience. Last thanks to a new member in my family who is my beloved and beautiful baby.

Abstract

Sentiment analysis or opinion mining is used to automate the detection of subjective information such as opinions, attitudes, emotions, and feelings. Many researchers spend long time searching for suitable papers for their research. Online reviews on papers are the essential source to help them. Thus, online reviews can save the researcher's time. It provides effort and paper cost. In this thesis, propose a new technique to analyze online reviews in the scientific research domain called: "Sentiment Analysis Of Online Papers" (SAOOP). SAOOP aims at supporting researchers and saving their time and efforts by enabling them to report the total evaluation for the papers. SAOOP includes main two evaluations for each research paper: Sentiment score and System score. Sentiment score which is an evaluation for the paper based on analyzing online sentiment reviews. System score is an evaluation for the paper based on topic domain 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 improves accuracy and solves several sentiment evaluation challenges. The proposed technique uses bag-of-words model for fitting the review structure which is short and relevant to a scientific topic domain. 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. Although SAOOP works on the sentiment analysis word level, it can deal with order of words and grammar in review sentences. 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 by using similarity and difference algorithms. SAOOP also classifies reviews based on the topic essential features and keywords. Each class of reviews has some features or keywords. This classification has a big effect on the meaning of sentiment polarity and sentiment score of the words. For example; title, field name, name of authors or their abbreviation and name of journals or conferences. 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 spam and fake detection, 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.

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 help and support researchers to retrieve the suitable papers for their research in a short time while saving their effort.

Finally, we evaluate the efficiency of the proposed technique by comparing it with two sentiment analysis techniques. The comparison terms are based on measuring accuracy, performance and the percentage understanding rate for parts of speech. The comparison depends on two different data sets: real data set which is split into two data sets with a training set and test set, and a verified data set which is a real set.

CHAPTER 1: INTRODUCTION

Chapter 1
Introduction

World Wide Web has become the most popular communication platforms to the public reviews, opinions, comments and sentiments. These sentiments and opinions about products, places, books or reseasch papers and to 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 [1]. 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 [2].Notably, a large fragment of WWW researchers make their content public, allowing researchers, societies, universities, and corporations to use and analyze data. According to a new survey conducted by dimensional research, April 2013: 90% of customer's decisions depends on online reviews [3]. According to 2013 Study* [4]: 79% of customer's confidence is based on online personal recommendation 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 mach their research direction.

1.1 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 [5] depends on two issues sentiment polarity and sentiment score. Sentiment polarity [5] is a binary value either positive or negative. On the other hand, sentiment score which relies on one of three models. Those models are Bag-of-words model (BOW) [6], part of speech (POS) [7], and semantic relationships [8]. BOW [6] model is the most popular for researchers and based on the representation of terms/ words, but it neglects

* http://searchengineland.com/

language grammar and words ordering. POS [7] tagging which is a grammatically tagging model especially verbs, adjectives and adverbs [9]. For example; (The book is not good.) declaring in (The/DT book/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 a semantic relationship method is the most complex method, which is based on the relationship between concepts or meanings for example; antonym, synonym, homonym etc.

Sentiment analysis (also known as opinion mining) 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 a text or online reviews for one topic based on sentiment classification levels, such as positive or negative. Existing analysis approaches to sentiment reviews can be grouped into four main categories: word level, sentence level, document level, and aspect/ entity level.

1.2 Problem Definition

Although there are hundreds of thousands of researcher, who write and read online papers daily. Analyzing scientific papers domain is hard, because there are several features that affect evaluating sentiment reviews. We face the difficulty of evaluation sentiments shows in that selective the suitable sentiment technique to understand part of speech of the language. The accurate meaning is a very important source in making a decision. Especially in research domain to save a long time in searching for the suitable paper and save efforts for researchers. Sentiment evaluation requires a huge lexicon for dealing with the sentiments and their polarities. Sentiment evaluation has several challenges including spam or fake reviews as well as the duplication of reviews. The implicit and explicit 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. All challenges have a bad effect on the understanding of reviews and the sentiment evaluation. There are a research gap between the sentiment challenges and sentiment evaluation.

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:

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

1.1 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. So 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 [10, 11] to improve the accuracy. It classifies the topic features or keywords into five classes (Topic, author, publishing date, citation number 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).

1.2 The System Score evaluation

The system evaluation score which cares of three parameters in the evaluation of the system to evaluate domain parameters namely place of publication whether journal or conference, citation number of the research paper and the publishing date of the research paper are the systems parameters that they have a methodology.

2. Comparative Evaluation:

We measure the efficacy evaluation through making a comparison between the proposed techniques with two sentiment analysis techniques. The comparison terms are based on the following three issues:

2.1 **Accuracy:** It is defined in terms of the systematic and random errors. Accuracy is also used as a statistical measure of how well a binary classification test correctly identifies or excludes a condition.

2.2 **Performance (F-Measure):** It is one of algorithmic efficiency to analyze and measure resources usage.

2.3 **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.

The comparison includes two different data sets: real data set and verified data set which is a real set around10.000 text reviews [12].

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 the enhancement model and how to evaluate sentiment and system score.

- **Chapter 5: Experimental Results:** presents the 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

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:

2.1 Related Work in Sentiment Analysis

The paper [13] 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. The researchers in [14] 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. It is most weakness illustrate in not handling some cases in multi-topic segments.

Mobile devices products reviews were analyzed in [15]. A Machine learning which ML system [7] investigated the classification accuracy of Naïve Bayes algorithm. In addition, the research made a judgment of the product quality and status in the market is advantageous. 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. The paper [16] analyzes 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, the research splits related work in two other classification [17]: 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/word 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. Essentially a text review is split into single sentences ("sentence-based") and words ("words-based") or very short texts from a single source.

The previous research on sentiment-based categorization of the input documents has implicated either the using models inspired mostly by cognitive linguistics [18] or the manual or semi-manual construction of discriminant-word lexicons [19]. For instance, in [20] 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%). Unfortunately, their proposed method entails building a discriminant-word lexicon using manual selection and tagging of words from several thousand messages.

The Corpus-based approaches inspect the incorporation with seed words based on large groups of text [20] or search for the context-dependent labels by considering the local constraints [21]. Alternatively, people have searched into investigating knowledge encoded in WordNet 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. The prior work in this trend through [22] inspected the impacts of adjective orientation and grad-ability on sentence subjectivity. The target told us about a given sentence is subjective or not judging from the adjectives appearing in that sentence.

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. Pang and Lee in [23] use a subjectivity detector to remove objective sentences from a given document. Then, using minimum cuts formulation, they integrate 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. They used a semantic relationship model for analyzing and evaluating online sentiments. Their technique provides multi-topic domain. But the sentiment requires wider supervised training and evaluation resources.

The Research on opinion mining on YouTube performed [24] for discussing how social media can be utilized to radicalize 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. Their work examines an approach is indeed fruitful. 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 (automatically) labeled data collected from the online websites, the researchers approached the related task of detecting a sentiment polarity in reviews via supervised learning approaches. Interestingly, our baseline experiments on this task, show that humans may not always have the best intuition for choosing discriminating words. While they did experiment with a set of different features in the previous research [25], 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/documents and evaluate sentiment scores.

2.2 What is Lexicon-Driven Methodology?

The paper utilized the MPQA [26] 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 correlation between the aggregate sentiment in tweets and the Gallups opinion surveys.

SentiStrength [27] is constructed as a lexicon-based algorithm which specifies a sentiment polarity (positive/negative) and identical strength value between 1 and 5 to a given text. Moreover the paper introduced\ list includes 298 positive and 465 negative terms annotated with polarity and strength values. SentiStrength is using the emoticons list, negations list and boosting words in making a decision operation. In order to deal with emphatic lengthening the authors present a three-step method for reducing words of the standard form. When making a comparison between various machine learning classifiers on MySpace comments by SentiStrength. The authors find that their proposed method executes better for classifying negative sentiment, but not suitable for positive sentiment.

The paper [28] proposed rule-based method for entity-level sentiment analysis in Twitter. They evaluated a sentiment score for each entity depending on its textual proximity to words from a sentiment lexicon. It also executed simple anaphora resolution by resolving pronouns to the closest entity in the tweet. The rule-based algorithm differentiates between demonstrative, imperative and interrogative sentences and can, among other things, handle comparative sentences, negation and but-clauses. For enhancing the recall of the proposed approaches, the researchers recognize extra tweets that are likely to be opinionated and train a support vector machine (SVM) to appropriate polarity labels to the contained entities.

The authors in [29] discuss the particular challenges of sentiment analysis in the domain of social media messages. Its purpose is a rule-based method with constructing a shallow linguistic analysis containing named entity extraction and event recognition. It works for producing a sentiment polarity and score for a given tweet. The research [30] also examined the approach of building a sentiment lexicon from Twitter data and detect value in creating domain-specific sentiment lexica. The authors developed the SentiWordNet [31] to include four Part-of-Speech (POS) tags namely adjectives, adverbs, verbs and nouns having 2 million words out of which 3% are adjectives. The sentiment polarity classification classifies each word in one from three scores positive, negative and objective such that sum of the score for each word sums to one.

The research [32] presented an approach for recognizing and analyzing sentiment and opinions. The process has four steps. First step was recognizing the opinion and understanding it. For defining the opinion they introduced an algorithm to classify a word as positive, negative or objective which was based on WordNet. It made a proposition that was to gather synonyms of a word with the same sentiment polarity as the source word. In order to avoid words with multiple meaning (dual nature) they applied an approach to recognize closeness of a word to each class (positive, negative, objective). For their presented approach for giving a high recall the initial seed list should be large enough and with wide variety of words.

The study [33] introduced a comprehensive study on the challenge of detecting sentiment polarity of words. It is considered bi-polar sentiment classification of words i.e. a word can be either positive or negative. They made semi-supervised label propagation in a graph for sentiment polarity detection of words. Each of these words exemplify node in the graph whose polarity is to be determined. They focused on three languages fundamentally English, French and Hindi but demand that their work can be extended to any other language for which WordNet is available. Much work has been done towards developing the subjective lexicon for English languages.

The bootstrapping approach is presented in [34] 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.

2.3 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 effects solutions for improving accuracy for text. We can minimize the 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. In the following, we discuss related work with sentiment challenges with respect the type of review and topic domain [10]:

1) **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 [11]. 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 sentiment\ts 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. There are a large number of duplicate reviews and many of them are clearly spam. For example, different user-ids posted duplicate or near duplicate reviews on the same product or different products. Duplicate detection is done using the shingle method with similarity score > 0.9.
- **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). With constructing a machine learning model to classify each review, i.e., to assign a probability likelihood of each review being a spam [35].

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 [36]. 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. Further some researchers used the sentiment information model to support online spam detection though applied on two Twitter datasets for examining the differences between spam and normal sentiments. This information modeled based on a graph Laplacian and incorporated into an optimization formulation. The experimental results demonstrate the effectiveness of the proposed framework as well as the roles of different types of information.

2) **Implicit and Explicit Negation:** Negation is the biggest challenge 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 movie", 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", 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.

The researchers' approach [37] uses for the two negative types modeling and representing negation in word based on n-gram feature space for data-driven Machine Learning based sentiment analysis. The explicit negation benefits from (i) high quality NSD methods like LingScope and (ii) modeling not only negation of word unigrams, but also of higher order word n-grams, especially word bigrams. Their approach is also easily extensible to other word n-gram weighting schemes aside from encoding pure presence or absence, e.g. weighting using relative frequencies or tf-idf. There is a paper presented that a system for identifying the scope of negation though utilizing the shallow parsing, which means of a conditional random field model informed by a dependency parser. The scope of negation detection is limited to explicit rather than implied negations within single sentences. A new negation corpus is presented that was constructed for the domain of English product reviews obtained from the open web, and the proposed negation extraction system is evaluated against the reviews corpus as well as the standard BioScope negation corpus.

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.

3) Domain-independence: This is the biggest challenge faced by sentiment analysis and opinion mining. There is a difference effect of the topic domain and multi-topic domain models in evaluation of sentiment analysis. The domain dependent challenges requires recognizing the nature of domain with its features and words. One features set may give very good performance in one domain, at the same time it perform very poor in some other domain.

- **Topic domain:** The researchers in [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 employs the local dependency among sentiments by their model. The research analyzed the sentiment of user messages (or posts) of an online cancer support community, the Cancer Survivors' Network (CSN) of the American Cancer Society. We identify whether a post is positive or negative based on the polarity of the emotion expressed in it and model the task of sentiment analysis as a binary classification problem. 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 the classifier with the help of one or even multiple domains that we call the source domains. Here it gets tough to identify which features from the source domains are related to which features in the target domains. Second challenge here is to apply the trained classifier on a domain that we haven't included in 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.

The proposed model for the researcher here [39] is that a probabilistic generative model of a word with a domain label, domain dependence/independence and a word polarity, and this can also judge the document polarity which can treat the differences between domains. This model can extract words with domain-dependent polarity, making it possible to create domain-dependent word polarity dictionaries for each domain. Parameter values can be learned with Gibbs sampling. They can increase data from source domains lead to an improved performance rate in target domains. Other experimental results also found that as the number of target domains increased, F-value decreased. The researchers proposed a system for analyzing online hotel reviews [14] to help in decision making and managing the quality controls. It analyzes German reviews. It has multi-topic/multi-polarity; the system would recognize the neutral e.g., "don't know" to "classify sentiment polarity that as neutral" and the multi-topic cases identified in their corpus. It is most weakness illustrate in not handling some cases in multi-topic segments. **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 conveys an objective or subjective assessment. Several solutions are possible, for instance training classifiers on translated corpora, using translated lexicons, building lexicons or corpora for targeted language. Recent experiments with automatic translation for sentiment analysis show that the performance of machine translation does not degrade the results too much. We refer the interested reader to [40] for a good overview of the topic.

4) **World knowledge:** Often world knowledge needs to be incorporated in the system for detecting sentiments. Consider the following examples: "He is a Frankenstein". Just finished Doctor Zhivago for the first time and all I can say is Russia sucks. The first sentence depicts a negative sentiment whereas the second one depicts a positive sentiment. But one has to know about Frankenstein and Doctor Zhivago to find out the sentiment. The main task in this approach is the construction of word lexicons that indicate positive class or negative class. The sentiment values of the words in the lexicon are determined prior to the sentiment analysis work. Lexicons can be created in different ways. It can be created by starting with some seed words and then using some linguistic heuristics to add more words to them, or starting with some seed words and adding to these seed words other words based on frequency in a text. SentiWordNet 3.0 is a publicly available lexical resource explicitly devised for supporting sentiment classification and opinion mining applications [41].

5) **Construct Sentiment Lexicon**

In addition, the construction of huge lexicons becomes an important obstacle in sentiment analysis. It is based on extracting features and keywords of topic or multi-topic domain. A text or sentence may have multiple entities. It is extremely important to find out the entity towards which the opinion is directed. Consider the following examples. Samsung is better than Nokia, Ram defeated Hani in football. The examples are positive for Samsung and Ram respectively but negative for Nokia and Hani.

- **Extracting features and keywords:** The proposed technique is based on extracting keywords and relevant features of each topic. In addition, to produce a solution for some words have many meanings and different sentiment values relevant to the topic. The proposed technique is based on Classification review of each domain features and keywords

- **Grouping synonyms:** Many times text contains different words having same meaning. So such word should be identified and group together for accurate classification. It is a difficult task to identify these words, as people often use different words to describe the same feature. For example, “voice” and “sound” both refer to the same feature in phone review.

We face also a problem but it appears with several challenges that’s called **bipolar sentiments:** There are some words and marks in text have bipolar values especially in various domain with various features and keywords. The problem faced the sentiments how to achieve the real and accurate meaning. The researchers presented a new approach for identifying the frequent bigrams where a word switches polarity, and to discover which words are bipolar to the extent that it is better to have them removed from the polarity lexica. Their results scores match human perception of polarity and bring enhancement of the classification results by their improved context-aware method. Their introduced approach improves the assessment of lexicon relying on the sentiment detection algorithms and can be further used to quantify ambiguous words.

2.5 Related work in web Scrapping/extracting

The data mining web term [42] 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. The web data mining technology is opening horizons on not just collecting data, but it is also raising a lot of concerns related to data security. 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. 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:

FMiner* which is a software for web scraping, data extraction, screen scraping, web harvesting, web crawling and web macro support for windows and Mac OS X. It integrates best-in-class features with an intuitive visual project design tool, in order to make your next data mining project a breeze. Simply choose the output format and record the steps on FMiner according to the target of extracting from a web site. FMiner's powerful visual design tool captures every step and models a process map that interacts with the target site pages to capture the information.

Import.io** grants for the users a desktop application to support them in scrapping all required data with the unlimited amount of web pages. The service cures each page as a potential data source to generate API from. Import.io will be a guide you through the process of creating the scraping matrix by constructing connectors (for navigation) or extractors (to pull out the needed data). The extracting processing takes around 24 hours to get the results. The user's data is private and they can schedule auto refreshments at any chosen period of time.

Easy Web extract Tool*** is visual screen scraper for extracting data for business purposes. This text extractor rips desired web content (text, URL, image, html) from WebPages with minimum effort. Easy Web Extract is using Web2Mine to be easy and fast data extraction. Easy Web Extract is perfect for exporting and scrapping text data into Excel (CSV), text, XML file, HTML formats, MS Access DB, SQL Script File, MySQL Script File, and HTTP submit form and ODBC Data source. One weakness is taking a long time into a project scrapping and extracting the URL. We use it in our proposed work to ease data extracting and web scrapping. In order to use it in our proposed technique and to evaluate sentiment score and system score of scientific papers. We discuss how to use it in chapter 4.

---

* Fminer web scrapping tool, http://www.fminer.com/

** Import.io web scrapping tool, http://enterprise.import.io/how-it-works/

*** Easy web extract tool, http://webextract.net/

2.6 Chapter Summary

This chapter surveys previous work in structured sentiment analysis. A common theme in much of this work is based on two approaches: the syntactic approach and the lexical approach. In general terms, the syntactic approach uses one of sentiment analysis models to evaluate the sentiments. The lexical approach, on the other hand, is an approach to evaluate the sentiment polarity with sentiment classification levels. We also discuss the related works in sentiment analysis and the challenges effects on evaluating the sentiments. Finally, we show how we can get the online sentiment with web crawling and scraping data online.

CHAPTER 3: BACKGROUND OF SENTIMENT ANALYSIS

Chapter 3

Background of 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. In the following sections, we will discuss sentiment analysis area and the factors affecting them.

3.2 What is Sentiment Analysis?

We present the sentiment and sentiment analysis definitions and the differences between them.

3.2.1 What are Sentiments?

Sentiments can be recognized as emotions, or as judgments, opinions or ideas prompted or colored by emotions or susceptibility or feelings [43]. 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 [11].

3.2.2 What is Sentiment Analysis?

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 [11]. 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 [8]. In sentiment analysis, studying the subjective language: language used to rapid a private situation in the context of a text or conversation. The research [44] 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").

3.2.3 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 [45]. 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 [43]. 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 (NLP), text analysis (TA) and 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 from marketing to customer service.

3.3 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 [1].

3.4 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:

- 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.
- 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.

- 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.

3.5 Sentiment Analysis (SA) & Natural Langauge processing (NLP):

We extract some important definitions for sentiment analysis, in the following:

3.5.1 Definitions

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

It is an area of the computer science, artificial intelligence, and computational linguistics interested in the interactions between computers and human (natural) languages. Intrinsically, NLP is related to the field of human–computer interaction

(HCI). Many challenges in NLP include natural language understanding, that is, enabling computers to deduce meaning from human or natural language input, and others involve natural language generation [46].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.

We need to discuss some very basic terminology to be useful in understanding the NLP models and techniques.

- **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.

For segmented languages such as English, the existence of whitespace makes tokenization relatively easier and uninteresting. However, for languages such as Chinese and Arabic, the task is more difficult since there are no explicit boundaries. Furthermore, almost all characters in such non-segmented languages can exist as one-character words by themselves but can also join together to form multi-character words.

- **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. After we have provided the essential terminology,

3.5.2 Some mutual NLP tasks:

- **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.
- **Machine Translation (MT)**: According to machine translation, the goal is to have the computer translate the given text in one natural language to fluent text in another language without any human in the loop. This is one of the most difficult tasks in NLP and has been tackled in a lot of different ways over the years. Almost all MT approaches use POS tagging and parsing as preliminary steps.
- **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” [47]. 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

3.5.3 Classifier:

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

3.5.4 Text Mining

One of the important definitions is **Text mining**, also referred to as text data mining, almost synonymous to text analytics, refers to the process of deriving high-quality information from 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), 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, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities).

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 [48]. **Text analysis** involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics. The comprehensive goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP) [46] 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 field concerned with the statistical or rule-based modeling of 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 and cognitive science, and applied computational linguistics focuses on the practical outcome of modeling human language use.

3.6 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.

3.6.1 Bag of Words

This model [6], which is the most used in supervised text classification, a document is represented as an unordered collection of words disregarding grammar and even word order. According to this model, during the training phase, a dictionary is built by dependent on training data and is then used to characterize between the positive and negative documents in the testing procedure.

For example if we have the two documents below:

1) Insurance is a very significant issue.
2) Insurance should be considered strictly.

The dictionary which is constructed based on BoW will be:

Dictionary= ,1:” Insurance”, 2:”is”,3:”a”, 4:”very”, 5:” significant”, 6:”issue”, 7:”should”, 8:”be”, 9:”considered”, 10:” strictly”-

Hence, the feature vector of each document has “10 dimensionalities” based on the constructed dictionary. As demonstrate in sentiment analysis discussion, word appearance is very informative (in contrast with word frequency in information retrieval). According to the nature of natural language, one word can express the authors’ attitude clearly while a sequence of words cannot.

For example, in the sentence below, only the words “like” and “not” show the polarity of the sentence while the whole sentence seems to have positive polarity.

3.6.2 Lexicon

In this research, instead of constructing a dictionary based on documents, we have used the WordNet dictionary [49]. WordNet is a large lexical database of English with 58058 words and 4 part of speech tags. Nouns, verbs, adjectives and adverbs are grouped into set of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. There are 3 chief reasons for selecting WordNet. First of all, the dictionary which was made based on “Aggregated Dataset” (aggregation of 6 small datasets) was too big, approximately 117660 words, and the constructed matrix based on it was practically unusable by MATLAB and PRTools. Since the verbs have different tenses in documents and nouns have different forms of singular and plural, the size of the dictionary has increased uselessly while in the WordNet dictionary words are in their lemmatized form and therefore, MATLAB and PRTools could take the advantage of constructed matrix. The research recommend that it pressed upon that the different tenses of verbs or singular or plural forms of nouns are not important in sentiment analysis while the appearance of them is informative. That is why it is a safe assumption to use lemmatized forms of words.

Second, since WordNet is a huge dictionary, it is more probable to have all the words as a source in comparison with the "Aggregated Dataset". Although in chapter 1, it was mentioned that corpus-based dictionary will be useful in opinion classification, due to the small size of the available dataset in term of the number of documents with societal theme, it would not be too far off if we say that WordNet dictionary will be of much more significant help. Finally, WordNet is a popular and available dictionary. The experiments conducted based on it, are corpus independent and therefore, the results can be generalized for other datasets as well.

For example; "I am going school", "I be go school", "There are 2 boys", and "There be 2 boy"

All the words are converted to their lower case form in terms of their letters, to be compatible with the WordNet words. Also all the numbers, punctuation marks and other elements except words with 2 or more letters are removed from dataset since they were not that much informative in comparison with other words. Therefore, terms such as 19, a, @ and words like that are removed from the datasets. For instance;

"For 4 years to come, corruption isn't the extreme", "for year to come corruption be not the extreme". After applying the pruning procedure on the dataset, documents are ready to be presented by the BoW model. Each dataset is described in a 2 dimensional matrix of documents and WordNet dictionary. The dimensionality of each dataset relies on BoW model.

Each document is checked based on the words in the WordNet lexicon and existence of each word in each document is shown by 1, otherwise it is 0. Hence, a row in this matrix will be a feature vector corresponding to a document. The resulting matrix has several zero columns which show that the head word of the column has not been used in any documents in the dataset at all. Thus, they are not informative and can be removed from the matrix.

Bag of Words is a supervised learning approach as the truth sense of word. This model trains a classifier by each word in each document. Consequently, it will give similar results for similar documents. For instance; the two sentences below are close to each other in terms of Euclidean distance from the Bow point of view and receive the same labels in testing procedure.

1) I do like football

2) I do like banana.

While the two sentences below have totally opposite polarity, but likewise their feature vectors are close to each other according to BOW model and mistakenly receive the same labels which is surely wrong.

1) I do like football

2) I do not like football

This sort of BOW mistakes convinced us to look for other feature models that are more accurate and take into account the details which BOW misses.

3.6.3 Part of Speech Tags

Part-of-speech tagging (POS tagging or POST), [50] also known grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context—i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph. Tagging the grammatical features of each word is also a very good strategy to improve accuracy ratios and detect useful patterns for classifications. For example, authors of subjective texts usually describe themselves in first person while verbs in objective texts are usually in the third person. 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.

3.6.4 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 [51] 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 [53].

3.7 The sentiment analysis problems

In this section, we will discuss the problems faced sentiment analysis shortly [54]. 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.

3.7.1 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 [55]:

3.7.2 Text analysis:

There are some problems faced in sentiment analysis. There is a different techniques in the text analysis as word by word, phrases, part of speech (based on grammatical tagging) or based on sentences. Based on the text analysis algorithms.

3.7.3 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) [56].

- **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.

3.7.4 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 Research and Technical challenges, as in the following:

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.

<u>The research type challenges includes six challenges [11, 54]:</u>

1. **Spam and Fake Detection:** Spam and fake challenge examine the duplicate or empty reviews and to reach real number of reviews. Another challenge with spam detection is the evaluation of some words or phrases about non-specific topic. Further, there is a review has sentiments or feelings but is not to for the domain.

• **For example:** *Review 1:"The movie is good" & Review2: "the horror movies 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.

2. **Implicit and Explicit Negation:** Negation is the biggest challenge in sentiment analysis [57]. This problem divides into two types: explicitly and implicitly negative [10, 58].

- **For example:** *"I do [not like^{+}] $^{-}$ this movie"*
- **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 movie [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 exam"*. In the last example we observe need to also refers to the negative polarity in the review.

3. **Domain-independence:** One of the essential problem faced by opinion mining. There is a difference effect of the topic domain and multi-topic domain models in evaluation of sentiment analysis. We need to recognize the nature of domain with its features and words.

- **For example:** *Review 1:* [*This is bad*] *& Review2: [This is a bad 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.

4. **Natural language processing overheads:** The natural language overhead like ambiguity, co-reference, Implicitness, inference, emotion detection, etc. created hindrance in sentiment analysis too.

A definition of co-reference and **anaphora resolution**: sentences like "I want that one!" and "Queen Elizabeth II is the Queen regnant of sixteen independent sovereign states, she is politically neutral and by convention her role is largely ceremonial" do not imply what "that", "Queen regnant", "she" or "her" refer to, which makes it even more difficult to define which emotion is expressed and who or what is the target.

Emotions and opinions can be expressed explicitly and implicitly. **Implicitness** is a challenge for computational systems, since even for humans it is not easy to identify and analyze these expressions correctly. This also applies for humor, sarcasm, irony, etc. Another problem is **inference**, the process of drawing conclusions by applying certain clues (logic, statistics, etc.) to observations or hypotheses. Inference has been a popular field of research, and applications such as expert systems and business rule engines have followed.

5. Entity identification & extracting features and keywords

A text or sentence may have multiple entities. It is extremely important to find out the entity towards which the opinion is directed. Consider the following examples. Samsung is better than Nokia Ram defeated Hani in football. The examples are positive for Samsung and Ram respectively but negative for Nokia and Hani.

- **For example:** The mobile sound is good.
- **Observation:** the sound is one feature from the mobile domain.

6. World knowledge & Pragmatics

Often world knowledge needs to be incorporated in the system for detecting sentiments. Consider the following example:

- **For example:** [*The author of this paper look like Einstein*]
- **Observation:** The world knowledge is important challenge as in the previous review. [*Einstein*] in the name of scientist this refers to positive polarity but it is very hard to understand with computer algorithms.

<u>The technical type challenges includes four challenges:</u>

7. **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.

8. **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.

9. **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.

10. **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.

3.8 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 [59]. 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 [60]. 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 [39] 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 insights.

3.8.1 Sentiment lexicons:

There are some sentiment vocabularies and annotated word lists:

• **SenticNet:** is as a framework, SenticNet consists of a set of tools and techniques for sentiment analysis combining common-sense reasoning, psychology, linguistics, and machine learning [38]. In this context, SenticNet is more commonly referred to as sentic computing, a multi-disciplinary paradigm that goes beyond mere statistical approaches to sentiment analysis by focusing on a semantic-preserving representation of natural language concepts and on sentence structure.

• **SentiWordNet:** a lexical resource for opinion mining. SentiWordNet [39] assigns to each synset of WordNet three sentiment scores: positivity, negativity, objectivity.

• **WordNet-Affect:** is an extension of WordNet Domains, including a subset of synsets suitable to represent effective concepts correlated with affective words. Similarly to our method for domain labels, we assigned to a number of WordNet synsets one or more effective labels (a-labels) [31]. In particular, the affective concepts representing emotional state are individuated by synsets marked with the a-label emotion. There are also other a-labels for those concepts representing moods, situations eliciting emotions, or emotional responses.

3.8.2 Sentiment Tools & Applications:

Three different techniques will be assessed and compared:

- Training a classifier using samples from the dataset
- Classify using generic lexicons
- A pre-trained state-of-the-art system

There are many sentiment tools that can support in the text analysis and evaluate sentiments, they use different algorithms with the same target. Table 3.1 shows a sample of these tools online.

Table 3.1: Sentiment analysis Tools and applications

Tool	Short name	URL
AlchemyAPI	alc	www.alchemyapi.com
Lymbix	lym	www.lymbix.com
ML Analyzer	mla	www.mashape.com/mlanalyzer/ml-analyzer
Repustate	rep	www.repustate.com

Semantria	sma	www.semantria.com
Sentigem	sen	www.sentigem.com
Skyttle	sky	www.skyttle.com
Textalytics	tex	core.textalytics.com
Text-processing	txp	www.text-processing.com

There is a comparison between the pervious tools in [63]. For evaluating the quality of 9 state-of-the-art commercial sentiment detection tools for approx. as mentioned in Table 3.1 on the dataset around 30,000 different short texts (tweets, news headlines, reviews etc.). The best tools have an accuracy of 75% for some document types (tweets), but the average accuracy over all documents is at best 60%. Surprisingly, the accuracy decreases if texts get longer, which is due to the decline in the ability to detect "other" sentiments. The comparison observation that existing sentiment corpora are prone to error, with error rates up to 15% per corpus [57].

3.9 Sentiment Analysis Research

3.9.1 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 [58]. 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.

3.9.2 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:

- **Document-level sentiment classification:** The task at this level is to classify whether a whole opinion document expresses a positive or negative sentiment [61]. 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.

- **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 [62], 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 [63] 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.

- **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.

3.10 Chapter Summary:

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 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 Online Papers (SAOOP)

Chapter 4

Proposed Technique:

Sentiment Analysis Of Online Papers (SAOOP)

Sentiment analysis interprets the subjective information from online reviews to implicit assessment based on score. The scientific domain has hundreds of thousands researchers who care about scientific research. These research takes a long time to select suitable papers for their research. Online reviews on papers are the essential source to help them. The reviews save reading time and papers cost. In this chapter, we present a new technique to analyze online reviews which is called: "Sentiment Analysis Of Online Papers" (SAOOP).

4.1 Proposed Technique (SAOOP) Overview

SAOOP is used in opinion mining [53] and evaluating sentiment analysis for online scientific reviews. This technique is based on analyzing online sentiments in word level. The goal of SAOOP is evaluating online scientific paper and relies on sentiment analysis on reviews and topic parameters is called: "system score". Our technique presents an enhancement Bag-Of-Words (BOW) model [49] to improve accuracy. This enhancement depends on a word weight instead of the term frequency of each word. The standard bag-of-words (BOW) model [4] is used for analyzing sentiments text such as document, opinion or review. The BOW model evaluates sentiments on word/term level. The enhancement also provides solutions for the two essential weaknesses [59] of Bag-of-words model. The major weaknesses of BOW model are that 1) It classifies sentiment polarities manually (not automatic) with creating manual lists of 'positive' and 'negative' words, 2). It has low accuracy, because it neglects text grammatically and ordering, and Thus BOW requires a huge lexicon with duplicate and repeated words.

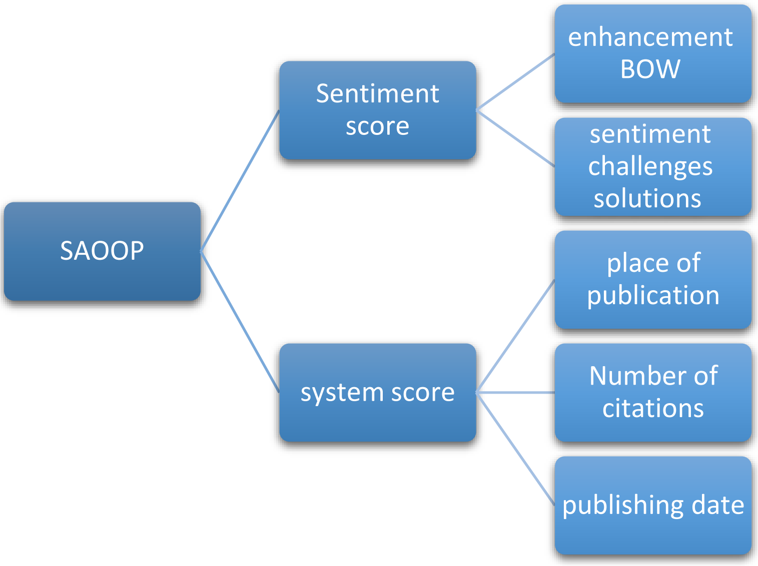

Figure 4.1: The proposed Technique (SAOOP) Evaluation Components

Figure 4.1 shows that SAOOP technique consists of two components sentiment score and system score with respect to topic parameter classification for evaluating online papers. Sentiment score depends on total reviews evaluation score. And system score which depends on the sum of total scores of three important parameters of paper (place of publication, number of citations paper and paper publishing date).

The proposed technique SAOOP evaluates a system score that relies on topic domain parameters. These parameters (place of publication, number of citations and publishing paper date) highly affect the assessment of paper content. In this research domain, there are three parameters have a significant impact in evaluation any paper. These parameters are (place of publication, publishing date and citation number of paper). Place of publication for the paper which refers the international published journal or conference. This place is responsible for evaluation the paper content and review not to repeat content with plagiarism to keep the copyrights. Each place of publication has impact factor and ranking tier which we depend on it in our proposed technique evaluation. Publishing date of the paper refers to the year which the paper published in. This date is an essential factor to know the hot areas of research. In addition, to show the number of reviews related to the date that can help us in evaluation. Finally, Citation number which is the number quotes of each paper. Further, we identify the relationship between the citation number and the date of the published paper. The hardest issue faces us in the evaluation topic parameters is that how to evaluate the logical meaning of parameters with attention to the relationship between them.

Place of publication is the place that can be found on the title page. The evaluation of the place of publication is essential in evaluation process because there are a highly ranked. The evaluation of conferences or journals is international evaluation and reflects the value of each paper. The citation is very important in the paper evaluation because it declares how many people use it in their researches and refers to it. Publishing date has two sides in the evaluation process. First: how to achieve the best paper in time, because if the paper is old it will be got some reviews but if it is recent, it is very hard to find the reviews on it. Second: there is a relationship between publishing date and the number of reviews. **For example;** if there are two papers in 2012, but the first one had 30 reviews and the second had just 7 reviews, it's meaning that the first had more effects even if the polarity of them is negative. So it's very difficult to evaluate the date of publishing score with maintaining the balance of meaning and effects of the date.

Further, review classification is substantial in evaluating sentiment analysis in our proposed technique. Review classification is based on features and keywords of the papers domain. We assume a five classes for reviews (Topic, Authors, publishing date, publishing place, and number of citation). Each class of reviews has some features and keywords and has a different effect on the evaluation score. **For example;** Review.1: "**The paper has old author in this field**" & Review.2: "**The paper is old**" in first and second reviews: the word [Old], and with blind lexical evaluation, they have the same evaluation [Negative]. But with our classification of reviews the first review refers to "Author" class and [Old] word has the [Positive] polarity because it refers the experience of the author in this field. But in the second review [Old] word will be [Negative] polarity because it refers to the "Publishing Date" Class of the paper. So we care of the classification of reviews and topic features in evaluation to achieve the nearest meaning of the reviews.

Researcher Guide which is a very helpful guide for beginner researchers, it introduces the hierarchy tree of the entry research point, and report the latest papers in this point and the highest rated papers based on SAOOP evaluation. And all papers related of this point. This guide depends on ACM library tree*. For example; the selected input research point is "Data mining" in Figure 4.2.

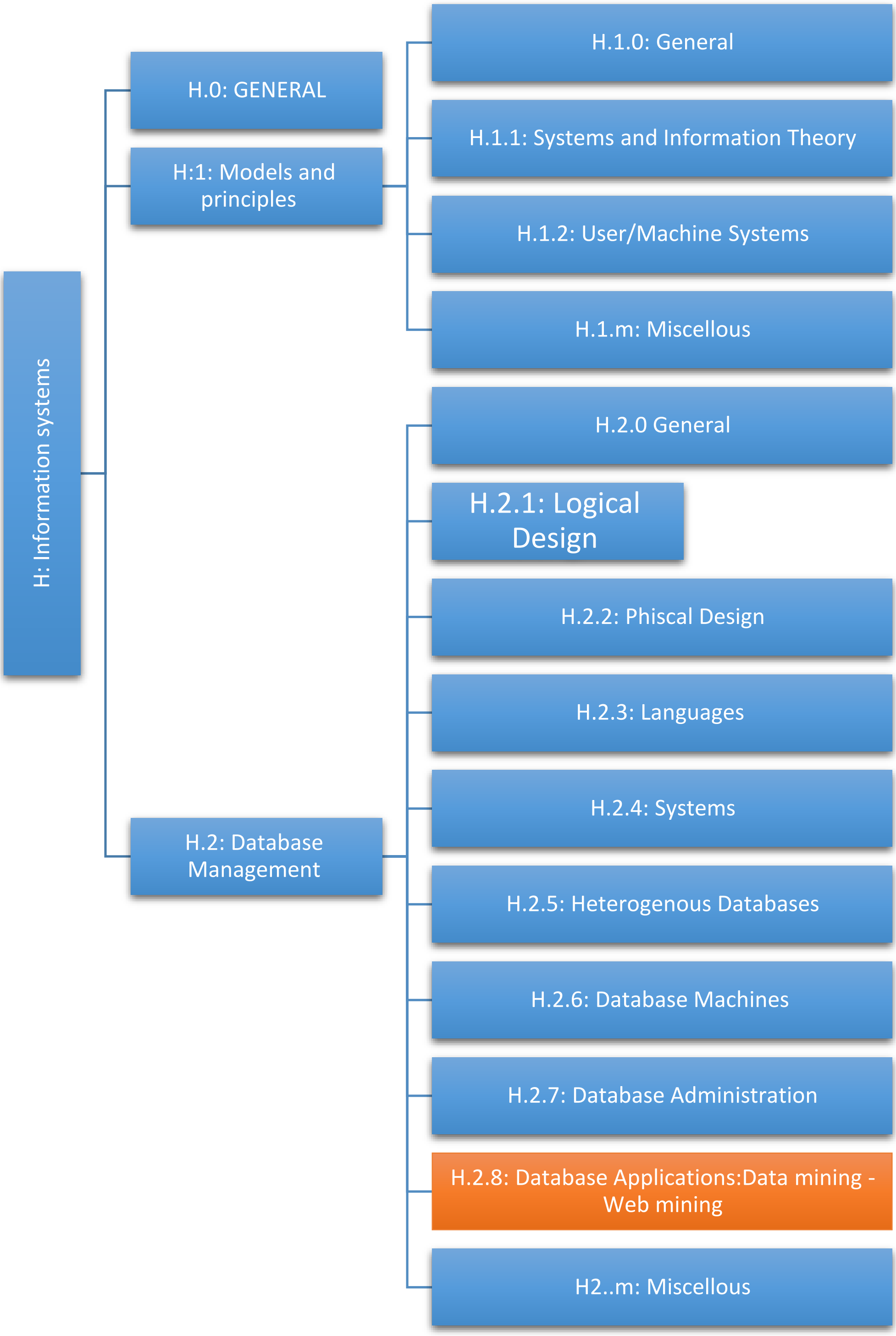

Figure 4.2: Researcher Guide based on ACM library

* http://www.acm.org/about/class/2012

4.2 Proposed Technique (SAOOP) Methodology

The methodology of the proposed technique summarizes in analyzing online scientific papers using the aggregation evaluation between the domain parameters evaluation which called: "system score" and the sentiment score of reviews which called: "sentiment score". Each analysis level has the specific approach to analyze and evaluation it and there are some challenges faces this evaluation. In the following, we discuss that how to evaluate each of them and the proposed SAOOP Algorithm:

4.2.1 Evaluate System Score

The introduced equation of the evaluation explain the priority of each parameter. The place of publication (PP) is the most important parameter. Then the citation (C) and publishing date (D) have the same importance. Because there is an inverse relationship between them, which declares in the next equation,

$$S(C) = \frac{current\ year - publish\ date\ year}{number\ of\ Citation} \qquad (1).$$

It's not true the highest citation number is new paper. Because whenever the paper is old, it becomes more vulnerable to Quote and the citation number increases. So the newly paper which has the big number of citation is better than the older one which has the same citation number. We can evaluate the total score of system score of domain parameters using the following equation,

$$V(SS) = \sum\left(\frac{S(pp)}{\lambda} + \frac{S(C)}{2\lambda} + \frac{S(D)}{2\lambda}\right) \qquad (2).$$

V (SS) refers to the total score of paper domain parameters. The score of publication place as known S (PP), the score of paper citation number as known S(C), and the score of paper publish date as known S (D). We assume λ is a constant equal 2, we divide into λ and 2λ to assign the priority of parameter in system evaluation. In the following chart, explaining of the flowchart of evaluation of the system score.

For example; one paper publishing in 2013 that's mean from 2 years and it has ten citations, not equal evaluation one paper publishing in 2005, that's mean from 10 years and it has ten citations. The first one is the highest score because the number of citations in shortly is high, we can predicate if the paper has the same time 10 years, it mostly has 50 reviews not 10 reviews such as the second paper. In other words, the first paper has 5 papers into each year but the second has 1 into each year. To evaluate score of publishing place conference we depend on ACM conferences tiers with our sample into computer science conferences, such as “VLDB: Very Large Data Bases is in the top tier: tier 1”, “ER: Intl Conf. on Conceptual Modeling (Conf. on the entity Relationship Approach)” is in next tier which is in lower tier: Tier 2, and “IDEAS: Intl Database Engineering and Application Symposium” is in a lower tier: Tier 3”.

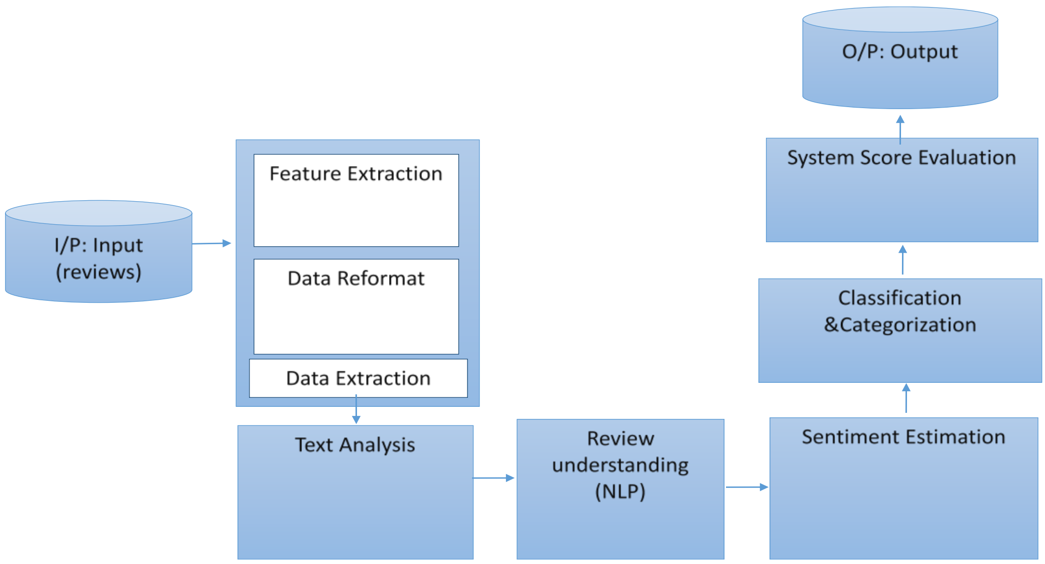

Figure 4.3: SAOOP Technique Overview

Figure 4.3 presents the proposed technique SAOOP overview. The sentiment analysis has six stages to get the evaluation score of each paper. These stages illustrate in web crawler [64], text analysis, natural language processing (NLP), Sentiment Estimation, classification and categorization stages, and evaluating system score. Classification and categorization stage: categorization refers the reviews classifications for one or more class from the five classes we assumed before (Topic, author, publishing date, citation number and place of publication). And the classification refers to sentiment classes level from [0, to 1] into five classes also [very negative, negative, neutral, positive, very positive]. And system score depends on evaluating a topic domain parameters. The output is the sentiment evaluation score of all reviews with all papers with caring of the number of reviews parameter. So our consequent result is ranking to each paper with the total score of sentiment and system scores with graphical reports of results.

4.2.2 Evaluate Sentiment Score

The proposed SAOOP assigns a sentiment polarity based on this approach, we considered the words weight replacing term frequency, by assuming each word has two values and polarity with our assumption equation,

$$V(w) = \sum(W(p) + W(n)) = 1 \qquad (3).$$

V (w) is the word value, W (p) refers to a positive value and W (n) refers to a negative word, the selection between positive or negative polarity Influenced by the meaning of words and each other polarity. But the sentence contains negative that differs in the word value. If the word is positive, convert to negative polarity and the negative score will be as in the equation,

$$V\ (w)\ = W(p) - 0.2. \qquad (4).$$

And if the word is negative, the score will be calculated by V (w) =$\mathrm{W(n)} + 0.2$. We select 0.2 because it’s suitable to our five sentiment class’s levels. Our proposed technique also creates papers ranks with calculating sentiment and measuring domain parameters. There is a difficult problem between large numbers of reviews and evaluating sentiment polarity of each one, it’s improper the most review number having assessment higher score.

For example; one paper publishing in 2013 this means from 2 years and it has twenty reviews, not equal evaluation one paper publishing in 2005, that’s mean from 10 years and it has twenty reviews. The first one is the top rated because the evaluation number of reviews in short time. In another example, one paper publishing from 2 years and it has twenty negative reviews, not equal evaluation other one publishing from 10 years and it has positive twenty reviews. The second one has maximum rated because the evaluation numbers of positive reviews are larger than the one, although the second is the oldest. So we mentioned before double trouble with reviews number and the relationship between date and other relation between sentiment polarity of reviews and number of reviews. That interprets difficulty of evaluation domain parameters.

4.2.3 SAOOP Algorithm

The goal of the proposed technique SAOOP is analyzing and evaluating the scientific paper online. This evaluation is based on the total score of sentiment analysis of online reviews and system score of topic domain parameters, By assuming,

$$P(TS) = \sum_{R=0}^{n} (T(SA) + T(SS)) \quad (5).$$

In equation.5, P (TS) refers to a total score of each paper, T (SA): is a total score of sentiment score of all reviews on each paper with caring of the number of positive reviews. In the next equation,

$$T(SA) = \sum_{i=1}^{n} \frac{P(SA(R))}{n}. \quad (6).$$

We calculate the total score of all reviews divides into the number of reviews. With respect deleting the spam or fake reviews to achieve the real number of reviews r and make n = r.

4.3 Enhancement Bag-of-words (BOW) Model:

By a study of analyzing the online scientific reviews, we identify the nature of review structure which is short and formal. So the BOW is the most suitable model to deal with these reviews but we need to improve the accuracy and avoid its weaknesses.

4.3.1 Enhancement Technique Pseudo code

Hence, we present an enhancement BOW model in the proposed technique (SAOOP) which is explained steps in the following: The input is online sentiment reviews, and the output is the sentiment scores of each word. Algorithms discusses the pseudo code of the enhancement.

Table 4.1: The Pseudo code of the Enhancement BOW model

Algorithm : Pseudo code of Enhancement BOW
Begin {
1. For each paper P do
2. Web scrapping data
3. Reformat data
4. Calculate number of sentiment S
Begin {
5. For each review R in P
6. Check and delete fake or spam sentiments reviews.
7. Get the real number R or reviews.
8. For sentence sent ∈ classification reviews data do.
9. For review category a ∈ A do, class='Topic', and

Score= 0.

10. For word w ∈ s do.
11. If O (w) > 0 then.
12. Remove stop and punctuation lists.
13. Convert all w into UPPER case.
14. Create a new lexicon for all positive and negative words.
15. Assume each word w has two values (positive & negative), and The total score of w equal 1,

$$\sum Wp + Wn = 1$$

Assuming each word has 2 values (W p=positive value, W n=negative value)

Begin {

16. If having explicit negative words (such as not)?
17. Check on the next word w to detect score.
18. We assume the negative value for positive word,
    And assume the positive value for the negative value for the negative word,

$$V\,(W) = W\,(N) - 0.2.$$

19. Each w has class from five sentiment classification levels (very negative, negative, neutral, positive, and very positive).

$$V\,(W) = W\,(N) + 0.2.$$

20. Detect sentiment score and polarity.
21. End If.

}

22. Else If, (having a word from second negative list such as never) {
23. Convert the polarity of the sentence by,

$$V\,(Sent) = S\,(Sent) * -1.$$

} End

24. End If.
25. Else If, (check on future words such as (wish, hope)) {
26. Check on next word and detect polarity and score (go to step17)}
27. Else If.
28. Use POS tagging to check on nouns.

Begin {

29. If w is noun?

30. If w is feature?
31. Detect sentiment score and polarity
32. End if
33. Else If, (w is keyword?).
34. Detect sentiment score and polarity.
35. Else If, (w is world knowledge).
36. Detect sentiment score and polarity.
37. Else if go to step 17.
38. End for.

    } End

39. Assign review classification of each S in R.

    Begin {

40. If s € review classes
41. Determine class.
42. End if
43. Else If class = ‘topic’
44. Compute sent score and polarity.

    } End

45. Calculate R (SA) is a total sentiment score of each review r.
46. End for
47. End for
48. Calculate T (SA) is a total sentiment score of each paper p. Calculate AVG (SA) is an average sentiment score of each p.

    T (SA): is the real total score of all reviews.

    r: is the number of real reviews without spam or fake reviews.

49. End For

    } End.

4.3.2 The Comparison

This section discusses a comparison between the standard bag of words and our enhancement bag of words. In order to show the effects and changes in enhanced technique and how it will be more effected in understanding reviews.

Table 4.2: Comparison between Standard BOW and the enhancement BOW

The Standard BOW Model	**The Enhancement BOW Model**
Manual positive and negative by hand and train a classifier model	Automated giving a sentiment polarity
Based on Term Frequency	Based on Word Weight
Each word has one sentiment score and one sentiment polarity	Each word has two values (positive and negative) and the total equal (=1) , and one sentiment polarity
Required a huge lexicon with duplication, repetition with putting all words available in lexicon.	Applying similarity and difference algorithm to create a miniature lexicon
If there are negative in the sentence just change polarity for each word and that not correct e.g positive_score = +, and negative_score = -positive_score. e.g great if it has value= 0.8 , not great has -0.8	More accurate evaluation by assuming five sentiment classification, and if negative effect before the word, e.g great = 0.8 will be not great= good = 0.6
Neglect grammar	Although it is based on word level, it can handle and understand grammar
focuses completely on the words as a blind	Cares words, expressions and several parts of the sentence and deal with grammar and ordering
Ignore ordering of words	Deal with ordering for words which illustrated in negative or expressions.
Have some limitations with low accuracy results and take a big time	Just can't deal with all phrases but it improves accuracy results and less time
Can't deal with bi-polar words or ambiguous words such as: "low" is positive in "low price" but negative in "low quality"	Can evaluate the bi-polar words based on the review classification and evaluate ambiguous words and knowledge words based on the hierarchal database model
The assessment does not reveal which topic features or characteristics of the product led to the positive or negative	Make a review classification based on topic keywords and characteristics

opinion, although this may be more crucial to the seller than the overall assessment	
If the review compares the item to other items, the bag-of-words approach is unable to distinguish references to the different items	It can compare between reviews and evaluate each one separately

Although we discuss a comparison in Table 4.2, we find the SWOT [65] analysis comparison is very useful to show strengths, weaknesses, opportunities and threats. SWOT analysis [66] became one of the most popular tools for strategic planning or making a decision. It can help in improving our models. Strengths are those features of the business which allow you to operate more effectively than your competitors. Weaknesses are areas capable of improvement. Opportunities identify any new opportunities for techniques. Threats can be external or internal, and are anything which can adversely affect the techniques. With applying SWOT analysis on the compared two algorithms, the results presents in Table 4.3 that discuss the weaknesses of the bow and how can handle them in the strengths of the proposed enhanced BOW model in a new SAOOP technique.

Table 4.3: Comparison between standard BOW and Enhancement Bow in SWOT Analysis

Algorithm **SWOT**	**The Standard BOW**	**The Enhancement BOW in SAOOP**
Strengths	- Ease to use - Using for small reviews - Topic domain - Deal with images, text, and documents	-Improve Bag of words and combine with POS tagging algorithm -categorize reviews - extract features - identify objects and evaluate it Applied KNN- Naïve base classifiers to measure accuracy. Graphic reports -Handle some sentiment analysis challenges Easy Clarity High accuracy

		Topic domain or any domain based on dictionary Memory ability Scale classification -1,0,1
Weaknesses	- Less accuracy - Manually dictionary - neglect grammar - neglect ordering - Don't deal with Numbers Questions - Fake or spam review -World knowledge User mention -Hash tags -Emotions	- Not fast enough -Don't deal with -Numbers -Questions
Opportunities	Automate algorithm High accuracy	-More fast -Arabic sentiment analysis for scientific papers -Create some lexicons to suitable with some domains
Threats	Binary words World knowledge Numbers Questions &User mention Hash tags & Emotions	Numbers (10/10, or 100%) Questions Words not splitting Emotions Deal with hash tags, user mentions and emotions.

4.4 Proposed Solutions for Sentiment Analysis Challenges

There are several sentiment analysis challenges as mentioned in chapter.3. These challenges rely on the review structure and topic domain. The reviews of a scientific papers domain are short and have some effective features. With scanning a real dataset of scientific reviews as on CiteuLike website*, we recognize the greatest impact challenges on these reviews are (Implicit and explicit Negative, Topic domain extracting features and keywords, generate a huge lexicon, World knowledge, and spam and fake reviews). Our proposed technique (SAOOP) produces solutions for the most significance sentiment analysis challenges to improve accuracy.

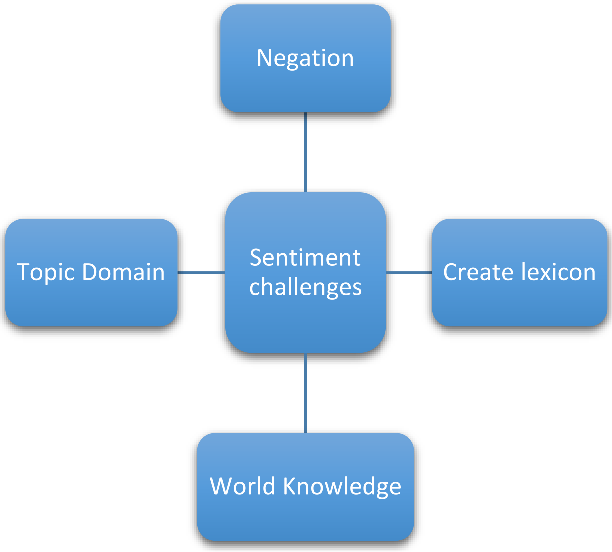

Figure 4.4: The boundary of sentiment challenges solutions

4.4.2 Topic domain independence

Domain-dependent has a difficult nature, it faces how to recognize the topic features, keywords and characteristics. There are fuzzy words that have several meanings with different sentiment scores and have bi-polar sentiment polarity relevant to the topic. Extracting keyword or features and evaluating them also are important problems. One feature set may give very good performance in one domain, at the same time it performs very poor in some other domain. We can produce a solution with a small scale with applying the proposed technique on one topic domain and examine domain parameters evaluation by categorization reviews. Because they also give

* www.citeulike.com

different meaning with the same word. Our research presents a technique to recognize topic nature automatically. It is based on extracting keywords and relevant features of each topic. In addition, to produce a solution for some words have many meanings and different sentiment values relevant to the topic. It’s based on classification review of each domain features and keywords by assuming a major five classes for reviews (Topic, place of publication, publishing date, citation number or authors).

For example; “IEEE is **[great +]** publication for your paper”, we can put IEEE is in a place of publication classification (based on feature name of publication) and the polarity is positive. “The publishing conference is **[great+]**”, this review refers to the place of publication classification (based on keywords) and the polarity refers to positive. In another example, “The paper publishing date is **[old-]**”, this reviews refers date classification (based on date attribute) and “Old” having the negative score. “The author is **[old^{-}]** in this field”, but we categorize the last review in author classification and it is meaning the author is expert in this field so “Old” will be had positive score.

It improves the sentiment score to be more accurate and fair. We assume some words have '0' value because it depends on classifications of each sentence of each review, we have a group of words having a polarity and score to related with the detect classification. Table 4.4 explains a review classification in scientific papers domain.

Table 4.4: Review classification in scientific papers domain

Topic	**Scientific papers domain**	
Topic classification reviews	1) Topic 2) Authors 3) Place of publication 4) Publishing date of the paper 5) The citation number of each paper	
Keywords	Topic	Word from title, abstract, or field or keywords
	authors	author, researcher,
	Place of publication	Conference, journal, place of publication
	Publishing date	The date or year or month of the publication

	Citation number	Citation, quotas
Terms or features of determined topic	Topic	Implicit meaning
	authors	Name of authors
	Place of publication	Place, names of journals or conferences
	Publishing date	Year, date, any word refers to time
	Citation number	The number of mentioned this paper in other research

4.4.3 Negation

Negation is the biggest challenge in sentiment analysis [57]. The new technique produces a solution to improve evaluation negative with the enhanced bag of words technique. This research handles the two techniques: explicitly and implicitly negative [9]. First: explicitly is deliberately formed and are easy to self-report and by keywords. Second implicitly is the unconscious level, are involuntarily formed and are typically unknown to us without any keywords of negative. In addition, the handling the negative meaning of some conjunctions such as "not only", and "But". The dual negative is the most important case which cares to achieve the total sentiment polarity. Reverses polarity of mid-level terms: great V.S not great. A method often followed in handling negation explicitly in sentences like:

"**I do [not like^{+}] $^{-}$ the paper**", is to detect the negative polarity because the word (not) and convert the sentence operator to negative. But this does not work for "**I do [not like^{+}] $^{-}$ this research but I [like^{+}] the field**". But still there can be problems. Another example, "I find the functionality of this new methodology [less $^{-}$] practical", this review refers to explicit comparative negative. "**This algorithm is [not great^{+}]$^{-}$]**", we handle in this review the positive and negative evaluation which declares in [not great! = bad] but [not great = good]. Implicitly negative such as "**This research is [very [complex $^{-}$] $^{-}$]**" this example doesn't have any negative keyword, but the meaning has negative and the polarity will be negative of this sentence.

There are sentences having keywords of negation and they don't have the negative polarity such as "**[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. We separate between “not only” and not because not only strengths the meaning (more positive or negative, I [wish $^{-}$] to work [harder $^{-}$]”. In the last review, our technique present future words e.g. wish refers to negative but we must check the polarity of the next sentence polarity because maybe changed the polarity depends on meaning.

4.4.4 Creation lexicon

Our proposed SAOOP yields an improvement over prior published bag of words built lexicons. We also provide an improvement in calculation technique used in reviews sentiment analysis. Our technique presents a solution to take care of grammar (which is one of limitation of Bag-Of-Words) and to save time in caring with the ordering of several words or expressions to create subsequences of terms. We produce two phases:

- **Phase 1. Data Preparation Phase**

Less number of words in vocabulary lexicon to fast search based on similarity and differences algorithms. We neglect verbs tenses or word formula (singular or plural), that’s meaning we neglect English grammar and syntax because of our comparison and differentiation with the infinitive verbs, and singular words with most letters similarity.

- **Phase 2. Lexicon Development Phase**

Evaluation words /terms: is based on enhanced bag of words: we don’t depend on term frequency. It is based on assuming each word has two values and the total of them equal 1. Each term has 2 polarities (+/-).

The proposed technique creates a new lexicon for sentiments reviews which is based on the hierarchal database model. This model is a data model where the data is organized like a tree. The structure allows repeating information using parent/child relationships: each parent can have many children, but each child only has one parent. All attributes of a specific record are listed under a feature type. The lexicon has words, prefixes and suffixes and hierarchal nouns to produce a solution for bi-polar words and evaluate topic features of reviews. We use Part-of-speech (POS) tagging model to recognize nouns in constructing the nouns tree in lexicon. The advantage of this hierarchical model of nouns is each parent can hold the same name of child with different value that supports us in evaluating bi-polar and fuzzy words or the topic features.

4.4.5 World knowledge requirement

The proposed technique presents a solution for Knowledge about worlds' facts, events, people are often required to correctly classify the text. Trying to achieve higher accuracy and get the evaluation for some neutral reviews. The World knowledge challenge solution is based on the hierarchical database of nouns. Hierarchal model between nouns to achieve the polarity, score, and meaning. Also to differ between them and keywords or features. In a hierarchical model [67], data is organized into a tree-like structure, implying a single parent for each record. A sort field keeps sibling records in a particular order. Hierarchical structures were widely used in the early mainframe database management systems. This structure allows one one-to-many relationship between two types of data. This structure is very efficient to describe many relationships in the real world; recipes, table of contents, ordering of paragraphs/verses, any nested and sorted information.

Semantic (hierarchal) relationships between nouns to achieve the polarity, score, and meaning. Also to differ between them and keywords or features. Consider the following example, "the author is a [**lion**$^{-}$] in this field", the previous review present negative polarity because it's a name of animal but in real evaluation it's a positive polarity. In the next review, "Bing is really [**Einstein**?]" evaluation sentiment analysis without world knowledge classifies above sentence as neutral, but it is an objective sentence because Einstein is the name of the famous scientist, so it refers a positive polarity also. It's very hard for software to understand that automatically. We create a huge lexicon database to contain the world knowledge especially related to researchers and the most common in the reviews. We assume values of the words based on the most common meaning. It depends on keywords and classification of reviews.

4.4.5 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.

4.5 Design and Implementation

4.5.1 Design

We have used three layer architecture for our proposed SAOOP tool. The top most layer is the presentation layer (GUI), which manages all the interaction to end user. The middle layer is the application logic layer which includes all the functionalities such as *text analyzer*, *sentiment classification*, *sentiment word evaluation techniques*, *lexicon* which are used to manage knowledge resources. The bottom layer is the database layer and contains the database for paper, paper Metadata, and review relation and sentiment words and prefixes shown in Figure 4.4:

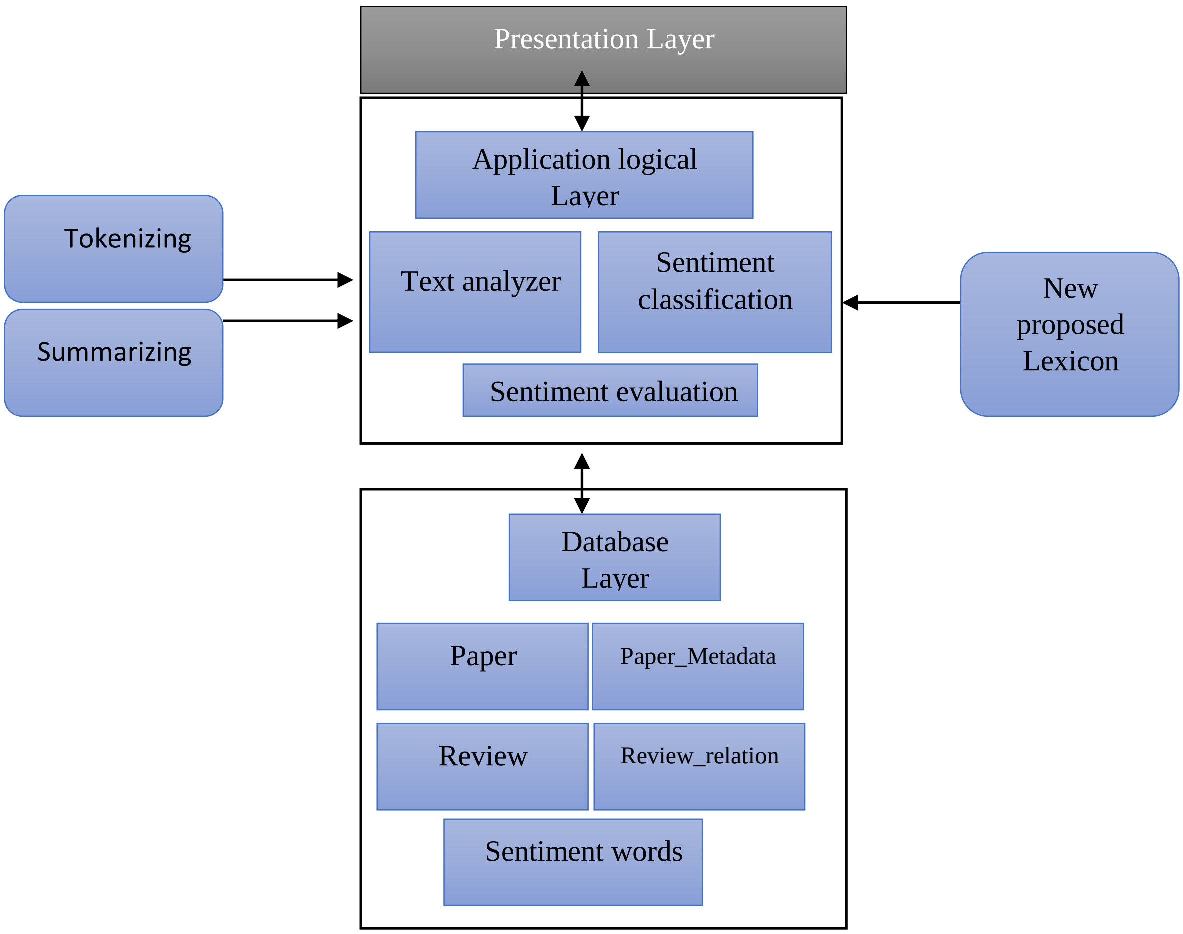

Figure 4.5: Proposed Design of Database

4.5.2 Implementation: The implementation of proposed SAOOP technique is constructed based on using C# programming language working on Microsoft visual studio 2010 platform. Our lexicon is based on SQL Server Management Studio 2008.

4.6 Chapter Summary

This chapter introduced a newly proposed technique for analyzing a scientific papers domain and evaluating online sentiments reviews in scientific research papers domain. It's called: "Sentiment analysis of online papers" (SAOOP). SAOOP uses a natural language processing, text analysis and opinion mining in the sentiment analysis evaluation. It presents an enhancement of bag-of-words model and creates a new miniature lexicon. The proposed technique uses bag-of-words model for fitting the review structure which is short and relevant to a scientific topic domain. The proposed technique solves two essential Bag-of-words weaknesses to improve accuracy. Not only SAOOP evaluates the score but also determines the polarity in one from five polarity classification levels. In addition, it can estimate the paper score based on evaluation some main parameters of papers. The SAOOP aims at supporting researchers and saving their time and efforts by enabling them to report the total score of the papers. SAOOP can achieve to the accurate meaning of the reviews from reviews categorization based on papers parameters.

CHAPTER 5: EXPERIMENTAL RESULTS

Chapter 5

Experimental Results

In this chapter, we discuss three experiments for evaluating the proposed technique efficiency. The **first** experiment shows how the proposed technique works to evaluate the online scientific paper. This evaluation relies on two parts: **1)** Enhancement the Bag-Of-Words (BOW) in text analysis and the sentiment score evaluation. **2)** Production of sentiment analysis challenges solutions based on review classification according to extract topic features and keywords. Respect, it also can evaluate paper based on essential topic domain parameters (place of publication, publishing paper date, and number of citations).

The **second** experiment is the comparison between the standard bag-of-words (BOW) model and our enhancement a bag-of-words model. This experiment aims at showing the effects and changes in enhanced technique and how it will be more effected in understanding reviews.

The **third** experiment presents that the efficiency evaluation of the proposed technique (SAOOP) with making a comparison between our proposed technique and two famous techniques in evaluating sentiment analysis. Namely "Natural Language Toolkit-Text processing" (NLTK) [36] and "recursive deep models for semantic" (NLPS) [8]. In this comparison, we compare the accuracy and performance results based on two datasets. We also compare with the effects and solutions of sentiment analysis challenges.

5.1 Experiment 1: SAOOP Case Study:

In this section, we discuss one case study for the proposed technique SAOOP. First, making web scrapping using EasywebExtract tool for extracting data of this input paper. The input is a link of paper. In Table 5.1, we discuss the data of web extraction.

Table 5.1: Extracting data for a case study

No.		Data
1.	**Paper link:**	http://www.citeulike.org/user/karelvdv/article/5943973
2.	**Name:**	EnaCloud: An Energy-Saving Application Live Placement Approach for Cloud Computing Environments
3.	**Field:**	Cloud Computing
4.	**Keywords:**	Cloud computing
5.	**Authors:**	Bo Li, Jianxin Li, Jinpeng Huai, Tianyu Wo, Qin Li, Liang Zhong
6.	**Place of publication:**	IEEE international conference in cloud computing
7.	**Number of Citations:**	4
8.	**Publishing Date:**	2011
9.	**Number of Reviews:**	10
10.	**Reviews**	1. Not only is this paper good but also having the best algorithm results. 2. Wonderfully, it has an old author in this field. 3. Li is Einstein researcher. Good work. 4. You need to improve your work by using some examples. 5. You need to improve your work by using some examples. 6. The topic is old. I wish to search in another point. 7. IEEE is a great publisher for your research. 8. I can't understand this paper because it's complex. 9. ---- 10. It's not great.

Then we reformate data of the extracting data. We also extract data of each review and the writer names. A writer who is a researcher writes commentaries on scientific papers online. Table 5.1 has a sample of sentiment reviews set for analyzing them. Assuming sentiment classification polarity has five classes (Very positive, positive, fair, negative, and very negative) and review topic classification divides into five classes (topic, publishing date, and place of publication, author, and citation). And there are cases handle with our enhancement BOW with similarity and differences algorithms to reduce generating lexicon. With handling five sentiment challenges based on required review structure (topic domain, spam and fake detection, create lexicon, world knowledge, and negation). Applying algorithm in the enhancement BOW for evaluating sentiment score with the next steps in Table 5.2.

Table 5.2: Steps of proposed technique

No.	Steps
1.	The input.
2.	Web scrapping/ Web extracting.
3.	Text analysis process. o Split into sentences.
4.	Check on class for each sentence in each review. o Classify each sentence in each review. o Assume SA score=0.
5.	Text analysis process (continue) o For each w in sentence (tokenization) o Remove stop list. o Remove punctuation list. o Convert all words into Upper case.
6.	Create reduced lexicon. o Assume each word has two values (positive and negative). o Convert word to infinitive o Construct the reduced lexicon.

7.	Check on words. ○ Check on word list. ○ Check on prefixes list. ○ Else, Check on Nouns (Features, keywords, noun).
8.	Check on Negative. ○ Check on explicit noun. ○ Check on Two negative level lists. ○ Detect sentiment polarity. ○ Detect sentence score.
9.	Check on Future words.
10.	Check on nouns.
11.	Assign sentiment classification for each review.
12.	Get the total score of sentiment score for each word. (Real total score and Average total score).

The results are declared in Table 5.3.

Table 5.3: Results of Case study

No.	Handle case	Sentiment analysis Challenge	Sentiment polarity	Review Class	Sentiment Score
1.	Expression+ Adjective+ Superlative	-Similar to Negative -Keywords Topic	+	Topic	2
2.	Adverb + Adjective	- Bi-polar word - Keyword Topic	+	Author	1.4

3.	Noun + Special Cases (need to)	- World knowledge - Name (noun) Topic	+	Author + Topic	1.4
4.	Adjective+ Verb+ Wish (Case)	- Implicit Negative - Special Case	-	Topic	-0.7
5.	Spam Review (Duplicate review)				
6.	Adjective	- Keyword Topic	-	Topic +Topic	-1.2
7.	Adjective	- Name Topic - keyword Topic	+	Place of publication	0.8
8.	Verb+ Adverb+ Adjective	- Explicit Negative - Implicit Negative	-	Topic	-1.4
9.	Fake Review (Empty review)				
10.	Negative+ Adjective	- Explicit Negative	+	Topic	-0.8

In the next table, we explain reviews after text analysis (removing stop list and punctuation list).

Table 5.4: Sentiment Score Evaluation

No. of Reviews	**Review after Stop and punctuation lists**					
1.	**Not Only**	**Paper**	**Good**	**But also**	**Best**	**Algorithm**

	Expression	Keyword Topic	In Word list	Expression	Superlative	Feature Topic
	+0.6	Non-Score	+0.6	Non-score	+0.8	Non-score
	Number of sentence =1					
	Sentiment Score= 0.6+0.6+0.8= 2 Sentiment polarity= Positive (+)					
2.	**Wonderfully**	**Old**	**author**	**Field**		
	In Word list	In Word list	Keyword Topic	Feature Topic		
	+0.8	+0.6	Non-score	Non-score		
	Number of sentence =1					
	Sentiment Score= 0.6+0.8= 1.4 Sentiment polarity= Positive (+)					
3.	**Li**	**Einstein**	**Research**	**Good**	**Work**	
	Name (noun) Topic	World knowledge Noun	Keyword Topic	In Word list	Feature Topic	
	Non-score	+0.8	Non-score	+0.6	Non-score	
	Number of sentence =2					
	Sentiment Score= 0.6+0.8= 1.4 Sentiment polarity= Positive (+)					
4.	**Need to**	**Improve**	**Work**	**Example**		
	Special case (-) polarity	In Word list	Feature Topic	Feature Topic		
	-1	+0.6	Non-score	Non-score		
	Number of sentence =1					

	Sentiment Score= -1 *0.7= -0.7 Sentiment polarity= Negative (-)			
5.	Delete Spam Review (Duplicate Review with the same writer and time)			
6.	**Topic**	**Old**	**Wish**	**Point**
	Keyword topic	In Word list	Wish (Special case)	Feature topic
	Non-score	-0.6	-0.6	Non-score
	Number of sentence =2			
	Sentiment Score= -0.6-0.6= -1.2 Sentiment polarity= Negative (-)			
7.	**IEEE**	**Great**	**Publisher**	**Research**
	Name (Noun) Topic	In Word list	Feature Topic	Feature Topic
	Non-score	+0.8	Non-score	Non-score
	Number of sentence =1			
	Sentiment Score= 0.8 Sentiment polarity= Positive (+)			
8.	**N't ➔Not**	**Understand**	**Paper**	**Complex**
	Negative	In Word list	Feature Topic	In Word list
	-1	0.6	Non-score	-0.8
	Number of sentence =1			
	Sentiment Score= -0.6-0.8= -1.4 Sentiment polarity= Negative (-)			
9.	Delete Fake review (Empty review)			
10.	**N't ➔ Not**		**Great**	
	Negative		In Word list	

	-1	0.8
	Number of sentence =1	
	Sentiment Score= -1*0.8= -0.8 Sentiment polarity= Negative (-)	

With our enhancement Bag-Of-Word model, we compute the total score of sentiment =2.1 shown in Table 5.4. Then the evaluation of total score of sentiment score= 2.1/8=0.2625, with the division of the real number of reviews=8. In addition, the evaluation of sentiment polarity is positive. Moreover, we evaluate system score of this paper, we mention this approach in Table 5.5.

Table 5.5: System Score Evaluation

No.	**Step**	
1.	For each paper P do.	
2.	Get Paper parameters values	
3.	Get publishing date	
	If it current year? (year= current date)	No
	If not current year? Yes	Year =2015 -2011= 4 S(D)= 0.8
4.	Evaluate citation score	
	With assuming levels into 10 regions.	S (C)= 0.9
5.	Evaluate place of publication S(PP)	
	Determine tier of place of publication	IEEE = Tier 1 with highest 1
	Detect PP score	S(PP) =0.8

6.	The output Evaluate V (SS), assuming λ is constant =2. V (SS) = $\sum \frac{S(PP)}{\lambda} + \frac{S(C)}{\lambda} + \frac{S(D)}{\lambda}$ = 0.8/2+0.9/4+0.8/4= 0.4+0.225+0.2=0.825.

The total score of system score= 1.0875 with polarity (positive).

5.2 Experiment 2: Comparison between Standard Bag-Of-Words (BOW) and the Enhancement BOW Model

The discussion of this experiment explains the comparison between the proposed technique and the standard BOW technique in online scientific papers domain. This comparison shows the accuracy results based on the real dataset. A real set (1000 reviews) from the CiteULike website in computer science branch. This comparison also discusses the challenges solutions impact on evaluating sentiment analysis.

Let in a group of reviews has positive sentiment (belong to class positive) and reviews have negative sentiment (belong to class negative). Then, in relation to class positive: TP FP TN FN.

$$precision = \frac{TP}{TP+FP} \qquad (5),$$

$$Recall = \frac{TP}{TP+FN} \qquad (6),$$

The accuracy equation declares in the next equation,

$$Accuracy\ (acc.) = \frac{TP+TN}{TP+TN+FP+FN} \qquad (7),$$

By reporting, all the measurement mentioned above by practical interpretation. The true positive rate or recall can be understood as the rate at which positive reviews are predicted to be positive (R), whereas the true negative rate is the rate at which negative reviews are predicted to be negative. 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).

With the examination of the percentage degree of different techniques accuracy [68] on text reviews content. In order to compute the accuracy of each technique, by calculating the intersections of the positive or negative proportion given by each technique. Table 5.6 presents the percentage of accuracy for the three compared techniques. With the results of Table 5.6, they illustrate differences between accuracy for the two models with showing results of recall, precision, and accuracy.

Table 5.6: Average results for all datasets

Metric	**BOW**	**SAOOP**
Precision	0.834	0.856
Recall	0.560	0.867
Accuracy	0.618	0.817

SAOOP gets a better results of accuracy (83.5 %) than standard (62%). So the enhancement bag of words increases the accuracy with around 20%.

In the next table, we discuss the comparison between the two algorithms standard BOW and proposed Enhanced BOW with SAOOP technique. Table 5.7 illustrates a comparison which depends on part of speech features, the goal of the algorithm, sentiment classification levels, the input type: documents, images or reviews, the data size of the input data, dataset scope, clarity, efficiency, memorability, simplicity [68]. The definition of memorability is the quality or state of being easy to remember or worth remembering, which can help to declare the relationship between the data and its features. The Algorithm will be clear if it is familiar and easy to use. The efficiency of algorithms depends on the accuracy results. The last issue simplicity of algorithm which is simple if it is concise to write down and easy to grasp. Simplicity” of an algorithm is affected by “cultural” factors: Means of presentation (notation, assumptions…etc.) and Previous knowledge of the reader.

Table 5.7: Difference between Standard BOW and Enhancement BOW models

No.	Algorithm	Standard BOW	Enhancement BOW
1.	Goal	Text analysis and give polarity for words in text	Evaluate sentiment score for reviews
2.	Sentiment classification	2 or three classes	5 classes
3.	Input type	Documents, text or images	Reviews
4.	Data size	Small number of texts or review	Large number of reviews
5.	Dataset	Any scope, refer topic domain to minimize dictionary	Topic domain is the best to minimize dictionary and can extraction features and entities
6.	Clarity	No	Yes
7.	Efficiency	No , less accuracy and manually dictionaries	Yes, high accuracy
8.	Memorability	No	Yes
9.	Simplicity	Yes	Yes

5.3 Experiment 3: Comparison among Proposed SAOOP Technique and Two Sentiment Techniques

The third experiment presents the comparison between our proposed technique (SAOOP) and two sentiment analysis techniques. In this comparison, we evaluate the efficiency of the proposed technique by measuring the accuracy and performance analysis based on two datasets. We also compare with the effects and solutions of sentiment analysis challenges.

5.3.1 Datasets

We use two different datasets: 1) real dataset: we spirit it into two datasets with training set (1000 text reviews) and test set (5000 text reviews), 2) verified dataset: a real set is with unknown evaluation around10.000 text reviews (including more than 5.000 positive words, 5.000 negative words).

- **Real dataset**

The first sample set is a sample of WWW.citeulike.com papers reviews and Metadata posted by computer science papers branch [69]. The comparison in real dataset in computer science scope including two parts: training data and test data. Training data is a set of data to evaluate sentiment around 1000 reviews, we know the values before. The second part is the test data: is a set of data to evaluate sentiment with hide class label around 5000 text reviews. CiteULike receives in excess of 200,000 distinct visits (defined by Google Analytics as a group of page views by a unique user with time out after 30 minutes inactive) monthly, with each visit originating an average of 2.77 page views [9]. Of that 200,000 around distinct users who have previously visited the site on multiple occasions. There are currently 505,402 items posted in the database (counting n people post the same article); 1,676,130 tags (counting n if there are 'n' tags applied to an article); and 130,548 distinct words used these numbers are growing exponentially. This sample set allows us to study the responses to noticeable past texts. In addition, the evaluation the improving of our techniques methodology in sentiment analysis and evaluate papers. SAOOP can handle ten cases to ease to understand text review accurately by CiteULike writers. SAOOP can care and evaluate of some English grammar to improve BOW model.

- **Verified dataset**

The second dataset which is called verified dataset is a real dataset of data we can't know the evaluation before. We have around 10.000 text reviews in this sample. It is splitting into two parts of verified data reviews as positive and negative. These datasets include a wide range of online papers texts reviews: general reviews. Our technique can evaluate sentiment score with the relationship of reviews categorization. With applying on this human-verified sample set [12], we fit to quantify the range with different sentiment analysis techniques can accurately evaluate polarity of text reviews.

NLTK can interpret less than 10% of all relevant reviews. In addition, we compare with the percentage of handling sentiment analysis challenges to high accuracy and performance of sentiment analysis of the three techniques of the text reviews. According to the comparison, SAOOP had a new solution for some sentiment challenges but NLPS and NLTK, they and can't produce a methodology to solve them expect some cases in negation but they have many logical errors. We examine the average result analysis of the two big dataset that spirited into three datasets, that illustrate the highest average results with the sentiment score of our proposed SAOOP technique then NLPS and the lowest one is NLTK Technique.

Finally, the summarization the results with the average of the three datasets (real and verified sets), we find the average of sentiment score of our proposed technique improve the results. Because of working binary analysis solutions of some important challenges and evaluate some technical cases in the text which have a problem in evaluation to be more accurate. In next section, we discuss the accuracy results of our comparison. In the next table, we present a sample example of coverage rate meaning of reviews in three techniques.

5.3.2 Comparison Measures

In order to define the evaluation of accuracy and performance (F-Measure) of the three techniques, we consider the following (shown in Table 5.8):

Table 5.8: Measurement Metrics Table

		Actual observation	
		Positive	Negative
Predicted expectation	**Positive**	**x**	**y**
	Negative	**z**	**w**

Let present True positive (x) was defined when a text was correctly classified as positive, False Positive (y) is a negative text which was classified as positive, False Negative (z) is a positive text but was classified as negative, and the last one True Negative (w) is a correctly classified as negative [65]. In order to compare and evaluate the techniques, we consider the following metrics, commonly used in information retrieval: true positive rate or recall: $R = x/(x + z)$, false positive rate or precision: $P = x/(x + y)$, accuracy: $A = (x + w)/(x + y + z + w)$, and F-measure (performance): $F = 2 \bullet (P \bullet R)/(P + R)$. We will in many cases simply use the F-measure, as it is a measure of a test's accuracy and relies on both the precision and recall [10].We report all the measurement mentioned above by practical interpretation. The true positive rate or recall can be understood as the rate at which positive reviews are predicted to be positive (R), whereas the true negative rate is the rate at which negative reviews are predicted to be negative.

The accuracy represents the rate at which the method predicts results correctly (A). The precision also called the positive predictive rate, calculates how close the measured values are to each other (P). We also use the F-measure to compare results, since it is a standard way of summarizing precision and recall (F). Ideally, a polarity identification method reaches the maximum value of the F- measure, which is 1, meaning that its polarity classification is perfect. The y-axis is a percentage of the understanding sentence rate.

5.3.3 Comparison Results

In order to facilitate understanding the advantages, disadvantages, and limitations of the various sentiment analysis techniques, we present comparison results among them.

1) Understanding of Word Coverage:

We begin by comparing the coverage of part of speech of English grammar sentence across the representative scientific reviews from CiteULike website. In the next Figure, we summarize the results of precentage of coverage the parts in reviews, the result for our proposed technique SAOOP and the NLPS technique result which is predicting the sentiment of reviews based on a recursive model , which explained in chapter 4. Our research divides the comparison into three groups that based on affiliated with certain. These groups are expressions, words, and negative.

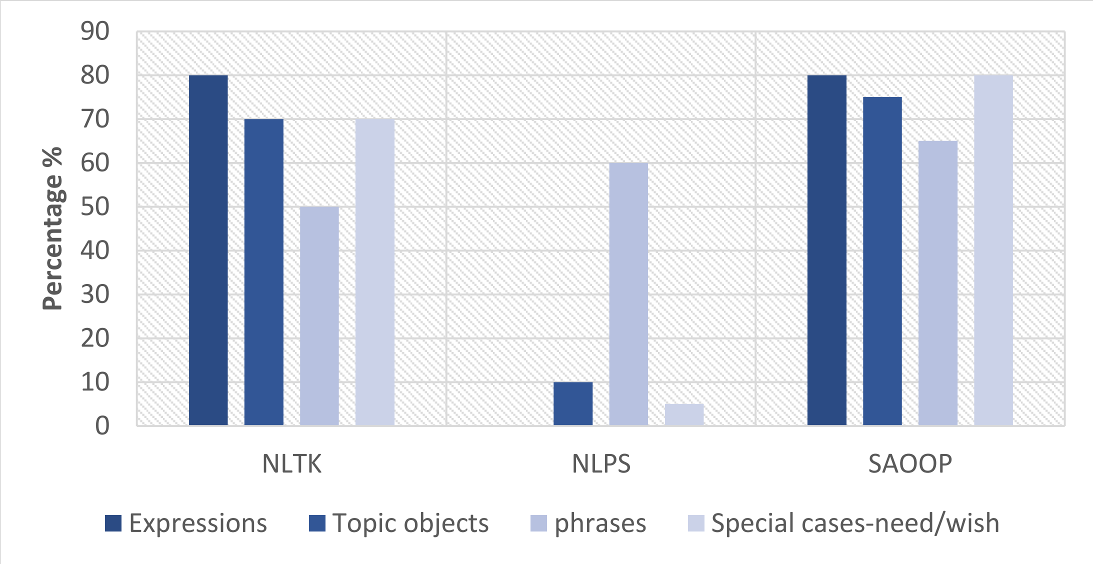

Figure 5.1A: Phrases and Expressions Comparison Group

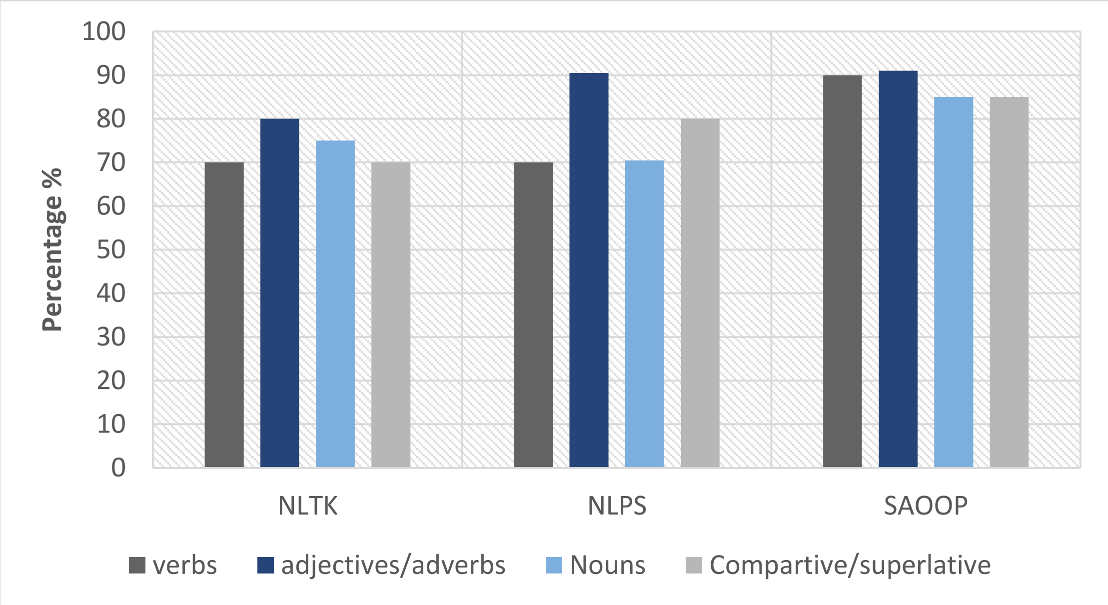

Figure 5.1 B: Words Comparison Group

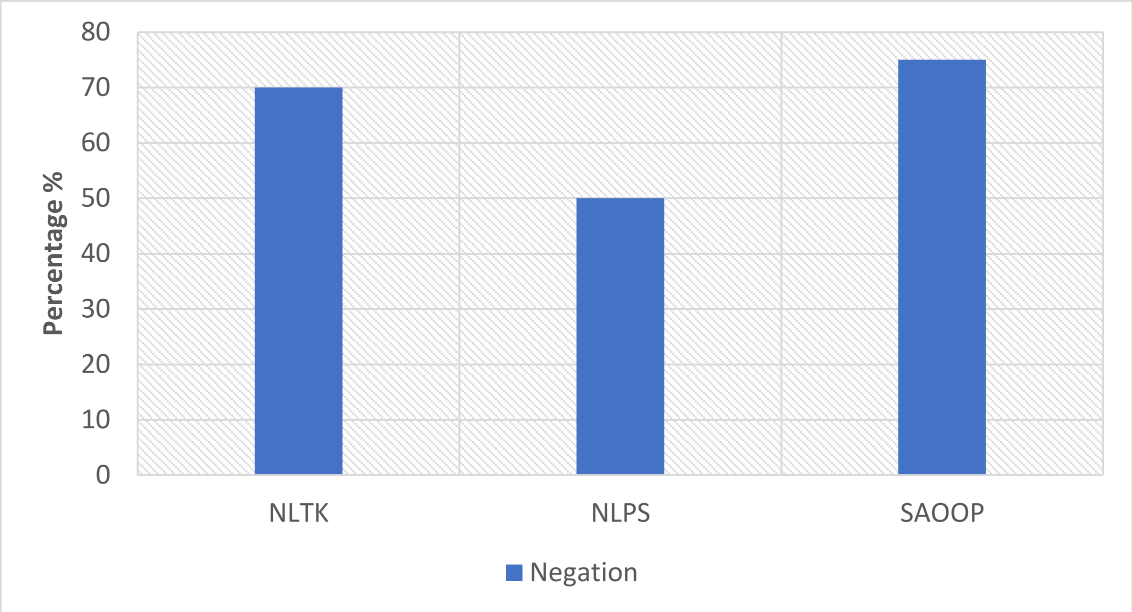

Figure 5.1 C: Negative Comparison Group

Figure 5.1: Correctness Part Of Speech Comparison among Sentiment Techniques

Although the proposed technique is based on the word level, it can care the ordering and grammar of sentences in most times. We present the results among the three techniques in evaluating the correctness percentage of achievement the coverage parts in sentence reviews. As shown in Figure 5.1, SAOOP has the highest understanding sentence coverage with 82.5 % with two datasets with three datasets samples, respectively, followed by NLPS we can't evaluate the total sentiment score but with detecting word by word polarity its percentage is 72%.

2) Sentiment Analysis Challenges:

This thesis introduces solutions for five sentiment challenges for improving accuracy. These challenges are chosen based on the review structure (short and formal) relevant to the topic domain. These challenges are Topic domain extracting features, bi-polar sentiments, implicit and explicit negative, world knowledge, and create lexicon. With respect to deal with the real sentiments through repetition or empty reviews that called: spam & fake reviews. In the following, the results comparison between the sentiment analysis challenges for the three compared techniques. The next comparisons discover the solutions can face and deal with the sentiment challenges. We conclude the results of the sentiment analysis challenges comparison for the three techniques in Figure 5.2.

Figure 5.2: Percentage of handling sentiment analysis Challenges for the three techniques

3) Accuracy Analysis

With the examination of the percentage degree of different techniques accuracy on text reviews content. In order to compute the accuracy of each technique, we calculated the intersections of the positive or negative proportion given by each technique. Table 5.9 presents the percentage of accuracy for the three compared techniques. For each technique in the first column, we estimate from the two datasets of reviews. Finding that some techniques have a high coefficient as in the case of SAOOP (82.5%), while others have least overlap such as NLTK (62%) and NLPS (70.2%). The last “column” of the table shows on average to what extent each technique agrees with the other two samples. The last “row” quantifies how other methods agree with a certain technique, on average.

Table 5.9: Average results for all datasets

Metric	SAOOP	NLTK	NLPS
Recall	**0.856**	**0.571**	**0.253**
Precision	**0.867**	**0.845**	**0.846**
Accuracy	**0.817**	**0.629**	**0.715**
F-measure	**0.846**	**0.665**	**0.729**

With the results of Table 5.10, they illustrate differences between accuracy and performance (F-measure) of the three techniques. Table 5.10 shows techniques recall, precision, Accuracy and performance.

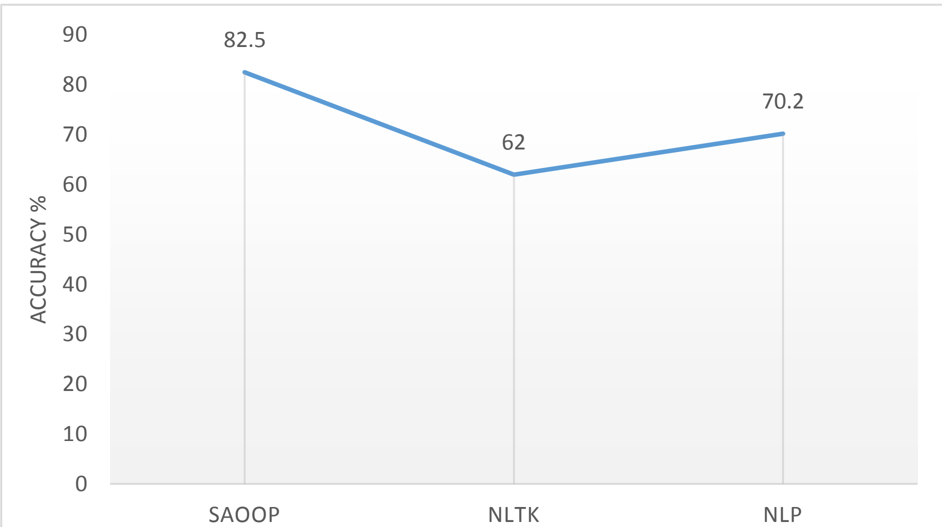

Figure 5.3: Differences between Accuracy of three techniques

Figure 5.3 is shown the accuracy results of them. In a summary, the result indicates that existing tools vary widely in terms of accuracy about sentiment score, with scores ranging from 60% to 80%.

Table 5.10: Percentage of Accuracy between Techniques

Dataset	SAOOP	NLTK	NLPS	Average
Training set 1.000	83.5%	62%	72%	**72%**
Test set 5.000	81.99%	61%	70%	**70.99%**
Real set 10.000	82.5%	60%	71.56%	**71.186%**
Average	**82.5%**	**61.514%**	**71.604%**	-

4) Performance Analysis

In this section, we show an evaluation of the performance of the three compared techniques in Figure 5.4. For comparing the performance results, Table 5.10 which gives the average of the results obtained for all datasets. For the F-measure, a score of 1 is ideal and 0 is the worst possible. The technique has the highest F-measure faces sentiment analysis challenges and covers ten cases of each text review (0.846), which has the highest sentiment accurate and understanding text coverage. The second rated technique in the understanding of F-measure is NLPS, which obtained a much higher coverage than understanding and challenges. It is important to note the problem in it that it can't be interpreted into of total score of the text review. We observe better performance on datasets that contain more expressed sentiment, such as text reviews (e.g., papers online).

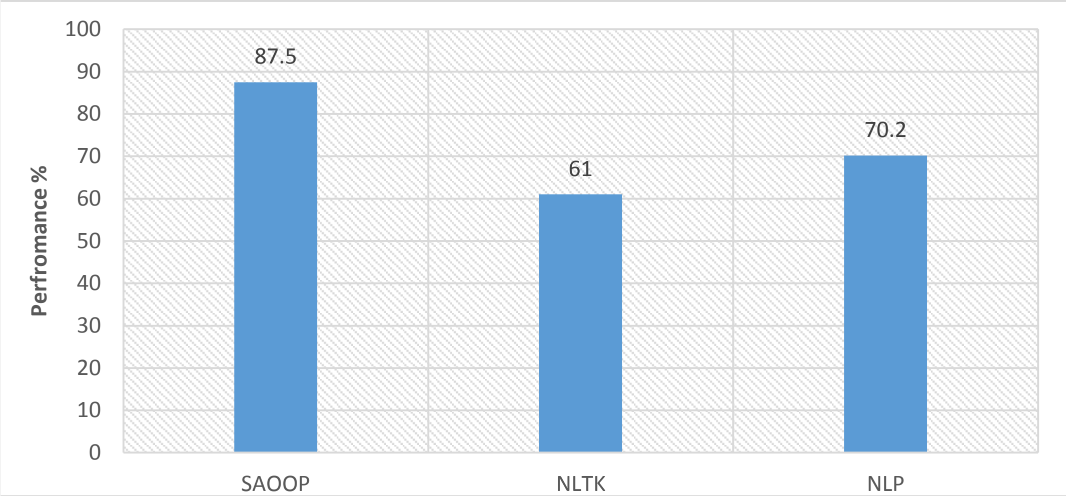

Figure 5.4: Differences among Performances of three techniques

5.4 Chapter Summary

The proposed technique SAOOP has been shown to effective, since it captures more contextual meaning based on word weight in the enhancement of BOW, resulting a classification accuracy of 82%. Our observation the precision and recall for each sentiment category separately, since the effect of our proposed technique has a significantly different impact on the negative and positive class. Although, the proposed technique increases the precision and recall of both the classes, we could observe a significantly higher improvement of precision and recall in dealing with sentiment challenges. This is a clear indication of the effectiveness of incorporating the impact of world knowledge, spam detection, and negation, by interesting the topic domain features and keywords and constructing the newly miniature lexicon. Although the proposed technique is based on the word-by-word model, it can understand some phrases as do not directly through caring with the classification of reviews.

CHAPTER 6:
CONCLUSION&
FUTURE WORK

Chapter 6

Conclusion & Future Work

The conclusion of our thesis presents a new technique for analyzing sentiment reviews. A proposed technique targets performing statistical and numerical analysis on online sentiments for scientific papers domain. It is based on the enhancement Bag-of-words (BOW) model. We present this proposed technique to improve accuracy for sentiment reviews and understand implicit and explicit meaning accurately. 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 becomes depends on it to achieve the efficient product. Although, there are hundreds of thousands of researcher, who write and read online papers daily, until now no research finds enough in this field. 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 domain based on sentiment analysis. This technique aims at supporting researchers in selecting the suitable 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 based on essential topic parameters evaluation. This technique is called sentiment analysis of online papers "SAOOP". It improves accuracy and understanding the online sentiment reviews.

The approach of evaluating the sentiments consists of creating the enhancement of Bag-of-words model and producing solutions for the essential sentiment challenges in this domain. 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 intrudes solutions for the most significance sentiment challenges for improving accuracy. These challenges are negative, spam & fake 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 that place of publication, publishing date, and a number of citation paper.

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 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 (70%) and SAOOP (82%). The performance results between three techniques which give the average of the results obtained for three datasets. The technique with the highest F-measure was faced sentiment analysis challenges is SAOOP (87.5%) and the NLPS is the second on with (71%) and the NLTK with (61%).

6.2 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. For future work, we think that more works need to be done with many online websites.
3. Improve our proposed technique to work on multi-languages in the scientific domain.

References

[1] Suman, D.R, & Wenjun, Z., “Social Multimedia Signals: A Signal Processing Approach to Social Network Phenonmena”, ISBN-13: 978-3319091167, Springer International Publishing Switzerland, 2015.

[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 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] 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.

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

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

[15] 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.

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

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

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

[19] 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.

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

[21] 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.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

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

[31] 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.

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

[33] 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.

[34] 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.

[35] Radulescu, C., Dinsoreanu, M., and Potolea, R., “Identification of spam comments using natural language processing techniques”, Intelligent Computer Communication and Processing (ICCP), 2014 IEEE International Conference on, Sepbtemebr 2014.

[36] 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.

[37] 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.

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

[39] 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.

[40] 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.

[41] 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).

[42] 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.

[43] 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.

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

[45] 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)

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

[47] 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.

[48] 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.

[49] Liu, B., "Sentiment Analysis: A Multi-Faceted Problem", IEEE Intelligent Systems, 2010.

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

[51] 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.

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

[53] 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.

[54] 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.

[55] Chenghua, L., & Yulan, H., "Joint Sentiment/Topic Model for Sentiment Analysis", CIKM '09 Proceedings of the 18th ACM conference on Information and knowledge management, Hong Kong, China, ACM 978-1-60558-512-3/09/11, 2009.

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

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

[58] 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.

[59] 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.

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

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

[62] 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.

[63] 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.

[64] 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.

[65] 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.

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

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

[68] 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.

[69] 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, 2014.

قسم نظم المعلومات
كلية الحاسبات والمعلومات
جامعة القاهرة

تحليل الأوراق العلمية على أساس تحليل المشاعر

إعـــداد

دعاء محيي الدين محمد حسين

رسالة الماجستير المقدمة في كلية الحاسبات والمعلومات

جامعة القاهرة

كجزء من متطلبات الحصول على درجة الماجستير في

في

نُـــظـم المعـــلـومـــات

تحـــت إشـــراف

د. هـــدى مخـتار
قسم نظم المعلومات
كلية الحاسبات والمعلومات
جامعة القاهرة

كـــلية الحـــاسبات والمعـــلومات
جـــامـــعـــة القـــاهـــرة
جـــمهورية مصر العـــربيـــة
فبراير 2016

يستخدم تحليل الآراء أو التنقيب عن الآراء للكشف عن المعلومات الشخصية للكاتبين مثل الآراء والمواقف والعواطف، والمشاعر بطريقة آلية. هناك مئات الآلاف يهتمون بالبحث العلمي وقد يستغرق منهم وقتاً طويلاً للبحث عن الأوراق المناسبة لأبحاثهم. وقد أصبحت أراء المستخدمين عبر الانترنت على الأوراق البحثية هي مصدر مهم من مصادر تقييم هذه الأوراق ومساعدة الباحثين. هذه الآراء والأستعراضات تساعد في حفظ الوقت والجهد وتوفير تكلفة شراء بعض الأوراق. في هذه الرسالة، نحن نقترح تقنية جديدة لتحليل الآراء والأستعراضات على الانترنت في مجال البحث العلمي. ويطلق عليه: "تحليل أراء المستخدمين عبر الأنترنت على الأوراق البحثية" (SAOOP). يهدف SAOOP في دعم الباحثين وتوفير وقتهم وجهودهم من خلال تمكينهم من تقرير التقييم الإجمالي للأوراق. يشمل تقييم هذه التقنية SAOOP: نوعين من التقييم لكل ورقة بحثية: (نسبة الثقة في آراء الناس عبر الأنترنت ، وتقييم أهم الخصائص التي تؤثرعلى الورقة البحثية). أولاً: نسبة الثقة في آراء الناس عبر الانترنت: هو تقييم للورقة على أساس تحليل الاراء والاستعراضات على الانترنت.ثانياً: تقييم أهم الخصائص التي تؤثر على أي ورقة بحثية وهم مكان النشر لهذه الورقة ،وتاريخ صدورها، وعدد الأستعانة بها في ابحاث أخرى.

يستخدم SAOOP طريقة المعالجة اللغة الطبيعية، وتحليل النص والرأي في تقييم تحليل الآراء عبر الأنترنت. SAOOP يقدم تقنية جديدة من خلال تطويروتعزيز لنموذج حقيبة من بين الكلمات (BOW) في تحليل الآراء. وتستخدم هذه الطريقة لتحسين فهم هذه الأراء و دقة المعلومات وحل العديد من التحديات تقييم الأراء. تستخدم هذه التقنية المقترحة نموذج حقيبة من بين الكلمات لتناسب تركيب بنية الرأي التي هي قصيرة وذات الصلة بموضوع مجال علمي. التقنية المقترحة تقدم حلاً لأهم مشكلتين في نموذج حقيبة من بين الكلمات: منخفضة الدقة وتستخدم نهج العمل اليدوي. طريقة تحسين وتعزيز نموذج حقيبة من بين الكلمات يعتبر هو النموذج الآلي لتحليل استعراض الآراء وتقدم تقييم لها. هذه التحسينات تهدف إلى زيادة دقة فهم الجمل والثقة في تحليل الآراء. على الرغم من أن SAOOP يعمل على تحليل الآراء على مستوى الكلمة، إلا أنه يمكنه التعامل مع ترتيب الكلمات والاهتمام بالنحو في الجمل هذه الاستعراضات. كما أنه يبني معجم مصغر جديد لتجنب المشاكل في النموذج الأصلي حقيبة الكلمات . هذا المعجم الجديد يمكنه تجنب التكرار والأزدواجية والبحث فيه بسهولة وبسرعة. المعجم المقترح يمكنه أن يتعامل مع الصفات، الأسماء، الأفعال والأحوال، والصفات، البادئات، والطبقات النحوية الأخرى إلى مرادف في مستوى الكلمة باستخدام التشابه والاختلاف الخوارزميات. كما يصنف SAOOP الأراء استنادا إلى الموضوع السمات الأساسية والكلمات الرئيسية. كل فئة من هذه الأراء لديها بعض الخصائص أو الكلمات الرئيسية. هذا التصنيف له تأثير كبير على معنويات معنى الأستقطاب والشعور النتيجة من الكلمات. على سبيل المثال: العنوان، اسم الكاتب، أو اختصار واسم مجلات أو مؤتمرات. هذه تصنيف تلخيصها في خمسة فصول: (مكان النشر، تاريخ النشر والمؤلفين، عدد الاقتباس واسم الموضوع). ويوفر SAOOP مستويات قطبية تصنيف الآراء مقسمة إلى خمس فئات. هذه الفئات هي (السلبية جداء، سلبي، محايد، إيجابياً، إيجابياً جداً). أنها تستخدم لتقدير قوة المشاعر القطبية.

تقدم تقنية الحلول المقترحة لمعظم تحديات الأراء أهمية دقة فهم الأراء و تحسين النتائج. ومن خلال عرض عدد كبير من الأراء على البحوث العلمية، علينا أن نحدد هذه التحديات. هذه التحديات هي الرسائل غير المرغوب فيها والكشف عن الآراء الوهمية، والنفي والمعارف العالمية، وخصائص مجال الموضوع، وخلق معجم ضخم. يمكن لهذه التقنية المقترحة أن تقيس العدد الحقيقي للأراء من خلال تحديد االآراء غير المرغوب فيها والأستعراضات وهمية. كما يمكن تقييم قطبية الكلمات والجمل فيما يتعلق بالنفي سواء كان صريحاً أو ضمنياً. كما يمكن استخراج الخصائص والكلمات المرتبطة بالموضوع لدعم تقييم الأراء وفهم غموض بعض الكلمات أو الكلمات ثنائية القطبية. علاوة على ذلك يمكن فهم المعرفة

والمعلومات العالمية مثل التعرف على أسماء العلماء الشهيرة. وبالتالي فإن التقنية المقترحة تنتج حلا مع نموذج قاعدة بيانات الهرمي في الأسماء للتعامل مع هذه المشكلة.

الجزء الثاني من التقييم الأوراق البحثية هو تقييم أهم الخصائص الخاصة بهذه الأوراق ويسمي نتيجة النظام. وتستند هذه النتيجة على الثلاث خصائص الرئيسية في مجال الأوراق العلمية. هذه الخصائص هي مكان النشر، عدد الاقتباس وتاريخ ورقة النشر. يعتبر مكان النشر هو مكان نشر البحث العلمي او الورقة العلمية سواء مجلة أو مؤتمر ما. ويٌعرض عدد الأقتباسات بعدد الأبحاث والأوراق العلمية التي تستشهد بهذه الورقة وتعتبرها أحدى مراجعها. أما تاريخ النشر فهو تاريخ نشر هذه البحث ويشير الى العام الذي نُشر فيه هذا البحث. هذا التقييم يعد ديناميكياً، فهناك علاقة عكسية بين تاريخ نشر الأوراق البحثية وعدد الأقتباسات لهذه الورقة. ويعد التقييم الكلي للأبحاث مساعد وداعم للباحثين للوصول لأفضل اختيار للأبحاث العلمية والأوراق المناسبة لأبحاثهم في وقت قصير مع توفير جهدهم.

التقنية المقترحة تقدم أيضاً عدة أمكانيات لدعم الباحثين مثل البحث عن طريق نتائج التقييم ونوع الموضوع البحثي، وإنشاء الرسوم البيانية على أساس تقييم الأوراق،وإنشاء دليل الباحث الذي يٌعد دليل مفيد جدا للباحثين المبتدئين. هذا الدليل يقدم شجرة من التسلسل الهرمي من نقطة البحث المدخلة ويبني تقارير عن البحوث التي أجريت مؤخرا في هذه النقطة والأوراق أعلى تصنيف على أساس تقييم SAOOP.

وأخيرا، فنحن نعرض طريقة لتقييم كفاءة التقنية المقترحة عن طريق مقارنتها مع اثنين من تقنيات تحليل الأراء. وتستند هذه المقارنة على قياس الدقة والأداء ومعدل فهم أجزاء الجملة في كل رأي مستعرض. هذه المقارنة تعتمد على مجموعتين من البيانات مختلفة: مجموعة البيانات الحقيقية التي تنقسم إلى مجموعتين من البيانات هما مجموعة التدريب ومجموعة الاختبار، والمجموعة الثانية هي مجموعة بيانات حقيقية.

إقــــــــــرار

أقر بأن هذا العمل لم يسبق قبوله لأي درجة علمية وليس مقدماً حالياً للحصول على أي درجة علمية أخرى. كما انه تم إسناد كل ما هو مقتبس في هذه الرسالة إلى مصدره الأصلي مع ذكر المصادر بوضوح تام.

اسم الطالب: دعاء محيي الدين محمد حسين.

التوقيع: .