A SCRUTINIZED STUDY ON THE ALORITHMS AND APPLICATION OF SENTIMENT ANALYSIS

[1]Dr. B.Jagadhesan, MCA. MBA. M.Phil., PhD, [2]R.Georgina Shefani.

[1]Associate Professor & Head of the Department of Computer Science,

Dhanraj Baid Jain College,

Chennai, India,

Email:bjagadhesan@gmail.com

[2]M.Phil Scholar Department of Computer Science Dhanraj Baid Jain College,

Chennai, India,

Email:georgina.shefani@gmail.com

Abstract

The availability of social media to the mass has enabled people to publicly voice their beliefs. Due to the increase in the number of brands and products available to the common man, there is a huge choice that is also available. Hence to sustain the business the product should fulfil the customer requirements and stand stable. Sentiment analysis has been of great use as a tool in analysing the emotions and opinions that will be used by the businesses to improvise and modify the product. Sentiment analysis is also called as opinion mining because in most of the cases it involves opinion of the people. This paper discusses the algorithms that are involved with sentiment analysis and its application on how it is being used.

Keywords: **Text Mining Algorithm; Knowledge Discovery; Applications; Information Extraction; Information Retrieval; Patterns**

I. Introduction

In the world of automation various fields have been implementing various methods to simplify the process. The main goal of text mining is the extraction of patterns .By using sentiment analysis we will be able to find the reactions or emotions of a particular product or brand through various modes .By finding out the pattern we will be able to modify or improvise the product or to launch more products to build a stronger business and to satisfy the requirement of the customers. This paper has listed the various algorithms that can be implemented for the extraction of patterns and its various applications.

II.Sentiment Analysis Algorithms

Sentiment analysis uses various Natural Language Processing (NLP) methods and algorithms, which we'll go over in more detail in this section. The main types of algorithms used include:

- Rule-based systems use a set of manually crafted rules to draw a conclusion.
- Automatic systems rely on machine learning techniques to analyse and learn from data.
- Hybrid systems combines both rule-based and automatic approaches.

A. Rule Based Algorithm

A rule-based system generally uses a set of human-crafted rules to help identify subjectivity, polarity, or the subject of an opinion. These rules may include various techniques developed in computational linguistics, such as:

- Extract the text
- Tokenization split the words in the sentence
- Remove the stop words.
- Exclude punctuation
- Running this pre-processed text against the sentiment lexicon will provide the inferred emotion

Consider the following example **'Product XX is very good'**

TABLE 1

Tokenization	**Pre-processing**
Product	
XX	
is	Stop word
Good	
.	Punctuation

TABLE 2

Product	Neutral
XX	Neutral
good	Positive

It gives a positive sentiment for the above sentence because of the word good. It found out that the word good is a positive lexicon hence it returned a positive sentiment for that sentence. The rule-based system works in way that it first defines two lists of polarized words, that can be a negative or a positive word. It then counts the number of positive and negative words that appear in the given comment or text, and according to that it returns whether it is a negative or a positive sentiment by comparing the most number of the words in the line. Rule based approach is weak because it is built in a way that it does not understand how a particular word appears on a sentence. Though u can add more rule to support the process the new result that is produced after adding the new rule changes the previous result totally.To tackle the real-world data which is complex we need to employ different lexicons at different levels in order to get the right outcome.

- VADER (Valence Aware Dictionary and Sentiment Reasoner): Widely used in analysing sentiment on social media text because it has been specifically improvised to analyse various sentiments expressed in social media

- Text Blob: Very useful Natural Language Processing library that comes pre-packaged with its own sentiment analysis functionality. It is also based on NLTK.

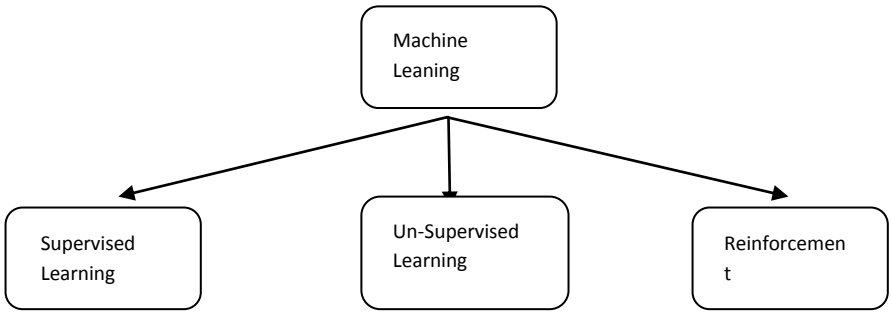
- Sentiwordnet: This is also built into NLTK. It is used for opinion mining. This helps in deducing the polarity information from the given problem instance

Pros	**Cons**
No prior training is needed	Specific to a given situation e.g. the Vader lexicon works well for social media hence customisation is not a choice
Execution is faster	The rules are applied on the dataset without considering the dataset, hence rules are independent of the text
Can work on smaller problem instance	Creating rules need the help of experts

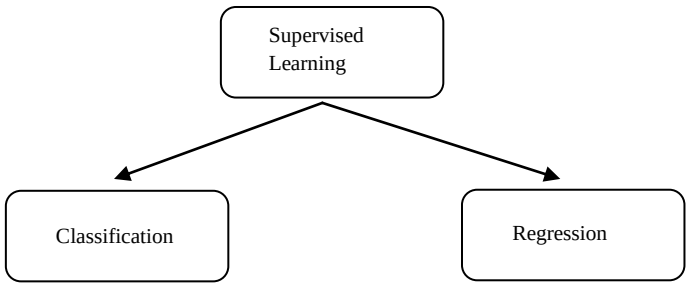
B. Machine Learning or Automatic Approaches

Automatic methods, contrary to rule-based systems, don't rely on manually crafted rules, but on machine learning techniques. A sentiment analysis task is usually modelled as a classification problem, whereby a classifier is fed a text and returns a category, e.g. positive, negative, or neutral.

Here's how a machine learning classifier can be implemented: Machine learning falls into three different categories.



1) Supervised Learning: The machines are trained on mathematical models with the input data that is pre-labelled. The trained the model is then tested against more data and this time the model generates the predictions. The difference must be less between the correct and incorrect predictions. Generally, the machine learning models can be better trained on a relatively bigger corpus of data.



- *Classification:* When the outcome is definite or predictive it is called classification. e.g.red or blue, positive or negative. The classification step usually involves a statistical model like Naïve Bayes, Logistic Regression, Support Vector Machines, or Neural Networks. There are many classification algorithms. They are

a) Naïve Bayes: a family of probabilistic algorithms that uses Bayes Theorem to predict the category of a text.

b) Linear Regression: a very well-known algorithm in statistics used to predict some value (Y) given a set of features (X).

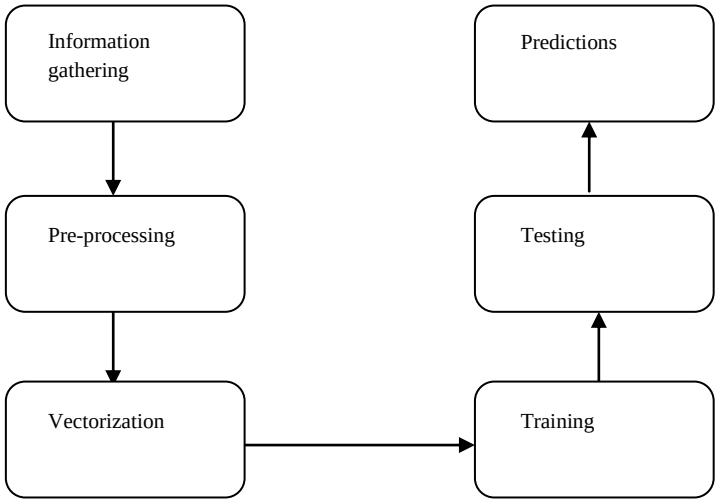
c) Support Vector Machines: The Support Vector Machine (SVM) algorithm is a popular machine learning tool that offers solutions for both classification and regression problems

d) Deep Learning: a diverse set of algorithms that attempt to mimic the human brain, by employing artificial neural networks to process data.

- *Regression:* When the outcome is continuous or prone to continuous changes it is called regression. e.g. weather forecasting or stock market.

2)Un-supervised Learning: Unsupervised learning is a type of machine learning that looks for previously undetected patterns in a data set with no pre-existing labels and with a minimum of human supervision

The steps followed in the machine learning process



- Information gathering: Building the corpus of the data that will be used in training the model
- Pre-processing: Removing the stop words and punctuation
- Vectorization: Converting the textual data into numeric is called vectorization of the text this is because machine learning does not understand text rather, they understand numbers
- Training: The numerically transformed data is split into training and testing sets. The training set is used to train the Machine Learning classifier by providing both the features and labels as inputs. There are multiple classifiers that can be experimented with and they are available in toolkits and libraries, out-of-the-box. The key point is to start exploring because there is not a single algorithm that fits everything, there are different classifiers suitable for different needs. Hence only after experimenting one can conclude on which is best for what types of analysis.
- Testing: During the testing period the model adapts and learn to associate a selected input to a corresponding output and it also checks for sample that are previously used in training. The text

input is first transferred into the feature vector and this is done by the feature extractor. To generate the model the feature vector and the tags are paired and fed into the machine learning formula. In the prediction method, the feature extractor reworks the text inputs into feature vectors. These feature vectors are then fed into the model, that generates fore fold tags (i.e. positive, negative or neutral).

- Predictions: After the testing has been completed it is time to employ the model and the give predictions with the dataset provided

C. Hybrid Approaches

Hybrid systems combine the desirable elements of rule-based and automatic techniques into one system. One huge benefit of these systems is that results are often more accurate. There are two major Sentiment Analysis methods. Let's look at both.

III. Applications Of Text Mining

There are many usage and application depending upon the scope in which it is applied.

A. Forecast and Anticipation of Crimes:

The internet is a mode through which millions and millions of people communicate with each other. By using the normal means, it would be difficult to point to a message that can be considered a threat, hence by using advanced text analysis software we can find messages from various communication sources on finding the threat alerts. Various law enforcement sources around the world employ these kinds of text analysis tools to prevent various terrorist attacks and unlawful activities

B. Risk Management:

Every organisation it be small or big needs be aware of any risks they are or will be facing in the near future. And because of this there is a huge demand in the recent years for risk analysis. Many organisations including banks are deploying these tools to help them decide on various thing like investing in the correct place, or even on the right person to provide loans with. The text mining technologies used by such high-end software absorb petabytes of data and present information in a consumable format. This helps in risk mitigation. Such software is helping financial institutions all over the world, to decrease their percentage of non-performing assets.

C. Knowledge Management:

Many big organisations like the healthcare manages a huge volume of data, and this amount if data keeps increasing every minute and building in physical storage to store all the information is a huge a=task and even though physical storage is possible sometimes it can be a huge mess when trying to retrieve a particular text or data. Because going through tons of files and folders are a huge task, When hospitals are considered they need to store the patient files and also they need to promptly retrieve them when needed, this would be difficult without the any help of a very good text analytics system that organises and manages the data and would allow easy access to the data according to the region,name,or any category we need.

D. Customer Care Services:

The use of text mining and natural language processing are being used in many customer care services. Times have changed from pressing one for recharge from using the same in banking for so many different services. The format changes to time to appear more humane. Many e-commerce services use this software to mimic human tones there by improving the interactions with the customer.

.

E. Scam Detection By Insurance Companies:

There has been a rise in the insurance fraud, and again text analytics that goes through the huge volume of data that has been stored and this saves a lot of time for the company officials because as soon as the software finds an error with the particular file or record it automatically flag the record, and this

marks it highly probable to determine the fraud. Even though the software alone is not a fool proof way to determine a fraud. It helps to direct human attention to the cases that are needed.

F. Personalized Advertising:

Digital marketing has grown to an enormous level and has seen great revolutions. It relates to the type of data that you type and search and it customises it for your purpose and this is the reason the product you were interested in amazon has suddenly appeared int your Facebook page. Companies show advertisement that has a high probability of you clicking on and that in turn can be converted into a sale

G. Business Intelligence:

Decision making is very crucial, more important when u need to answer somebody on why u took a particular decision and how that decision will bring out a positive aspect to the growth of the company. Text mining gives you that extra edge that supports your gut feeling. It might be easy for a team to go through tons of documents regarding strategic information but it might come handy when u only verify related documents so that it saves u from the petabytes of data u have on the same and help draw a better conclusion.

H. Content Enrichment:

Creating a content is the one thing that an artificially created bot cannot do still. However, it can collect tons of information that is related to the topic that you are currently working on. It can also collect the latest news and the articles that are most viewed and this helps in making a calculative guess on how the articles should be formed with the tons of information from the pre-existing documents available helps you to create informative content.

.

I. Spam Filtering:

E-mails have become the official mode of communication in most of the organisations. However, with this significant rise in the spam folder. Out of ten emails we get at least nine are spam. Spam not only occupies space but also is the main entry point of viruses and scams. Many companies are striving hard to filter spam by using many text analytics to create a healthier experience.

IV. Conclusion

Text mining is an emerging sphere of data mining that is used to gather knowledge of the huge mass of data that may be in any form such as comments, feedbacks, reviews etc. Now sentiment analysis is the proceeding trend that uses data from social networking sites to gather data and draw a conclusion. In this paper various algorithms that are employed with sentiment analysis are discussed for efficient and accurate text mining. By employing the various algorithms associated with text mining one can derive meaningful patterns that can be useful for the development of various businesses and along with that the applications of text mining in various environment and businesses are also explained. We could find that almost every business directly or indirectly employs sentiment analysis is every way possible.

References

[1] S. M. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, Text mining: predictive methods for analyzing unstructured information. Springer Science and Business Media, 2010.
[2] S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, “Data mining techniques and applications–a decade review from 2000 to 2011,” Expert Systems with Applications, vol. 39, no. 12, pp. 11 303–11 311, 2012..
[3] K. Sumathy and M. Chidambaram, “Text mining: Concepts, applications, tools and issues-an overview,” International Journal of Computer Applications, vol. 80, no. 4, 2013..
[4] N. Padhy, D. Mishra, R. Panigrahi et al., “The survey of data mining applications and feature scope,” arXiv preprint arXiv:1211.5723, 2012.
[5] Juan Jose Garcia Adeva and Rafael Calvo, “Mining Text with Pimiento”, University of Sydney.
[6] Rashmi Agrawal, Mridula Batra, "A Detailed Study on Text Mining Techniques", IJSCE, ISSN: 2231-2307, Vol. 2, Issue-6, January 2013.