Sentiment Analysis: Algorithmic and Opinion Mining Approach

Meet Photographer

B. Tech Student, Computer Science and Engineering, Malla Reddy Engineering College, Hyderabad, Telangana, India.

Abstract - *The Machine Learning has many applications which is proven by its vow, in which there is no doubt. By reading a particular book a person can get a idea of what it is about. But in a case of digital information like Blogs, reviews of websites, etc he can't be assured that this information is 100% right. Then in such situations, a field of Sentiment Analysis plays an efficient role about understanding the questions like what do people feel about a certain topic? Did they understand the topic? etc. In this paper, we will discuss how we can analyse the sentiments and the latest techniques that are developed to face the challenges of working with emotional-text.*

Key Words: **Sentiment Analysis, Opinion Mining, Classification, Clustering, Genetic Algorithm.**

1. Introduction

Sentiment Analysis is a broad concept of text classification tasks where we are served with a list of phrases and we are supposed to tell if the sentiments, opinions, and speculations, behind that is positive, negative or neutral. Sentiment analysis can also be known as Opinion mining due to the significant volume of opinions.

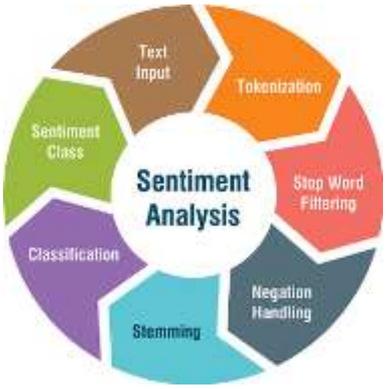

Fig: Structure of Sentiment analysis

From the point of view of machine learning, this task is nothing else but a supervised learning task. Sentiment analysis is the process of identifying and detecting the emotions of the subjective information using the natural language processing and text analysis.[1]

1.1 Example

Consider the statements:

a) "Sinzu saw Strawberry".

Which expresses a sentiment of Sinzu towards strawberry, but it does it doesn't indicate anything about it. We cant say about the sentiment of this statement.

b) "Sinzu hates strawberry".

Which expresses a sentiment of Sinzu towards strawberry, but it does not mean it is false, because the sentiment is negative. Likewise, not all objective sentences are *false*.

C) "Sinzu loves strawberry".

Which expresses a sentiment of Sinzu towards strawberry, but it does not mean it is true, because the sentiment is positive. Likewise, not all objective sentences are true.

Sentiment analysis is the process of identifying and detecting the emotions of the subjective information using the natural language processing and text analysis. From the point of view of machine learning, this task is nothing else but a supervised learning task.

1.2 Types of Questions Arise at the time of Sentiment Analysis

- Is this product review positive or negative?
- Is this customer satisfied with my hotel service?
- On twitter, what will be reactions of people regarding my posts?

Fig: Sentiment Analysis

1.3 Areas of Sentiments Analysis implications

- Extraction of Information.
- Question answering.
- Summarization
- Online shopping because most of the customers purchase products based on the reviews and price.

1.4 Goals of Sentiment Analysis

Because of the complexity that occurs due to the problem i.e. expressions, emotions that are used in text, Sentiment Analysis includes several separate tasks. These are generally combined to produce specific information about the opinions found in text. This section provides an overview of the following tasks.

The first task is the opinion detection, that can be viewed as classification of texts as objective. Usually opinion detection depends on the examination of adjectives used in sentences. For example, polarity of “this is amazing movie” can be examined easily by looking to the adjective.[2]

2. Opinion Mining Approach

In Sentiment Analysis, Opinion Mining plays a crucial role in suggestions. There are two ways to predict an opinion. They are Direct opinion and Comparison.

Direct Opinion: This approach gives direct opinion based on the query. As for example, ”I don’t like this book” – directly gives negative opinion.

Comparison: This approach doesn’t give direct opinion instead shows comparisons between similar objects. For example, “I liked the last book more than this” – compares the two books and specifies that the last book was better than this book.[3]

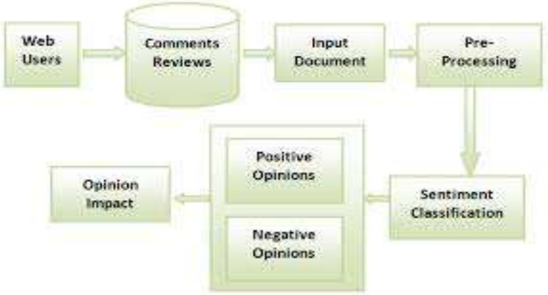

Fig: Work Flow of Opinion Mining

Opinion Mining is generally referred as identifying, extracting, and studying the subjective information provided by the statement using text analysis, Natural Language Processing, etc. We can say that opinion feature extraction is a sub-process of Opinion Mining.

The process in Opinion Mining is divided as follows:

Tokenization is the process used to split up the sentence into tokens by removing the delimiters like white spaces, comas, etc. Stemming removes the excess phrases and reduce the relevant tokens to the single type. Normalization is a process like punctuation that has English texts to be published in both higher and lower case characters and which turns the entire sentence into lowercase or uppercase.

2.1 Feature extraction phase consists of it feature types

1. It identifies its type of features used by opinion mining, feature selection i.e. it is used to select good features of opinion classification, then

2. Feature weighing mechanism i.e. weights each feature for better recommendation and

3. Reduction mechanisms i.e. features of optimizing its classification process.

2.2 Types of feature in opinion mining

1) Term frequency - The presence of a term in the document carries a specific weight age.

2) Term co-occurrence - features which occurs repeatedly like uni-gram, bi-gram or may be n-gram.

3) POS Information - Part of speech (POS) tagger is used to partition POS tokens.

4) Opinion words - Opinion words are the words which expresses positive i.e. good or negative i.e. bad feelings.

5) Negations - Negative words (not or not only) shift sentiment orientation of the sentence.

6) Syntactic dependency - It is represented by a parse tree, it contains the word dependency based features.

2.3 Structure of Opinion Mining

Opinion Mining is also called as sentiment analysis i.e. a process of finding user’s emotions or opinion towards a product or an article. Opinion mining concludes that whether the user’s intension is positive, negative or neutral about a product, article, event, etc. Opinion about the text in reviews, comments, blogs, etc contains subjective information related to the topic. Reviews classified as positive or negative review. Opinion mining and summarisation process involve three steps, first is Opinion Retrieval, Opinion Classification and the last is Opinion Summarization.

2.4 Data Retrieval

This is the procedure of collecting the review text from the review sites. Different review websites may contain different reviews for the products, movies, hotels, etc.

Technique such as Web Crawler can be employed for collecting the review data from many sources and to store them in a database. This step involves retrieval of reviews, blogs and comments that are given by users.

2.5 Opinion Classification

The next primary step that is included in sentiment analysis is a classification of review data. For a Given review document M = {M1….. M1}, a predefined category set, K =

{positive, negative}, sentiment classification is to classify each of the type in M, with a label expressed as in K. The approach involves, classifying the review of text into two types of form namely positive and negative.

2.6 Opinion Summarization

The last step is the Summarization of opinion is a most important character in the opinion mining process. Summary of reviews of the data provided should be based on the features or subtopics that are in the reviews. Many works should be done on summarization of the product reviews. The opinion summarization is the process which involves the following approaches. Feature based summarization is a type of summarization that involves the finding of the frequent terms i.e. the features that are appearing as in many reviews. The summary is analysed by selecting those sentences that contain particular featured information. Characteristics present in review text can be analysed by using the Latent Semantic Analysis (LSA) method. Term frequency is a count of the term occurrences in a particular document. If a term has a higher frequency then it means that this condition is more important for the summary presentation. In many product reviews, certain product features that come out frequently and is associated with user's opinions about it. It is the architecture of Opinion Mining, that says how the input should be classified on various steps to summarize its reviews.[3]

2.7 Basic Tools of Opinion mining

These are the tools which are used in determining the emotions or expressions used by the users in the text in the form of sentences or phrases.

2.7.1 Red Opal

This tool is used by the users to determine the properties or features based reviews of the products. The ratings in which we can see on the products in online shopping websites is being calculated by this software on the basis of the reviews provided by the users. And this ratings are provided on the screen through the means of internet connectivity.

2.7.2 Review Seer Tool

This tool is used to do work related to the aggregation sites which helps to collect the positive and negative sentiments of the particular product based on its features. For this task it uses the Naïve Bayes classifier approach. Then at last the result is displayed as a simple understandable sentimental sentence.

2.7.3 Opinion observer

This is a kind of opinion mining system which that is used to analyze and compare the different opinions on cyber space by using the contents generated by user. This system illustrate the result in graph format clearly showing the opinion of a product feature by its feature. It uses a Word Net-Exploring method to give prior polarity.

2.7.4 Web Fountain

Base Noun Phrase (BNP) Beginning definite heuristic approach is being used here for the extraction of the product features. Development of the simple web interface can also be possible.

The second task is Polarity Classification and Arranging. After the completion of first task our goal is to classify the opinion as one of two opposite sentiment polarities i.e. positive or negative opinion. Mostly, this research is done on the product reviews.

The above mentioned task can be done on several levels like Term, Phrase, Sentence, or at Document level. Here the process is cyclic i.e. the output of one level can be given as the input for other higher layers. As for instance, the result of sentiment analysis of phrases may be supplied to evaluate the sentences and then paragraphs and finally to the documents. Different techniques are available for different levels. Techniques using either n-gram classifiers or lexicons most probably work on term level whereas the Part-Of-Speech tagging technique is used for the phrase and the sentence analysis. Heuristics are frequently used for the generalization of the sentiment to document level.

2.8 Techniques Used in Opinion Mining

The data mining algorithms may be classified into different approaches as Supervised, Unsupervised and Semi-supervised algorithms. Supervised algorithm approach works with set of examples of known labels. Unsupervised approach aims to obtain the similarity of attributes value in the dataset without knowing the values of labels of the example. And the Semi supervised approach is being used in the examples when the dataset is a combination of both the labelled and the unlabelled examples.

3. Algorithmic Approach

Major data mining techniques which are used to gather the knowledge and information are: classification, clustering, association rule mining, genetic algorithm, neural networks, data visualization, fuzzy logic, decision tree and Bayesian networks.

Some of them are explained as follows:

3.1 Classification

Classification is of the type of Supervised technique where every instances belongs to specific and class it is indicated by values of the class attribute or any other special goal attribute. The categorised values are taken by a goal attribute where each attribute belongs to a corresponding class. These different parts which exist in each example are the set of predictor-attribute values and the goal attribute value.

In classification technique, mining function can be classified as set of tasks such as training and test set. In training phase, the model that is used for an effective classification would be formed by the training set and in the testing phase, the model would be evaluated on a test set. The main goal of classification algorithm is that to improve a predictive accuracy in training a model. A hybrid approach of Naive Bayes with Genetic Optimization technique, used to generalize the result as well as comparatively give better result compared to Naive Bayes approach and Support Vector Machine approach. The algorithms and other approaches which are being discussed includes the following:

- Naive Bayes Classifier Approach
- K-Nearest Neighbour
- Support Vector Machines

3.1.1 Naive Bayes Classifier Approach

First technology is the Naive Bayes classifier algorithm which is based on Bayes classification theory. The technique classifies text according to the particular feature of text. This value of particular feature is dependent on a probability of class variables.

Naïve Bayes theorem prepares the system efficiently follow the supervised learning strategy with respect to probability reasoning. The Naive Bayes classifiers, have worked, to solve many of the complex real world conditions. An important and effective benefits of the algorithm is require a small amount of the training data to evaluate parameters like means, variances for text classification. For predicting the future events Bayesian Reasoning is used to applied to make the decision and the inferential statistics which will deals with the probability of inference rule. Probability Rule, according to the Naive Bayes theorem, which are as follows –

P(h/D)= {P(D/h) P(h)}

Where, P(D/h) - Probability of D under given h

3.1.2 K-Nearest Neighbor

The K-Nearest Neighbor Algorithm which is being widely used-for classification, regression and also for non-parametric method. In N-Dimensional space, each attribute is pointing to trains sample with N-dimensional numeric attributes. When the unknown sample is being given to the K-Nearest Neighbour Algorithm, it search for pattern of space for the K-training samples that are very closer to an unknown samples. The Euclidean distance which determine the property of “closeness” measure. When the KNN approach is applied to value, should be appropriate and effectiveness of the approach mostly depends upon the value.

3.1.2.1 Advantages of K-NN Algorithm

It is Robust even in case of large dataset used with noisy training data.

Building of model is easy, efficient and inexpensive.

It can be widely used for Multi-Class Model Classes and for objects used with Multiple-Class labels.

3.1.2.2 Application of KNN Algorithm

In the areas of agriculture, banking for loan management, climate forecasting, medical, news, and for user training purpose

3.1.3 Support Vector Machines

SVM was introduced by-Guyon, Boser and Vapnik, widely used for classification, pattern recognition and regression. SVM has the capability to classify the dimensions or the size of input space. SVM acquires major advantages because of High Generalization Performance with prior knowledge. The Goal of SVM is find the best classification-function, even it aims to differentiate between the members of two classes in training the data. SVM needs to classify given patterns correctly which can maximize the efficiency of SVM Algorithm. SVM use the Vector Space Model (VSM) to separate samples into different classes, viz. done by the learning process of Support Vector Machine. The 3 types of learning process i.e. used in SVM - Supervised, Unsupervised and Semi-Supervised Learning.

3.1.3.1 Advantages of SVM algorithm

It provides greater benefits of text classification when high-dimensional spaces are used.

More prediction accuracy and Better interpretation of the inheritance of data.

It has good ability in learning without depending on dimensions of feature space.

3.1.3.2 Application of SVM algorithm

Used in many problems like Text categorization, as for example in web searching, email filtering, etc.

SVM is used in detecting the breast cancer.

Used in testing and validating the bacterial image.

3.2 Clustering

The clustering is an unsupervised technique that i.e. used to perform natural grouping of the instances. Clustering is a method of dividing data into different groups that too with the similar objects. In clustering each and every group of similar object or data in any respect is called cluster, which differs from the objects of other clusters. Clustering

Algorithm are used for the data compression. The few algorithms of clustering are as follows:

- K-Means Clustering Algorithm
- Self Organized Map(SOM) Algorithm

3.2.1 K-Means Clustering Algorithm

K-Means Clustering Algorithm is the most popular clustering i.e. widely used in most of applications and falls under the Partitioning Algorithms which aims in constructing various patterns and it evaluates them by using few applicable criteria. With a given collection of data, different clusters are formed having an unique characteristics. When the number of n objects was to be group into K clusters, k cluster centre has to be initialized.

3.2.1.1 Advantages of K-means clustering algorithm

K-Means Algorithm provides an appropriate result, when handling with the large data set viz. distinct as well as well separated.

It is used moreover in various no. of applications i.e. image processing, unsupervised neural networks processing, pattern recognition, etc.

3.2.1.2 Applications of K-means clustering algorithm

It is used in Acoustic Data to understand the speech by converting waveforms into the specific category, ML and also in data mining.

Used in segmentation of Colour-based image.

3.2.2 Self organized Map (SOM) algorithm

Self organized Map is one of the type of Artificial Neural Network (ANN) i.e. Unsupervised learning methodology viz. introduced by professor Kohonen so that it is also known as Kohonen's Self-Organizing Map. It is mostly used in Vector Quantization and is used to detect features that may inherits to the problem and thus known as Self-Organizing Feature Map. The SOM consists of several components known as neurons or nodes. This each node will be assigned a specific weight in output space which reflects the cluster content.

3.2.2.1 Advantages of SOM algorithm

Since SOM using the Unsupervised Learning Method, it doesn't need any human interference except the input data.

Used in Vector Quantization and can be applied for comparing the variety of maps with different sizes.

3.2.2.2 Applications of SOM algorithm

SOM is used in Speech recognition, representation for the spectra of different speech samples and in voice analyse applications.

SOM is used to identify the sleep ECG by using cluster of decisive data and to monitor ECG signal with 2-D display effect for the trajectory.

3.3 Genetic Algorithm

Genetic Algorithm is an optimized technique which is derived from the Darwin's Principle. It gives an Adaptive Procedure for the survival of first Natural Genetics. GA-maintains the number of potential solutions of candidate problem which can be termed as individuals, by the manipulation of these individuals with the help of genetic operators like Crossover, mutation, Selection.

4. CONCLUSIONS

To get the solution of any type of problem the main hectic work is dataset which becomes the key factor. Once the dataset is selected then based on it any kind of mining algorithm can be explored. Then the further issue is of selecting the approach i.e. on the base of dataset and an application we can select Supervised, Unsupervised Approach or the combination of duo i.e. classification and clustering algorithm for accurate result.

In this paper, there is a discussion of few algorithms which are widely used to extract emotions i.e. Sentimental Analysis such as Naive Bayes Classifier, KNN, Support Vector Machine, K-means clustering and Artificial Neural Network.

References

1) https://www.datacamp.com/community/tutorials/simplifying-sentiment-analysis-python

2) Https://www.researchgate.net/publication/283954600_Sentiment_Analysis_An_Overview_from_Linguistics

3) dfad8c1bf88b0afc716758c77d533ded7dd0.pdf

4) V4I10-0386.pdf

5) V6I2-0128.pdf

BIOGRAPHY

Meet Photographer is pursuing his B. Tech degree in Computer Science and Engineering from Malla Reddy Engineering College (Autonomous), Hyderabad, India. His current interests include Natural Language Processing.