

Emotion Prediction from Text through Sentiment Analysis

Muhammad Waqas CIIT/SP16-MCS-076/VHR
Azam Saleem CIIT/SP16-MCS-022/VHR



January 2018

Department of Computer Science
**COMSATS INSTITUTE OF INFORMATION
TECHNOLOGY**
VEHARI – PAKISTAN

Submission Form for Final-Year

PROJECT REPORT



PROJECT ID		NUMBER OF MEMBERS	2
TITLE	Emotion Prediction from Text through Sentiment Analysis		
SUPERVISOR NAME	Mr. Waseem Akram		

MEMBER NAME	REG. NO.	EMAIL ADDRESS
Muhammad Waqas	CIIT/SP16-MCS-076/VHR	kbwaqas@gmail.com
Azam Saleem	CIIT/SP16-MCS-022/VHR	chazamsaleem@gmail.com

CHECKLIST:

Number of pages in this report	59
I/We have enclosed the soft-copy of this document along-with the codes and scripts created by myself/ourselves	YES / NO
My/Our supervisor has attested the attached document	YES / NO
I/We confirm to state that this project is free from any type of plagiarism and misuse of copyrighted material	YES / NO

MEMBERS' SIGNATURES

Supervisor's Signature

Submission Form for Final-Year
PROJECT REPORT



Approval Letter

It is certified that this work, entitled “Emotion Prediction from Text through Sentiment Analysis” submitted by “*Muhammad waqas and Azam Saleem*” is hereby approved as Partial Fulfilment for the award of Degree of “*MCS*”.

External Examiner

Internal Supervisor

Mr. Waseem Akram

Dr. Muhammad Rafique Mufti

In-Charge

Department of Computer Science,
CIIT Vehari.

Department of Computer Science
**COMSATS INSTITUTE OF INFORMATION
TECHNOLOGY**
VEHARI – PAKISTAN

Declaration

“No portion of the work referred to in the dissertation has been submitted in support of an application for another degree or qualification of this or any other university/institute or other institution of learning”.

MEMBERS' SIGNATURES

Acknowledgements

In the name of ALLAH, the most kind and most merciful.

I would like to thank to my friends and parents who help us directly or indirectly who kept backing me up in all the times, both financially and morally...

I would also like to thank to Dr. Muhammad Rafiq Mufti (H.O.D of Computer Science) and my Supervisor Mr. Waseem Akram (Research Associate) for his guidance and encouraging us to work hard and smart. I have found him very helpful while discussing the optimization issues in this dissertation work. His critical comments on my work have certainly made me think of new ideas and techniques in the fields of optimization and software simulation.

We are grateful to the ALLAH Almighty who provides all the resources of every kind to us, so that we make their proper use for the benefit of mankind. May He keep providing us with all the resources, and the guidance to keep helping the humanity.

DEDICATION

**To Almighty ALLAH and the Holy Prophet
Muhammad (P.B.U.H)
&
Our Loving Family**

Abstract

Human Emotions plays a key role in decision making in human life. Detecting and understanding human emotions in textual form is challenging task. This research work will focus on emotional analysis to predict human emotion.

.

Table of Contents

I

APPROVAL LETTER.....	1-1
DECLARATION.....	1-2
ACKNOWLEDGEMENTS.....	1-3
ABSTRACT.....	1-5
1 INTRODUCTION.....	1-9
1.1 INTRODUCTION.....	1-10
1.1.1 Area of Research.....	1-10
1.2 OBJECTIVE.....	1-11
1.3 PROBLEM DESCRIPTION.....	1-11
1.4 PROJECT SCOPE.....	1-11
1.5 FEASIBILITY STUDY.....	1-11
1.5.1 Risks Involved:.....	1-11
1.6 SOLUTION APPLICATION AREAS.....	1-11
1.6.1 Politics.....	1-12
1.6.2 Marketing intelligence.....	1-12
1.6.3 Blogs analysis.....	1-12
1.7 TOOLS/TECHNOLOGY.....	1-12
1.7.1 Tools.....	1-12
1.7.2 Language.....	1-12
1.7.3 Dictionary.....	1-12
1.7.4 Toolkit.....	1-12
1.7.5 Operating System.....	1-12
1.8 EXPERTISE OF THE TEAM MEMBERS.....	1-12
1.9 MILESTONES.....	1-13
2 LITERATURE REVIEW.....	2-14
2.1 TECHNIQUES OF SENTIMENT CLASSIFICATION.....	2-15
2.1.1 Machine learning approach.....	2-15
2.1.1.1 Supervised learning.....	2-16
2.1.1.1.1 Probabilistic classifiers.....	2-16
2.1.1.1.1.1 Naive Bayes Classifier (NB).....	2-16
2.1.1.1.1.2 Bayesian Network (BN).....	2-17

2.1.1.1.1.3	Maximum Entropy Classifier (ME).....	2-17
2.1.1.1.2	Linear classifier	2-18
2.1.1.1.2.1	Support Vector Machines Classifiers (SVM)	2-18
2.1.1.1.2.2	Neural Network (NN).....	2-19
2.1.1.1.3	Decision tree classifiers	2-20
2.1.1.1.4	Rule-based classifiers	2-21
2.1.1.2	Weakly, semi and unsupervised learning	2-21
2.1.1.3	Meta classifiers.....	2-22
2.1.2	<i>Lexicon-based approach</i>	2-24
2.1.2.1	Dictionary-based approach.....	2-25
2.1.2.2	Corpus-based approach	2-25
2.1.2.2.1	Statistical approach.....	2-27
2.1.2.2.2	Semantic approach	2-28
2.1.2.3	Lexicon-based and natural language processing techniques.....	2-29
2.1.2.3.1	Discourse information	2-30
2.1.3	<i>Other techniques</i>	2-32
3	METHODOLOGY	3-34
3.1	DATA PRE-PROCESSING	3-35
	<i>Need of Text Pre-processing in NLP</i>	3-35
3.1.1	<i>Cleaning</i>	3-36
3.1.1.1	URL Removing	3-37
3.1.1.2	Stop word	3-38
3.1.1.3	Duplication Removing	3-38
3.1.1.4	Case Sensitive	3-38
3.1.2	<i>Tokenizing</i>	3-38
	Challenges in Tokenization	3-38
	Isolating:.....	3-39
	Agglutinative:.....	3-39
	Inflectional:	3-39
3.1.3	<i>Stemming /lemmatizing</i>	3-39
	Stemming.....	3-39
3.1.3.1.1	Problem in Stemming	3-40
	Lemmatizing.....	3-40
3.1.4	<i>Parts-Of-Speech (POS)-Tagging</i>	3-40
	What is tagging?	3-40
	What is Parts-Of-Speech Tagging?.....	3-40
	Parts Of Speech tagger	3-41
	Tag Set.....	3-41
3.2	DATA ANALYSIS	3-42
3.2.1	<i>Synsets</i>	3-42
3.2.2	<i>Dictionaries</i>	3-43

Wordnet	3-43
Emotion.xml	3-43
3.2.3 Negation	3-43
3.3 FLOW CHART	3-44
4 RESULT	4-45
5 REFERENCES	5-51

1 Introduction

1.1 Introduction

Language is known to be a powerful tool to communicate and convey information and also a means to express emotion. Emotion identification is currently being studied in neuroscience, psychology and cognitive science, computer science and behavioral science. It is important to develop a method to recognize the automatic emotion and predict human emotion response based on the information available.

In our work we consider a textual input as a source of information.

The textual communication will allow us to state directly what we feel. Our social nature incites us to share our feelings with other. People use blogs to express their category of emotion thoughts, feeling, opinions and experience with each other¹.

However there is no agreement on which categories should be regarded as the most representative. There could be more than one category of emotional word (multi-category).

Multi category emotion recognition can be formulated as follow:

Given the input text, Detect which emotion category are expressed, If there is no emotion that marked as a neutral, If there is emotion score them for the determination of category and it also tell about the positive and negative emotion.

The main challenge in the automatic emotion prediction is the accuracy and the quality of the output. We can express our emotion two ways: first one “this was one of the happiest day.” in this we explicitly stating feeling. Second “I passed my exam.” Implicitly describing what happened with us. But here in my project I have worked on explicit emotion.

The text which we get may be unstructured. We have to use any technique to make unstructured data into structured form. Text mining is the technique of data mining use to organize the text to make it meaning full by using some data structure (parsing, addition of some derived linguistic features). Text mining includes sentiment analysis, text-categorization, text-clustering, document summarization.

1.1.1 Area of Research

Sentiment analysis is the new emerging area of research.

It is mainly focus on the knowledge discover and Information retrieval from text. The goal of sentiment analysis is to make computer able to detect and express emotion².

Natural language processing techniques have been applied to automatically identify the information content in text.

1.2 Objective

The main objective of my research is the prediction of emotions. These emotions are very helpful in decision making. We will take data by multiple sources like media or either by simple input. We will able by using this system to predict the people emotions and their thought like he/she has positive or negative thoughts.

1.3 Problem Description

During the conversation with someone thorough machine, sometimes we are not able to understand the emotions of other person which can lead us to misunderstanding. In this research we are going to predict the emotions of other human through machine.

1.4 Project Scope

- This research completely surrounded in human emotion prediction
- It help us for better human emotion prediction
- Also helpful for better decision making

1.5 Feasibility Study

Our research is feasible one example is, a large organization cannot read all reviews of customer, and they can use emotion prediction by the machine to make graphs for decision making.

1.5.1 Risks Involved:

Accuracy

1.6 Solution Application Areas

We are going to predict human emotions which are very helpful in decision making.

Examples are given below:-

1.6.1 Politics

As is well known, opinions matter a great deal in politics. Some work has focused on understanding what voters are thinking.

1.6.2 Marketing intelligence

Helpful in business for decision making and enhance our business.

1.6.3 Blogs analysis

Analyze blog sentiment about movies and correlate it with sale.

1.7 Tools/Technology

1.7.1 Tools

- Anaconda
- Spyder

1.7.2 Language

- Python 3.6

1.7.3 Dictionary

- Wordnet
- Emotiona.xml

1.7.4 Toolkit

- NLP Toolkit

1.7.5 Operating System

- Windows upward

1.8 Expertise of the Team Members

Text mining Muhammad Waqas

NLPToolkit Azam Saleem

Python 3.6 Muhammad Waqas, Azam Saleem

1.9 Milestones

Phase	Name	Starting Date	Completion Date
Introduction	Muhammad Waqas Azam Saleem	01-07-2017	24-07-2017
Literature Review	Muhammad Waqas	01-08-2017	10-09-2017
Methodology 1) Text refinement 2) Part of speech Tagging 3) Sentiment analysis 4) Emotion prediction 5) Negation handling	Muhammad Waqas Azam Saleem	30-09-2017	18-11-2017
Result	Muhammad Waqas Azam Saleem	25-11-2017	5-01-2018

2 Literature Review

2.1 Techniques of sentiment classification

There are three techniques of Sentiment Classification which are machine learning approach, lexicon based approach and hybrid approach³. There are famous algorithms of Machine Learning (ML) applied by Machine Learning approach (ML). It uses features of linguistics. The lexicon-based approach depends on sentiment lexicon, a known collection of precompiled terms of sentiment. It is further divided into two approaches one is dictionary based approach and other is corpus-based approach, these approaches use the methods of statistical or semantic to find the polarity of sentiment. The hybrid approach is the combination of both approaches. This approach plays a very common key role in various methods with sentiment lexicons.

ML approach used by text classification methods can be divided into two learning methods supervised and unsupervised. A large number of documents of labelled training make used by supervised methods. When these documents of labelled training are difficult to find then unsupervised methods are used.

Lexicon-based approach depends on opinion lexicon finding which used in text analysing. This approach has two methods. First is dictionary based approach and the second is corpus-based approach. First approach depends on finding seed words, and after that searches the synonyms and antonyms dictionaries of these words. The second approach have opinion words seed list and then in a large corpus it finds the other words of opinion to help in finding opinion words with orientation of specific context. This is done by two methods statistical and semantic.

2.1.1 Machine learning approach

ML approach is used to solve the SA as a problem of regular text classification that uses the features of syntactic and linguistic, this approach is based on famous ML algorithms. Definition of the Problem of Text Classification: there is a record of training set $D = \{Y_1, Y_2, \dots, Y_n\}$ where each record is labelled to a class. The model of classification is related to the labels of the class in the underlying record's features. For an unknown class instance which is given the model is used in the prediction of label of a class for it. When to an instance only one label is assigned this is hard classification problem. And when to an instance a probabilistic value is assigned this is referred to as soft classification problem.

2.1.1.1 Supervised learning

The methods of supervised learning based on the existence of the documents of labelled training. Supervised classifications have many types in literature.

2.1.1.1.1 Probabilistic classifiers

For classification probabilistic classifiers use mixture model. Mixture model consider each class of the mixture as a component. Generative model is the mixture of each component and that generative model for that component provides the probability of sampling a particular term. These classifiers are also known as generative classifiers. There three most famous classifiers will discuss in coming sections.

2.1.1.1.1.1 Naive Bayes Classifier (NB)

The most commonly used and simplest classifier is the Naïve Bayes Classifier. This model for a class computes the probability of posterior which is based on the word's distribution of document. Naïve Bayes model for ignoring the word's position in document work with the features of BOW. Byes theorem is used for the probability prediction and a feature set is given belonging to a particular label.

$$P(\text{label/features}) = (P(\text{label}) * P(\text{features/label})) / P(\text{features})$$

For a label the $P(\text{label})$ is prior probability. Prior probability for given feature is being classified as a label is $P(\text{feature/label})$. For the occurrence of a given feature se the prior probability is $P(\text{feature})$. The above equation could be written for independent features as follows:

$$P(\text{label/features}) = (P(\text{label}) * P(f_1/\text{label}) * \dots * P(f_n/\text{label}) / P(\text{features})$$

Kang and Yoo proposed an improved NB classifier⁴, this classifier was purposed to solve the tendency problem for the classification of positive accuracy 10% approximately higher appearance than the classification of negative accuracy appearance. When two classes express their accuracy as an average values this creates an accuracy of average decreasing problem. This algorithm with the reviews of restaurant narrowed the positive and negative accuracy gap as compared to NB and SVM.

2.1.1.1.1.2 Bayesian Network (BN).

Features independence is the basic assumption of NB classifier and the other assumption is that to assume fully dependent of all features. This assumptions leads us to model of Bayesian Network and the acyclic graph is directed by this whose dependencies of conditions are represented by edges and variables are represented by nodes. For variables and their relationships BN is a complete model. For a model the probability of complete joint distribution over all the variables is specified. BN complexity of computation is expensive in Text mining. Due to this reason this is used so less⁵.

In the consideration of problem of real world Hernandez and Rodriguez use BN ⁶ in which there are three different target variables to characterized the attitude of the author. Classifiers of multi-dimensional network are purposed by them. To exploit the potential relationships between them different target variables are joined in the same task of classification. Classification of multi-dimensional framework is extended to the domain of semi-supervised in order to take advantage of huge unlabelled amount of information available in this context.

2.1.1.1.1.3 Maximum Entropy Classifier (ME)

The maximum classifier is also known as conditional exponential classifier. This use encoding to convert the feature of label sets to vectors. Which is then used for the weight calculation for each feature, and then to determine the most likely label for a feature set that can be combined. Set of X {weights} is used to parameterized this classifier, which is used in combining the joint features generated from X {encoding} feature-set. Each C {(featureset, label)} pair to a vector is encoding map. The following equation is used to compute the probability of each label:

$$P(fs/label) = (\text{dotprod}(\text{weights}, \text{encode}(fs, \text{label}))) / (\sum(\text{dotprod}(\text{weights}, \text{encode}(fs, l)) \text{ for } l \text{ in labels}))$$

The ME categorization module was used by Kaufman ⁷ for evidence Offers between all language pairs with a small amount Training data. Other auto-developed tools Extract parallel data from non-parallel languages or need a lot of training data. Their results show that the ME classifier can produce Useful results for almost every language pair. It can Allows you to create parallel attachments for many new languages.

2.1.1.1.2 Linear classifier

Taking into account $\bar{X} = \{x_1 \dots \dots x_n\}$ - normalized frequency of words of the document, Vector $\bar{A} = \{a_1 \dots \dots a_n\}$ is the vector of a linear coefficient is the same as the dimension Feature space, b is a scalar; linear output the predictor is defined as $p = \bar{A} \cdot \bar{X} + b$ which is the result linear classifier. This is the result linear classifier. Predictor p is a separate hyper plane between different classes. There are several types of linear classifiers; among them, support vector machine (SVM)^{8 9} this is a form the classifier attempts to determine Linear delimiters between different categories. Two among of these are discussed in the following subsections.

2.1.1.1.2.1 Support Vector Machines Classifiers (SVM)

The basic principle of SVM is the definition of linear separators Find a space that best separates different classes. Text data is ideal for SVM classification, because the rare nature of the text, in which some functions do not matter, but they tend to correlate with each other and usually organized into linearly separable categories¹⁰. Support vector machines can build nonlinear solutions in the original image Compare data instances and object space non-linearly Product's internal space, class can be linear separation Use hyper planes¹¹. SVM is used in many applications, among these applications Rating based on its quality. Chen Zeng¹² used two multi-classes SVMs Method: SVM unidirectional and stand-alone multi-machine SVM assessment evaluation. They propose a way Evaluate the quality of information in product reviews Consider the problem of classification. They also adopted it Information Quality (IQ) structure used to search for information a set of functions. They work on digital cameras MP3 comments. Their results show that their method can be accurate

Lee and Lee¹³ used SVM as the polarity of sensations classification. Unlike the binary classification problem, they claim is the subjectivity of opinion and the reliability of expression this must be taken into account. They proposed a structure Provide a compact digital summary of your comments Platform micro blogging. They identify and extract items Mentioned in the comments related to user requests, then use SVM rating advice. They work Twitter posts for their experiment. They found this consideration the user's credibility and opinion subjectivity is crucial Collect micro

blogging comments. They proved it Its mechanism can effectively find market intelligence (MI) Support decision-makers by establishing monitoring The system tracks different aspects of external comments Real-time business.

2.1.1.1.2.2 Neural Network (NN)

Neural networks are many of their basic units of neurons. Enter the neuron is represented by the over line \bar{X}_i vector, which is the frequency word in the i -th document. Have a Weight A is associated with each neuron used to calculate the function of its input $f()$. Linear function Neural Network: $\pi_i = A \bar{X}_i$. In binary classification this question assumes that the class X_i tags are marked the class is given by y_i and the sign of the prediction function π_i label.

Multilayer neural networks are used for nonlinear boundaries. These multiple layers are used to induct several segments linear limit for approximation Closed area related to a particular category Exit Neurons in the diet of neurons of the early layer behind the layer. The learning process is complicated, because Errors must be retro-multiplied to different levels. For text data, there is an implementation of NN Found in ^{14 15}.

Support vector machine and compare the artificial experience neural networks proposed by Morass and Valiati ¹⁶ on the analysis of moods at the document level. They did it this comparison is associated with a broad and successful SVM It is used in artificial neural networks, whereas artificial neural networks rarely concern as a way to learn from emotions. They discussed this both requirements, the results model and the background Method for obtaining better classification accuracy. They also use popular monitoring methods to select functions and weigh the standard evaluation context in the traditional BOW model. Their experiments show that artificial neural networks provide better results for SVM, with the exception of some unbalanced data contexts. They tested three reference datasets for watching movies, GPS, cameras and books on amazon.com. They proved that statistically significant differences in the review of ANN film critics outperform SVM. They confirmed some potential limitations of both models, which are rarely discussed in the AS literature, such as the cost of running SVM and RNA in training. They proved that the use of information Gain (an inexpensive method of selecting functions in terms of calculation) can reduce the computational effort of ANN and SVM without significantly affecting the accuracy of the classification obtained.

Support vector machines and neural networks are also used to classify personal relationships in biographical texts proposed by van de Camp and van den Bosch¹⁷. They regard the relationship between two people (one is the topic of biography and the other one is mentioned in this biography) as positive, neutral or unknown. Their case studies are based on historical biographic information that describes people within a specific area, region and time frame. They show that their classifier is these relationships can be labelled above the majority of the class's baseline score. They found that training sets that contain relationships of multiple people produce more satisfying results than sets that focus on one particular entity. They show that SVM and monolayer NN (1-NN) algorithms have the highest scores.

2.1.1.1.3 Decision tree classifiers

The decision tree classifier is provided the training data space is hierarchically decomposed, where the data is segmented using attribute value conditions¹⁸. A condition or predicate is the presence or absence of one or more words. The separation of the data space is performed recursively, until the leaf nodes contain a minimum number of records for this purpose classification. There are other types of predicates that rely on the similarity of documents to correlate a set of terms that can be used to further separate the document. A different type of segmentation is the segmentation of one attribute that uses a specific word or expression for segmentation on a specific node in the tree¹⁹. Multi-element estimates based on similarity perform checks using common text documents or clusters and the similarity of documents with these word clusters. Discrimination based on the discriminant uses discriminant analysis, for example the Fisher discriminator, for resolution²⁰. Decision trees in the categorization of the text make small changes that usually have standard packages, such as ID3 and C4.5. Lee and Jane²¹ used the algorithm C5 of the C4 algorithm. It depends on the concept of the tree; Hu and Lee²² offer a way to extract the term "t" and the keyword "its context using the spanning tree (MST) structure." Therefore, they developed a so-called model for describing the theme of the classification of feelings. In its definition, "thematic terms" represent certain aspects of a particular object or object specific to a domain. They introduced the automatic extraction of thematic terms based on domain terminology engines. They then use these extracted terms to distinguish the subject from the

document. This structure conveys emotional information. Their method is different from the usual algorithm of the machine learning tree, but in fact can learn basic positive and negative knowledge.

Yan and Bing proposed a graphical approach²³. They proposed a method of communication that combines the characteristics of internal and external proposals. The characteristics of the two proposals are documentary evidence and documentary evidence. They said that the definition of the emotional orientation of the commentary is not only the nature of the sentence itself. They studied the area of the chambers and compared their methods with uncontrolled and controlled methods (NB, SVM). Their results show that the presented method showed better results than any method, without using the characteristics of external proposals and exceeding other previous representative methods.

2.1.1.1.4 Rule-based classifiers

In classifiers based on rules, the data space is modelled by a set of rules. The left side is a condition for a set of characteristics expressed in disjunctive normal form, and the right side is a class label. Conditions are present. The absence of a term is rarely used, because it is not informative in scattered data. There are a number of criteria for the generation of rules, at the training stage all rules are built according to these criteria. The two most common criteria are support and trust²⁴. Support is the absolute number of instances in the training dataset that are relevant to the rule. Confidence refers to the conditional probability that the right side of the rule is executed, if the left side is satisfied. Some combined rules algorithms were proposed in²⁵. Decision trees and decision rules tend to encode rules for physical space, but decision trees tend to be implemented using a hierarchical approach. Quinlan²⁶ studied the problem of decision trees and decision rules in a single structure; because a particular path in a decision tree can be thought of as a rule that classifies text instances. The main difference between decision trees and decision rules is that DT is a strict hierarchy of data space, whereas rule-based classifiers allow overlapping in the solution space.

2.1.1.2 Weakly, semi and unsupervised learning

The main purpose of text categorization is to classify documents into a specific number of predefined categories. To accomplish this, as noted above, a large number of marked

training documents are used to oversee learning. In text categorization, creating these labelled training documents is sometimes difficult, but it's easy to collect unlabelled documents. Unsupervised learning methods overcome these difficulties. There are many research papers in this area, including the work of Ko and Seo²⁷. They proposed a way to classify documents into sentences and categorize each sentence for each category and sentence similarity measure using a list of keywords. The concepts of weak monitoring and local monitoring are used in many applications. Youlan and Zhou²⁸ proposed a strategy to provide weak monitoring at the feature level, not an example. They get the initial classifier by incorporating the previous information extracted from the existing sentiment lexicon into the affective classifier model learning. They refer to the previous information as labelled features and use them directly to limit the prediction of the model to unlabelled instances using the generalized expected criteria directly. In their work, they are able to identify field-specific polar words, making it clear that the polarity of a word may vary from field to field. They worked on film reviews and datasets on multidomain feelings from IMDB and amazon.com. They have shown that their approach is better than other methods of classifying poorly controlled feelings, and they are suitable for any task to classify a text if there is relevant prior knowledge. Xianghua and Guo²⁹ also use an uncontrolled approach to automatically discover the aspects discussed in Chinese social journals and how they feel in all aspects. They use the LDA model to detect global multidimensional topics from social changes and then extract local topics and related sentiments based on the sliding window context. In the conclusion of the text. They extracted social journals from a series of blogging data (2000-SINA) and a dictionary (300-SINA Hownet). They showed that their approach achieved good results exchange results and helped to improve the accuracy of SA. It also helps to develop a wide range of topics and feelings. There are other uncontrolled approaches that rely on the semantic orientation of PMIs or the similarity of words and polar prototypes using lexical associations, semantic spaces, and distributions of PMIs

30

2.1.1.3 Meta classifiers

In many cases, researchers use one or more classifiers to test their work. One of them is the work proposed by Lane and Clarke³¹. They proposed an ML approach to address the

issue of positive or negative proactive documents in media analysis. The imbalance in the distribution of positive and negative samples, changes in reporting over time, and effective training and assessment procedures for models are the challenges they face in meeting their goals. They used the three data sets the company created to analyze the media. They assessed the document in two ways: they revealed the existence of the beneficiaries and assessed negative and positive responses to the positive and positive aspects. They use five different types of functions to create plain-text records. They tested some classifiers to find the best classifiers (SVM, nearest neighbours KN, NB, BN, DT, student rules, and others). They have shown that allocating accounting categories to training data can increase productivity, but NB may be affected.

Rui and Liu ³² studied the application of ML algorithm in Twitter streaming data. In their work, they examined whether word of mouth (WOM) affects movie sales and assesses dynamic models of panel data. They use NB and SVM for classification. Their main contribution is the classification of tweets, taking into account the unique characteristics of Twitter. They differ in the opinions of the consumer (who has not yet purchased the product) and the consumer (who buys the product). They participated in reviews of reference movies and Twitter data. They collected WOMT WOM data using Twitter API data and Box-OfficeMojo.com. Their results show that WOM has a significantly greater impact on product sales for Twitter users with large numbers of followers than Twitter followers with fewer followers. They found that pre-consumer WOM had more impact on movie sales than WOM before spending. In another article, many classifiers were compared after applying the classifier to the Markov model. It's about capturing dependencies between words and providing a dictionary that improves the predictive performance of several popular classifiers. This was proposed by Bai ³³, who introduced a two-step prediction algorithm. In the first phase, his classifier explores the conditional dependencies between words and encodes them in a labeled acyclic directed graph to represent mutable emotions. In the second phase, he used meta-heuristics to improve his algorithm. Mutual verification accuracy. He has been online to comment on two collections of IMDBs, and his three online news collections compare his algorithms with SVM, NB, ME and others. He said his approach identifies a small set of prognostic signs compared to other approaches and gives better predictions in the direction of the sensation. His conclusion shows that feelings are

captured by the conditional dependencies between words, as well as by keywords or high-frequency words. The complexity of his model is linear in number of samples.

Controlled and uncontrolled ways can be combined. This was done by Valdivia and Kamala. They suggest using metaclassifiers to develop a polar classification system. They worked with the Spanish film critic team to translate their parallel teams into English (MCE). First, they generate two separate models using both of these cases, and then apply machine learning algorithms (SVM, NB, C4.5, etc.). Second, they integrated the SentiWordNet-aware chassis into English cases and created a new uncontrolled model using semantic positioning. Thirdly, they combine three systems using a metaclassifier. Their results have surpassed the results of the use of individual cases and showed that their approach can be considered as a good strategy for classifying polarity in the presence of parallel corps.

Walker and Anand³⁴ used ML classifiers to rank positions. The position is defined as a global position held by the person in relation to the object, idea or position³⁵ For example, the location is similar to the viewing angle or viewing angle, and may be considered an identifier for placing the "side" of the speaker. Agree or disagree with the political decision Walker and Anand³⁶ assessed the position that people occupy and applied it to the political debate. They used convincement's 104 bilateral debates to discuss 14 different topics and to try to determine the speaker's position or attitude. Their main task is to determine the potential contribution to the discussion of the characteristics of horizontal categorization of contextual dialogues. The main role of the context is to compare the context less result to the context, where only five pairs of topics go from one context to another context. They use SVM, NB and a rule-based classifier to classify. Based on their discussion of the subject, they used the sensory, subjective, dependent, and conversational features to get a better classification on a singular basis.

2.1.2 Lexicon-based approach

There are many opinion-based employees employing many emotions the words of positive feedback are used to describe some descriptive statements States, whereas negative feedback is used to express something unwanted states. There are also opinion sentences and idioms which together are called the lexicon of opinion. There are three

main approaches to compile or collect the list of opinion words. The manual approach takes time and is not used alone. It is usually combined with the other two automated ones Approaches as a final check to avoid the errors that have arisen of automated methods. Both automated approaches are presented in the following subsections.

2.1.2.1 Dictionary-based approach

^{37, 38} presented a key-based strategy a small set of opinion words are manually collected with leading events. Then this whole thing grew through research in the well-known WordNet ³⁹ or synonym ⁴⁰ corpus their synonyms and antonyms. Newly discovered words are Add to start list and start next iteration. Iteration When no new word is found, the process stops. After the process is over, you can do a manual check to Remove or correct the error. The verb-based approach is a big loss which failed to find the words of opinion with domain and Context-specific guidelines. Qiu and He⁴¹ used a dictionary approach based on the identification of sentiment sentences in the context Advertisement. They proposed an advertising strategy improve the relevance of ads and the user experience. They used syntactic dictionary of analysis and sentiment and proposed a rule based approach to address the extraction of thematic words and the attitude of consumer's identification in the extraction of advertising keywords. They worked on web forums of [automotvieforums.com](http://www.automotvieforums.com). There the results showed the effectiveness of the project approach advertising to extract keywords and select ads.

2.1.2.2 Corpus-based approach

A corpus-based approach helps to solve the problem of finding opinion words in a context-specific orientation. Its method depends on the syntactic pattern or pattern, appearing with the start list of opinion words to find other opinion words in the big corpus. One of these methods was represented by Hatzivassiloglou and McKeown ⁴². They started with a list of seed opinion adjectives and used them with a set of language restrictions to identify others adjective words of opinion and their orientations. Limitations are for connections like AND, OR, BUT, EITHER-OR. For example, the AND conjunction says that adjectives connect generally have the same orientation. This idea is mentioned the sentiment consistency, which in practice is not always

consistent. There are also adversative expressions such as, but however, those are listed as changes of opinion. To determine whether two adjectives are identical or different Orientations, learning is applied to a large body. Then, the links between the adjectives form a graph and the grouping is complete in the graph to generate two sets of words: positive and negative. The conditional random field method (CRF)⁴³ was Used as a method of training the sequence to extract opinions expression. It was also used by Jiao and Jia⁴⁴ for discriminate the polarity of feelings with a multi-line model corresponding algorithm. Their algorithm is used in Chinese reviews online. They created a lot of emotional dictionaries. They worked in online automotive, hotel and computer magazines. There the results showed that their method achieved high performance. Xu and Liao⁴⁵ used a two-level CRF model uncommitted interdependencies for extracting comparative relationships. This was done using complex dependencies between relationships, essences and words, and un-fixed interdependencies among relations. Their goal was to a graphical model for extracting and visualizing comparative relationships between the products of customer feedback. They published results as comparative maps of the relationship to support decision-making in business risk management. They worked on a mobile client reviews amazon.com, epinions.com, blogs, SNS and emails. Their results showed that their method can extract more accurately than other methods, and their comparative map of relationships is potentially a very effective tool support enterprise risk management and make decisions. Taxonomic approach for extracting the level of functionality opinions and outline them in the taxonomy of Cruz and Trojano⁴⁶. This taxonomy is a semantic representation stubborn parts and attributes of the object. Their main goal was the OM-oriented area. They identified a set of domain-specific resources that capture valuable knowledge about how people express an opinion on this area. They used resources that automatically a set of annotated documents. They worked on three different domains (helmet, hotel and car reviews) of epinones.com. They compared their approach with other independent domains their results proved the importance of domain in order to build an exact withdrawal of opinion because they led to increased accuracy with respect to approaches that are independent of the field. Using a corpus-based approach alone is not as effective as a dictionary-based approach because it is hard to prepare a huge corpus to cover all English words, but this approach has one major advantage that

can help find areas and contexts Specific opinion words and their use of domain corpora. Corpus-based methods use statistical methods or semantic methods, as shown in the following subsections:

2.1.2.2.1 Statistical approach

Find patterns of co-occurrence or Opinions about seeds can be obtained using statistical methods. This can be done by removing the rear polarities, using the coincidence of adjectives in the body, proposed by Fahrni and Klenner⁴⁷. You can use all indexed documents on the Internet as a case for building a dictionary. This overcomes the problem of inaccessibility a few words if the enclosure used is not large enough⁴⁸. The polarity of the word can be identified by studying the appearance of a word in large annotated corpus texts⁴⁹. If the word occurs most often among positive texts, so its polarity is positive. If this occurs more often among negative texts, its polarity is negative. If it has equal frequencies, so this is a neutral word. Similar words of opinion are often found together body. This is the main observation that the current state methods are based on. Therefore, if two words appear together often in the same context they are likely to have of the same polarity. Therefore, the polarity of the unknown word can be determined by calculating the relative frequency Joint appearance with another word. This can be done using PMI⁵⁰. Statistical methods are used in many applications related to be approved One of them is the detection of manipulation of critics while driving a random statistical test called the Runs test. Hu and Bose⁵¹ expected a style of writing comments will be random due to the different origin of customers, if the reviews were indeed written by customers. They worked on reviews of books on amazon.com and found about 10.3% of products reviews of manipulations. Latent semantic analysis (LSA) is a statistical approach which is used to analyse the relationship between a set of documents and the conditions specified in these documents for to create a set of relevant models associated with documents and terms⁵². Cao and Duan⁵³ used LSA for find the semantic features of the reviewed review texts influence of various functions. The purpose of their work is to understand why some critics get a lot of usefulness votes, while others almost do not get a vote at all. So Instead of providing a useful level for exams that do not have they investigated the factors that determine the number of a useful voice that receives a specific exam (including yes and

no votes). They worked on software Comments on CNET Download.com. They showed that the semantic features are more influential than other characteristics, influencing how useful notifications of voting are accepted.

Semantic orientation of the word - the statistical approach used with the PMI method. There is also an implementation a semantic space called Hyperspace Analog to Language (HAL) which was proposed by Lund and Burgess⁵⁴. Semantic space the space in which words are represented by dots; position each point with each axis is somehow connected with meaning of the word. Xu and Peng⁵⁵ developed an approach based on HAL, called the analogue to language of Hyperspace (S-HAL). In their model, information about the semantic orientation of words is characterized by vector space, the classifier was formed to identify the semantic orientation of terms (words or sentences). The hypothesis was tested by the method of deducing the semantic orientation from PMI (SO-PMI). Their approach created a set of weighted functions based on surrounding words. They worked on the news pages and used the Chinese case. Their results showed they surpassed SO-PMI and demonstrated advantages in modelling of semantic orientation characteristics in relation to with the original HAL model.

2.1.2.2.2 Semantic approach

The semantic approach gives a feeling value directly and is based on different IT principles resemblance between words. This principle gives similar results meaning of feelings, to close the words semantically. WordNet for The example provides various types of semantic relationships between the words used to calculate the polarity of the senses. WordNet it could also be used to get a list of mood words, iteratively Expand the original set using synonyms and antonyms and then determine the polarity of feelings for a stranger the word is the relative number of positive and negative synonyms this word⁵⁶.

A semantic approach is used in many applications for construct a lexicon model to describe verbs, nouns and Adjectives for use in SA as a work presented by Max and Vessel⁵⁷. Their model describes the detailed subjectivity relationships between actors in a sentence that expresses separately report for each actor. These relations of subjectivity

are labelled with information on the identity of the relationship owner and orientation (positive or negative) attitude. Their model included semantic categorization corresponding categories for SA. It provided the means for identification relationship holder, the polarity of the relationship, and description of emotions and feelings of different participants participating in the text. They used Dutch WordNet in their Job. Their results showed that subjectivity and sometimes the subjectivity of the subject can be reliably identified.

The semantics of electronic content WOM (eWOM) is used for consider the eWOM content analysis proposed by Pai and Chu⁵⁸. They have drawn positive and negative assessments, and helped consumers make decisions. Their method can be used as a tool to help companies better understand the valuation of products or services and, accordingly, translate these views into business intelligence for use as a basis for improving products / services. They worked in Taiwanese magazines Fast-food. Their results showed that their approach is effective in conducting eWOM assessments related to services and products.

Semantic methods can be mixed with statistical methods perform the SA task as a work presented by Zhang and Xu⁵⁹ who used both methods to find the weakness of the product on the Internet opinion. Their seeker of weakness revealed features and group explicit characteristics using the morpheme-based method for identification comments of functional words. They used Howmet measure of similarity for frequent and infrequent characteristics that describe the same aspect. They identified Implicit functions with a choice based on collocation statistics PMI. They grouped the characteristic words in relevant aspects, applying semantic methods. They used the SA method on the basis of a proposal for determining polarity of each aspect in the proposals, taking into account influence of adverbs degree. They can find shortcomings product, because it was probably the most unsatisfied customer reviews or the most unsatisfactory aspect compared with reviews of products from their competitors. Their results showed good results of weakness researcher.

2.1.2.3 Lexicon-based and natural language processing techniques

Methods of natural language processing (NLP) sometimes used with a lexicon approach

to find syntax structuring and assistance in the search for semantic relations⁶⁰. Moreo and Romero⁶¹ used NLP as a pre-treatment before using their vocabulary-based algorithm SA. Their proposed system consists of autofocusing Sensitive module and sensory analysis module to evaluate the opinions of users on the topics in the news that use a taxonomy dictionary specifically designed for news analysis. Their results were promising in scenarios where the language prevails.

SA, presented by Caro and Grella⁶² was based on a deep analysis of NLP proposals, using analysis of dependence as a pre-treatment step. Their algorithm SA was based on the concept of spreading a feeling that suggested that each linguistic element as a noun, a verb, etc., can have an intrinsic value of the feeling that is spreading through the syntactic structure of the analysed sentence. They introduced a set of syntax rules intended to cover a significant part of the sense of feeling expressed by the text. They proposed a data visualization system in which they It is necessary to filter certain data objects or to contextualize data, so that only information related to the user's request, shown to the user. To do this, they submitted contextual method of visualizing opinions by measuring distance, in textual estimates, between the query and polarity of words contained in the texts themselves. They expanded their algorithm by computing the context polarity evaluation. Their approach has proven highly effective after by applying it to the educational building of 100 restaurant reviews.

Min and Park⁶³ used NLP from a different perspective. They used NLP methods to determine time and time expressions with extraction methods and ranking algorithm. Their proposed indicators have two parameters that fix temporary expressions associated with the use of products and products legal entities in different periods of purchase. They identified important linguistic parameters for parameters through experiment with the survey data of the bypass, using NLP methods. They worked on reviews of amazon.com products. Their results showed that their indicators were useful and free from unwanted bias.

2.1.2.3.1 Discourse information

Importance of speech in SA has recently grown. Information about the speech can be found either among the proposals, or among the proposals of the same sentence. The

abstract of emotions at the discourse level is studied. In^{64 65}. Asher et al.⁶⁶ used five types of rhetorical relations: Contrast, correction, support, result and continuation with information on the moods attached to the annotation. Somasundaran et al.⁶⁷ proposed a concept called the structure of opinions. Components of opinions are opinions and are relations between their targets⁶⁸. They improved their work and studied the choice of design in the simulation of the speech scheme improve the classification of feelings⁶⁹.

The theory of rhetorical structure (RST)⁷⁰ describes how divide the text into intervals, each of which represents a significant part text. Heerschop et al.⁷¹ proposed a structure who prepared the SA document (in part) on the basis of the document speech structure that was obtained by applying RST to level of supply. They suggested that they could improve performance of the classifier of feelings by dividing the text into important and less important text fragments. They used vocabulary for the classification of film critics. Their results showed increasing the accuracy of SC in comparison with the baseline that makes do not take into account the structure of speech.

A novel, uncontrolled approach to detecting an intra-proposal level of speech relations to eliminate ambiguity of polarity was presented by Zhou et al.⁷². First, they determined speech scheme with speech limitations based on polarity on the RST. Then they used a small set of basic sentences to collect a large number of speech were transformed into semantic sequential representations (SSR). Finally, they adopted an uncontrolled method for generating, Weigh and filter new SSRs without call phrases to recognize discursive relations. They worked on the data of Chinese training. Their results showed that the proposed methods are actually recognized defined speech relations and achieved significant improvement.

Zirn et al.⁷³ a fully automatic frame is presented For SA, a fine at the sub-object level, combining several lexicons of feelings and surroundings, as well as speeches reports. They use Markovian logic to integrate polarity estimates various lexicons feel the use of information about relations between neighbouring segments. They worked on product reviews. Their results showed that the use of improved accuracy characteristics of polarity predictions Accuracy reaches 69%.

The usefulness of RST in ranging large-scale polarity blog articles were reviewed by Chenlo et al.⁷⁴. They applied methods at the phrase level to select the key phrases that

have passed general feeling on the subject of blog posts. Then they The RST analysis applied to these fundamental phrases for guidance the classification of their polarity and, thus, evaluation of the polarity of the document in relation to subject. They found that bloggers tend to express their more clearly in the design and assign segments of text, not in the heart of text Himself. Their results showed that RST provided valuable information on the structure of the discourse of texts that can be it is used for more accurate classification of documents in terms of their respected feelings in multi-threaded blogs.

2.1.3 Other techniques

There are methods that cannot be roughly classified as ML Approach or approach based on lexicon. Formal conceptual analysis (FCA) is one of these methods. FCA was invited Ville⁷⁵ as a mathematical approach used for structuring, analysis and visualize data based on the notion of duality, called Galois Theory⁷⁶. Data consists of a set of objects and its characteristics are structured in formal abstractions, called formal concepts. Together they form an ordered lattice concept partial order relations. The lattices of concepts are realized by identifying the objects and their respective attributes for defined domain, called conceptual structures, then the relationship among them. Blurred formal concept the analysis (FFCA) was designed to eliminate uncertainty and fuzzy information. It is successfully applied in various applications of the information field⁷⁷.

FCA and FFCA were used in many SA applications as presented by Li and Tsai⁷⁸. In their work, they proposed a classification structure based on FFCA, for the conceptualization of documents in a more abstract form of concepts. They used the training examples to improve the random results caused by ambiguous terms. They used FFCA to form a classifier using concepts instead of documents to reduce obscurity. They worked on a benchmark (Reuters 21578) and two sets of polarity data on the opinion of the movie and the eBook opinion. Their results showed excellent performance in all data sets and proved its ability to reduce sensitivity to noise, as well as its adaptability in inter-domain applications.

Kontopoulos et al.⁷⁹ also used FCA for assembly domain model of ontology. In their work they proposed Use of ontological methods for a more effective sense Analyze

Twitter messages, breaking every tweet in the totality of aspects pertaining to the subject. They worked on in the field of smartphones. Their architecture gives more detailed analysis of postal opinions on a specific topic because it distinguishes the characteristics of the domain and assigns their corresponding estimates.

Other sensory analysis systems at the concept level were developed recently. Mudinas et al.⁸⁰ anatomy of pSenti - analysis of moods at the conceptual level a system built into the intellectual development based on vocabulary and learning-based approaches. Their system reaches higher the accuracy of the classification of the polarity of the senses, as well as the feeling the detection of force in comparison with systems based on pure vocabulary. They worked on two sets of real-world data (CNET software reviews and movie reviews IMDB). They have surpassed the proposed hybrid approach to the state of the art such as SentiStrength.

Cambria and Havasi introduced SenticNet 2 in⁸¹. They developed SenticNet 2; semantics are generally available and emotional resource for the extraction of beliefs and the analysis of feelings; to overcome the cognitive and emotional gap between data of natural language at the level of words and feelings at the level of the concept transferred by them. Their system was built by means of seismic computing, which is a new paradigm that uses both artificial intelligence and Semantic Web. They showed that their system can be easily integrated into real-world applications to effectively combine and compare and unstructured information.

Systems of sense analysis at the conceptual level were used other applications, such as e-health. This includes patients' opinions analysis⁸² and crowd testing⁸³.

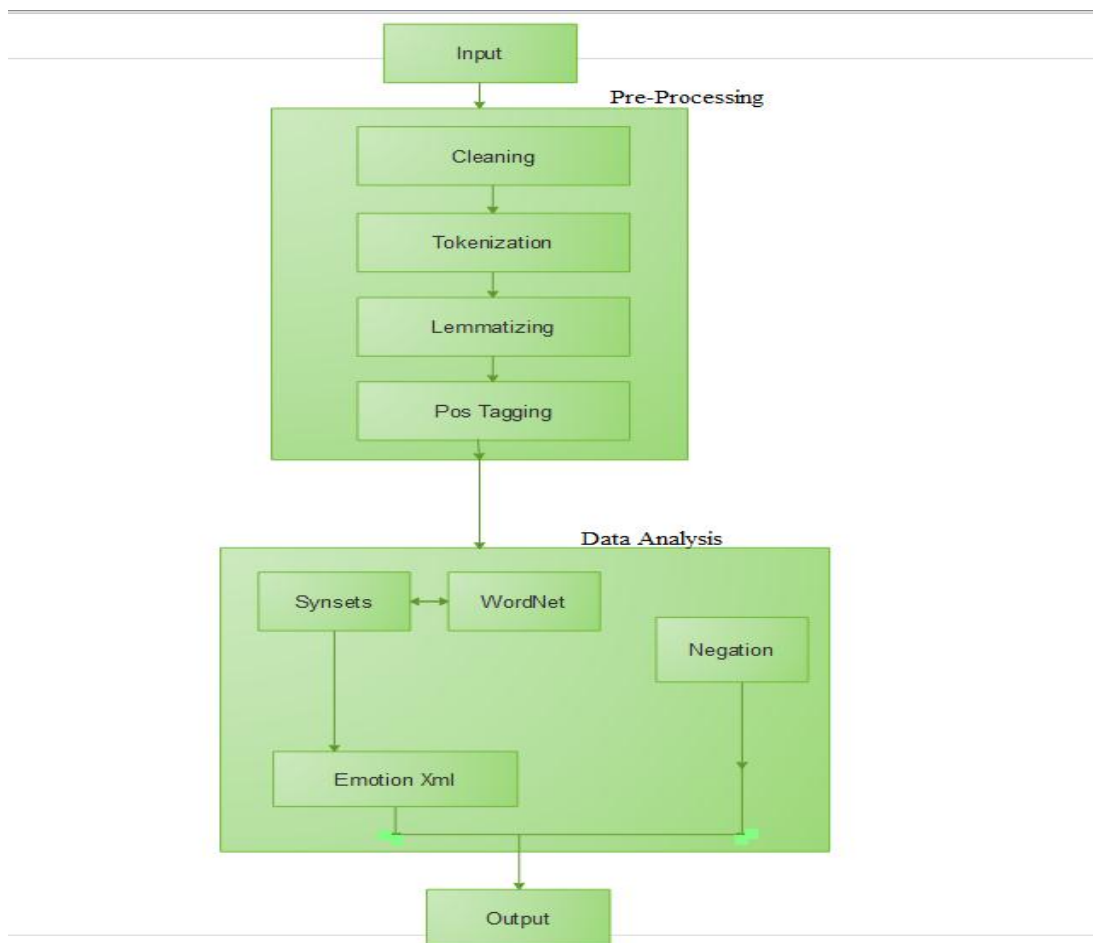
3 Methodology

3.1 Data Pre-processing

Text pre-processing is an important part of NLP, since the sentences, words and characters are determined at this stage. This basic step proceeds to all further processing steps, from analysis and tagging components, such as Tokenizing, stemming and part-of speech taggers.. Texts are pre-processed through these groups of activities. Because the text data often contains some special formats like number formats, date formats and the most common words that unlikely to help Text mining such as prepositions (with, up etc.), articles(a, an etc.), and pro-nouns(he, she, it etc.) can be eliminated.

Need of Text Pre-processing in NLP

- To decrease indexing size of the Text.
- Stop words are not useful for searching
- Stemming used for matching the similar words in a text document



3.1.1 Cleaning

Cleaning is the process of removing the waste parts which are not useful for retrieving information from any process. These waste or useless parts are just time consuming in the process. These are the characters such as punctuation marks, hyphens (-), and brackets etc. which are removed in this section and replace with white spaces.



3.1.1.1 URL Removing

This part also consists on removing the URL which is not necessary for processing because it has no exact meaning and we cannot predict any emotion from URL.

3.1.1.2 Stop word

Stop words are most of the words which are frequently used in sentences to join the words. These words are essentially meaningless. These words are in high frequency which cause obstacle during the processing and understanding the content. Stop words like 'and', 'this', 'are' etc. these words are not useful in classification, so we must remove these words before processing. This process is very important to reduce the size of data and increase the efficiency of the system.

3.1.1.3 Duplication Removing

This part is also necessary to remove the words which are repeatedly used in the text. By removing the words we can reduce the data size and process fastly. This improves the performance of the system.

3.1.1.4 Case Sensitive

All text must be in lower case or upper case but preference is given to lower case text reading.

3.1.2 Tokenizing

The process of breaking stream of text into words, symbols, phrases or other meaningful part called tokens. The main purpose of the tokenization is to identifying the meaningful words in a sentence. This list of tokens further used as input for next processing such as text mining. In computer science tokenization is useful for lexical analysis.

Textual data is in bulk of character format in the starting. Only the word data set is required for retrieval of information from any process. The main use of tokenization is to identifying the meaningful keywords.

Challenges in Tokenization

Tokenization has some challenges which are depend on the type of language. For Example

- I. English and French are used white space between words to make the tokens.

- II. Languages such as Thai and Chinese do not have clear boundaries of word. . Tokenizing unsegmented sentence require additional information. Structured of word also affected on the processing of tokenizing. Structure of languages can be grouped into three categories

Isolating:

Words do not divide into smaller units.

Example: Chinese

Agglutinative:

Words divide into smaller units.

Example: Japanese, Tamil

Inflectional:

Boundaries of words are not clear and ambiguous in terms of grammatical meaning.

Example:

Latin

Here we are using NLTK (Natural Language Tool Kit) word_tokenize to tokenize the text which split the sentence from the white spaces. Before making token data must not contain the stop word, URL and duplication.

3.1.3 Stemming /lemmatizing

Stemming

The process of reducing a word into its stem is called stemming. The most famous example is the Porter stemmer, introduced in the 1980's and currently implemented in a variety of programming languages. This is also called converting word into its root.

Example

The words fish, fishes and fishing all stem into fish, which is a correct word.

3.1.3.1.1 Problem in Stemming

The words study, studies and studying stems into studi, which is not an English word. Most commonly, stemming algorithms (Porter stemmer) are based on rules for suffix stripping.

Stemming is also having over stemming problem which split the single word into pieces. Smile this word split into ➔s m i l e

Lemmatizing

Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word. In order to achieve its purpose, lemmatization requires knowing about the context of a word, because the process relies on whether the word is a noun, a verb etc.

For example, a lemmatiser should map gone, going and went into go

3.1.4 Parts-Of-Speech (POS)-Tagging

What is tagging?

Tagging is the assignment of descriptors to the given tokens. The descriptor is called tag. The tag may indicate one of the parts-of-speech, semantic information, and so on.

What is Parts-Of-Speech Tagging?

The process of assigning one of the parts of speech to the given word is called Parts Of Speech tagging. It is commonly referred to as POS tagging. Parts of speech include nouns, verbs, adverbs, adjectives, pronouns, conjunction and their sub-categories.

Example:

Word:	Paper,	Tag:	Noun
Word:	Go,	Tag:	Verb
Word:	Famous,	Tag:	Adjective

Note that some words can have more than one tag associated with. For example, chair can be noun or verb depending on the context.

Example:

1. He's been chosen to chair the task force on school violence.
2. Chair having four legs and a back for one person

Parts Of Speech tagger

POS tagger is a program that assigns the tag to the word.

Taggers use different kinds of information:

Dictionaries, lexicons and rules etc. Dictionaries have categories of a particular word.

For example, run is both noun and verb. Tagger solves this ambiguity by probabilistic information.

Taggers identify the tag which is best available in the list. In the natural language decision making about tags is difficult due to its complex nature.

Tag Set

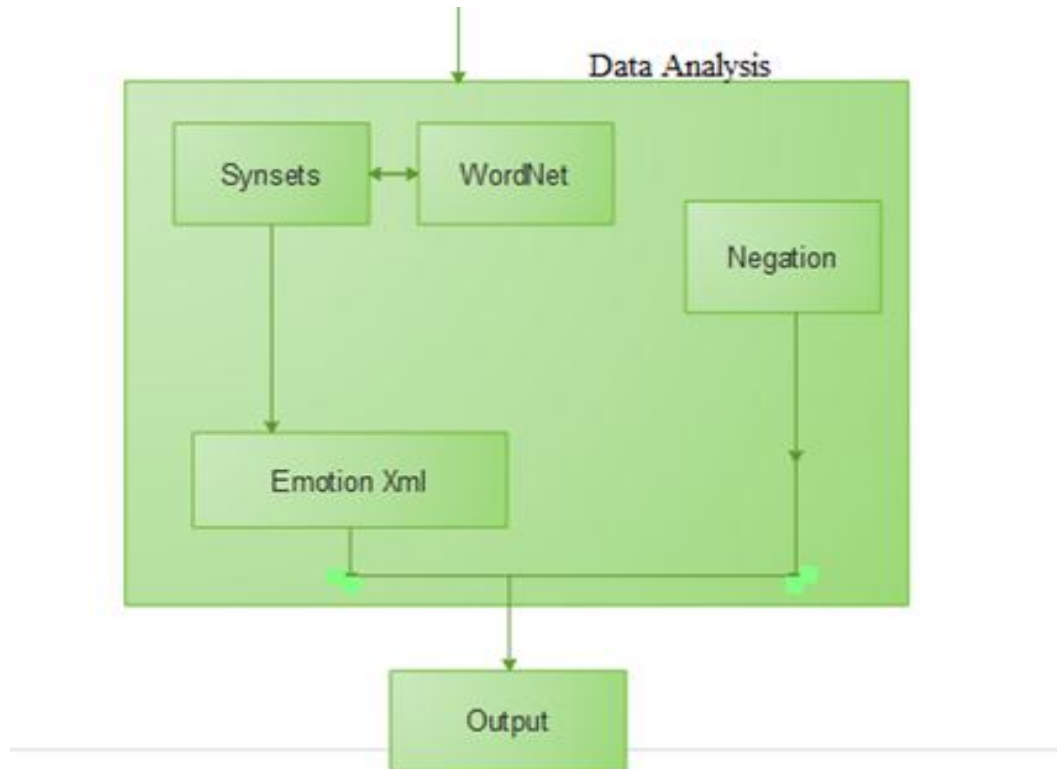
Tag set is the set from which tagger chose the tag for words.

Example:

Tag	Meaning	English Examples
ADJ	Adjective	new, good, high, special, big, local
ADV	Adverb	really, already, still, early, now
NN	Noun	year, home, costs, time, Africa
VER	Verb	is, say, told, given, playing, would
.	punctuation marks	. , ; !
NNP	Pronoun	he, their, her, its, my, I, us

3.2 Data Analysis

It is the main part of emotion prediction. The data which we received from pre-processing examine and pass through the some state to verify that either it containing emotion are not. Before going in to process we need to know about few thing which are bellow.



3.2.1 Synsets

Synsets mean synonym which are having similar meaning of word. We can get synsets by using Wordnet dictionary. To get the synsets we pass word with part-of-speech. By passing POS-tag similar word retrieves which can replace the original word.

Wordnet. Synsets ("happy", "a")

Here in this example “happy” is the word and “a” is the POS-tag which is passing to dictionary.

[Felicitous, glad]

These words can be used instead of happy.

3.2.2 Dictionaries

Wordnet

Wordnet is a dictionary which consist the English word. Nouns, verb, adverb, adjectives are grouped into sets. This contains these word and their synonyms. Wordnet have sematic relation of word. Words are finding explicitly based on their similar meaning.

Emotion.xml

Emotion.xml is the dictionary which contains the set of word which are collected manually with right orientations. Emotion.xml is the dictionary which contains the set of word which are collected manually with right orientations. It consist small set of opinion words. This dictionary will increase in size by searching. The newly found words are appended to the seed list. The iterative process breaks when no new words are found. After the process is completed, manual inspection can be carried out to remove or correct errors.

3.2.3 Negation

Negation is the important part of emotion prediction. If negation is present in the text then the meaning of that sentence is completely changed.

Here is the list of few negation words which can change the meaning of sentence.

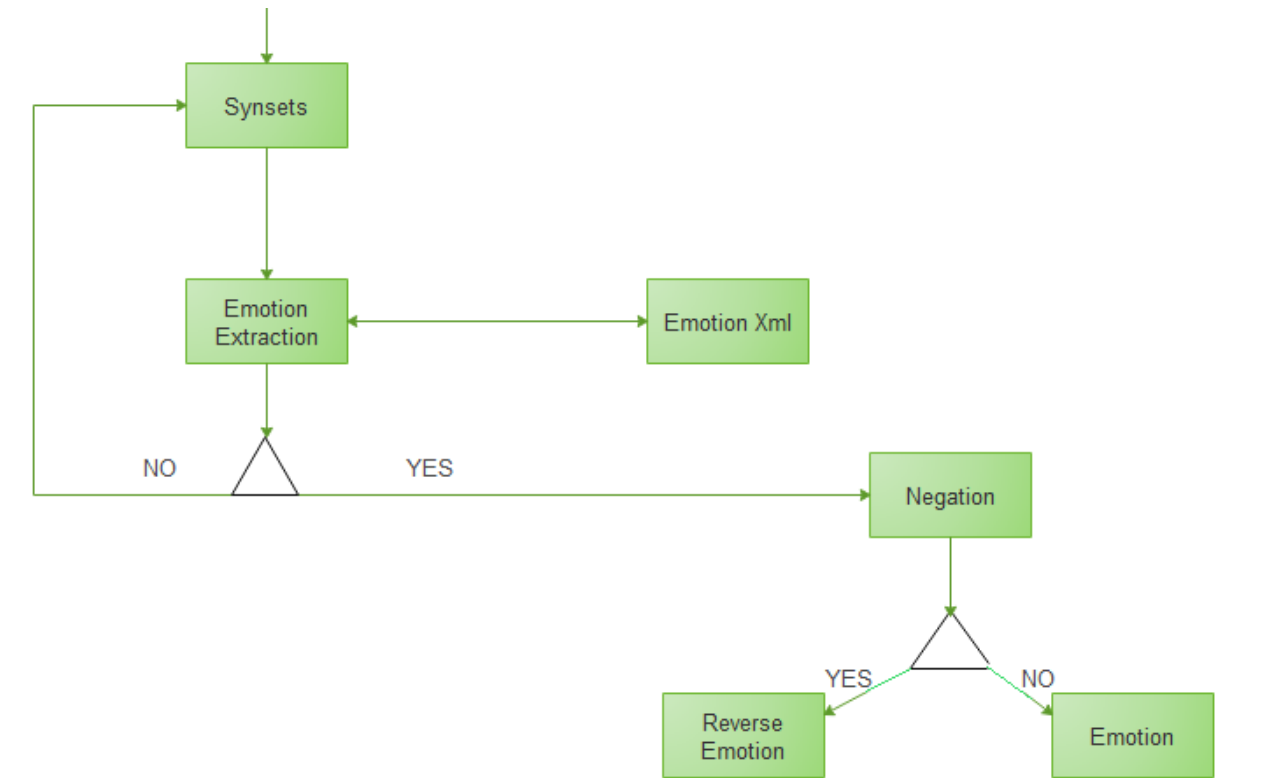
No	Not	didn't	hasn't	doesn't	wouldn't	couldn't
isn't	don't	never	haven't	can't	aren't	nobody
Wasn't						

Example

- I. I am not happy (not negates happy).
- II. Nobody gives a good performance in this movie. (Nobody negates good).

3.3 Flow chart

Synsets are defined in previous section. When the synsets are made then emotion extraction starts and we compare synsets with the emotion.Xml and fetch the appropriate emotion for the synsets. If the emotion extraction fails, then again move to synsets step and if emotion found then go to the negation step. Then we check the negation of emotion, if negation found then print reverse emotion otherwise print actual emotion.

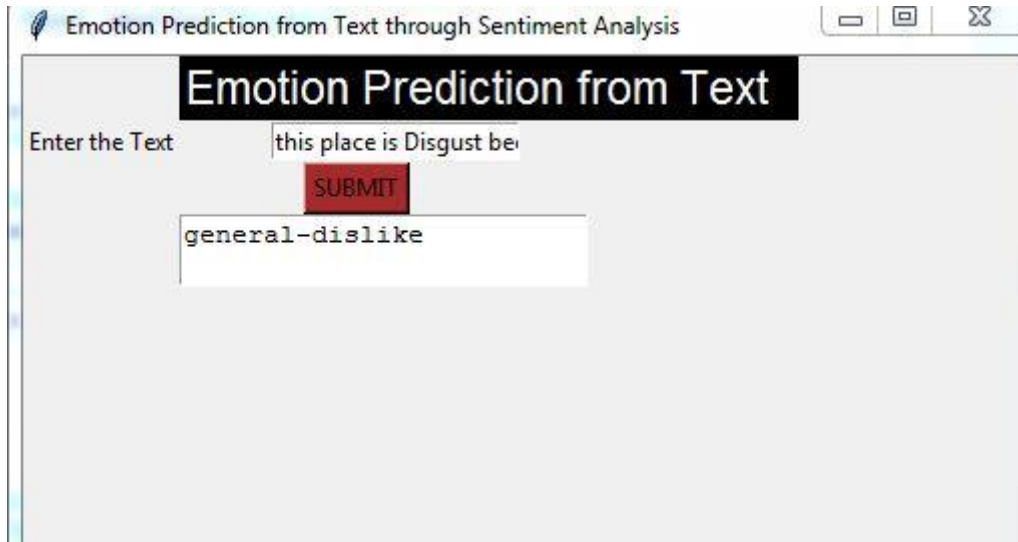


4 Result

Following sentences are passed to get the result.

1. This place is disgust because there is very dirt.

By the human judgment this is negative sentence and the result of this sentence is also negative.



2. I am amorousness Going on trip \$?123*.

Emotion Prediction from Text through Sentiment Analysis

Emotion Prediction from Text

Enter the Text

i am amorousness | Goi

SUBMIT

love

3. **I believe if you keep your faith, you keep your trust, you keep the right attitude, if you're grateful, you'll see God open up new doors. Read more at:**
[https://www.brainyquote.com/quotes/joel_osteen_579036?src=t_positive".](https://www.brainyquote.com/quotes/joel_osteen_579036?src=t_positive)

Emotion Prediction from Text through Sentiment Analysis

Emotion Prediction from Text

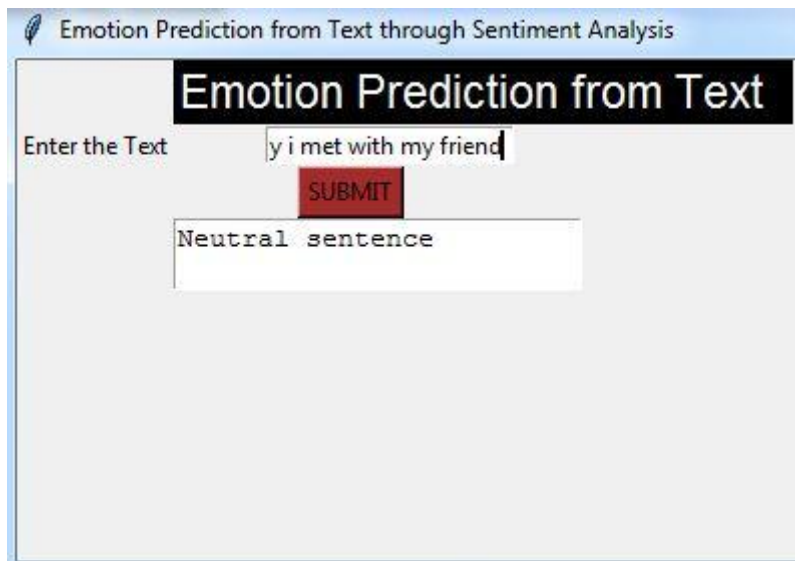
Enter the Text

I believe if you keep yo

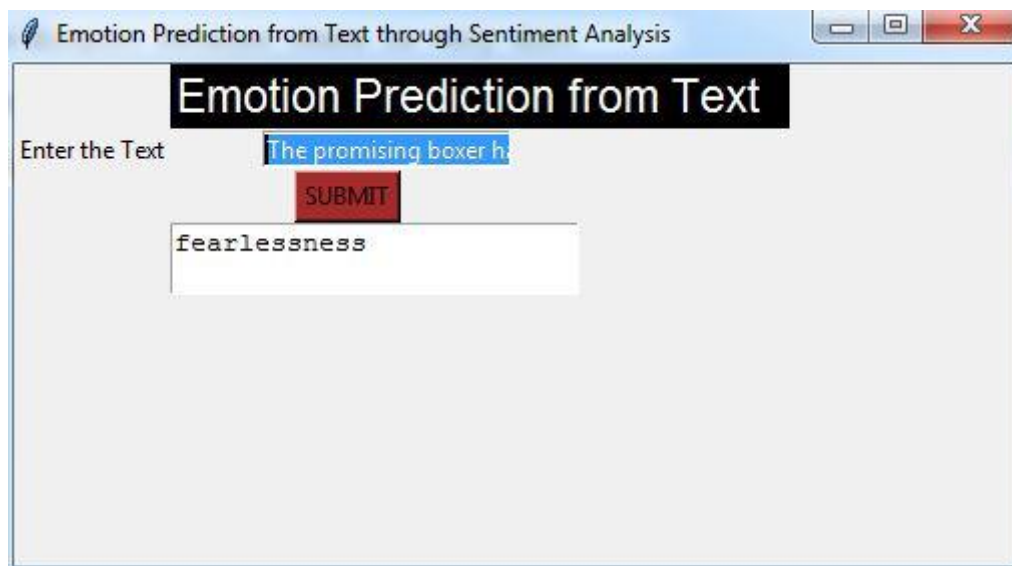
SUBMIT

fearlessness

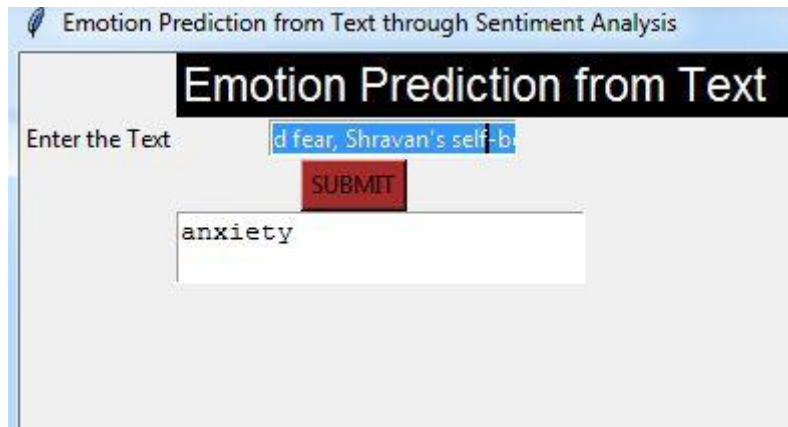
4. today i met with my friend.



5. The promising boxer has confidence and faith in his abilities.



6. Used to thriving on people's insecurities and fear, Shravan's self-belief is perceived as defiance.



Emotion Prediction from Text through Sentiment Analysis

Emotion Prediction from Text

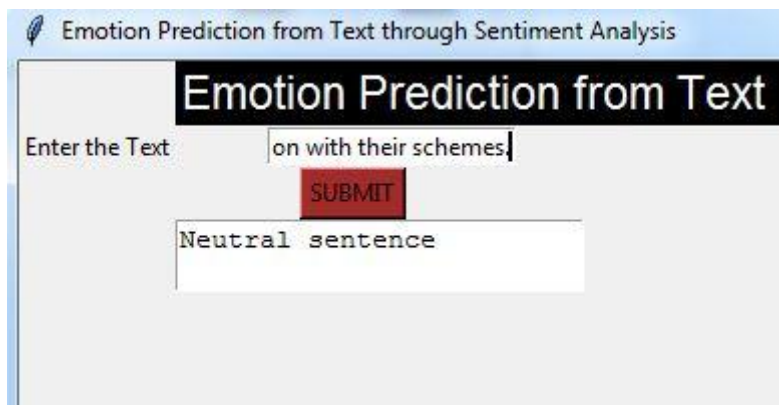
Enter the Text

d fear, Shravan's self-b

SUBMIT

anxiety

7. They dupe their parents and make vacation plans and get caught for them too, but they carry on with their schemes.



Emotion Prediction from Text through Sentiment Analysis

Emotion Prediction from Text

Enter the Text

on with their schemes

SUBMIT

Neutral sentence

8. do you think he will be happy with his presen.

Emotion Prediction from Text through Sentiment Analysis

Emotion Prediction from Text

Enter the Text

do you think he will be

SUBMIT

happy

5 References

¹ T. Wilson, J. Wiebe, and R. Hwa, "Just how mad are you? finding strong and weak opinion clauses," in *aaai*, vol. 4, 2004, pp. 761–769

² B. Agarwal and N. Mittal, "Prominent feature extraction for review analysis: an empirical study," *Journal of Experimental & Theoretical Artificial Intelligence*, pp. 1–14, 2014

^[3] Diana Maynard, Adam Funk. Automatic detection of political opinions in tweets. In: *Proceedings of the 8th international conference on the semantic web, ESWC'11*; 2011. p. 88–99.

^[4] Kang Hanhoon, Yoo Seong Joon, Han Dongil. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. *Expert Syst Appl* 2012;39:6000–10.

^[5] Aggarwal Charu C, Zhai Cheng Xiang. *Mining Text Data*. Springer New York Dordrecht Heidelberg London: Springer Science+Business Media, LLC; 2012.

^[6] Ortigosa-Hernández Jonathan, Rodríguez Juan Diego, Alzate Leandro, Lucania Manuel, Inza Inaki, Lozano Jose A. Approaching sentiment analysis by using semi-supervised learning of multi-dimensional classifiers. *Neurocomputing* 2012;92:98–115.

^[7] Martín-Valdivia María-Teresa, Martínez-Cámara Eugenio, Perea-Ortega Jose-M, Alfonso Ureña-López L. Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. *Expert Syst Appl* 2013.

⁸ Cortes C, Vapnik V. *Support-vector networks*, presented at the Machine Learning; 1995.

⁹ Vapnik V. *The nature of statistical learning theory*, New York; 1995.

¹⁰ Joachims T. Probabilistic analysis of the rocchio algorithm with TFIDF for text categorization. In: *Presented at the ICML conference*; 1997.

¹¹ Aizerman M, Braverman E, Rozonoer L. Theoretical foundations of the potential function method in pattern recognition learning. *Autom Rem Cont* 1964;8:21–37.

¹² Chin Chen Chien, Tseng You-De. Quality evaluation of product reviews using an information quality framework. *Decis Support Syst* 2011;50:755–68.

¹³ Li Yung-Ming, Li Tsung-Ying. Deriving market intelligence from microblogs. *Decis Support Syst* 2013.

¹⁴ Ruiz M, Srinivasan P. Hierarchical neural networks for text categorization. In: *Presented at the ACM SIGIR conference*; 1999.

¹⁵ Ng Hwee Tou, Goh Wei, Low Kok. Feature selection, perceptron learning, and a usability case study for text categorization. In: *Presented at the ACM SIGIR conference*; 1997.

-
- ¹⁶ Moraes Rodrigo, Valiati João Francisco, Gavião Neto Wilson P. Document-level sentiment classification: an empirical comparison between SVM and ANN. *Expert Syst Appl* 2013;40:621–33.
- ¹⁷ van de Camp Matje, van den Bosch Antal. The socialist network. *Decis Support Syst* 2012;53:761–9.
- ¹⁸ Quinlan JR. Induction of decision trees. *Machine Learn* 1986;1:81–106.
- ¹⁹ Lewis David D, Ringuette Marc. A comparison of two learning algorithms for text categorization. *SDAIR* 1994.
- ²⁰ Chakrabarti Soumen, Roy Shourya, Soundalgekar Mahesh V. Fast and accurate text classification via multiple linear discriminant projections. *VLDB J* 2003;2:172–85.
- ²¹ Li Y, Jain A. Classification of text documents. *Comput J* 1998;41:537–46.
- ²² Yi Hu, Li Wenjie. Document sentiment classification by exploring description model of topical terms. *Comput Speech Lang* 2011;25:386–403.
- ²³ Zhao Yan-Yan, Qin Bing, Liu Ting. Integrating intra- and interdocument evidences for improving sentence sentiment classification. *Acta Automatica Sinica* 2010;36(October'10).
- ²⁴ Liu Bing, Hsu Wynne, Ma Yiming. Integrating classification and association rule mining. In: Presented at the ACM KDD conference; 1998.
- ²⁵ Medhat W, Hassan A, Korashy H. Combined algorithm for data mining using association rules. *Ain Shams J Electric Eng* 2008;1(1).
- ²⁶ Quinlan JR. Induction of decision trees. *Machine Learn* 1986;1:81–106.
- ²⁷ Ko Youngjoong, Seo Jungyun. Automatic text categorization by unsupervised learning. In: Proceedings of COLING-00, the 18th international conference on computational linguistics; 2000.
- ²⁸ He Yulan, Zhou Deyu. Self-training from labeled features for sentiment analysis. *Inf Process Manage* 2011;47:606–16.
- ²⁹ Xianghua Fu, Guo Liu, Yanyan Guo, Zhiqiang Wang. Multiaspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon. *Knowl-Based Syst* 2013;37:186–95.
- ³⁰ Read J, Carroll J. Weakly supervised techniques for domainindependent sentiment classification. In: Proceeding of the 1st international CIKM workshop on topic-sentiment analysis for mass opinion; 2009. p. 45–52.
- ³¹ Lane Peter CR, Clarke Daoud, Hender Paul. On developing robust models for favourability analysis: model choice, feature sets and imbalanced data. *Decis Support Syst* 2012;53:712–8.
- ³² Rui Huaxia, Liu Yizao, Whinston Andrew. Whose and what chatter matters? The effect of tweets on movie sales. *Decis Support Syst* 2013.
- ³³ Bai X. Predicting consumer sentiments from online text. *Decis Support Syst* 2011;50:732–42.
- ³⁴ Walker Marilyn A, Anand Pranav, Abbott Rob, Fox Tree Jean E, Martell Craig, King Joseph. That is your evidence?: Classifying stance in online political debate. *Decis Support Syst* 2012;53:719–29.

-
- ³⁵ Somasundaran S, Wiebe J. Recognizing stances in online debates. In: Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP; 2009. p. 226–34.
- ³⁶ Walker Marilyn A, Anand Pranav, Abbott Rob, Fox Tree Jean E, Martell Craig, King Joseph. That is your evidence?: Classifying stance in online political debate. *Decis Support Syst* 2012;53:719–29.
- ^[37] Hu Minging, Liu Bing. Mining and summarizing customer reviews. In: Proceedings of ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD'04); 2004.
- ^[38] Kim S, Hovy E. Determining the sentiment of opinions. In: Proceedings of international conference on Computational Linguistics (COLING'04); 2004.
- ^[39] Miller G, Beckwith R, Fellbaum C, Gross D, Miller K. \WordNet: an on-line lexical database. Oxford Univ. Press; 1990.
- ^[40] Mohammad S, Dunne C, Dorr B. Generating high-coverage semantic orientation lexicons from overly marked words and a thesaurus. In: Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP'09); 2009.
- ^[41] Qiu Guang, He Xiaofei, Zhang Feng, Shi Yuan, Bu Jiajun, Chen Chun. DASA: dissatisfaction-oriented advertising based on sentiment analysis. *Expert Syst Appl* 2010;37:6182–91.
- ^[42] Hatzivassiloglou V, McKeown K. Predicting the semantic orientation of adjectives. In: Proceedings of annual meeting of the Association for Computational Linguistics (ACL'97); 1997.
- ^[43] Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of International Conference on Machine Learning (ICML'01); 2001.
- ^[44] Jiao Jian, Zhou Yanquan. Sentiment Polarity Analysis based multi-dictionary. In: Presented at the 2011 International Conference on Physics Science and Technology (ICPST'11); 2011.
- ^[45] Xu Kaiquan, Liao Stephen Shaoyi, Li Jiexun, Song Yuxia. Mining comparative opinions from customer reviews for competitive intelligence. *Decis Support Syst* 2011;50:743–54.
- ^[46] Cruz Fermín L, Troyano José A, Enríquez Fernando, Javier Ortega F, Vallejo Carlos G. Long autonomy or long delay? The importance of domain in opinion mining. *Expert Syst Appl* 2013;40:3174–84.
- ⁴⁷^[47] Fahrni A, Klenner M. Old wine or warm beer: target-specific sentiment analysis of adjectives. In: Proceedings of the symposium on affective language in human and machine, AISB; 2008. p. 60–3.
- ^[48] Turney P. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of annual meeting of the Association for Computational Linguistics (ACL'02); 2002.

-
- [49] Read J, Carroll J. Weakly supervised techniques for domainindependent sentiment classification. In: Proceeding of the 1st international CIKM workshop on topic-sentiment analysis for mass opinion; 2009. p. 45–52.
- [50] Turney P. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of annual meeting of the Association for Computational Linguistics (ACL'02); 2002.
- [51] Hu Nan, Bose Indranil, Koh Noi Sian, Liu Ling. “Manipulation of online reviews: an analysis of ratings, readability, and sentiments”. *Decis Support Syst* 2012;52:674–84.
- [52] Deerwester S, Dumais S, Landauer T, Furnas G, Harshman R. Indexing by latent semantic analysis. *JASIS* 1990;41:391–407.
- [53] Cao Qing, Duan Wenjing, Gan Qiwei. Exploring determinants of voting for the “helpfulness” of online user reviews: a text mining approach. *Decis Support Syst* 2011;50:511–21.
- [54] Lund K, Burgess C. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behav Res Methods* 1996;28:203–8.
- [55] Tao Xu, Peng Qinke, Cheng Yinzhaoh. Identifying the semantic orientation of terms using S-HAL for sentiment analysis. *Knowl- Based Syst* 2012;35:279–89.
- [56] Kim S, Hovy E. Determining the sentiment of opinions. In: Proceedings of interntional conference on Computational Linguistics (COLING'04); 2004.
- [57] Maks Isa, Vossen Piek. A lexicon model for deep sentiment analysis and opinion mining applications. *Decis Support Syst* 2012;53:680–8.
- [58] Pai Mao-Yuan, Chu Hui-Chuan, Wang Su-Chen, Chen Yuh- Min. Electronic word of mouth analysis for service experience. *Expert Syst Appl* 2013;40:1993–2006.
- [59] Zhang Wenhao, Hua Xu, Wan Wei. Weakness finder: find product weakness from Chinese reviews by using aspects based sentiment analysis. *Expert Syst Appl* 2012;39:10283–91.
- [60] Bolshakov Igor A, Gelbukh Alexander. *Comput Linguis (Models, Resources, Applications)* 2004.
- [61] Moreo A, Romero M, Castro JL, Zurita JM. Lexicon-based comments-oriented news sentiment analyzer system. *Expert Syst Appl* 2012;39:9166–80.
- [62] Caro Luigi Di, Grella Matteo. Sentiment analysis via dependency parsing. *Comput Stand Interfaces* 2012.
- [63] Min Hye-Jin, Park Jong C. Identifying helpful reviews based on customer's mentions about experiences. *Expert Syst Appl* 2012;39:11830–8.
- [64] Asher N, Benamara F, Mathieu Y. Distilling opinion in discourse: a preliminary study, presented at the COLING'08;2008.
- [65] Somasundaran S, Wiebe J, Ruppenhofer J. Discourse level opinion interpretation, presented at the Coling'08; 2008.
- [66] Asher N, Benamara F, Mathieu Y. Distilling opinion in discourse: a preliminary study, presented at the COLING'08; 2008.
- [67] Somasundaran S, Wiebe J, Ruppenhofer J. Discourse level opinion interpretation, presented at the Coling'08; 2008.

-
- [68] Liu B. Sentiment analysis and opinion mining. Synth Lect Human Lang Technol 2012.
- [69] Somasundaran S, Namata G, Wiebe J, Getoor L. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In: Presented at the 2009 conference on Empirical Methods in Natural Language Processing (EMNLP'09); 2009.
- [70] Mann W, Thompson S. Rhetorical structure theory: toward a functional theory of text organization. Text 1988;8, 243–28.
- [71] Heerschop B, Goossen F, Hogenboom A, Frasincar F, Kaymak U, de Jong F. Polarity Analysis of Texts using Discourse Structure. In: Presented at the 20th ACM Conference on Information and Knowledge Management (CIKM'11); 2011.
- [72] Zhou L, Li B, Gao W, Wei Z, Wong K. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In: Presented at the 2011 conference on Empirical Methods in Natural Language Processing (EMNLP'11); 2011.
- [73] Zirn C, Niepert M, Stuckenschmidt H, Strube M. Fine-grained sentiment analysis with structural features. In: Presented at the 5th International Joint Conference on Natural Language Processing (IJCNLP'11); 2011.
- [74] Chenlo J, Hogenboom A, Losada D. Sentiment-based ranking of blog posts using rhetorical structure theory. In: Presented at the 18th international conference on applications of Natural Language to Information Systems (NLDB'13); 2013.
- [75] Wille R. Restructuring lattice theory: an approach based on hierarchies of concepts. In: I. Rival, Reidel, Dordrecht-Boston; 1982, p. 445–70.
- [76] Priss U. Formal concept analysis in information science. In: Presented at the annual review of information science and technology; 2006.
- [77] Li S, Tsai F. Noise control in document classification based on fuzzy formal concept analysis. In: Presented at the IEEE International Conference on Fuzzy Systems (FUZZ); 2011.
- [78] Li Sheng-Tun, Tsai Fu-Ching. A fuzzy conceptualization model for text mining with application in opinion polarity classification. Knowl-Based Syst 2013;39:23–33.
- [79] Kontopoulos Efstratios, Berberidis Christos, Dergiades Theologos, Bassiliades Nick. Ontology-based sentiment analysis of twitter posts. Expert Syst Appl 2013.
- [80] Mudinas Andrius, Zhang Dell, Levene Mark. Combining lexicon and learning based approaches for concept-level sentiment analysis. Presented at the WISDOM'12, Beijing, China; 2012.
- [81] Cambria Erik, Havasi Catherine, Hussain Amir. SenticNet 2: a semantic and affective resource for opinion mining and sentiment analysis. In: Proceedings of the twenty-fifth international florida artificial intelligence research society conference; 2012.

[82] Cambria Erik, Benson Tim, Eckl Chris, Hussain Amir. Sentic PROMs: application of sentic computing to the development of a novel unified framework for measuring health-care quality. *Expert Syst Appl* 2012;39:10533–43.

[83] Cambria Erik, Hussain Amir, Havasi Catherine. Towards crowd Validation of the UK National Health Service. Presented at the Web Science Conf, Raleigh, NC, USA; 2010.