SENTIMENT ANALYSIS USING MACHINE LEARNING

By

Sharan Chhugani
18BCE354

ACKNOWLEDGEMENT

I would like to show my sincere gratitude and appreciation to all those because of whom I was able to take up this title as my seminar as it gave me a new horizon to look at. Special thanks to Dr. Ankit Thakkar , my guide for this seminar without whom delivering quality work in this given period of time would not have been possible. Also I show my appreciation to the faculties who helped me with areas that I was facing a difficulty in.

ABSTRACT

In this report we first understand what sentiment Analysis does and how. The report is divided into four sections. We go through all the pre-processing techniques, feature and its extraction techniques and then techniques to classify data such as sentences, documents etc on the basis of their polarity.

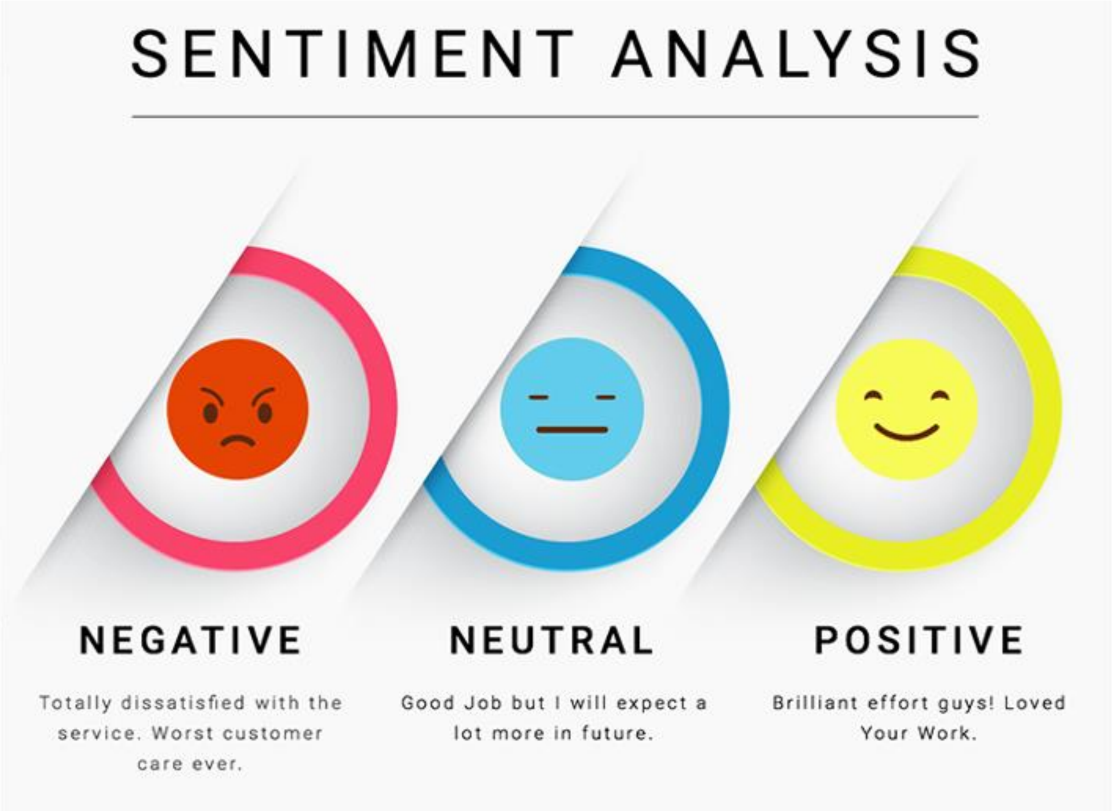
CONTENTS

CHAPTER 1

INTRODUTION

1.1 General

Sentiment Analysis in a broad view is the analysis of textual information which can be feedbacks, reviews, statements, expression or even simple conversations with the use of Natural Language Processing, Machine Learning, CNN and algorithms.


SENTIMENT ANALYSIS
NEGATIVE
Totally dissatisfied with the service. Worst customer care ever.
NEUTRAL
Good Job but I will expect a lot more in future.
POSITIVE
Brilliant effort guys! Loved Your Work.

The analysis is a step by step process to go through each and every area, expression and tone of the context and then give the result if it is positive, negative or neutral.

Sentiment Analysis is in very brief the analysis of sentiment present in conversations and statements to create an aggregate review of a particular product, service, business or event.

How it Works:

Most sentiment prediction systems work just by looking at words in **isolation**, giving positive points for ***positive words*** and negative points for ***negative words*** and then summing up these points.

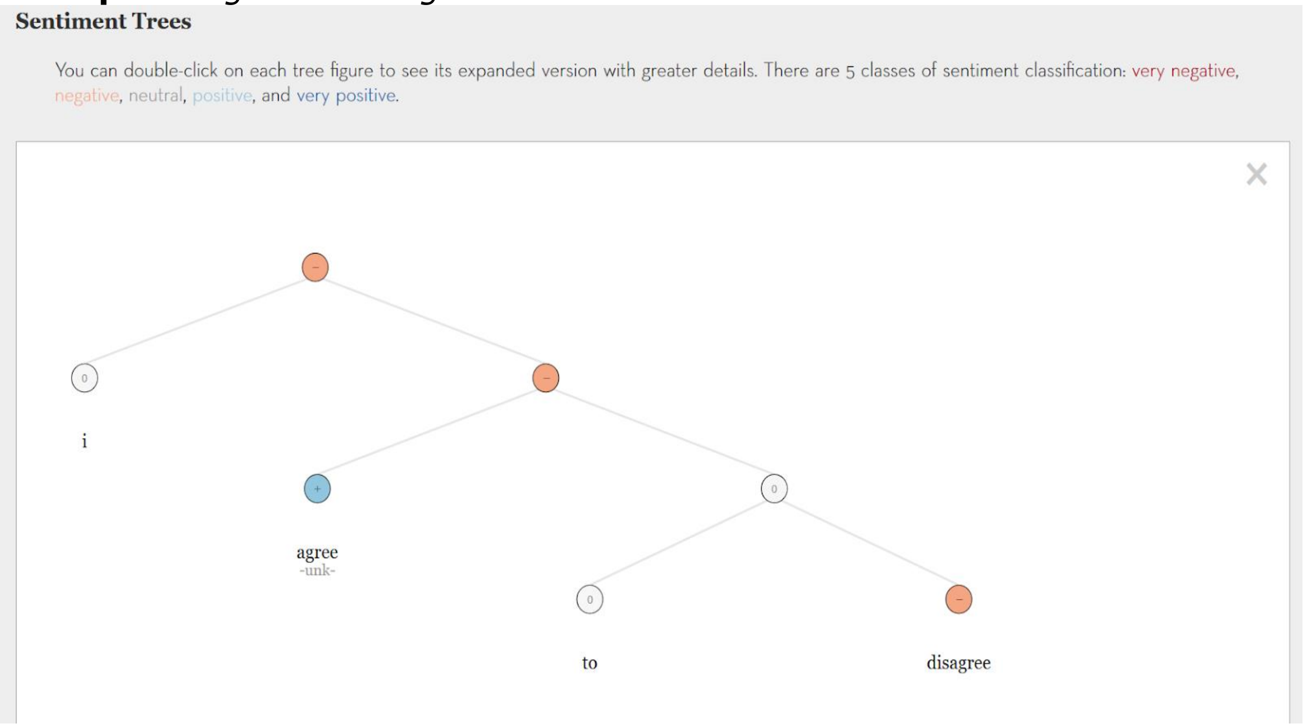
It computes the sentiment based on how words **compose the meaning of longer phrases** instead of the order of the words.

This way, the model is not as easily fooled as previous models. For example, the model learned that *funny* and *enjoyable* are positive but the following sentence is still negative overall

> ***"The video was neither funny, nor enjoyable"***

It was actually trained on a Dataset by Stanford
https://nlp.stanford.edu/sentiment/treebank.html

Example: *I Agree to Disagree*



Sentiment Trees

You can double-click on each tree figure to see its expanded version with greater details. There are 5 classes of sentiment classification: very negative, negative, neutral, positive, and very positive.

From the above example you can see that "**agree**" is a positive term while "**disagree**" is a negative term and the self-learning model was able to tell that it was overall a **negative statement.**

SA counts on four tasks: Data Acquisition, PreProcessing, features extraction and sentiment classification or visualization and summarization of results.

1.2 Scope of Study

This study covers the basic understanding of sentiment analysis, a generic overall process of its working, its applications, areas of problems in sentiment analysis, the pre-processing techniques available, features and feature extraction techniques, techniques to classify data on basis of polarity and brief description of machine leaning techniques.

CHAPTER 2
Literature Survey

2.1 General

The literature survey for this report comprised of papers, renowned articles and authentic lecture videos. These sources explained the concept and need of sentiment analysis. Survey Papers and articles that covered the techniques for the above mentioned purposes were taken into consideration.

CHAPTER 3
Brief Explanation of Sentiment Analysis

3.1 Overview

The analysis is a step by step process to go through each and every area, expression and tone of the context and then give the result if it is positive, negative or neutral.

Sentiment Analysis is in very brief the analysis of sentiment present in conversations and statements to create an aggregate review of a particular product, service, business or event.

3.2 Working of the Analysis

Majorly all the algorithms for classifying the sentiment of the statement work just by looking at words in isolation, giving positive points for positive words and negative points for negative words and then summing up these points.

It computes the sentiment based on how words compose the meaning of longer phrases instead of the order of the words.
This helps the model to tell the difference when a statement only looks negative and when a statement actually is negative or positive. For example, the model learned that agree and disagree are positive and negative words respectively but the following sentence is still negative overall.
"I agree to Disagree"

SA counts on four tasks: Data Acquisition, Preprocessing, features extraction and sentiment classification or visualization and summarization of results.

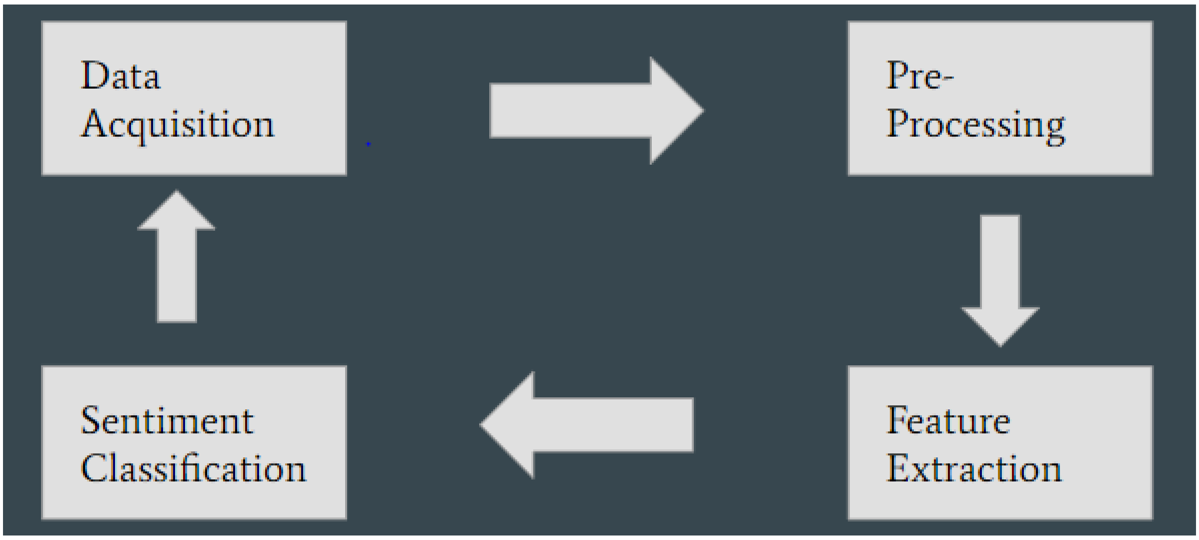


Figure 3.1

CHAPTER 4

Preprocessing Techniques

Pre-Processing techniques help in clearing or filtering the text in order to remove the noise present in the form of unnecessary punctuations, misspelled words and slangs, typographical error and abbreviations.

4.1 Tokenization

This is the most common pre processing technique which is used by almost all researchers. It is defined as "a kind of lexical analysis that breaks a stream of text up into words, phrases, symbols, or other meaningful elements called tokens"[2]. It is "a task for separating the full text string into a list of separate words" [2].

4.2 Remove Unicode strings and noise

All the non English like characters and Unicode (standard 16 bit code for each character) are removed from the dataset which was developed by crawling millions of web pages by using regular expressions. This helps in getting rid of the noise present in the dataset.

4.3 Replacing URLs and user mentions

This is majorly used when the data is extracted from social media. These texts contain user mentions and urls which are normally of no use to the further process. There are several ways to deal with them. Some researchers remove them completely while other simply replace the @ containing string with a particular string(e.g. AT_USER) and remove the # symbol and let the string of the url be as it is.

4.4 Replacing slang and abbreviations

The slangs (informal and typically restricted words and phrases) and abbreviations (shorthand words) present in the text cannot be interpreted as it is so they are replaced by their meaning or their full forms. A lookup table can be used to map these to their appropriate replacements.

4.5 Replacing contractions

The contractions such as don’t and shouldn't are replaced with strings as “do not” and “should not”. One essential reason for this is not to lose the meaning of the word after tokenization (as it will separate it as two words) and another is the need of “not” for other pre-processing techniques.

4.6 Removing numbers

Numbers present in the text are removed since they don't preserve a sentimental value. However this should be performed after the slangs and abbreviations are removed or replaced.

4.7 Removing punctuation

This is a very common technique in pre-processing where all the punctuations as ‘!’, ‘?’, ‘,’, ‘.’ are removed and text is converted to plain text. Though at times due to removal of punctuations leads to unavailability of the intensity of the sentiment.

4.8 Replacing repetitions of punctuation

To solve the above problem, this particular technique is used before the one mentioned above. When a particular punctuation occurs more than once continuously it is replace with a word that preserves the intensity and whose meaning is well understood by the classifier.

Eg!!! Is replaced by “multiExclamationMark”

4.9 Replacing negations with antonyms

In this method the “not” in searched for and the word following to it is taken into consideration. If a particular antonym is found for that word from WorldNet, the string is replaced by it.

4.10 Lowercasing

This is a common technique used by almost all researchers. All the capital letter are converted to lowercase. It helps in reducing the noise and makes the tokenisation easy. When similar words appear, it becomes easier to combine them.

4.11 Handling capitalized words

Lowercasing removes all the capital letters ignoring the fact that when the entire word is in capital it signifies high intensity sentiment. To preserve them, this technique is used before lowercasing. Words with more than one letter capital are searched for and they are replaced with a new word which contains a particular annotation as its prefix.

This helps in recognizing that the word was all capital.

E.g. HAHAHA will be replaced by ALL_CAPS_HAHAHA.

4.12 Removing Stop Words

The words which possess high frequency i.e which are present in major of the documents are said to be stop words as they don't need to be analysed. These stop words are removed from the text.

4.13 Replacing elongated words

Words which have one particular letter more than one time continuously that are meaningless but signify intensity are called elongated words. These words should be replaced by their original words so that the classifiers don’t treat them as different words.

4.14 Spelling Corrector

There is a high possibility of misspelled words in the text. These might lead to inaccurate classification hence these errors should be rectified by auto corrector tools.

4.15 Part-of-Speech (POS) tagging

A part-of-speech label is assigned to each word in the text such as verb (vb), noun (nn), adverb (av) etc to identify and filter only those words that fall into required category for the application. After this technique only the most important words are provided as input to the classifier.

4.16 Lemmatization

In this technique the morphological words are looked for and their inflection is removed to make it as its base form as found in the dictionary. E.g. caught is transformed to catch

4.17 Stemming

The ending letter of the words is removed to transform the word to their root form or stem. This allows us to merge them and reduce the dimension.

E.g. morning, morn

E.g. wishing, wish

4.18 Handling negations

The words that simply imply negation are searched for in the text and the prefix NEG_ is appending to the next word. This makes the identification of the negative words easier for the classifier.

CHAPTER 5

Features and Feature Extraction

5.1 Types of Features

5.1.1 Basic features used in word polarity

1. Sentiment words
2. Emotion icons
3. Exclamation marks
4. Negation words
5. Intensity words(very, really etc)
6. elongated words (e.g. goooood)

7. Unigrams, bigrams, n-grams (spell check, word breaking, prediction etc).
8. Unigrams and Bigrams with Position
9. Adjectives.

5.1.2 Features based on subjective sentence occurrence statistics

Multiple occurrences of the subjective words in a particular review are noted. Their frequency is calculated.

5.1.3 Sentence-level features

All the reviews are pruned to keep only the sentences that are more useful for sentiment analysis. For pruning, thresholds are set separately for every sentence level feature. Sentences with length of at most 12 words are accepted as short and sentences with absolute purity of at least 0.8 are defined as pure sentences.

5.2 Types of Features Extraction Techniques

5.2.1 Delta-tf-idf (term frequency–inverse document frequency) weighting of word polarities.

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t,D)$$

The tf–idf is the product of two functions, term frequency and inverse document frequency.
Term frequency is *raw count* of a term in a document.

The inverse document frequency is a method to know exactly how much information the word provides.
Log (N/(d(- D1 : d = dj))

The aim is to obtain points or frequencies which are context-dependent that may replace the polarities coming from

SentiWordNet. This might help in better differentiation of sentiments.

5.2.2 Bag of Words

Bag of words is a modal which calculates the frequency of the words from the entire corpus of documents. It then creates a matrix of the sentences and the words. The entries are based on if the word is present in that particular sentence. However not all words from all the sentences are taken into consideration, the “n” words with the highest frequency are taken while the others are discarded. This is how the BOW model is prepared.

CHAPTER 6

Problems related to open mining

6.1 Subjectivity Classification

The problem is to differentiate between subjective and objective statements. Subjective sentences consist of opinions and views that show the feeling of the reviewer .While Objective sentences are confined to facts.

E.g.:- "But now we, as a pathologists, need more objective measures because symptoms, to a certain degree, are subjective." (from Times)

S = {s1,..., sn} (set of statements in the document D)

S = Ss U So where

Ss = set of Subjective sentences in the document D

So = set of Objective sentences in the document D.

6.2 Sentiment Classification

It has two different approaches. If the no of classes in which the document is categorized is 2(positive, negative) it is Binary Sentiment Classification and if that is 5(strongly negative, negative, neutral, positive, strongly positive) , it is Multi-Class Sentiment Classification.

6.3 Word Sentiment Classification

Word Sentiment Classification techniques construct semantic datasets which preserve orientation (i.e. positive or negative) of the entire lexicon manually or semi- manually. These techniques further help on document sentiment classification. To classify the semantics orientation of the word or phrase mostly predefined datasets are used.

6.4 Document level Sentiment Classification

From the previous classification of data from the corpus, the document is classified and categorized on the basis of its polarity. This method includes of feature extraction and training classifier.

6.4.1 Feature Extraction

6.4.1.1 Lexical Filtering

1 Based on Hypernym as provided by WorldNet

2 Based on Part of Speech tags.

6.4.1.2 Appraising adjective

6.4.2 Training Classifier

Various techniques like Naive Bayes, maxima Entropy, Support Vector Machine, K-Nearest Neighbor.

6.4.3 Single Domain and Multiple Domains

Algorithm in Multiple Domain

-Train a base classifier from the old labelled data

-Choose a few unlabeled data that are informative to the new domain and label it

-Retrain the Classifier with these special examples.

6.5 Opinion Extraction

Two major areas for opinion Extraction are opinion-bearing word and opinion holder.

-Opinion bearing seed words are collected.

-Expand these Opinion bearing seed words.

-Extract Sentences and phrases from the input document and Opinion bearing words are extracted.

CHAPTER 7

Sentiment Classification Techniques

Sentiment classification techniques are majorly divided into the following categories

7.1 Lexicon Based

Lexicon Based approach is relies on a sentiment lexicon. A sentiment Lexicon is a dataset of tuples consisting of words and their sentiment (polarity).

There are two underlying approaches

-Dictionary Based

-Corpus Based

7.2 Machine Learning Based

Many machine learning algorithms are used as they provide maximum accuracy. They can further be divided into two parts.

-Supervised

-Unsupervised

7.3 Hybrid

These methods combine both the variations i.e. Lexicon based ad machine learning based.

The focus of this paper will be more on machine learning techniques.

CHAPTER 8

Machine Learning Techniques for Sentiment Classification

Supervised learning Methods

Supervised learning methods rely on the predefined labelled document and datasets provided to the classifiers.

8.1 Probabilistic Methods (generative classifiers)

8.1.1 The Naive Bayes Probability

It creates a frequency table and calculates the prior probability (P(c)) by (Nc/N) for all classes where,

Nc - total count of Particular class in training set

N - total count of class in the training set.

Then compute the Conditional Probability for each word and against all the classes.

P(W/C)= (Count(W,C) + 1) / (Count(c) + |V|).

Where

W - Is the word

C - Is the class

Count (W, C) is the count of that particular word in the class

Count(c) - is the total number of words of the data that belong to that class

|V| - is the total num of words in the entire dataset

Finally count the posterior Probability by
P(Class)=(x1*x2*...*xn) * P(C)
Where
X1, x2...xn - are conditional calculated probability of each word against that class.

The class corresponding to the highest resultant value is said to be the class of the sentence or the document.

8.1.1.1 Multinomial Naive Bayes

Almost works the same as naive bayes, while the only difference is the denominator. Instead of all words we only consider unique words.

8.1.1.2 Bernoulli Naïve Bayes (BNB)

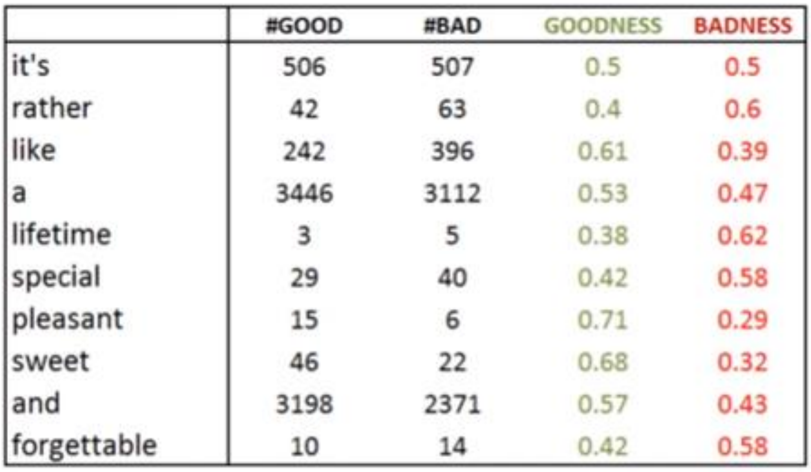
Naive bayes use bayes theorem. Bernoulli is an alternative to that. It also takes into consideration the words that are not present in the document. For the words that are present, the frequency is 1 and words that are not, the frequency is 0.

8.1.1.3 Gaussian Naive Bayes

We first compute the prior probability of each class. Then we calculate a Gaussian for all the features. The Gaussian has two parameters, the mean and the variance. The Gaussian is calculated for all the features and against all the classes. The Gaussian integers of all the features in one class are multiplied and the model is ready. When a new entry is to be classified its Gaussian is calculated for all the features against all the classes and the one with the highest probabilistic value is said to be the resultant class.

8.1.2 Maximum Entropy

This approach has at times proven to be better than naive bayes'.



	#GOOD	#BAD	GOODNESS	BADNESS
it's	506	507	0.5	0.5
rather	42	63	0.4	0.6
like	242	396	0.61	0.39
a	3446	3112	0.53	0.47
lifetime	3	5	0.38	0.62
special	29	40	0.42	0.58
pleasant	15	6	0.71	0.29
sweet	46	22	0.68	0.32
and	3198	2371	0.57	0.43
forgettable	10	14	0.42	0.58

Figure 8.1

The goodness and badness of each word is calculated on the basis of the dataset and the sentiment of the particular review is calculated.

The system works well for straight sentences but with complicated ones and with hidden meanings the system fails at times.

8.1.3 Support Vector Machines

They have very beautifully outsmarted naive bayes technique.

The basic idea behind the training procedure is to find a hyper plane, represented by vector w~. The hyper plane is found by multiplying two or more vector planes which represent different features. This hyper plane divides the features in form of vectors. The hyper plane is found in such a way that the distance between the margin from each document vector and the hyper plane is same and maximum.

The process then comes down to a simple optimization problem. That is if the document belongs to the positive or negative class.

Classification of test instances consists simply of determining which side of w~'s hyperplane they fall on.

8.1.3.1 Linear SVC (LSVC)

It is the fastest and simplest SVM technique. It tries to find a linear plane to differentiate between classes.

CHAPTER 9
Areas Touched by Sentiment Analysis

Currently there are several Applications of Sentiment Analysis

9.1 Social Media

People leave details of their opinions, thoughts and sentiments on social media platforms on various topics, products, entertainment factors and debatable events. To automate the analysis of such data, the area of Sentiment Analysis has emerged in the field of Social Media.

9.2 Business Products

Social media sentiment analysis for B2C and B2B businesses monitors and analyzes this large dataset and generates insights into the consumer mindset, something that businesses have needed since a long time.

9.3 Responsive AI

Google duplex AI is now able to make calls on user request, and have a conversation with the Shopkeeper, restaurant owner, cab service, salons etc. It understands the nuances and pronunciations of the receiver and develops responses accordingly. Its exception handling is quite remarkable.

9.4 Document Categorization

Documents present on the internet were by far being bifurcated in the basis tags attached to them but sentiment analysis has also made it possible to categorise them on the basis of particular sentiments present in them and make the user search more efficient.

9.5 Summarization

Taking feedbacks in natural language format and then summarizing it on the basis of the sentiment it preserves is the key feature for business tools and to increase the brand value.

9.6 Movie Reviews

On -line collection of reviews most of the time summarize their content in form of machine understandable indicator like stars.

CHAPTER 10

Summary and Conclusion

10.1 Summary

To summarize the entire report, it is a detail insight in the area of Sentiment Analysis. For any beginner who wishes to know the subject will find the base idea, sub domains, applications, the entire process of implementation and all the basic techniques that are required to understand the subject to its core. The entire concept and all the possible versions are beyond the reach of the report but an effort to provide vital information and a good start for any beginner has been made.

10.2 Conclusion

The breath of the topic Sentiment Analysis is very wide than one would imagine. This report is an effort to go through the major topics underlying Sentiment Analysis and the techniques used for it.

References

1. "*Thumbs up? Sentiment Classification using Machine Learning Techniques*" by Bo Pang, Lillian Lee and Shivakumar Vaithyanathan published on July 2002.
2. "*A survey on opinion mining and sentiment analysis: Tasks, approaches and applications*" by Kumar Ravi, published by Knowledge-Based Systems in June 2015.
3. *"A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis*" by Symeon Symeonidis, Dimitrios Effrosynidis, Avi Arampatzis, published on June 2018

4. "*A Feature Extraction Process for Sentiment Analysis of Opinions on Services*" by Henrique Siqueira and Flavia Barros
5. "*A survey on sentiment detection of reviews*" by Huifeng Tang, Songbo Tan, Xueqi Cheng published in 2009
6. "*Sentiment Analysis Algorithms and Applications: A Survey*" by Walaa Medhat, Ahmed Hassan and Hoda Korashy, published in april 2014.
7. "*Sentiment and Sarcasm Classification with Multitask Learning*" by Navonil Majumder, Soujanya Poria, Haiyun Peng, Niyati Chhaya, Erik Cambria, Alexander F. Gelbukh , Published in ArXiv 2019
8. "*A survey on opinion mining and sentiment analysis: Tasks, approaches and applications*" by kumar Ravi, published in November 2015.