ISSN 2319 – 7595
Volume 8, No.2, March - April 2019
International Journal of Information Systems and Computer Sciences
Available Online at http://warse.org/IJISCS/static/pdf/file/ijiscs35822019.pdf
https://doi.org/10.30534/ijiscs/2019/35822019

Sentiment Analysis in Product Reviews using Natural Language Processing and Machine Learning

Kuncherichen K Thomas[1], Sarath P Anil[2], Ebin Kuriakose[3], Neema George[4]
MLMCE, India, Kuncherichen13@gmail.com
MLMCE, India, sarathambalapatt18@gmail.com
MLMCE, India, ebinkuriakosek@gmail.com
MLMCE, India, neema.george@mangalam.in

ABSTRACT

Lately, we have seen a twist of online web based business sites. It shows an extraordinary chance to share our surveys and evaluations for different items we buy. Looking to the rating can't the only one help a client to get an outline about the item rather the most ideal route is to peruse the audits about the item. Be that as it may, at that point a fascinating issue comes up. Imagine a scenario where the quantity of surveys is in the hundreds or thousands. Which comprise of 10 to 15 pages at that point it's simply not possible to experience each one of those surveys because of wastage of time and exertion. Here comes the significance of audits. To mine profitable data from audits to comprehend a client's inclinations and make a precise end pivotal. In this work, we propose a sentiment based rating expectation technique to take care of this issue.

Key words: Sentiment Analysis, Opinion Mining, Stemming, rating prediction, VC dimension, TFIDF.

1. INTRODUCTION

In the seasons of today, the world is running with Ecommerce stores surrounding us. About all business stages practically are E-commerce store. With simple access to the Internet all over and learning about the technique, the market for Ecommerce has blasted to radiant statures in the ongoing past. There are various parameters which add to characterize the achievement and believability of an Ecommerce store. Be that as it may, one critical factor in elevating the reputation, standard and assessment of an Ecommerce store is Product Reviews. Product Reviews furnish an Ecommerce store with one of the most valuable resources available i.e. Customer Feedback.

One imperative assignment for the Ecommerce store is to keep up its reputation in the online market. Naturally, it requires a ton of effort to pick up that reputation however it costs only very little to lose it: Product Reviews are the most ideal approaches to keep up their series of wins. Item Reviews and criticisms have changed the amusement for online market since web has turned into a very common thing. The Product Reviews are the components which decide the trustworthy relationship of the customer with the store – they help construct dependability and trust and tell the potential buyer the item substantially more obviously and the perspectives that separate it from whatever is left of the items somewhere else [7]. An Ecommerce store which has a decent collection of customer reviews for the items demonstrates the wide reputation among customers. Presently reviews about an item plays important role on decision process for e.g., the client will just purchase the item by reading the reviews composed by the clients .by that he get clear thought regarding the determination and effectiveness of the items particulars and subtleties given by the organization to their items. However in down to earth circumstance half of the highlights that producer tells about the item won't be real. Therefore just legitimate clients who utilize that item can enlighten the precise insights about the product. Here comes the significance of reviews. Presently we see wastage of cash in purchasing bad Items because of the absence of legitimate rating expectation system. The presentation of semantic analysis on reviews tackles the above problem. Users premium is steady just in brief period. so client subjects from surveys can be delegate for e.g. if there should arise an occurrence of a versatile phone, different individuals have distinctive ideas. a few people focus on camera, where some focus on battery reinforcement thus on. They all have customized territory of enthusiasm for the item. The importance of sentiment analysis comes here. Sentiment analysis otherwise known as opinion mining is the process of determining the emotional tone behind a series of words [5]. Sentiment analysis is extremely useful in online e-commerce sites to monitor the reviews it allows us to gain an opinion about the product. Using sentiment analysis on product reviews helps us to extract the emotional tone towards the product. Through natural language processing and machine learning .product reviews in e-commerce sites are written in natural languages such as English. This technique is used to figure out the sentiment or emotion associated with the underlying text. So if you have a piece of text and you want to understand what kind of emotion it conveys, for example, anger, love, hate, positive, negative, and so on you can use the technique sentimental analysis

2.FRAMEWORK

The proposed framework of the research work is conducted in different modules

A. Input Data collection

Data are collected either by Data scraping or by downloading sample of online reviews [2] which is collected from the e-commerce sites. Data scraping is used to get real time data from e-commerce sites.

B. Sentiment analysis

Sentiment analysis is the automated process of understanding an opinion about a given subject from written or spoken language. Sentiment analysis is also called as opinion mining which is an area that includes natural language processing by which it extracts the opinion that is hidden in the text [6]. There are three attributes in extracting an expression

a) **polarity-** what kind of polarity customer express in his review they can be positive negative or neutral

b) **subject-** the thing that is being talked about

c) **Opinion holder-**the customer who express the opinion about a product through reviews

Presently, sentiment analysis is a subject of extraordinary premium and improvement since it has numerous functional applications. Since openly and secretly accessible data over Internet is continually growing, an expansive number of writings communicating feelings are accessible in review sites, discussions, web journals, and social media. With the assistance of supposition examination frameworks, this unstructured data could be consequently changed into organized information of popular sentiments about items, administrations, brands, legislative issues, or any subject that individuals can express conclusions about [2]. This information can be valuable for business applications like showcasing examination, advertising, item surveys, net advertiser scoring, item input, and client administration.

Opinion

The information in the text can be generally classified into two-facts and opinions. Where facts are the objective expressions and opinions are subjective expressions which consist of user sentiments, feelings towards the product.

Like other NLP problems the sentiment analysis also can be categorized into a classification problem where two sub problems must be resolved-

They are:

Subjectivity classification-classifying the sentence into subjective or objective

Polarity classification- classifying the sentence opinion into positive, neutral and negative

In an opinion, the element the content discussions about can be an item, its segments, its aspects, its characteristics, or its highlights. It could likewise be an item, an administration, an individual, an association, an occasion, or a subject. As an example, take a look at the opinion below:

"The battery life of this mobile phone is excessively short."

A negative feeling is communicated about an element (battery life) of a substance (mobile phone).

Direct vs. Comparative Opinions

There are two sorts of opinions: direct and comparative. Direct conclusions give a sentiment about a substance straightforwardly, for instance:

"The sound quality of mobile phone A is poor." This direct opinion states a negative sentiment about mobile phone A.

In comparative feelings, the opinion is communicated by contrasting a substance and another, for instance: "The sound quality of mobile A is better than that of mobile B."

Sentiment Analysis Scope

Sentiment analysis can be applied at different levels of scope:

- **Document level** sentiment analysis obtains the sentiment of a complete document or paragraph.
- **Sentence level** sentiment analysis obtains the sentiment of a single sentence.
- **Sub-sentence level** sentiment analysis obtains the sentiment of sub-expressions within a sentence.

Type of sentiment analysis

There are different types of sentiment analysis where in this system we propose a combination of fine grained sentiment analysis, emotion detection, and aspect based sentiment analysis

Fine-grained Sentiment Analysis

Here instead of looking just general opinions we are further moving very precisely to the opinion mining. Instead of taking positive, neutral and negative opinions can consider the following categories:

- Very positive
- Positive
- Neutral
- Negative
- Very negative

Also can use star representation as for very positive opinion we put 5stars and for very negative option we put 1 star.

Emotion detection

Emotion detection aims at detecting emotions like, happiness, frustration, anger, sadness etc. in the reviews. Just like mining the opinion from the review emotions also has its importance to form precise sentiment about a product.

Aspect-based Sentiment Analysis

In this type of sentiment analysis, not only talking about the sentiment of the review but also points about which particular aspect or feature of the product to which we gives an opinion. For e.g. - "the battery life of the mobile phone is too short". Here the sentence is expressing a negative opinion about the mobile phone, but more precisely, about the battery life, which is a particular feature of the mobile phone.

Working of sentiment analysis

There are many methods and algorithms to implement sentiment analysis systems, which can be classified as:

- **Rule-based** systems that perform sentiment analysis based on a set of manually crafted rules.
- **Automatic** systems that rely on machine learning techniques to learn from data.
- **Hybrid** systems that combine both rule based and automatic approaches.

In the proposed system we use a combination of both rule-based and automatic system which is called Hybrid system.

Rule-based Approaches

Usually, rule-based approaches define a set of rules in some kind of scripting language that identify subjectivity, polarity, or the subject of an opinion.

The rules may use a variety of inputs, such as the following:

From the given set of words our primary aim is to extract relevant information out of it. For this we use a technique called tokenization, where the plain text is converted into tokens or words. Different methods to extract the tokens are – using regular expressions and by using pre-trained model.

E.g. for converting a sentence of words into tokens are

Sentence: "The movie was awesome with nice songs"

Once you extract tokens from it you will get an array of strings as follows:

Tokens: ['The', 'movie', 'was', 'awesome', 'with', 'nice', 'songs']

Next step is stop words removal, all the words present in the plain text are not important some are common grammatical words to maintain the grammar of the sentence. Here our aim is to find the emotion behind the text.in that perspective some of the words like "is, was, were, the, so" etc. are not important. The method to remove such stop words are by storing such sop words in a file or dictionary and compare the extracted tokens with them. If any matching occurs remove such words. For e.g.-

Sentence: "The movie was awesome with nice songs"

After stop words removal: ['movie', 'awesome', 'nice', 'songs']

Stemming

This is the process where the words are reduced into its base form. For e.g.-car, cars, car's, cars' => car (stem or root word)

In our sentiment analysis our main aim is to extract the relevant main or root words only therefore we do stemming.

N-grams

A single word can convey the meaning of the text, sometimes a group of words. For e.g.-word "good" in perspective of online shopping conveys the meaning that 'having the required qualities or has high standard'. But "not good" changes the meaning completely and "not good" is exact opposite of "good". If we only extract single words from text then in the e.g. shown before that is "not good", then 'not'and'good' would be two separate words and the entire sentence predicted as positive by the classifier .This is the case that comes in unigram. However when classifier chooses (bigram) that is taking two words in one token it would take two words "not good" together and the classifier will convey exact sentiment of that text. Therefore for training our models we can use uni-gram or bi-gram or even n gram where n-words per token.

Sentence - The movie was awesome with nice songs

Uni-gram - ['The', 'movie', 'was', 'awesome', 'with', 'nice', 'songs']

Bi-grams - ['The movie', 'was awesome', 'with nice', 'songs']

Tri-grams-['the movie was', 'awesome with nice', 'songs']

Bag of words

Bag of words utilizes a basic methodology whereby we first concentrate the words or tokens from the content and afterward push them in a pack (fanciful set) and the central matter about this is the words are put away taken care of with no specific request. In this way the insignificant nearness of a word clinched is of principle significance and the request of the event of the word in the sentence just as its linguistic setting conveys no esteem. Since the bag of words gives no significance to the request of words you can utilize the TF-IDFs of the considerable number of words taken care of and place them in a vector and later train a classifier (naïve Bayes or any other model) with it. When prepared, the model would now be able to be bolstered with vectors of new information to anticipate on its sentiment. Now we have a bag of words which contain only required information which is filtered. After this NLP techniques implement machine learning algorithms to carry out predictive analytics.

Automatic Approaches

Automatic approaches rely on machine learning techniques. The sentiment analysis problem is actually a classification problem where from a input text we classify the sentiment of the text into positive, negative or neutral.

In the training process (a) using supervised learning the model is fed with the input text and results in corresponding sentiment output (tag) based on the test samples used for training. The feature extractor converts the text input into a feature vector. Pairs of feature vectors and tags (e.g. positive, negative, or neutral) are fed into the machine learning algorithm to generate a model, where in the prediction process (b), the feature extractor is used to convert unseen text inputs into feature vectors. These feature vectors are then fed into the model, which generates predicted tags (again, positive, negative, or neutral).

Feature Extraction from Text

The initial phase in a machine learning classifier is to change the content into a numerical representation, as a rule a vector [8]. Generally, every part of the vector speaks to the recurrence of a word or expression in a predefined dictionary (for example a dictionary of spellbound words). This procedure is known as feature extraction or text vectorization and the traditional methodology has beanbag-of-words or bag-of-ngrams with their frequency.

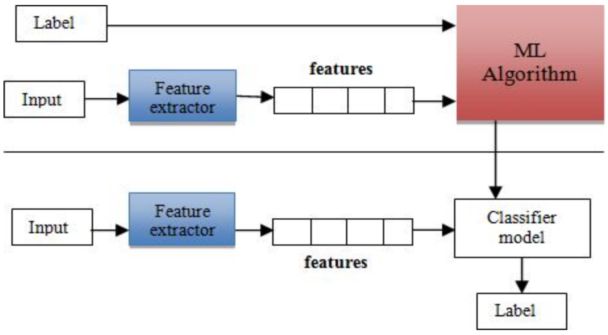

Figure 1: feature extraction process

Classification Algorithms
The classification step usually involves a statistical model like Naïve Bayes, Logistic Regression, Support Vector Machines, or Neural Networks

Sentiment Analysis Metrics and Evaluation
There are many ways in which you can obtain performance metrics for evaluating a classifier and to understand how accurate a sentiment analysis model is. One of the most frequently used is known as cross-validation.
Precision, recall, and accuracy are standard metrics used to evaluate the performance of a classifier.

Web Crawling
Web Crawler likewise named as "spider" or "web robot" is essentially a program that peruses World Wide Web and read its pages and other data in deliberate and robotized way so as to make sections for web indexes like Google, Yahoo records. This process is called Web crawling or spidering.
Fundamentally web crawler begins with a rundown of URL's to visit, and bring them as seeds. As crawler visits these URL's, it finds every one of the hyperlinks and data in that URL. URLs from outskirts are recursively visited one by one and in transit it duplicates and spares all the data from it. This present data's are mainly stored as it can be reviewed, read and archived from the live web. Along these lines it rapidly makes a trip starting with one page then onto the next and soon it gets spread over the web.

3.RELATED WORKS

In the following, we quickly survey some significant attempts to this paper.

Information Analytics has empowered clients to disentangle the covered up patterns in data [1]. Big data gives knowledge on customer behavior which can be utilized to settle on educated choices. A normal shopper is producing both organized and unstructured information which is changing business sector decision making.

Sentiment analysis is a series of methods, techniques, and tools about Detecting and extracting subjective information, such as opinion and attitudes, from language [4]. There have been different approaches for recognizing item includes from unstructured client reviews. In machine learning based approach, product features are assumed to be noun or noun phrases, so they are tagged and candidate product features are extracted by applying some Machine learning algorithms.

The following are the various classification models which are selected for categorization: Naïve Bayesian, Random Forest, Logistic Regression and Support Vector Machine.

Support vector machine (SVM) is the optimal margin classifier based on the Vapnik Chervonenkis dimension of statistical learning hypothesis and the structural risk minimization theory, which was first proposed by cortex Vapnik in 1995.compared with other algorithm, it has better preferences in the sample example, nonlinear and high dimensional pattern recognition problem.as the supervised classification method, support vector machine is generally utilized in word sense disambiguation ,test programmed classification ,data filtering in the field of natural language processing.

This work, tackles the extraction process, by breaking down the surveys dependent on product features. The key module of this framework is the product feature extraction module, which extracts item includes from unstructured reviews. Another algorithm is which separate item includes utilizing the blends of dependencies. Stanford dependency parser is utilized to recognize conditions in a sentence. For discovering supposition of review sentence, Stanford deep analyzer is utilized. A review matrix is built, which is utilized to discover significance and polarity of item feature.

4.CONCLUSION

In this work, we have presented a sentiment based rating prediction and recommendation model which is for predict the rating of products from user reviews. The goal is to give a feature based feeling of a substantial number of customer reviews of an item sold on the web. In this approach, we fuse sentiment similarity, interpersonal sentiment influence, and item reputation similarity into a unified matrix factorization framework to achieve the rating prediction task.

In our Future research, we will investigate complex strategies for opinion and product feature extraction, just as new classification models that can address the arranged names property in rating prediction and also, we can enhance the sentiment lexicons to apply fine-grained sentiment analysis.

REFERENCES

[1] S. Erevelles, N. Fukawa, and L. Swayne, "**Big data consumer analytics and the transformation of marketing**," Journal of Business Research, vol. 69, no. 2, pp. 897–904, 2016.
https://doi.org/10.1016/j.jbusres.2015.07.001

[2].P. Russom et al., "Big data analytics," TDWI best practices report, fourth quarter, pp. 1–35, 2011.

[3]Wang, H.; Lu, Y.; Zhai, C. Latent aspect rating analysis on review text data: A rating regression approach. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 25–28 July 2010; pp. 783–792.
https://doi.org/10.1145/1835804.1835903

[4] J. Narayanan R, Liu B, Choudhary A (2009) Sentiment analysis of conditional sentences. In: Proceedings of the 2009 conference on empirical methods in natural language processing
https://doi.org/10.3115/1699510.1699534

[5] J. Huang, X. Cheng, J. Guo, H. Shen, and K. Yang, “Social recommendation with interpersonal influence,” in Proc. 19th Eur. Conf. Artif. Intell., 2010, pp. 601–606.
[6] T. Kawashima, T. Ogawa, and M. Haseyama, “A rating prediction method for e-commerce application using ordinal regression based on LDA with multi-modal features,” in Proc. IEEE 2nd Global Conf. Consum.Electron., 2013, pp. 260–261.
https://doi.org/10.1109/GCCE.2013.6664818
[7] B. Wang, Y. Min, Y. Huang, X. Li, and F. Wu, “Review rating prediction based on the content and weighting strong social relation of reviewers,” in Proc. Int. Workshop Mining Unstructured Big Data Using Natural Lang. Process., 2013, pp. 23–30.
https://doi.org/10.1145/2513549.2513554
[8] Bafna, Kushal, and DurgaToshniwal. ”Feature based summarization of customers reviews of online products.” Procedia Computer Science 22 (2013): 142-151.
https://doi.org/10.1016/j.procs.2013.09.090