Republic of Iraq
Ministry of Higher Education and
Scientific Research
University of Diyala
College of Science
Department of Computer Science

Sentiment Polarity Identification Framework of Tweets

A thesis
Submitted to the Department of Computer Science\ College of Sciences\ University of Diyala as a Partial Fulfillment of the Requirements for the Degree of Master in Computer Science

By

Sanaa Hammad Dhahi

Supervised By

Assist. Prof. Dr. Jumana Waleed Saleh

2020 A.D. **1442 A.H.**

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

﴿ اللَّهُ نُورُ السَّمَاوَاتِ وَالْأَرْضِ مَثَلُ نُورِهِ كَمِشْكَاةٍ فِيهَا مِصْبَاحٌ الْمِصْبَاحُ فِي زُجَاجَةٍ الزُّجَاجَةُ كَأَنَّهَا كَوْكَبٌ دُرِّيٌّ يُوقَدُ مِنْ شَجَرَةٍ مُبَارَكَةٍ زَيْتُونَةٍ لَا شَرْقِيَّةٍ وَلَا غَرْبِيَّةٍ يَكَادُ زَيْتُهَا يُضِيءُ وَلَوْ لَمْ تَمْسَسْهُ نَارٌ نُورٌ عَلَى نُورٍ يَهْدِي اللَّهُ لِنُورِهِ مَنْ يَشَاءُ وَيَضْرِبُ اللَّهُ الْأَمْثَالَ لِلنَّاسِ وَاللَّهُ بِكُلِّ شَيْءٍ عَلِيمٌ ﴾

صدق الله العظيم

سورة النّور\ آية35

Supervisors' Certification

We certify that this thesis entitled*"* ***Sentiment Polarity Identification Framework of Tweets"*** was prepared by ***"Sanaa Hammad Dhahi"*** under our supervisions at the University of Diyala Faculty of Science Department of Computer Science, as a partial fulfillment of the requirements needed to award the degree of Master of Science in Computer Science.

Supervisor :

Signature:

Name: *Assist. Prof. Dr. Jumana Waleed Saleh*

Date: / / 2020

Approved by University of Diyala Faculty of Science Department of Computer Science.

Signature:

Name: Assist. Prof. Dr. Taha M. Hassan

Date: / / 2020

Head Computer Science Department

Dedication

I would like to dedicate this work to:

To my candle that light my life

My Mother and Father.

To My husband Ahmed.

For his unlimited love, his supported,

His patience and Encouragement for me,

About everything in my life.

He is my first and last success.

To My children's Adam and Ayham.

To My Brothers and Sister.

To All My Friends.

Acknowledgements

All my thanks first of all are addressed to Almighty ***Allah***, who has guided my steps towards the path of knowledge and without His help and blessing; this thesis would not have progressed or have seen the light.

My sincere appreciation is expressed to my supervisor ***Asst. Prof. Dr. Jumana Waleed Saleh*** for providing me with ideas, inspiration and continuous support me during the period of my study.

I am extremely grateful to all members of Computer Science Department of Diyala University for their general support.

Finally, I would never have been able to finish my thesis without the help from ***friends***, and support from ***my family*** and ***husband***.

Thank you all!

Sana'a Hammad

Linguistic Certification

This is to certify that this thesis entitled ***"Sentiment Polarity Identification Framework of Tweets "***, prepared by ***"Sanaa Hammad Dhahi"***at the University of Diyala / Department of Computer Science, is reviewed linguistically. Its language was amended to meet the style of the English language.

Signature:

Name:

Date: / / 2020

Scientific Amendment

I certify that the thesis entitled ***"Sentiment Polarity Identification Framework of Tweets"*** *presented by* ***"Sanaa Hammad Dhahi"*** *has been evaluated scientifically; therefore, it is suitable for debate by examining committee.*

Signature:

Name :

Date : / / **2020**

Abstract

In recent years, Twitter becomes a source of extracting information and knowledge for both individuals and organizations, where opinions and ideas of the users are sharing and exchanging in the form of texts called tweets, about everything that concerns people's daily lives. Therefore, sentiment analysis concerns analyzing people's feelings and classification of these opinions into negative or positive.

In this thesis, an efficient twitter sentiments classification framework has been built to increase the accuracy and decrease the error rate that may be occur in the classification process. A proposed framework consists of three main stages: pre-processing, feature extraction and classification of sentiment stage. In the feature extraction stage a set of (14) features were extracted which includes (13) features statistical were extracted from the tweet itself, and the feature number (14) is a semantic feature was extracted by using Document to Vector technique (Doc2Vec) was computed in order to increase the accuracy of the sentiment classification. In this thesis, two types of a common classifier (Naïve Bayes and Support Vector Machine) were used.

The proposed framework has been tested by using three twitter dataset (Sentiment140, SS-Tweet and STS-Test). The results indicate that the accuracy rate of Naïve Bayes using sentiment140 dataset is 94% and when using SS-Tweet dataset the accuracy rate is 75%, and when using sentiment140 dataset as train and SS-Tweet or STS-Test as test the accuracy rate is 87%,and when Support Vector Machine algorithm is used, the accuracy rate using sentiment140 dataset is 94% and when using SS-Tweet dataset the accuracy rate is 79%, and when using sentiment140 dataset as train and SS-Tweet ,STS-Test as test the accuracy rate is 77% , 84%, respectively.

Table of Contents

List of Abbreviations

Abbreviation	Description
AI	Artificial Intelligence
BOW	Bag of Word
CBOW	Continuous bag of words
CC	Coordinating conjunctions
CM	Confusion Matrix
CSV	comma-separated values
Doc2Vec	Document to vector
ESA	Explicit Semantic Analysis
IC	Information Content
IMDb	Internet Movie Database
IR	Information Retrieval
KNN	K-Nearest Neighbor
LCS	Least Common Subsume
LSA	Latent Semantic Analysis
MAP	Maximum Posteriori
ME	Maximum Entropy
ML	Machine Learning
NB	Naïve Bayes
NGD	Normalized Google Distance
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NN	Neural Network

Continue of List of Abbreviations

Paragraph2Vec	Paragraph to vector
PMI-IR	Point wise Mutual Information - Information Retrieval
POS	Part Of Speech
PV-DBOW	Distributed Bag of Words version of Paragraph Vector
PV-DM	Distributed Memory Model of Paragraph Vectors
RBF	Radial Basis Function
SA	Sentiment Analysis
SC	Sentiment Classification
SG	Skip-Gram
SGD	Stochastic Gradient Descent
SPTI	Sentiment Polarity of Tweets Identification
SS-Tweet	Sentiment Strength -Tweet
STS	Stanford Twitter Sentiment
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
TM	Text Mining
URL	Uniform Resource Locator

List of Figures

List of Tables

List of Algorithms

Chapter One

General Introduction

Chapter one

General Introduction

1.1 Introduction

Computational linguistics is the science which combines linguistics and artificial intelligence for automatic Natural Language Processing (NLP). In many natural language processing applications such as sentiment analysis (SA) the text similarity measures are used and these measures also use in some domain that is related to text mining. The similarity measure process considered an important task which has high effect in many applications that is dealing with text such as: text summarization, information retrieval, document classification and other applications. The methods used to determine similarity among texts are based on lexical matching is a simple method, it focuses on analyze the share words among texts and detect the degree of similarity among texts based on the words number which appearing in both documents and that match to the lexical style. Lexical methods are easy to implement, but it is poor in reflecting the relationship among words that have a similar meaning, such as words that have the same root or synonymous to each other, co-occurrence words may appear in the longer texts, but it may slightly in short texts or even scarce [1] [2] [3].

There are many applications in natural language processing such as sentiment analysis require specifying the semantic similarity. The concept of semantic similarity can be interpreted as a group of different words which have the similar meaning. Many areas of text mining use the concept of semantic similarity which is an important aspect in natural language processing [4].

1.2 Overview of Social Media

The growth and increase of social media has exploded the publicly accessible text created by users on the internet. This information that created by user can be used to supply insights into people's feelings and also, blogs, online forums and comments on social networking sites like Facebook, Instagram and Twitter can all be considered as a social media. Social media can get millions of people's opinions about a particular topic and it has become an increasing important source of information [5].

On the other side, peoples are more ready and glad to share things about their life's, their experiences and thoughts with the entire world via social media. People share their events by expressing their opinions and clarifying their comments on things that happen in society. The way for people to share their knowledge and sentiments with community through social media pushes companies to gather extra information about their companies and products and know the extent of their reputation among people and thus make important decisions to continue their business effectively [6].

The increased use of social media has made the SA take an important role in discovering people's opinions through written languages and focusing on detecting the polarity of sentiments if they are (positive, negative or neutral) towards a specific topic. For example, a political party may wish in determining whether or not people support their political process [7].

1.3 Overview of Semantic Similarity

Mainly, the texts can be similar in two lexical and semantic methods. Lexical method uses the idea of traditional matching to calculate the distance among text documents, similarity increases when two text documents contain the same character's sequence, this method always fail to find true similarity degree while semantic similarity method refer to texts are similar if they contain similar meaning in both, used in same context [8].

The nature of semantic similarity is to simulate the ability of individuals into comparing the texts. Semantic similarity is the measure of the distance across texts or a group of words and the determination of distance depends on the similarity on its meaning or semantic content [9].

Semantic similarity is a method that widely used in the language understanding, it measures how two texts (X, Y) are similar based on the meaning of them. Many types of semantic measures have been suggested to compute the semantic similarity, which range from semantic network-based metrics and distributional similarity metrics models, which are depend on learning from large text sets. Generally, semantic similarity methods can be categorized into two groups: knowledge based methods using lexical databases (manually created) and corpus based methods (using statistical methods) [2].

1.4 Related Works

Many researchers have been done to deal with sentiment analysis, to deal with the problem associated with natural language processing some of researchers use semantic similarity concept in SA which can improve the results of the tweets classification and others employ different techniques to classification positive and negative opinion from text. Here is a review of a number of these works.

- **A. Barhan and A. Shakhomirov (2012) [10]:** They proposed up a model which can extract from Twitter data the sentiment polarity of tweets. The features extracted were words containing emotional symbols and n-gram. The results show that the Support Vector Machine (SVM) performance is better than the Naïve Bayes (NB). SVM in combination with unigram feature extraction is the best performing method, which obtaining a precision of 81% and a recall of 74%.
- **P. Bellot et al. (2013) [11]**: They proposed to use many features for sentiment analysis in micro-blogging such as unigram, domain specific, DBpedia, WordNet

and Sentiwordnet features are using with SemEval 2013 dataset. The experimental result showed that add the above features able to improve the F-measure accuracy 2% with Support vector machine and 4% with Naïve Bayes.

- **G. Gautam and D. Yadav (2014) [12]:** They presented a semantic WordNet synonym analysis approach for SA in twitter dataset. This method depends on examining the semantic synonym similarity between training datasets and words in the testing, when it is found this similarity, it will be replacing the words in the testing dataset with their synonyms in the training dataset. The experimental result showed that the Naive Bayes Classifier (NB) obtained the accuracy 88% which is the best result as compared with other classifiers such as Maximum Entropy (ME) and support vector machine(SVM).
- **D.Zhang et al. (2015) [13]:** They focused on semantic features among words instead of lexical features and two tools are used to classify the Chinese comments texts are Word2Vec and SVM ^{perf}. The results of the proposed method to classification sentiment reached to 90% accuracy.
- **A.Tripathy et al.(2016) [14]:** They attempted to classify the reviews of movies using several classification algorithms like Naive Bayes(NB), Maximum Entropy(ME), Stochastic Gradient Descent (SGD**)** and Support Vector Machine (SVM) then using n-gram feature with these algorithms which are applied on dataset of the IMDB. They noticed that results obtained by applied the above classifiers are 86%, 88%, 85% and 88% respectively.
- **K. Kavitha and Ch. Suneetha (2017) [15]:** They focused on simplifying the rapid detection of sentimental contents. The K-Nearest Neighbor (KNN) and NB classifier are used for sentiment classification of the movie reviews. The experimental results show the NB yielded results 83% and KNN result 72%.

- **O.Araque et al.(2018) [16]:** They proposed an approach of using lexicons of sentiment , that is depend on measure the semantic similarity between lexicon of vocabularies and words in text .A proposal approach consists of a SA which use embedding based representations as well lexicon based semantic similarity as features. The results of the proposed method to classification sentiment reached to 89% accuracy.
- **A.Oussous et al.(2019)[17]:** They proposed an approach to detect the best model for classification the polarity. They showed that unigram is the best for classification the polarity also, they noticed removing the stop words decrease the performance of the classifiers that are used for classifying the sentiments. The experiment results prove that combining between more than one classifier gives the best results with accuracy:86% and presion:89%.

Table (1.1): Related works summarizations

NO.	Year	Author	Data sets	Feature set	Technique	Accuracy
1	2012	A. Barhan and A. Shakhomirov	Twitter messages	n-grams and emoticons	SVM, NB	recall: 74% precision : 81%
2	2013	P. Bellot et al.	SemEval 2013	unigram, Domain specific, DBpedia, WordNet and Senti-features	SVM, NB	adding features has improved the F-measure accuracy 2% with SVM and 4% with NB
3	2014	G. Gautam and D. Yadav	product reviews based on twitter data.	unigram and POS	NB,SVM and ME	88%
4	2015	D. Zhang et al.	Chinese comments on clothing products	Lexicon based and POS	SVM^{perf}	90%

5	2016	A.Tripathy et al.	IMDb	n-gram	NB, ME, SVM and SGD	NB:86% ME:88% SVM:88% SGD:85%
6	2017	K. Kavitha and Ch. Suneetha	movie reviews	POS tagging, unigram, bigram	K-NN, NB	NB: 83% K-NN: 72%
7	2018	O. Araque et. al.	Twitter related: sentiment140 SemEval2014, Vader and STS Gold. Movie reviews: IMDb,PL04 and PL05	Sentiment lexicon and Word embeddings	semantic similarity and lexical metrics	89%
8	2019	A. Oussous et al.	40k Arabic tweets	N-gram	SVM,NB and ME	accuracy:86% presion:89%

1.5 Problem Statement

Literature survey shows that the most studies of the sentiment analysis in twitter is depended on traditional(lexical)features such as part of speech, negation, hashtag, etc., to identify the sentiment polarity. Therefore, the main problem of this thesis is to build a sentiment polarity identification framework using semantic similarity features, in order to help the analysts of data in a huge company to making them able to deal with the general opinions and measure it accurately. For this reason, the analysts of texts (tweets) needs an efficient technique gives an analysis accurately to help them taking the accurate decision about any topic.

1.6 Aim of Thesis

The aim of this thesis is to design and implement a sentiment polarity identification framework of tweets able to accurately classification tweets into positive and negative in twitter by using Naive Bayes and Support Vector Machine algorithms, in order to obtain high accuracy to help the opinion's analysts to prevent the errors while identifying and classifying the sentiment from different datasets and the user also can make direct decisions about any movie, product, service, etc. Without the need for individual reviews, through a combination features provided by the proposed framework developed to achieve this purpose.

1.7 Thesis Outline

This thesis is structured around five chapters, including chapter one, it contains the following chapters:

Chapter 2: Theoretical Background

Presents semantic similarity methods, word embedding, tweet preprocessing, sentiment analysis applications and levels, sentiment analysis classification and accuracy measuring.

Chapter 3: **The Proposed Framework**

Presents the detail of the proposed framework and explains the practical stages of this framework.

Chapter 4: **Experiments and Results**

Includes the experimental results obtained from applying the proposed framework, the evaluation of classifiers techniques on the dataset.

Chapter 5: **Conclusions and Suggestions for Future Works**

Presents conclusions, discussions and suggestions for future works.