SENTIMENT POLARITY CLASSIFICATION OF TRENDY TOPICS

Thesissubmitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering
in
Software Engineering

Submitted By
MansiSethi
(801231016)

Under the supervision of:
Dr. ShaliniBatra
Assistant Professor, CSED

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR UNIVERSITY
PATIALA – 147004

June 2014

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "*Sentiment Polarity Classification of Trendy Topic*", in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Software Engineering* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of *Dr. Shalini Batra* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Mansi Sethi
Signature:

(**Mansi Sethi**)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

(**Dr. Shalini Batra**)
Assistant Professor
Computer Science and Engineering Department

Countersigned by

(Dr. Deepak Garg)
Head
Computer Science and Engineering Department
Thapar University
Patiala

(Dr. S. K. Mohapatra)
Dean (Academic Affairs)
Thapar University
Patiala

Acknowledgement

No volume of words is enough to express my gratitude towards my guide **Dr. ShaliniBatra**, Department of Computer Science and Engineering, Thapar University, Patiala, who has been concerned and has aided for all the materials essential for the preparation of this thesis report. She has helped to explore this vast topic in an organized manner and provided me with all the ideas on how to work towards a research oriented venture.

I am also thankful to **Dr. Deepak Garg**, Head of Department, Computer Science and Engineering Department and and**Ms. DamandeepKaur**, P.G. Coordinator, for the motivation and inspiration that triggered me for the thesis work.

I would also like to thank the staff members and my colleagues who were always there at the need of the hour and provided will all the help and facilities, which I required, for the completion of my thesis work.

Most importantly, I would like to thank my parents and the almighty for showing me the right direction out of the blue, to help me stay calm in the oddest of the times and keep moving even at times when there was no hope.

MansiSethi

Abstract

Twitter is becoming increasingly mainstream which make adequate preparation for valuable user-generated information by sharing contents. Extracting topics trends from twitter has drawn a lot of attention in recent epoch. Twitter data or tweets is often growing rapidly with accelerating speed, which poses remarkable challenge to existing topic extracting models and polarity classification. This thesispresents a novel approach to find the sentiment polarity classification in trendy topics. First the topics trends are determined and then the distribution of word is used to represent semantics of Twitter streams grouped in certain time span to calculate the semantic relatedness of Twitter streams to Wiki topics. The sentiment polarity classification of trendy topics is done by obtaining a vector of weighted nodes from the WordNet graph. Final estimation of the polarity is done in SentiWordNet using weights calculated from WordNet graph. The proposed method provides a supervised solution independent of domain. It has been experimentally evaluated that the proposed approach significantly improves the clarity of topic trends.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The invention of the web leads to the breaking down of the bridge between the users. Web has changes from static repository to dynamic one in which user can communicate all type of information. Due to this information exchange, different websites are practicing in the field of finding the users opinions. One explores the idea of relating topics together simulating the human process, by trying to understand the meaning and relations of information.

With the advent of Wikipedia, a collection of human knowledge base, a world of understanding is presented in a sequence of semantic representations. Since any article can be composed of a combination of several topics, chunk of information can be represented by a group of Wikipedia topics, which would relate to a subset of typical human communications (Twitter). Thus, one can classify a text or article into a topic or mixed topics by the relatedness of words that are normally distributed in any Wiki topic.

After determining the topic trends a new strategy is used for evaluating the polarity classification of the posts which expresses the sentiments of the users. This difficulty of classification of polarity is solved by linking a random walk analysis over the WordNet graph with the SentiWordNet scores. Several experiments have been done for validating the technique and analysis has been provided based on the real time tweets. The Figure 1.1 shows how the data is processed in the approach.

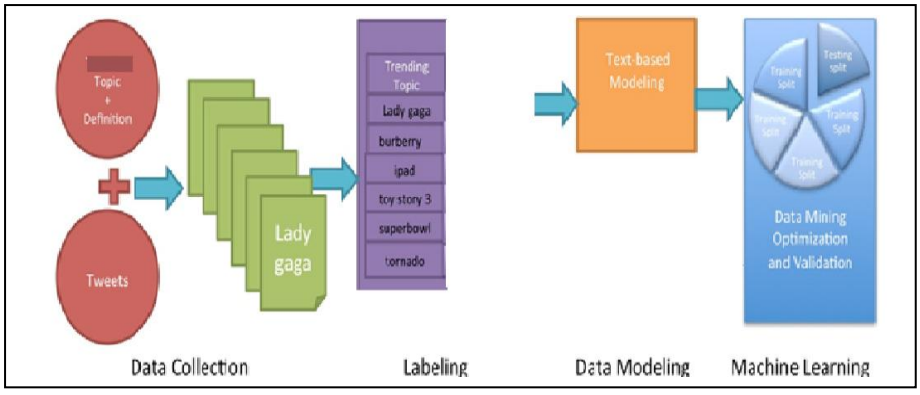


Figure 1.1: Overview of data processing

1.1. Extracting Topic Trends

In the earlier times, research had suggested that information processing system require accession to vast amounts of domain specific knowledge, in order to process natural language in human conversations or articles, just similar to what human beings would need to get the picture of the content of articles to provide a summary or a subject. A corpus of data is taken in Wikipedia in which one can easily categorize a textual matter into a topic. The Wikipedia contains pages that are extremely related to each and every individual topic, with a topic being an event, a thing, facts, a person, etc. The level of interest for a topic is associated to the degree of involvement of a topic in human discussions or conversations.

How human beings do condensed an article? According to their experience and background knowledge base, human beings condense the specific wording of a topic or an article. In contrast, information processing system will face difficulty in order to understand the contents of human conversations such as twitter streams or micro blogs.
Computing machine would require to process access and vast amounts of data and deduction of context, domain specific knowledge base and common sense is a highly large problem to handle. Apart of the processing, machine would need to learn in a same way to what a person would need in terms of education, background, speech, etc. such as the contents of articles, and interpret or summarize them into topics or categories in order to understand the information.

A distribution of words for a given topic is slightly unique or close to unique for each topic, and closely related topics would have similar distribution of words. These word distributions then would look in a same manner in human conversations, and matching them both would instruct us to relate a topic to a conversation.

1.1.1. Twitter

Twitter is a micro blogging website which has recently become very popular among the internet community in which individual have the ability to send updates in the form of 140 character long messages. Users update short messages called Tweets. Human beings often share their personal views and ideas on various topics, discuss present issues and

write about various different events of their lives. It is favored because there is no economic or political limitations and easily accessible by millions of people in different countries. Micro blogging platforms have become a rich pool of global opinions and sentiments as the number of users rise as shown in Figure 1.2.

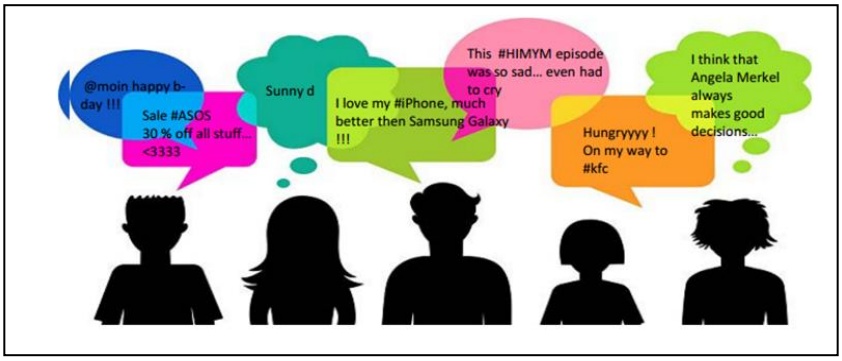


Figure 1.2: Users Opinion (Tweets)

Work done by [1] has shown that 19% of all tweets take mention of a brand of which 20% showed sentiment to the brand justifying it as the electronic Word Of Mouth [1] in the consumer market.

Twitter serves as a best platform due to its large user base from various socio cultural zones. Twitter contains a large number of twitter streams or tweets with millions added every day. These tweets can be easily collected which makes it easy to build a large training set.

1.1.2. Wikipedia

Wikipedia is perhaps one of the largest examples of open collaboration on the Web today. Its size, diversity, and the richness of the data records make it an ideal object of study for those wishing to investigate any aspect of collaborative work. One important aspect is the processes surrounding the removal of irrelevant content, and the relation this has to the retention of new collaborators.

Wikipedia, the 6th most visited web site in the world [2], is entirely community-created. The English-language version of the encyclopedia has 4.1 million articles and over 18 million registered users.

Like any other encyclopedia, Wikipedia has rules for what topics it deems worthy of inclusion. The site is very clear as to what it is not: Wikipedia is not a blog, a social network, or a directory, and it is definitely not "an indiscriminate collection of information". It is an encyclopedia, albeit one with more available space than a traditional one. Though it welcomes articles on topics not generally included in a print encyclopedia, it still maintains standards of notability to ensure that all topics covered are fundamentally *encyclopedic*. Namely, the "topic has received significant coverage in reliable sources that are independent of the subject" and that are easily verifiable. Additional specific notability guidelines are voluminous and extremely detailed, with policies governing the notability of academics, books, and organizations, among others. The ease with which articles can be created necessitates a method for removing the many unencyclopedic articles that are bound to result. Among Wikipedia's policies, therefore, are many governing the process of article deletion, a complex system that attempts to streamline the deletion process as much as possible while maintaining the site's collaborative ethos.

1.2. Sentiment Polarity Classification

The purpose of sentiment analysis is to determine the inclination or attitude of' a communicator through the contextual polarity of their speaking or writing. Their attitude may be reflected in their own judgment, emotional state of the subject, or the state of any emotional communication they are using to affect a listener or a reader. Simply put, it is trying to determine a person's state of mind on the subject they are communicating about. This information can be mined from texts, tweets, blogs, social media, news' articles or comments.

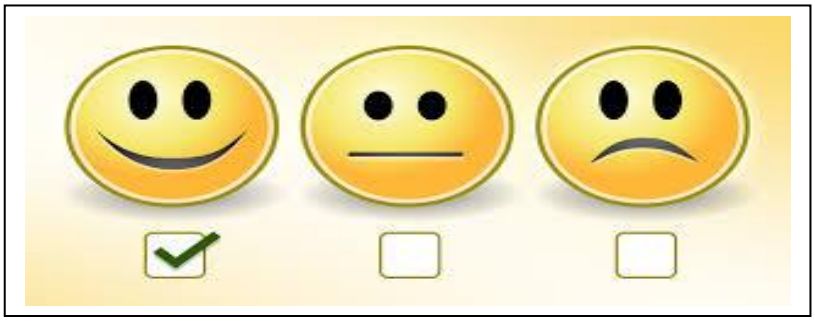


Figure 1.3: Sentiment Analysis

Sentiment Classification, a sub topic of Sentiment Analysis, is the study of computationally finding whether a text is negative or positive as shown in Figure 1.3. In sentiment classification, using machine learning, a classifier is required to be trained on a well labeled training set. This is called supervised learning. However, owing to its nature and the tweets that can be collected, it is a challenging task to label a training data set of such magnitude manually.

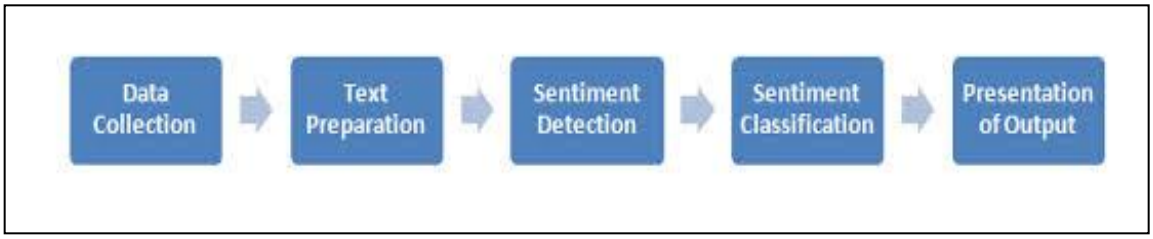


Figure 1.4: Flow diagram of Sentiment Polarity Classification

The Figure 1.4 shows the sequence of how the Sentiment Polarity classification takes place.Sentiment classification performed on user generated textual content finds many applications. For corporate, it can help in business decisions by knowing how users feel about a service or a product. Social organizations and political parties can collect feedback about their programs and legislation. Artists, musicians, cricketers, dancers and other entertainment icons can reach out for their fans and access the quality of their work. This broadly serves as an automatic polling system helps in relieving any manual intervention.

1.2.1. WordNet

WordNet is an electronic lexical reference system for English, designed in accordance with psycholinguistic theories of the organization of human lexical memory. In the form of an electronic database, this novel lexical reference system for English is being developed. Its design derives from psychological and linguistic theories about how lexical information is organized and stored in the memories of people who know English well and speak it fluently. The success of this experimental system would demonstrate the adequacy of the theories from which it derives, but even if those theories must be revised or replaced, the lexical database that is being developed in order to test them will be adaptable to a variety of practical applications. WordNet, supplemented on-line by

machine-readable dictionaries and made available via a multi-window workstation, can be profitably incorporated into any task that is facilitated by easy access to lexical information.

A concept is represented by a set of synonyms that can be used, in appropriate contexts, to express it; other semantic relations are represented by labeled pointers between the related concepts. WordNet will test the adequacy of current ideas about the structure of the lexicon by testing whether a realistically large sample of the English lexicon can be represented in this way.

The use of synonym sets is both an innovative and an expedient approach to dictionary design. Standard dictionaries develop uniform semantic representations for all the lexical items in English by systematizing the writing of sense definitions or by determining a set of linguistic primitives that constitute the meaning of lexical items. WordNet circumvents the writing and systematizing of sense definitions by representing concepts as relations among words arranged in a "vocabulary matrix", a giant network coding various relations by means of connections between words. It simply looks along a given row of the vocabulary matrix, notes all the words that can be used to express the same concept, and then substitutes that synonym set for the statement of the concept. If one accesses the dictionary by way of the horizontal word list, one gets a view of the polysemy of a word (all the different concepts that the word can be associated with). On the other hand, if one accesses the matrix from the vertical concept list, one gets a row containing all the different synonymous words that express a given concept.

1.2.2. SentiWordNet

SentiWordNet is an automatically generated lexical resource that assigns to each synset of WordNet a triplet of positivity, negativity and objectivity scores.

SentiWordNet is a general (or global) lexicon, i.e., its scores are deemed to be of general application regardless of the specific domain of the text which contains the terms to which the scores are associated. The hypothesis on global applicability of scores to terms may not hold on all the possible uses of terms in any domain, but from a practical perspective, it should hold for a large part of the cases. In SentiWordNet this hypothesis holds stronger because the application of scores to each distinct sense of a term allows to

discriminate a good number of otherwise sentiment-ambiguous cases, e.g., orange in the sense of the color or in the sense of the fruit.

For classification to be effective, every term has to be represented as having more positive inclination or negative inclination. Here, the SentiWordNet plays its role. The SentiWordNet is a document resource which contains a list of English terms which have been attributed a score of positivity and negativity. SentiWordNet provides this information which is extracted and matched to produce an overall score and hence prediction of the expression expressed in the document.

SentiWordNet is made up of tens of thousands of words, there meanings, part of speech represented and the degree of positivity and negativity of the word, ranging from 0 to 1. These words were all derived from the WordNet database, which is a database of english words and their meanings where terms are organized according to semantic relations or meanings. These words are all grouped by their synonyms into what is called synsets.

In its practical use, SentiWordNet can be considered a resource that provides a basic, wide-coverage, sentiment knowledge into the application in which it is used. Domain-specific applications will then likely pair it with a domain-specific resource in order to improve the precision on domain-specific terms.

SentiWordNet is an automatically generated resource and, as for any other automatic labeling process based on machine learning. With the release of SentiWordNet the related Web interface has been restyled and improved in order to allow users to submit feedback on the SentiWordNet entries, in the form of the suggestion of alternative triplets of values for an entry.

Generally, there could be 3 phases of sentiment classification: First involves extracting texts and words from a document being reviewed; Secondly, matching those words to a lexical resource, in order to determine the subjectivity score features of all terms in the document and generating a data set of these terms; Finally, using a classifier to analyze the sentiment of the document by training the classifier with the extracted data sets and testing the classifier to generate the desired knowledge.

1.3. Structure of the Thesis

The rest of the thesis is organized in the following order:

Chapter 2 - Provides literature review of extracting topic trends using Twitter and Wikipedia and sentiment polarity classification using WordNet and SentiWordNet.

Chapter 3 – This chapterdefines the problem and methodology used to solve it.

Chapter 4 - Gives a detailed introduction about Explicit Sentiment Analysis technique used to find the topic trends and WordNet graph to find the sentiment polarity classification using SentiWordNet.

Chapter5 – Provides the experiment performed for extracting the topic trends and to find its sentiment polarity and the results achieved.

Chapter 6 - Conclusion of thesis and suggestions for future work has been incorporated in this chapter.

Thesis concludes with references and publication.

Chapter 2

Literature Review

Studies in analysis of sentiment in 2002 to tackle the problem of Sentiment Analysis found lot of attention and interest amongst the research community for this field. Two major approaches were considered: the one created on the usage of Machine Learning(ML) methods and exemplified [3], and other based on the application of linguistic analysis [4]. These techniques are referred as supervised and unsupervised learning by many authors. The current technique of hybrid learning [5] exploits the advantages of both the approaches. A divergence is created in the study of long texts in Sentiment Analysis in which essential to conclude whether the data is objective or subjective before computing the polarity where one can assume that sole need is to calculate its polarity and text is considered as an opinion. A wide study on analysis of Sentiment can be found in [6].

The review of the related work section has been provided in the sequence of approach followed in the proposed method:

2.1 Extracting Trendy Topics

The topic modeling and semantic analysis in social media are widely used to understand textual data. Sentiment Analysis, content filtering, interest modeling, event tracking and many other applications are facilitated by social media. Zhao et al. [7] analyzed the differences in topics between traditional media and Twitter using Twitter streams - Latent Dirichlet Allocation (LDA) for examining short messages.

In [8] and [9] author tried to model the topics unceasingly over real time by conducting topic modeling of temporally sequenced documents in Twitter.

Chang *et al*. [10] used extensive human experiment to validate the significance of topic models. Probabilistic topic modeling is used to produce a latent topic demonstration due to its ease of use and its unsupervised analysis of text. It is a widespread tool for this type of computations.

Hoffman *et al*. [11] have demonstrated an efficient way of fitting streaming documents into 100-topic model.

Ritter *et al*. [12] demonstrated the analysis of large volume of tweets used in construction of a data-driven and conversational agent in order to learn the conversational structure using unsupervised topic model approach.

Markovitch and Gabrilovich[13] showed an innovative method where the concepts are derived from the concept space and wiki knowledge using Explicit Semantic Analysis (ESA). This approach uses co-occurrence of data to evaluate the connection of the twitter conversational space to wiki knowledge space.

Ramage*et al*. [14] use Labeled Latent Dirichlet Allocation (LDA) to studied microblogs into 5 main categories of topics, namely: substance, style, status, social, and other characteristics.

2.2 Polarity Classification

Sentiment analysis provides the scholars the aptitude to compute the sentiments in online data and analysis of sentiments is very useful for researchers. To perceive sentiment in text automatically the researchers have established algorithms in the field of sentiment analysis [15]. Some algorithms assigns a global polarity to a text, while some classify the objects (objective or subjective) and the other classify the polarity of opinions conveyed, such as movie reviews [15]. The linguistic analysis, lexicon-based and full-text machine learning techniques are the three commonly known techniques of sentiment analysis. The typical text features are the sets of all words, word pairs, and word triples originate in the texts. To predict their overall sentiment polarity, the algorithms that are trained go through the same features in the texts [16]. The list of words and their existence within texts are precoded to compute the final polarity of text in the lexicon approach [17]. In practice, algorithms often employ numerous approaches together with different enhancements, for instance pre cleaning the features explored, and techniques to handle with variations in text over time.

In addition to opinion polarity classification there are few algorithms which also detect the strength of sentiment in the text [15] and also some for casual online tweets [18]. These algorithms work on the foundation that human beings can distinguish among mild

and strong emotions in data. To indicate the strength of any opinion perceived, sentiment strength algorithms assign a numerical value to data.

In Twitter, there is the vast variety of topics and large amount of data present on which users expresses. In [19], researcher provides the first study on the polarity classification in tweets. A supervised classification study was conducted by the researchers in English. The emoticons in tweets were used by the researchers to distinguish amongst negative and positive twitter streams. In [20], the validity of this approach was verified. The researchers produced a corpus of negative twitter stream, with negative emoticons ":(", and positive twitter stream with positive emoticons ":)" in the twitter search APIs. For the classification of polarity in Twitter data streams, is used to study which classification algorithm approach are appropriate and which features are fit for the corpus. Naive Bayes, Maximum Entropy and SVM are the algorithms analyzed for polarity classification in [21]. These three algorithms provide the better results with the different features. They come up with some remarkable conclusions in their study, such as: In Twitter for the polarity classification, the POS-TAGS are not able to provide valuable information. Then represent tweets the simple use of unigrams provides better results when these results are compared to those attained in the polarity classification in large data. Finally, the grouping of bigrams and unigrams can improve the results as compared to simple use of unigrams.

In Twitter, a hybrid method is one of the very interesting studies performed by [21] that are applied for the classification of polarity

2.2.1 Modelization of a tweet

The use of emoticons have various controversial studies as a valid corpus of Twitter streams, to train a supervised sentiment classifier, Davidov et al. [22] used 50 hashtags and 15 emoticons as sentiment labels using K Nearest Neighbor (KNN) algorithm. The results obtained are promising and most importantly the experiments are verified by human juries.

On the other hand, for sentiment detection Twitter uses features like raw data (n-gram) representation for building a model in many of the studies. However, the use of syntactic features of the twitter streams and the inclusion of some information can improve the

result generated, as in [23]. On the basis of this previous work, SVM and General Inquirer are used as the study the target-dependent opinion classification of twitter streams by Jiang et al. [24]. The sentiments of the tweets are then classified as neutral, negative or positive.

According to the study conducted in the area of Sentiment Analysis in Twitter the various feature to be considered into account are represented by [25]. The study was conducted on the tweets labeled manually and also on the reduced group of corpus.

In [26], used reducing features an unsupervised method for Sentiment Analysis and their method are based on the Latent Dirichlet Allocation (LDA) methodology. The twitter streams are represented following the vector space model and using the TF-IDF metric to weight the terms after cleaning the corpus. The authors apply their proposal for the reduction of features once all the tweets are represented.

2.2.2 Random Walk Algorithm

A Random Walk is an accurate validation of a sequence that comprises of taking consecutive random steps. The Random Walks algorithms in Sentiment Analysis is based on the notion that if the process starts at a given word before hitting a word with a diverse polarity it is more likely to hit other word with the same semantic orientation. To follow this approach WordNet is needed as a lexical knowledge base, which especially in Sentiment Analysis (SA) and in Natural Language Processing (NLP) is used.

WordNet is a lexical knowledge base in which English terms are grouped into synsetsand a synset is an abstract depiction of a conception [27]. The labeled synonyms are the several terms that are originate in a synset. A polysemic is termed as when a term is related to more than one or several synset ids.

Furthermore, according to their syntactic role, synsets are interrelated with a variation of connections such as:

• **Nouns**: hypernyms, holonyms, hyponyms, coordination, and meronyms.

• **Verbs**: hypernyms, troponyms, coordination and entailment.

• **Adjectives**: related nouns, similar to and participle of verb.

• **Adverbs**: They inherit their structure from that of the adjectives and are defined in terms of the adjectives they are derived from.

A rich graph is constructed from this configuration. Figure 2.1 shows part of the WordNetsubgraph for the term 'Window', showing how numerous nodes are inter-related. To construct this graph, some hyponyms, hypernyms and synonyms have been used. This resource is a perfect applicant for analysis of graph algorithms because of the graph nature of WordNet. The graph nature of WordNet is used for experimentation by several researchers.

To compute the positive or negative polarity of words a Google PageRank [28] is used which is another form of Random Walk algorithm. In [29], the objective of the authors is to use a modified version of the original formula of PageRank in order to obtain a semantic ranking version of WordNet, to preference the random walk to negative as well as positive terms.

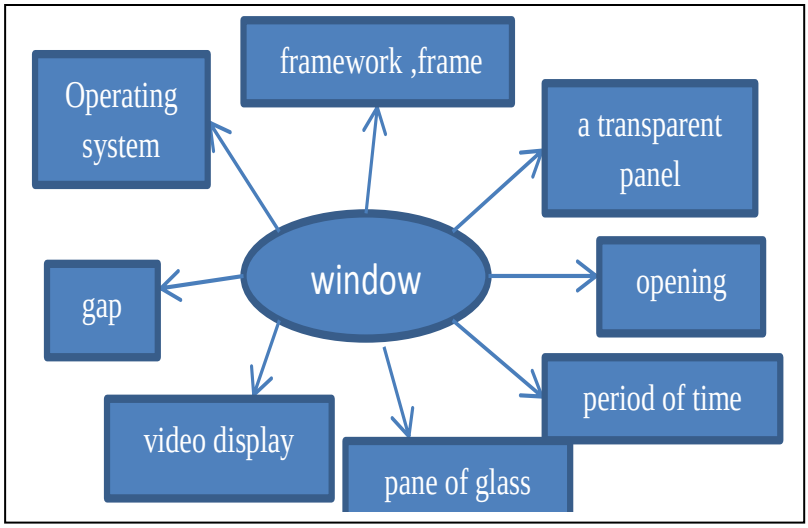


Figure 2.1: WordNetsubgraph for term 'Window'

2.2.3 SentiWordNet

SentiWordNet [30] is a lexical source on the basis of the well-renowned WordNet. SentiWordNet provides important information correlated to opinion inclination using synsets. A synset presents a "concept" that is disambiguous and also it is the elementary information in WordNet. The synsets are used as nodes over the lexical graph for most of the relations. The concepts of "negativity", "positivity" and "objectivity" is represented as a set of three scores for every synset by using SentiWordNet (the afterwards being

evluated from the two preceding ones). Therefore according to its subjectivity and polarity, every conception is weighted.

As SentiWordNet provides a field-independent resource for obtaining certain data about the amount of emotional charge of its concepts, this resources has been used by the sentiment mining community [31].

Table 2.1: Sample entries of SentiWordNet

Synsets	**Positive**	**Negative**	**Related Terms**
04587648	0.000	0.000	Window#1
00011665	0.125	0.375	Too-greedy#1 overgreedy#1
04670022	0.625	0.000	Reliableness#1 reliability#1
04125021	0.625	0.000	Safe#1

Table 2.1 offers an excerpt of SentiWordNet entries. Some terms receive a neutral weight within the resource that may be computed in certain circumstances. For example, term 'Window'in SentiWordNet is fully neutral.

Recently, in Twitter Sentiment Analysis [32] SentiWordNet has been also used. In this the researchers develop a method named Micro-blogSentimentAnalysisSystem (MSAS) to determine consumers sentiments. Firstly, the method implements a classification of subjectivity to differentiate objective and subjective twitter stream. The second step is to compute the sentiment polarity of the tweets. These twitter streams are gathered by an aggregation method of the polarity values of each and every term of the twitter stream. Then the researchers use SentiWordNet to obtain the semantic value, so if the twitter stream has more negative words then it is classified as negative, and if it has more positive words, the twitter stream is considered as positive.

Chapter 3

Problem Statement

3.1 Research Gaps

Presently the major focus of research in the area of sentiment polarity classification of trendy topics is in the efficiency of prediction of results and generating the exact outcome is still a difficult task. Although prevalent techniques have been successful in generating the expected results, but exact predictions of topic trends and polarity classification are important issues which still need a lot of consideration.

Based on the various challenges some of the research gaps have been identified:

1. **Efficient Corpus for extracting Topic trend**s: In majority of research related to topic trends, the corpus of twitter data streams are considered for topical analysis. But only twitter data streams could not predict the efficient results, as no particular topic can exactly be taken out just from the tweets. So for this some other corpus should be attached with twitter streams for performing topical analysis.
2. **Effective prediction of Sentiment Polarity:**To predict the sentiment polarity, the syntactic structure is used, but these syntactic structures are not efficient enough to predict the correct polarity of various words. So, a novel the approach is which should semantic analysis along with syntactic structures.

3.2 Problem Statement

The emergence of Web 2.0 has drastically altered the way users perceive the internet, by improvingcollaboration, interoperability and information sharing. With these emerging trends, it is important to know the current scenarios of the world i.e. a set of factors that can affect the consequences of an action.

The scenario analysis includes the process of analyzing or extracting the current topic trends and then analyzing its decision by considering alternative possible outcomes such

as positive, negative and objective in case of classification of sentiment analysis. It is designed to see the consequences of an action under different set of factors.

Scenarios should be feasible enough to provide an accurate picture of outcomes.Many scenario analyses use three different scenarios: neutral case, negative case and positive case. The expected scenario is the neutral case: if all things proceed normally, this is what the expected outcome will be. The positive and nagative cases are scenarios with more and less favorable conditions, but they are still confined by a sense of feasibility.

The process of scenario analysis is used to provide the pausible idea of what might happen, it is not used to identify the exact conditions, it is just used to approximate the results. An extraction of topic trends (using the Wikipedia topics and Twitter datasets) of the present world gives a better understanding of the current situation of the topic. Extracted topics when analyzed with the WordNet graph for sentiment polarity classification,using SentiWordNetwill give the correct picture of outcomes whether it is positive, negative or objective.

3.3 Objectives

• To study, analyze and explore existing techniques for polarity classification of the trendy topic in micro blogs.

• To propose a technique for extracting the trendy topics from microblogs and combining with one of the human knowledge base popular corpuses for analyzing the semantic polarity of these topics.

• To calculate the semantic polarity of the extracted topic on the proposed technique.

3.4 Methodology

The step-by-step methodology to be followed in analyzing scenarios is given below:

- Wikipedia topics and Twitter datasets are the corpuses used for extracting the latest topic trends.
- Random Walk Algorithm is applied using WordNet graph for removing the disambiguation of the words used in the tweets of the trendy topics.

- SentiWordNet is used to provide the positive and negative score to each synset id provided by WordNet graph.
- Once the scores are generated, polarity classification is done.
- The proposed technique has been tested for different trendy topics for calculating the polarity.

Chapter 4

Sentiment Polarity Classification of Trendy Topic

4.1 Introduction

Sentiment Polarity Classification of trendy topic is a way of analyzing the present scenario of the emerging world, telling us about all the pros and cons of the present topics or scenarios. It helps in examining possible alternative outcomes and by this predict the possible future events. It can certainly be considered as a main method of projection and does not the exact picture of the future. Current work pesents the practical approach for detecting the topic trends, the two important datasets are used :Twitter datasets and Wikipedia topics to recognize the topics of present day scenario. Once a topic is discussed in twitter streams it will definitly be serached on wikipedia, so the count of wikipedia topics and the number of times topic discussed in twitter streams helps us to extract the present day topic trends.

How people percept the topic and what are the outcomes of that topic is the important question that arises. In order to know the one exact picture or trend of that topic, polarity classification plays an important role. WordNet graph play an important role in removing the ambiguity of the words by providing different synset id for each word. Then SentiWordNet is used to get the positive and negative score of synset ids generated.This entire process helps in estimation of the final polarity of the trendy topics.

4.2 Solution Design

The Figure 4.1 provides the overall view of the proposed method.

4.2.1 Flow Diagram of Proposed Work

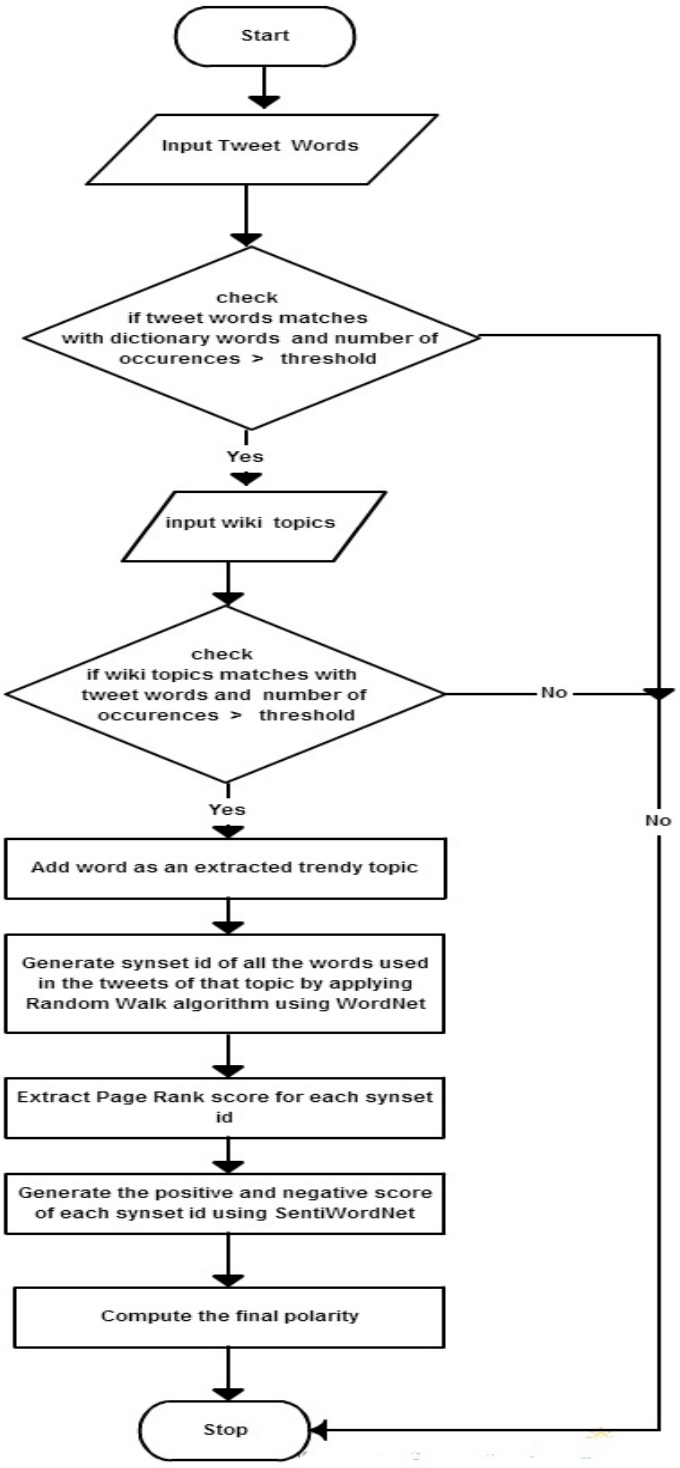


Figure 4.1: Flow Diagram of Proposed Work

4.2.2 Algorithm of Proposed Work

Input: Trendy tweet words

Output: Polarized value of semantic tweet word

```
For each word 'w' in tweets do
    Match with dictionary words
    If match 'w'=true
        tweet words = tweet word +1,
    End if
End for
For each word 't' in Wikipedia topics do
    Match with dictionary words
    If match 't' = true
        Wiki words = wiki words +1
    End if
End for
For each 't' in tweet words do
    Count the no. of occurrences of word in 't'
    If occurrences > α (matching threshold)
        Add word to extracted topic
End For
For each word of trendy topic tweets
    Apply random walk algo
    get synset id using WordNet
      For each synset id
          Extract the pagerank score
          Extract the positive and negative score using  sentiwordnet
          Compute the result for each word
      End For
End For
Compute the average of result of each word
Estimate the polarity
```

4.3 Extracting Topic Trends

4.3.1 Overview of the Proposed Model

The model proposes a technique to detect topics and their connectedness by extracting human beings curated topics from Wikipedia, and then accomplishing semantic analysis over both the twitter data and page content for those topics to correlate tweets or twitter stream to topics. A directed graph is generated for values of tweets amount produced per topic, based on what topics have in similar on their related conversations on twitter. Grouping that topic graph would allow to cluster the topics related by their usage.

The overview of the system proposed is as depicted in figure 4.2.

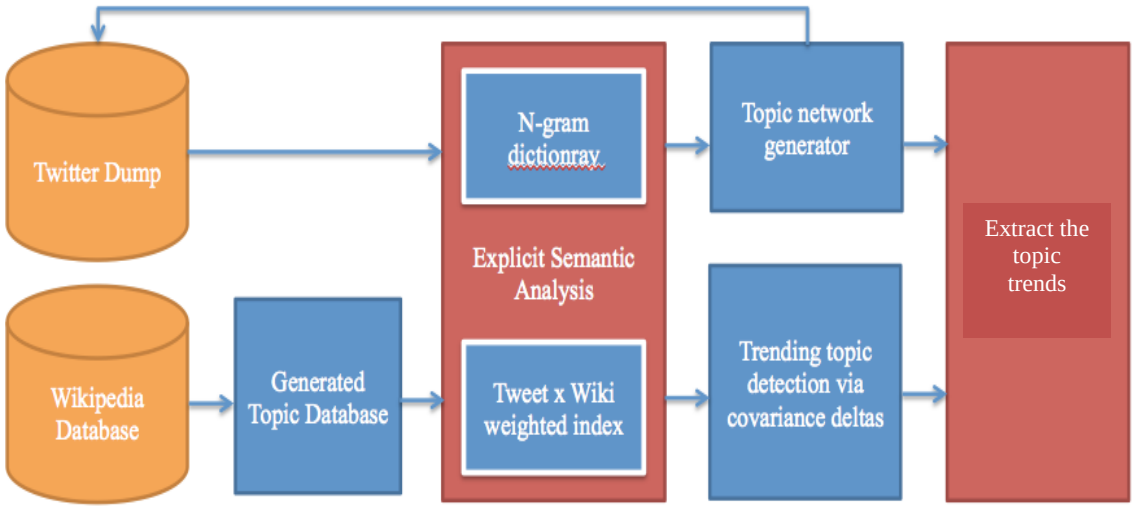


Figure 4.2: Architecture diagram of topic extraction method

4.3.2 Topic Formulation based on Wikipedia.

Wikipedia is a tremendous source of topics since information is neatly organized into a one topic per page, including separable pages for each meaning of vague keywords. Another stimulating concept is that events, people, companies, subjects, places, etc. are each packed into a distinct page closely defining a topic. Given this structure, go through

the modest approach of making each title of a disambiguated page from Wikipedia a distinct topic for our program.

4.3.3 Data Dumps and indexing.

A database dump for the Corpus of Wikipedia metadata composed from the Wikimedia archive. It consists of all the categories, pages, and links, with nearly 32.93 million records. Topics are ranked and indexed by their view count, and to simplify data analysis, the top articles built on ranking are crawled, detect the concrete page content for the topic.

Stanford SNAP's dataset collection by Jure Leskovec [33], comprising of approximately 476 million twitter stream has been used as the corpus of twitter streams.

4.3.4 Lexical Analysis and collection of n-gram dictionary

The group of a dictionary of words to match tweets and wiki pages along with n-gram dictionaries has been compiled from this list consists of approximately 7400 words by Alex Davies [34], explicitly created for analyzing Twitter, In making the most use of the data it comprises of significant common words.

Two matrices are generated with the total of n-grams found in each dataset, one with counts on tweets or twitter streams by day, and other counts for each wiki page.
The values of the tweet matrix (p_i^m) have been normalized corresponding to day 'm' and word 'i', as the tweet sample count differs for each day, and this was calculated as:

$$p_i^m = \frac{\#\ of\ times\ \ word\ \ i\ appeared\ \ \ on\ day\ m}{\sqrt[2]{\sum_k (\#\ of\ times\ \ word\ \ k\ appeared\ \ \ on\ day\ \ m)^2}} \quad (1)$$

The wiki matrix was normalized by corresponding topic n to words i, to adjust for length of page content and calculated as:

$$q_i^n = \frac{\#\ of\ times\ \ word\ \ j\ appeared\ \ \ on\ topic\ \ n}{\sqrt[2]{\sum_k (\#\ of\ times\ \ word\ \ k\ appeared\ \ \ on\ topic\ \ n)^2}} \quad (2)$$

4.3.5 Generating relevance using Explicit Semantic Analysis to link tweets to wiki pages.

In order to evaluate the matrices generated for both Wikipedia and twitter Explicit Semantic Analysis has been performed. Output ($p(topic_n|day_m)$ generated is:

$$p(topic_n|day_m) = \sum_{k=1}^{i} p(word_k|day_m) * p(word_k|topic\ _n) \quad (3)$$

$$= \sum_{k=1}^{i} p_i^m * q_i^n \quad (4)$$

where i is the total number of words in the lexicon space that are analyzed. This generates a new matrix of Tweet Days × Topics where the value for each is the lexical match of both datasets.

The word count generated for each value indicates how much tweets were in twitter for a given topic on each day.

4.4 Sentiment Polarity Classification

The flow diagram for Sentiment Polarity classification is as shown in Figure 4.3. This shows the description of how polarity classification takes place in following steps:

Figure 4.3: Steps of Sentiment Polarity Classification

4.4.1. Twitter Corpus

Although polarity study on micro blogging is one of the most interesting and significant task, there are very scarce permitted resources. One of the most useful resources for experimental analysis of the proposed methodologies for Sentiment Analysis is the twitter corpus used worldwide by the scholars. One of the major research efforts in this

area is categorization of the data, i.e. one must know which document express a negative opinion and which of them a positive one. It is a well- known fact that in microblogging sites like Twitter a boundless amount of data is issued every second, so the expansion of a descriptive manual labeled corpus is hard, inefficient and not reasonable for a small cluster of people. In [35] for Sentiment Analysis experimentations authors provide a better and illustrative Twitter corpus. One of the major drawbacks is that the researchers did not share Twitter Corpus as there is a limitation of the terms and services of Twitter.

A labeled corpus is generated for the evaluation of supervised polarity study approach. The main feature is that the size of the microblogs must be 140 characters so that users can easily express their feelings and thoughts in lesser words.

On contrary, in twitter the language used has some different characteristics which decided not to be involved for this study.

These characteristics are:

1. Retweets
2. Mentions
3. Links
4. Hashtags

Before feeding the twitter streams as a raw data to the sentiment analyzer there is a need to preprocess the tweets. For cleaning of the twitter data the following filters are applied:

1) **New lines removal**: Remove all the new line symbols as the users write their messages in two or more different lines.
2) **Emoticons with opposite senses**: Sometimes users write both positive and negative smileys in the same single line.For example:

   > @Harry Styles I have all day to try get a tweet off you:) when are you coming back to dublin i missed you last time, I was in spain:(

   Then the tweet has both positive and negative emoticons which ambiguous the impact of polarity in the message so remove this types of emoticons.
3) **Letters Repeated**: To highlight the messages users often repeat the letters several times. For example:

Blood drive todayyyy!!!!!:) Everyone donateeeee!!

So this is a problematic situation as when there is a repetition of letters it will make the word dissimilar. To remove this problem, remove the count of occurrences. When there are more than two times of occurences take only two occurrences. The above example would be converted into:

blood drive todayy:) everyone donatee!!

Figure 4.4 show the procedure followed in order to create our Twitter corpus.

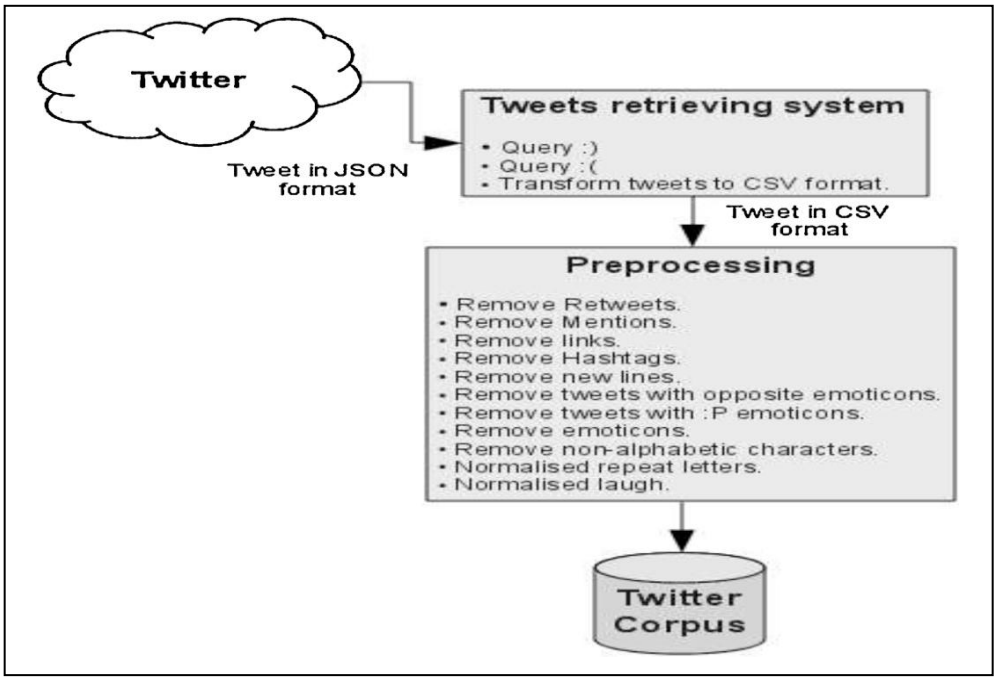


Figure 4.4: Twitter Corpus Workflow

4.4.2 Description of Polarity Computation

For polarity classification most of the proposed methods calculate a value of negativeness or positiveness. A neutrality value is produced by some of the twitter stream. Consider the following measurement of classification of polarity (which is very common): a real value in the interval [−1, 1] would be adequate. A positive polarity expressed in the tweet if the value is above zero, instead correspond to negative polarity when the value is below

zero. If the value closer to the zero, the more neutral will be the post. Therefore, a method for classification of polarity could be representing as a function *p* on a text *t* such that:

$$p : t \rightarrow \mathrm{R}$$

such that $p(t) \in [-1, 1]$.

4.4.2.1 Calculating the final estimation

It is significant that the final calculation points to analogous values, as a grouping of SentiWordNetscores with random walk weights. To this end, the weights associated after the random walk procedure to synsets are normalized so vectors of "concepts" sum up the unit as extreme value. The final polarity score is obtained by the product of this vector with related SentiWordNet vector of scores, as expressed in Equation (6).

$$p = \frac{r.s}{|t|} \tag{6}$$

where pis the score computed, **r** is the vectorcomputed by the random walk algorithm of weighted synsets of the tweet text over WordNet, **s** is the vector of polarity scores from SentiWordNet, and *t* is the set of conceptsderived from the tweet.

The final polarity score computed using Equation (6) is the average of the product of the random walk algorithm provided weights and difference of positive and negativeSentiWordNetscores, and, as revealed in Equation (6). So we can say that polarity (p) can be defined as:

$$p = \frac{\sum_{\forall s \in t} rw_s * \left(swn_s^+ - swn_s^-\right)}{|t|} \tag{7}$$

where *s* is a synset in the tweet or twitter stream *t*, $rw_s$is the synsets weight when using WordNet random walk is used, $\mathrm{swn}_s^+$and $\mathrm{swn}_s^-$are positive and negative scores for the synset s retrieved from SentiWordNet.

4.5 Implementation Details

4.5.1 Trendy Topic Extraction

Based on the ESA matrix generated, trending topics are extracted from the twitter dataset and Wikipedia topics. Three corpuses have been used for experimental purposes: dictionary words, tweets and Wikipedia topics. Threshold value is provided by the user to get refined results and maximum accurate topic and the default threshold is 1.

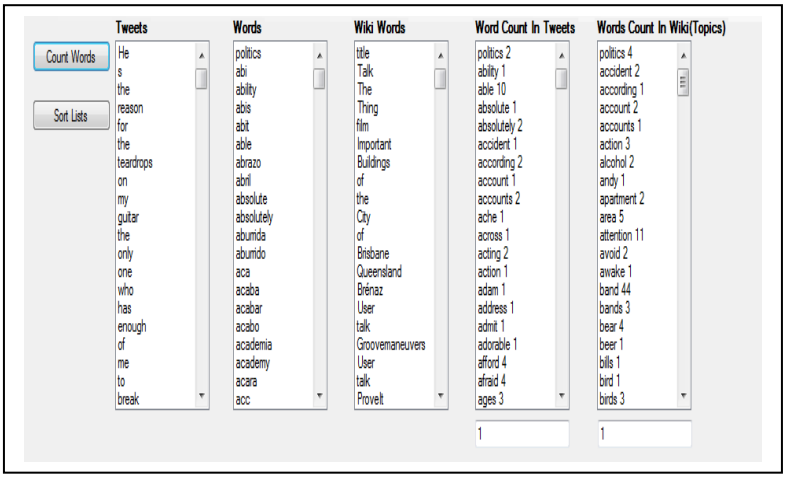


Figure 4.5: Extraction of trendy topic

The Figure 4.5 shows the trendy topic with the number of times of their occurrences.

4.5.2 Polarity Classification

Let the trendy topic is ‘windows’, so initial step is to extract all the tweets of windows and classify whether this topic is popular because of its positive impact or negative impact. The experiment not only classifies the trendy topics but it will also classify the polarity of non-trendy topic or any topic.

This example considers one tweet and shows how the results are computed or evaluated. The tweet or twitter stream to be processed is:

> Using Linux and loving it - so much nicer than windows... Looking forward to u wysiwyg latex editor! :)

The text to apply in the following step is shown after cleaning the tweet:

Using Linux and loving it so much nicer than windows Looking forward to using the wysiwyg latex editor

In this one needs a synset id, which is provided by WordNet. The synset id by WordNet for the words used in the tweets is shown:

using → 0115887-v
loving → 01463965-a
nicer → 00984333-a
windows → 04587648-n
Looking → 02133435-v
forward → 00075442-r
using → 01158872-v
latex → 15006118-n
editor → 10044879-n

After getting the synset id now for disambiguation process pagerank values are required. The pagerank value of the twitter stream is:

01158872-v:0.003819
01463965-a:0.004263
00984333-a:0.004060
04587648-n:0.005603
02133435-v:0.002388
00075442-r:0.013473
01158872-v:0.003819
15006118-n:0.007688
10044879-n:0.011033

Weighing the SentiWordNet polarity score of each synset id with its PageRank value or score is the final/last step. The difference between positive and negative scores is received as a result of the polarity score. The calculation for the above Pagerank value is:

[01158872-v] (0.000 - 0.000) * 0.003819 → use#1...
[01463965-a] (0.750 - 0.000) * 0.004263 → loving#1...
[00984333-a] (0.000 - 0.375) * 0.004060 → nice#4...
[04587648-n] (0.000 - 0.000) * 0.005603 → window#1...
[02133435-v] (0.000 - 0.000) * 0.002388 → look#2...
[00075442-r] (0.000 - 0.000) * 0.013473 → forward#3...
[01158872-v] (0.000 - 0.000) * 0.003819 → use#1...
[15006118-n] (0.000 - 0.000) * 0.007688 → latex#1...
[10044879-n] (0.000 - 0.000) * 0.011033 → editor#1...

The method assigns the tweets of trendy topic to the positive class as the final polarity score is 0.000186.

Chapter 5

Implementations and Result

Three inputs have been considered for analysis of extraction of trendy topics:

1. Twitter stream or tweets (Twitter datasets)
2. Wikipedia Topics
3. Dictionary data or words

A comparative study of these three input files has been used to analyze the extraction of trendy topics. The twitter stream is refined and every word in a tweet is taken as a single tweet word as "tweets". Then these words are compared with every dictionary word, the match words are extracted and taken as total word count in tweets. Further, tweet words are match with the Wikipedia topic words then finally the matched words are extracted and considered as a trendy topic if they successfully accomplish both the threshold conditions. The initial threshold condition is 1.

The structure of the topic extraction interface is as shown in figure 5.1:

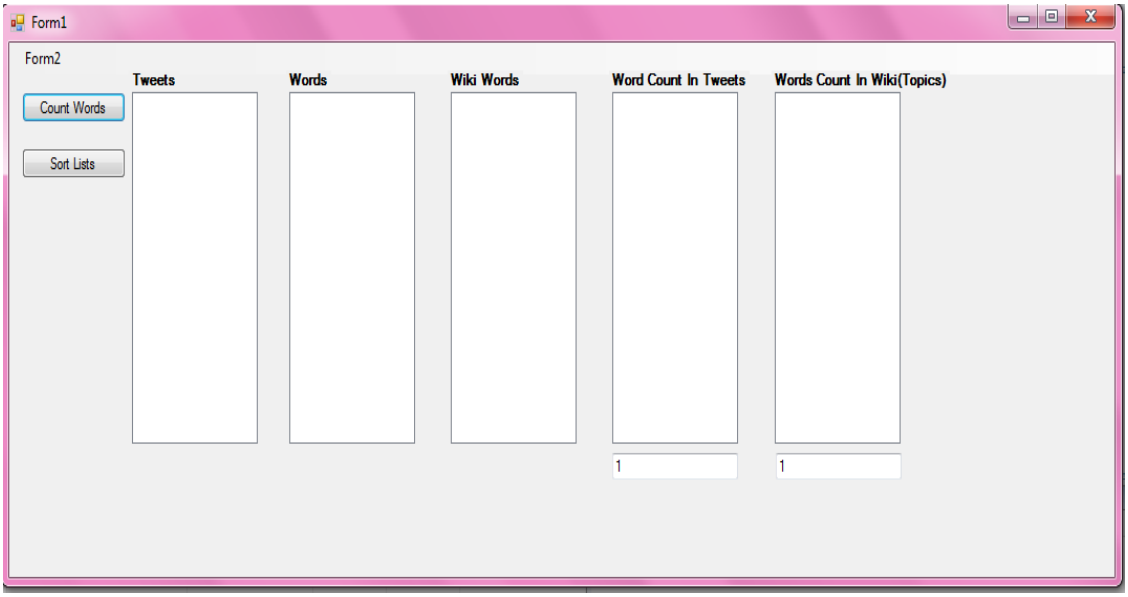


Figure 5.1: Structure of topic extraction interface

By clicking on "Count Words" it will input all the three files and count the number of occurrences then output the list of topics extracted from the input files.

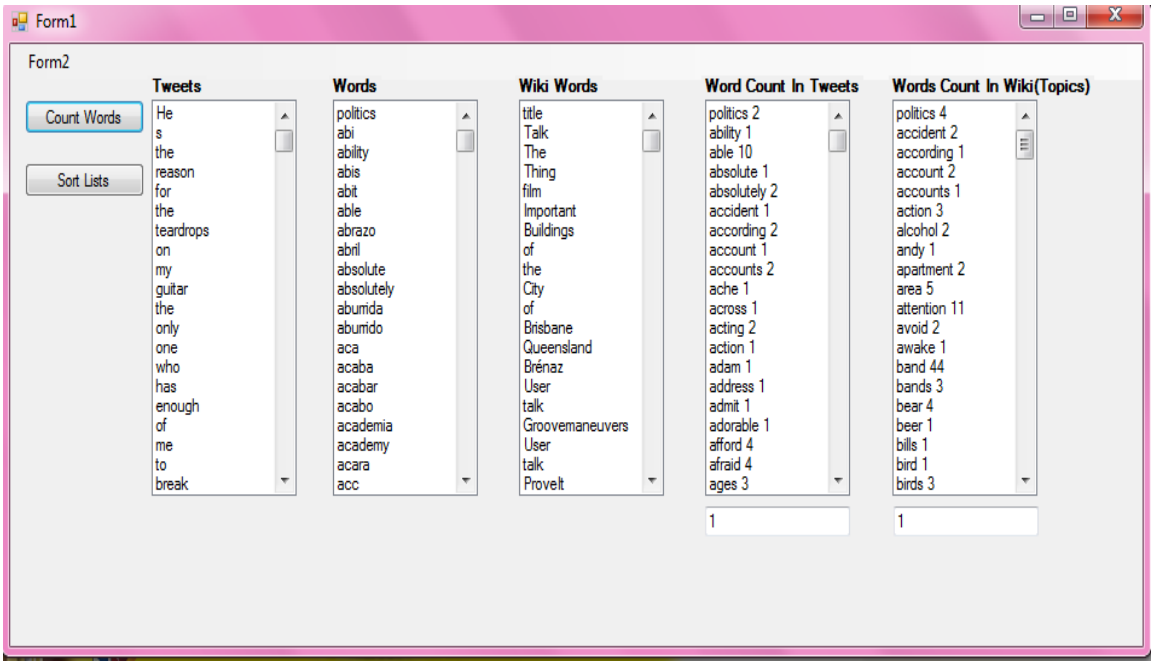


Figure 5.2: Trendy Topic Extraction

In figure 5.3, one can also give the threshold value which can which will refine the search and provide the more accurate topic trends. This helps in extracting more trendy topics.

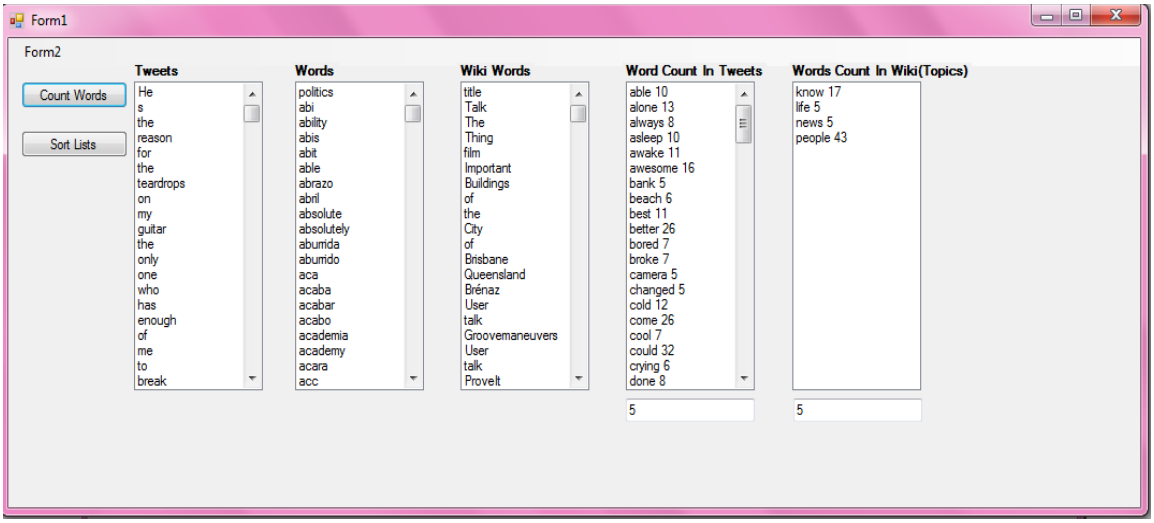


Figure 5.3: Topic Extraction using threshold values

The list of all the words such as tweet words, dictionary words, Wikipedia topic words and extracted words all these can be sorted in alphabetical order. By clicking on “sorted list”, the entire lists are available in alphabetical order.

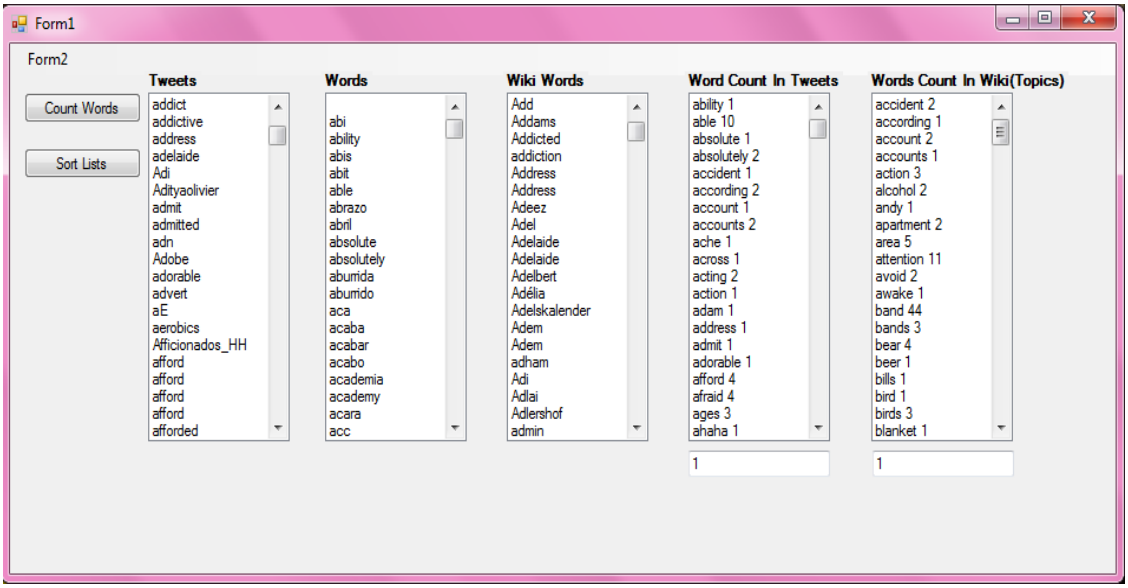
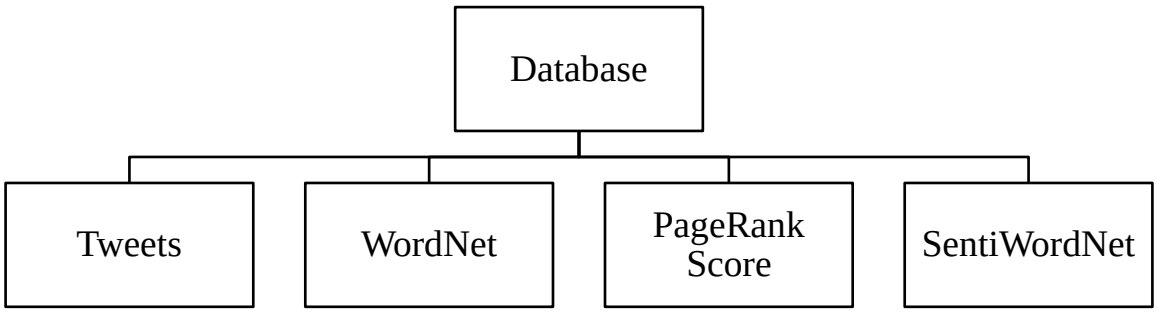


Figure 5.4: Sorted List

As seen in Figure [5.4], provides us with sorted list of extracted trendy topics.

Now to know the sentiment polarity classification of these extracted topics the four tables are used as a database. They are:

1. Tweets
2. WordNet
3. PageRank Score
4. SentiWordNet



The data of the tables are split into various numbers of files. To gather the large number of data of the files, the following program is used as shown in figure 5.5.

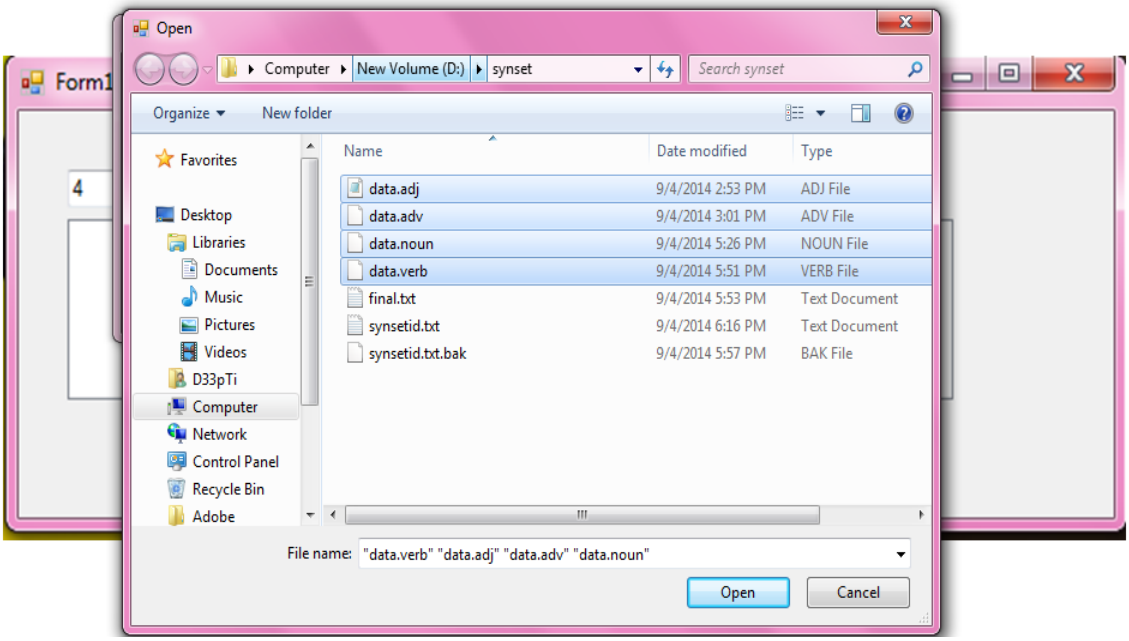


Figure 5.5: Upload the files

In this WordNet table consists of four types of words i.e. noun, verb, adjective, adverb contains in different files. These files will easily be gathered with the help of program as shown in Figure [5.5].

When the files are uploaded, the output file will be saved in the desired mention location and it will also show the total number of files combined as shown in figure 5.6.

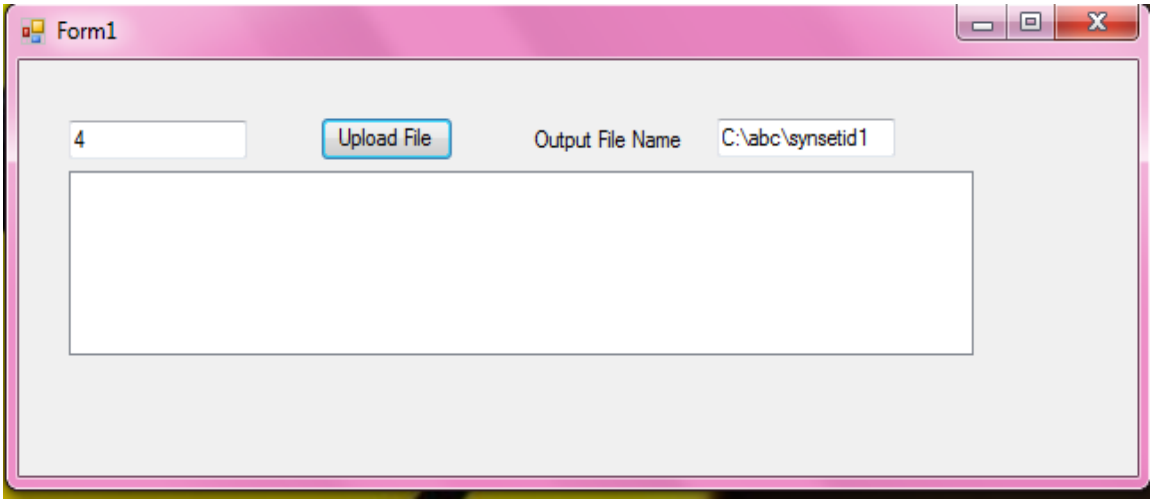


Figure 5.6: Combining the data

As shown in figure 5.6, location of the generated output file is shown i.e. the directory where the output file will be stored is mentioned along with the number of files uploaded.

Initially the table of Tweets is considered. This table provides the words used in the tweets of extracted topic.

Table 5.1: Tweets

ID	TWEETS
51	"@localtweeps Wow, tons of replies from you, may have to unfollow so I can see my friends' tweets, you're scrolling the feed a lot. "
52	our duck and chicken are taking wayyy too long to hatch
53	"Put vacation photos online a few yrs ago. PC crashed, and now I forget the name of the site. "
54	I need a hug
55	"@andywana Not sure what they are, only that they are PoS! As much as I want to, I dont think can trade away company assets sorry andy! "
56	@oanhLove I hate when that happens...
57	"I have a sad feeling that Dallas is not going to show up I gotta say though, you'd think more shows would use music from the game. mmm"
58	Ugh....92 degrees tomorrow
59	Where did u move to? I thought u were already in sd. ?? Hmmm. Random u found me. Glad to hear yer doing well.
60	"@BatManYNG I miss my ps3, it's out of commission Wutcha playing? Have you copped 'Blood On The Sand'?"
61	just leaving the parking lot of work!
62	The Life is cool. But not for Me.

Next the WordNet table is used to provide the synset id for each word used in tweets. These synset ids help us to remove the ambiguity of words. The Table 5.2 is as shown:

Table 5.2: WordNet

ID	WORD
223864	scrofulous
224041	unlovely
224135	unsightly
224254	bellied
224367	big-bellied
224465	bellyless
224561	banded
224711	unbanded
224812	belted
224967	banded
225096	belt-fed
225174	beltlike
225275	unbelted
225394	beneficent

Once the sysnet ids are generated (Table 5.2) the PageRank score is used. PageRank score of each of each word on the basis of its synset id is given in the Table 5.3.

Table 5.3: PageRank Score

ID	SCORE
02186338-a	.150775
02183611-a	.107604
00024073-r	.0536007
04424418-n	.0496565
00007846-n	.00302033
00931467-v	.0030184
06482401-n	.00205897
02200035-a	.0015648
14429985-n	.00121548
04928903-n	.0011718
02676054-v	.00109088
13597585-n	.000936668
13745420-n	.000866701
00106921-r	.00082903

Finally table containing the positive and negative score of the word using SentiWordNet is generated. It provides the score of each word using its synset id byWordNet as shown in Table 5.4.

Table 5.4: SentiwordNet

ID	POSSCORE	NEGSCORE
147397	0	0
147528	0	.125
147659	0	.625
147734	.625	0
148078	.625	.125
148642	.625	0
148852	.125	0
149120	.625	0
149262	.625	0
149461	.25	0
149686	.125	.25
149861	.25	.625

Table 5.5 provides the summarized computation of each word in the form of a view. This view is created using all tables generated in the previous steps. It computes the estimation of each word as shown in Table 5.5.

Table 5.5: Calculated Score

Synset id	Pos score	Neg score	PageRank score	Calculate
3553	0	0	.0000827547	0
130347	0	.25	.000416696	-.000104174
130347	0	.25	.000416696	-.000104174
130347	0	.25	.000416696	-.000104174
130347	0	.25	.000416696	-.000104174
130347	0	.25	.000416696	-.000104174
673766	.125	0	.0000714286	.000008928575
334996	0	0	.0000802188	0
334996	0	0	.000186293	0
334996	0	0	.000186293	0
26137	0	0	.00155726	0
26137	0	0	.0000654657	0

Now sentiment polarity is computed using an interface shown in figure 5.7.

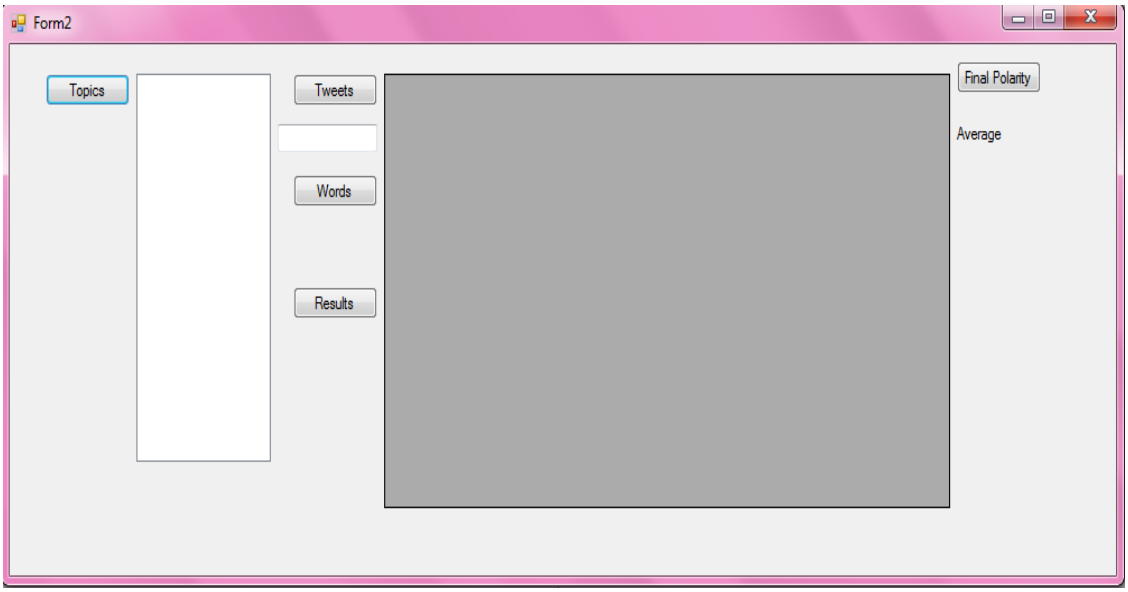


Figure 5.7: Interface for Computing Polarity

Figure 5.8 shows that by clicking on the button “Topics” all the trendy topics which are extracted (as shown above Figure 5.2) become input (by default).

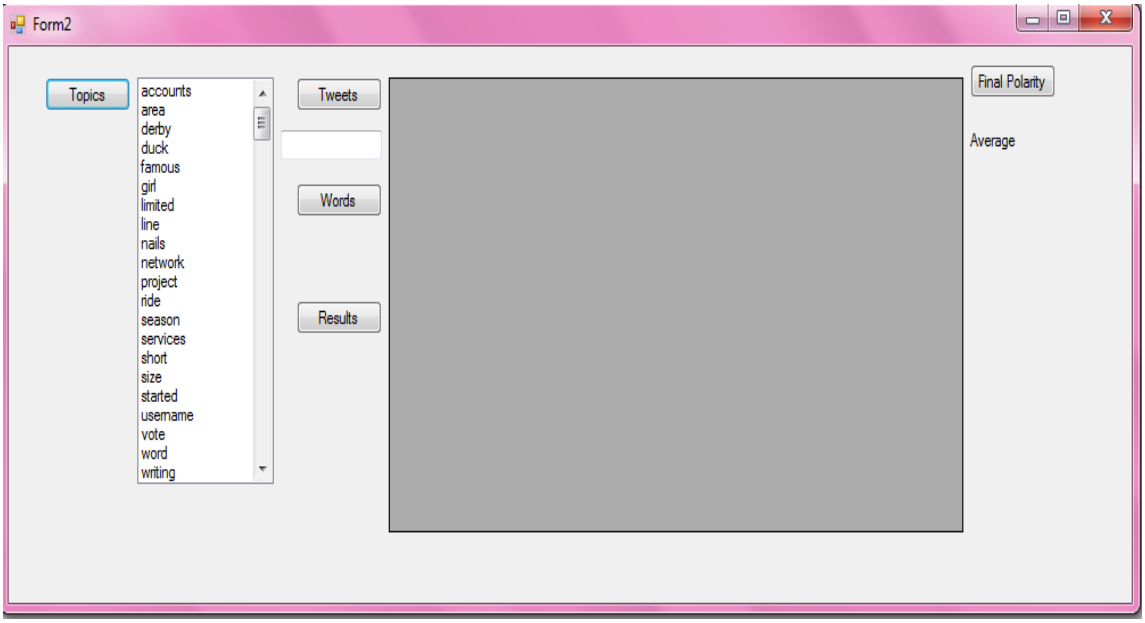


Figure 5.8: Trendy Topics

Once the user selects a particular topic from the listed topics, it become the trendy topic for which one has to calculate the final sentiment polarity.

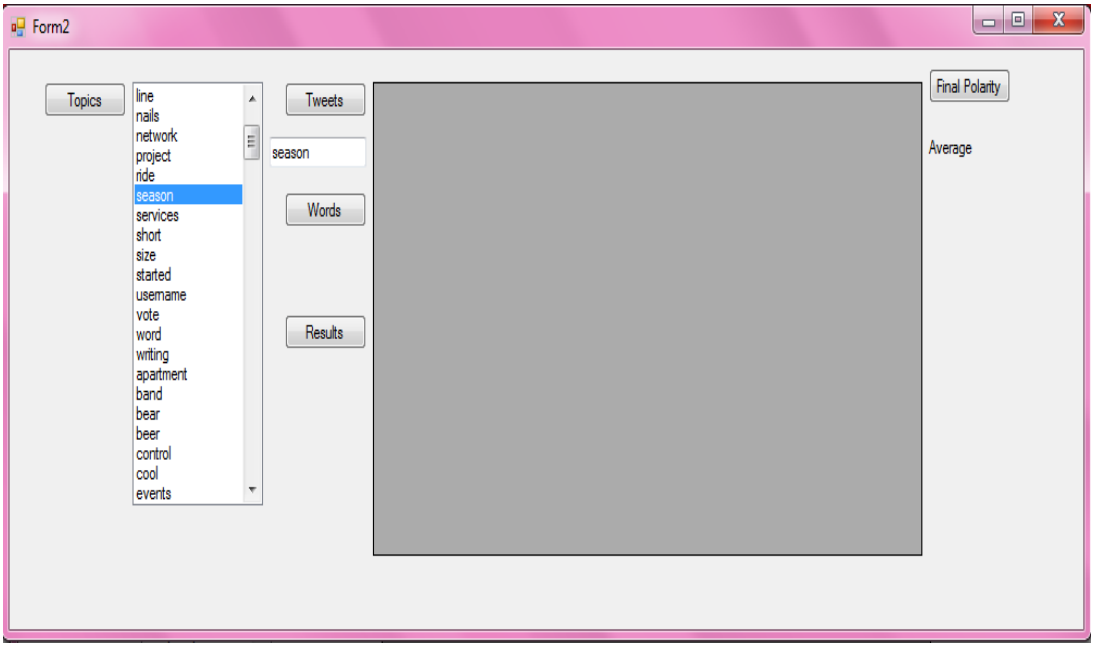


Figure 5.9: Inputted Trendy Topic

“Tweets” button as showm in Figure 5.9 will count the total number of tweets present in the database of the extracted topic. By clicking on this button it will display the total number of tweets of the extracted trendy topic as shown in figure 5.10.

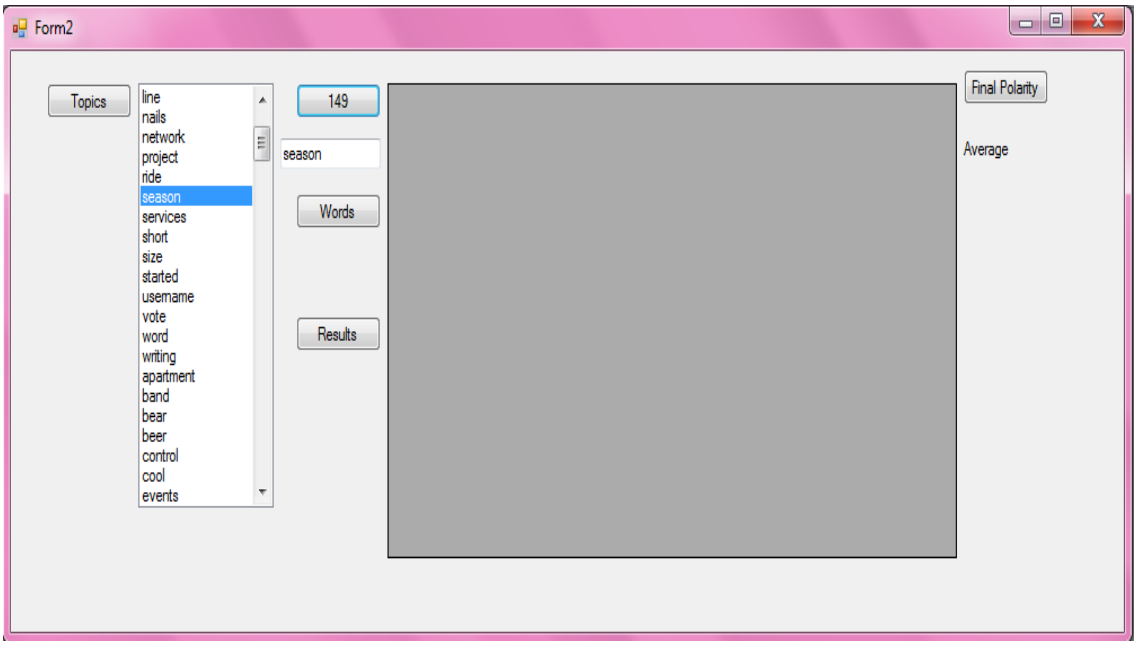


Figure 5.10: Number of Tweets

After getting all the tweets of the selected trendy topic, the total number of words in all the tweets of that topic are extracted. So when one clicks on the "Words" button, it will create a file of all the words in the database.Then these words are used for the further computations.

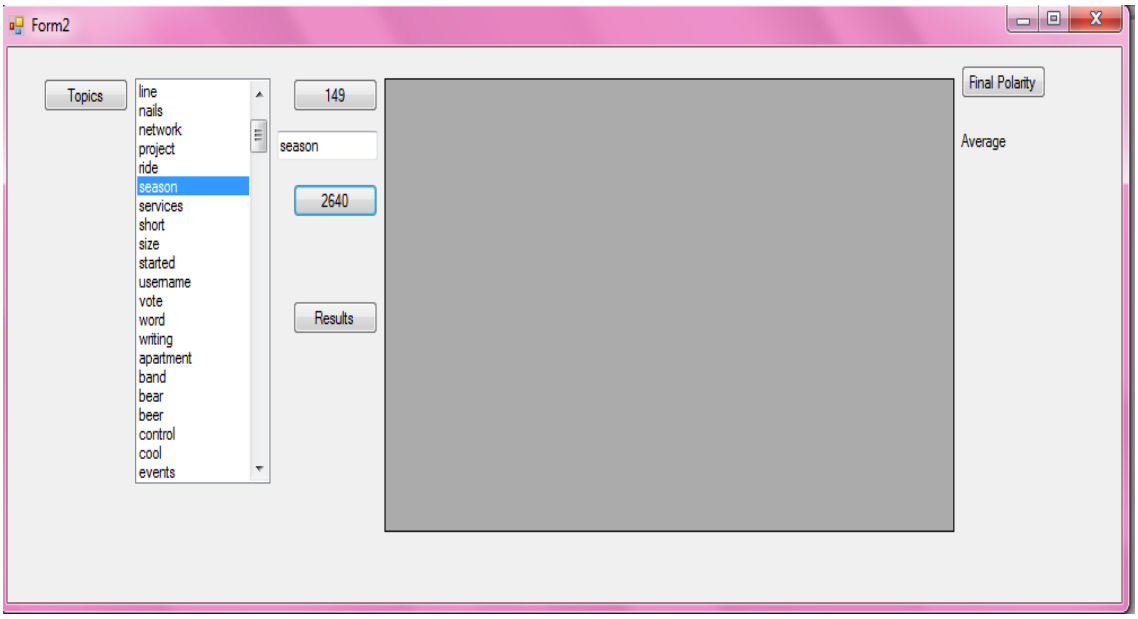


Figure 5.11: Number of Words

Now all the words obtained (as shown in above figure 5.11) are used to obtain the synset id of each word using the WordNet table shown above in Table 5.2 . Then using those

synset ids its PageRank score is found from PageRank Table 5.3, positive and negative score of a word is extracted from SentiWordNet (table 5.4) . The final polarity of each word is computed by clicking on the "Result" button as shown in the figure 5.12.

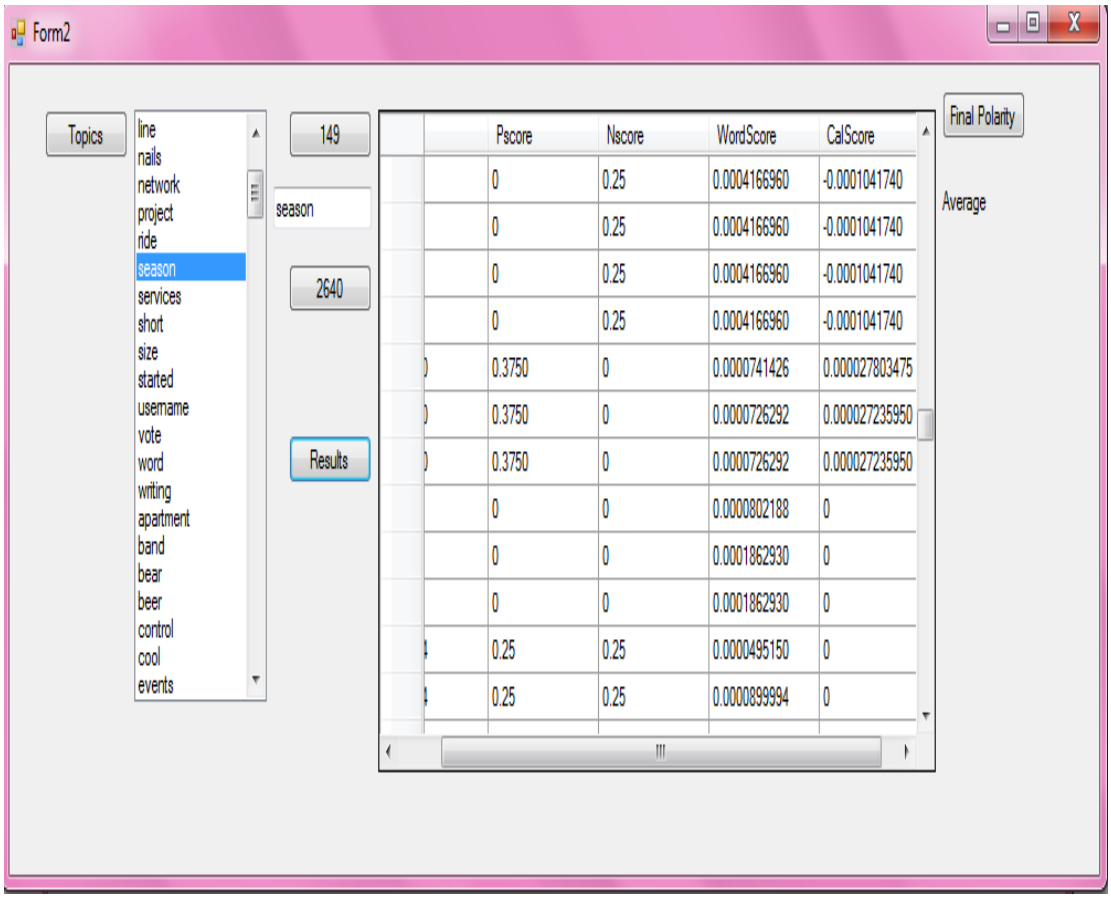


Figure 5.12: Computed Score of each word

Now, when one clicks on the button "Final Polarity" it will provide the average of the result of all the tweet words present in the tweets of extracted topic. The final polarity estimation is: if the average of the words calculated score is greater than zero then the sentiment polarity of the word is positive i.e. the extracted topic has its positive impact and when the average is less than zero then the sentiment polarity is negative i.e. the extracted has its negative impact and if the computed score is zero then the topic has its neutral impact.

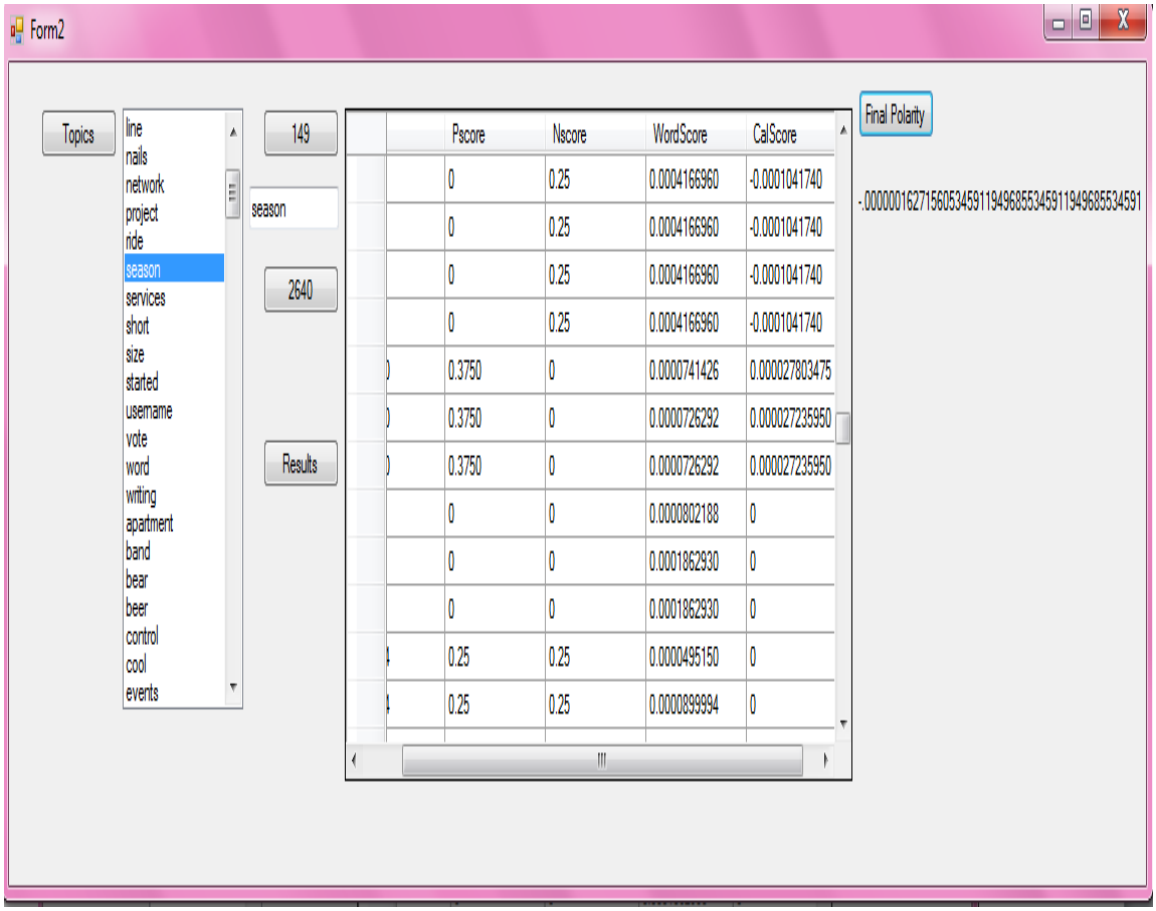


Figure 5.13: Final Polarity

As seen in the Figure 5.13, for topic ‘season’ the final polarity estimated is less than zero which clearly concludes that the topic has its negative impact so, with the final polarity score is -0.000000162 word ‘season’ is assigned the negative class.

From this method we can also compute the sentiment polarity of any topic. In this one just has to enter the name of the topic in the textbox as shown in figure 5.14.

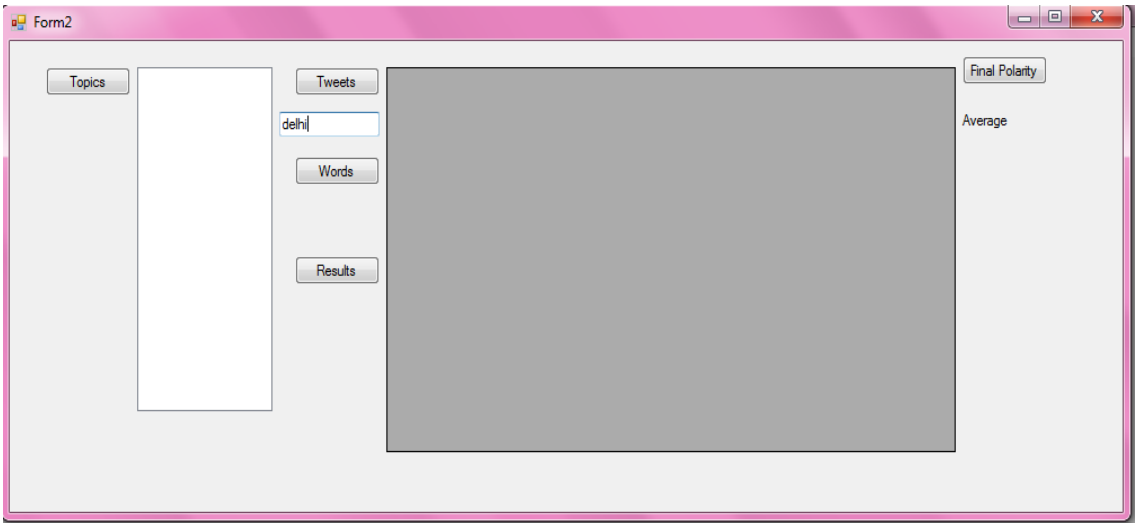


Figure 5.14: Anonymous Topic

By clicking on “Tweets” button it will count the total number of tweets of that topic in the database as shown in figure 5.15.

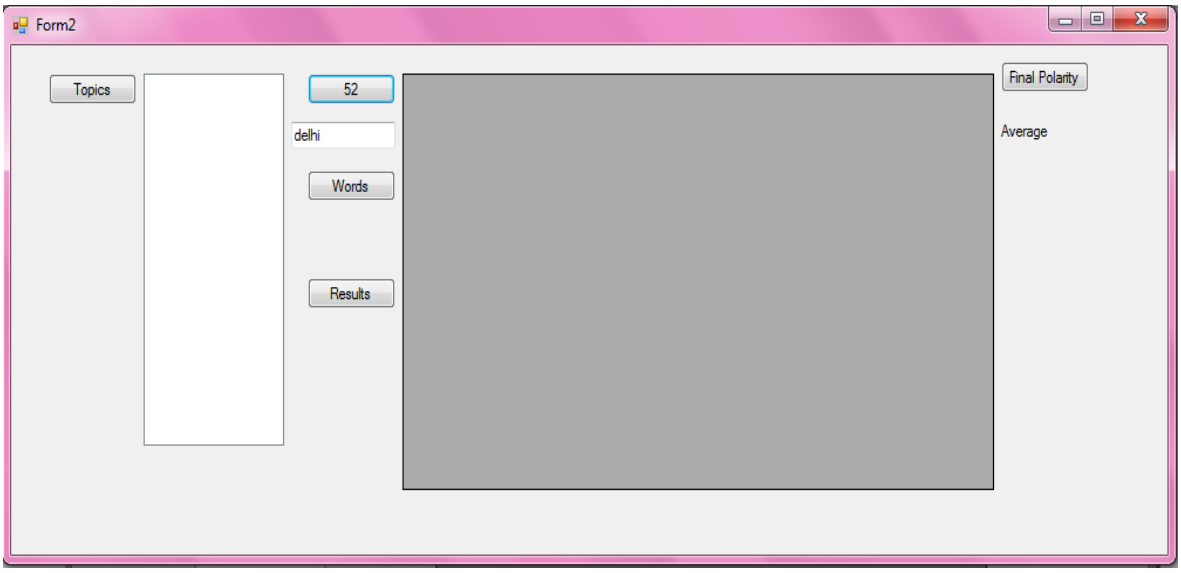


Figure 5.15: Tweets extracted

Now,step by step procedure shown in the above figures it will compute the results as shown in the figure 5.16.

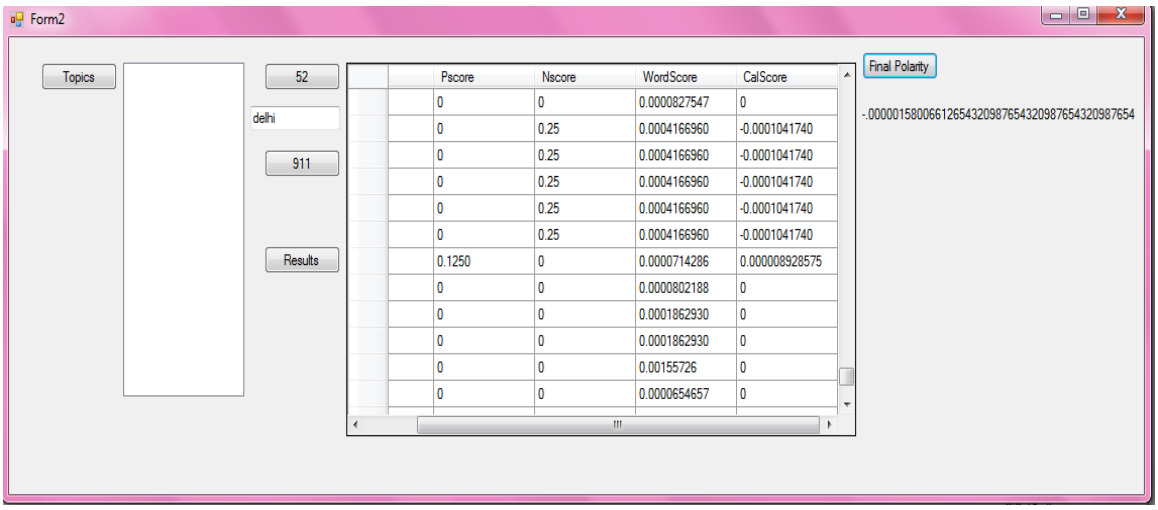


Figure 5.16: Computed Polarity

In the end we can conclude that the final sentiment polarity can be computed for any topic using the proposed approach.

Chapter 6

Conclusion and Future Scope

6.1. Conclusion

Wikipedia pages can be used as a mean of gathering topics as an alternative to generating topics through methods like bag of words. Also, correlating the amount of twitter streams or tweets that are related to each topic, by matching the word count on both the Wikipedia page content and tweets, generates an accurate depiction of the level of interest on a given topic.

In our experiment firstly, the topic trends are extracted using twitter dataset and Wikipedia topics, then this current topic is classified with the help of WordNet, SentiWordNet and PageRank value and then after that the final polarity is computed.
The biggest advantage of our experiment is that entire implementation has been done on actual available data and not on any assumed data.Sentiwordnet is used to get the positive as well as negative score of each word using synset id obtained by WordNet.

In terms of performance, it is possible to build a structure similar to a SVM based supervised approach by combining synsets weights provided by random walk algorithm with the SentiWordNet that gives the data with polarity scores. The dependence on the domain and need for a training corpus where the model was obtained is the disadvantages associated with supervised approach from which our solution is a general approach that does not suffer from this.

6.2 Summary of Contributions

- The datasets used for extracting the trendy topic is not only the Twitter, in this Twitter and Wikipedia both are considered for extracting trendy topics as this increase the efficiency of our results.
- Polarity Classification Analysis is done with both semantic and syntacticWordNet.

6.3 Future Scope

- The usage of words in Twitter is highly dynamic, it makes sense to improve topic trends by utilizing dynamic adjustments; analyzing twitter streams along shorter periods of time. Furthermore, analyzing data in real time would allow detection of a trending topic, not only for massive amounts of data as in search volume, but also on smaller spaces like small groups of people or conversations.
- A selection of connections could be done more carefully is of great interest as for building the graph all WordNet relations are taken into consideration so far. A polarity weights for individual terms and a graph to connect terms are the two main resources of the proposed solution. On the choice of these two resources any possible variation could be worth exploring.
- One major experiment could be undertaken in the assessment of the system for supervised polarity classification as a multi lingual solution, by means of WordNet in various different languages, like the MultilingualCentralRepository is found in the Spanish WordNet.

References

[1] Jansen, B.J., Zhang, M., Sobel, K., &Chowdury, A., Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology, 60(11), 2169–2188 , 2009.

[2] J. Giles, Internet encyclopaedias go head to head, *Nature,* vol. 438, no. 7070, pp. 900-901, 2005.

[3] Pang, B., Lee, L., Vaithyanathan, S., Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of theConference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79–86, 2002.

[4] Turney, P., Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the Association for Computational Linguistics (ACL), pp. 417–424, 2002..

[5] Prabowo, R., Thelwall, M., Sentiment analysis: a combined approach. Journal of Informetrics 3, 143–157, 2009.

[6] Pang, B., Lee, L., Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2, 1–135, 2008.

[7] Xin Zhao, Jing Jiang, JianshuWeng, Jing He, Lim Ee-Peng, Hongfei Yan, and Xiaoming Li., Comparing Twitter and traditional media using topic models. In Proceedings of the 33rd European Conference on Information Retrieval, 2011.

[8] Wang, Xuerui and Andrew McCallum, Topics over time: a non-markov continuous-time model of topical trends. In Tina Eliassi-Rad, Lyle H. Ungar, Mark Craven and DimitriosGunopulos, editors, *KDD*, pages 424–433. ACM. ISBN 1-59593-339-5, 2006.

[9] NoriakiKawamae, Predicting future reviews: sentiment analysis models for collaborative filtering. In Proceedings of the fourth ACM international conference on Web search and data mining, WSDM ’11, pages 605–614, 2011.

[10] Jonathan Chang, Jordan L. Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei, “Reading Tea Leaves: How Humans Interpret Topic Models,” *NIPS, page 288--296.Curran Associates, Inc.*, 2009.

[11] Matthew D. Hoffman, David M. Blei, and Francis R. Bach, "Online Learning for Latent Dirichlet Allocation," NIPS, page 856-864. Curran Associates, Inc., 2010.

[12] Alan Ritter, Colin Cherry, Bill Dolan, "Unsupervised Modeling of Twitter Conversations," Microsoft Research.

[13] Gabrilovich E. ,Markovitch S., "Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis",Proceedings of the 20th International Joint Conference on Artificial Intelligence, page 6—12, 2007.

[14] Daniel Ramage, Susan Dumais, Dan Liebling, "Characterizing Microbogs with Topic Models," AAAI, 2010.

[15] Pang, B., & Lee, L., Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 1(1–2), 1–135, 2008.

[16] Pak,A., &Paroubek, P., Twitter as a corpus for sentiment analysis and opinion mining. In Proceedings of LREC 2010 (pp. 1320–1326). Paris: European Language Resource Association, 2010.

[17] Taboada, M., J. Brooke, J., Tofiloski, M., Voll, K. and Stede, M. Lexicon-based methods for sentiment analysis. Computational Linguistics 37, 267–307,2011.

[18] Neviarouskaya, A., Prendinger, H., & Ishizuka, M., Textual affect sensing for sociable and expressive online communication. Lecture Notes in Computer Science, 4738, 218–229, 2007.

[19] Go, A., Bhayani, R., Huang, L., Twitter sentiment classification using distant supervision. In: CS224N Project Report, Stanford, pp. 1–12, 2009.

[20] Read, J., Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In: Proceedings of the ACL Student Research Workshop. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 43–48, 2005.

[21] Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., Liu, B., Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis.Technical Report HPL-2011-89. HP, 2011.

[22] Davidov, D., Tsur, O., Rappoport, A., Enhanced sentiment learning using twitter hashtags and smileys. In: Proceedings of the 23rd International Conference on

Computational Linguistics: Posters. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 241–249, 2010.

[23] Barbosa, L., Feng, J., Robust sentiment detection on twitter from biased and noisy data. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 36–44, 2010.

[24] Jiang, L., Yu, M., Zhou, M., Liu, X., Zhao, T., Target-dependent twitter sentiment classification. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies – Volume 1. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 151–160, 2011.

[25] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R., Sentiment analysis of twitter data. In: Proceedings of the Workshop on Language in Social Media (LSM 2011). Association for Computational Linguistics, Portland, OR, pp. 30–38, 2011.

[26] Hernández, S., Sallis, P., Sentiment-preserving reduction for social media analysis. In: Proceedings of the 16thIberoamerican Congress Conference on Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Springer-Verlag, Berlin, Heidelberg, pp. 409–416, 2011.

[27] Fellbaum, C. (Ed.), WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA, 1998.

[28] Page, L., Brin, S., Motwani, R., Winograd, T., The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. StanfordUniversity, 1999.

[29] Esuli, A., Sebastiani, F., Pagerankingwordnetsynsets: an application to opinion mining. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Association for Computational Linguistics, Prague, Czech Republic, pp. 424–431, 2007.

[30] Baccianella, S., Esuli, A., Sebastiani, F., SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Chair, N.C.C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (Eds.), Proceedings of the Seventh International Conference on Language

Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta, 2010.

[31] Denecke, K., Using SentiWordNet for multilingual sentiment analysis. In: Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24thInternational Conference on Data Engineering Workshops, pp. 507–512, 2008.

[32] Chamlertwat, W., Bhattarakosol, P., Rungkasiri, T., Haruechaiyasak, C., Discovering consumer insight from twitter via sentiment analysis. Journal of Universal Computer Science 18, 973–992, 2012.

[33] J. Leskovec. Snap.stanford.edu. Stanford Large Network Dataset Collection.

[34] A Davies. Alexdavies.net. A word list for sentiment analysis on Twitter.

[35] Saˇsa, P., Osborne, M., Lavrenko, V., The Edinburgh twitter corpus. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 25–26, 2010.

Publication

Published/Accepted

[1] MansiSethi and Dr.ShaliniBatra, "Real-Time Twitter Sentiment Analysis and Popularity Contest", The IEEE *International Conference on Computer Vision and Pattern Recognition***(**ICCVPR),at Tunisia, May 2014.

Communicated

[1] MansiSethi and Dr. ShaliniBatra, "Sentiment Polarity Classification of Trendy Topics in Twitter and Wikipedia Using SentiWordNet", The 3rd IEEE *Int. Conference on Advances in Computing, Communication and Informatics* (ICACCI) at Galgotias, Greater Noida, September, 2014.