

Sentiment Analysis of Text using Rule Based and Natural language Toolkit

Madhav Singh Solanki

Abstract— Analysis of sentiment or opinion is one of the computational studies of the human kind's written opinions, feelings, attitudes, emotions. It is one of the most active fields of research in the area of natural language processing and text mining in the past few years and it has been popular for its wide range of applications because opinions are central to almost all human activities and are key drivers of our behavior. Part of the reason for the lack of the previous study was that in digital forms there was little opinionated text. So it's no surprise that the field coin starts and grows swiftly. Coincide with the web social media. In fact, research has expanded beyond computer science into management science and social sciences, given its importance to the enterprise and society in general. This paper shows the sentimental analysis process by Natural Language Toolkit and Python Libraries to find the hidden meaning in the unstructured data..

Keywords— Sentiment analysis; machine learning; opinion mining; Python; Text Analytics.

I. INTRODUCTION

Sentimental analysis is commonly used to discover the concealed meaning and unseen expressions in the data format which are positive, negative or neutral. While text analytics is employed to analyze unstructured text data to extract related data and seek to convert the unstructured text data into some useful business intelligence data. Therefore, if the sentimental analysis is applied to the text analytics, the concealed meaning and expression of text data is extracted out in a positive, negative or neutral form and then transmuted properly into a specified structured text data format. We not only get specified data from the unstructured data in sentiment analysis for the operation of text analysis but also here we get the emotions embedded behind the text. By adapting sentimental analysis to text analytics on text database, currently trending topics can be explored and also about their positive and negative public influence on public or organization. For example, if someone runs a restaurant and straight away the word "spoiled" reflected in customer feedback reviews through the application of sentimental analysis of this text analytics data, customer's particular negative emotions are recognized that may also affect sales in that restaurant as well. So we can get an emotion analysis of text analytics data that may help us get a precise brand situation in the marketplace and what people deliberate about the product and assistance provided by respective organizations.

Occasionally, the sentimental analysis of the text analytics of their database product reviews and ratings have been attuned by some e-commerce companies. If there are

some products that receive negative feedback at the same time, that product will be identified through this process and subsequently removed from the portal or sent for modification purposes according to negative feedback. In order to minimize product and website branding, some competitors produce and post fake reviews on the website. The proposed methodology also identifies and manipulates such fake reviews posted on web portals like shopping portals.

By deploying sentiment analysis for text analysis, can we scan competitors for a particular product or service like this if people prefer the competitor website to purchase for a particular product? It is also possible to obtain such analytics from this proposed method. This process produces real-time reports on product selling and purchasing which is very useful to make a cost-related resolution to higher management people as many customers are very cost-sensitive on the market.

II. RELATED WORK

A paper [1] proposed a model and created a template: a binary process to classify the feeling into favorable and unfavorable groups and a three-way process of classifying a feeling that is good, bad and neutral. They used three designs for the test: a unigram model, a model which was centered on the function and a model centered on the tree kernel. They used twitter information in the form by which every tweet is designated as favorable, bad, neutral or junk. They used Dictionary of Language Affects (DAL) and expanded it via Word Net for previous polarity classification. Terms with a polarity below 0.5 were regarded as negative, above 0.8 positive and neutral. For all their studies, vector support machines were used and they obtained fivefold cross validation. Another paper [2], positive tweets were not considered for monitoring information in their studies. Only beneficial and bad tweets were used as a dataset. Another researcher in a paper [3] proposed the technique of extraction of views from an internet site and restriction of the assessment of feelings. The paper [4] combined all the difficulties faced by marketers while employing emotion analysis as another option that triangulates qualitative and quantitative approaches by accumulating real-time data. The Overview of various opinion mining methods was suggested by the paper [5]. , another cited paper [6] disclosed about various solutions for identification of feeling present in the text. The published paper [7] developed a strategy through the combination of rule-based ranking, controlled teaching, and

Revised Manuscript Received on August 05, 2019.

Madhav Singh Solanki, Department of Computer Science and Engineering, Sanskriti University, Uttar Pradesh, India. (E-mail: sanpubip@gmail.com)

computer teaching to improve efficiency. Yet another paper [8] introduced an application of sentimental analyzes and problems.

III. SENTIMENT ANALYSIS

Text communication is one of the day-to-day conversation's most popular forms. In our daily routine, we chat, post, tweet, share status, email, write blogs, share opinions and feedback. All of these activities generate a significant amount of text that is inherently unstructured. It is essential to analyze vast quantities of data in the area of the online marketplace and social media in order to understand the opinion of people. Natural language processing (NLP) allows the computer to interact naturally with the people. It helps the computer understand and derive meaning from the human language. NLP applies to a number of problems: recognition of speech, language translation, classification of documents information extraction. Analysis of film reviews is one of the classic examples of a simple NLP Bag - of - words model on film reviews NLP is not a new field (natural language processing) and neither is machine learning. But the merging of the two fields is quite contemporary and only vows to advance. This is one of those hybrid applications that everybody comes across every day (with a budget smartphone). Take into the account, for example, "keyboard word suggestion," or intelligent auto-completion; these are all the by-products of NLP and Machine Learning amalgamation, and of course, they have become the inseparable parts of our lives.

Especially, analysis of the sentiment or classification of feelings, opinion, sentiments is in the broad category of tasks of text classification where a sentence or a list of sentences is given as input data and the classifier should be able to detect whether that feeling is positive, negative or neutral. Sometimes the third attribute does not retain a binary classification problem. Recent tasks also consider emotions such as "somewhat positive" and "somehow negative"

A. Text classification

One of Text Mining's important tasks. It is an approach that is supervised. Identifying a given text category or class such as a blog, book, web page, news articles, and tweets. It has various applications in today's computer world such as spam detection, task categorization in CRM services, product categorization on e-shop websites, and website content classification for a search engine

B. Feature Generation using Bag of Words

We have a set of texts and their respective labels in the text classification problem. But we cannot use text directly for our model. You need to convert this text into certain numbers or numbers vectors. Bag-of-words (BoW) model is the simplest way to extract characteristics from the text. BoW converts text within a document into the matrix of word occurrence. This model concerns whether words have occurred in the document or not.

C. Sentiment analysis algorithms

For implementing sentiment analysis systems, numerous techniques and algorithms are available that can be

classified as rules-based systems which conduct sentiment analysis based on a set of manually designed rules. Automatic data learning systems based on the techniques of machine education. Hybrid systems that combine rules-based approaches and automatic approaches.

C.1 Rule-based Approaches

Rule-based approaches usually defined a number of rules in a certain type of scripting language, which indicates subjectivity, polarity or an opinion. There are various inputs available in the rules, such as conventional NLP techniques such as stemming, tokenization, voice tagging, and parsing. A basic example of these techniques is two lists of polarized words, e.g. negative words such as bad, worse, ugly, etc., and positive words such as good, best and beautiful.

Given a text:

The number of positive words in the text is counted. In the text, count the number of negative words. When the number of positive word appearances is higher than that of negative word appearances, a positive feeling gives the contrary a negative feeling. Return-neutral, otherwise. This system is very naive because it doesn't look at the combination of words in a sequence. More advanced processing may be possible, but such systems become very complex very quickly. Since it is necessary to add support for new terms and vocabulary, though new rules are hard to maintain. Additionally, the addition of new rules could lead to undesirable results as such systems require substantial investment in the manual tuning and observance of the rules.

D. Types of sentiment analysis

There are many types of sentiment analysis and sentiment analysis instruments ranging from polarity systems (positive, negative, neutral) to emotional detection systems (scary, happy, sad, etc.) or intentions identifying systems (e.g. not interested).

A. Fine-grained Sentiment Analysis

Systems also offer different polar flavors in determining whether a particular feeling, such as wrath, sadness or anxiety, is associated with positive or negative feelings. (i.e. feeling negative)

1. Very positive
2. Positive
3. Neutral
4. Negative
5. very negative

Usually, this known as fine-grained analysis of feelings.

For example, Very Positive=5 stars and Very Negative= 1 star could be mapped in the review to a five-star rating, for example, Some systems also offer different polar flavors in determining whether a particular feeling, such as wrath, sadness or anxiety, is associated with positive or negative feelings. (i.e. feeling negative)

B. Emotion detection

Emotion detection is designed to recognize the emotions like bliss, frustration, wrath, sadness and alike. Many

emotion detection systems use lexicons and complex algorithms for machine learning (i.e. the word lists and the emotions they convey). One of the disadvantages of using lexicon resorting is that the way people express their emotions varies enormously. Some words that usually express anger (e.g. good shit is in your product or his hotness kills me) can also express happiness.

C. Aspect-based Sentiment Analysis

In general, you may be interested in assessing the feeling in subjects, for example in products, not only in whether people speak about the product with a positive, neutral or negative polarity but also in what specific aspects or characteristics of the product people speak of. This is what an aspect-based analysis means.

D. Intent analysis

Intention analysis basically identifies what people want to do with a text rather than what they say. For example "I've been in the hold for 20 minutes. I want to know how to replace a cartridge." "Can you help me to complete this form?" "I want to know how to replace customer service." "The above complaints are not difficult to detect in the first, second and third texts, but machines have problems in identifying them. Sometimes the intended action may be inferred from the text, but at times contextual knowledge is required to be inferred."

E. Multilingual sentiment analysis

Multi-lingual sentiment analysis can be hard and time taking work. Normally a lot of pre-processing is necessary and a number of pre-processing resources are used in multilingual sentiment analysis. Most of these resources are available online (e.g. lexicons of sentiment), but a large number of others must be generated (e.g. translators or noise algorithms). The use of the resources available requires a lot of coding expertise and can take a long time.

D. Sentiment Analysis Applications

1. Social Applications

Sentiment analytics has been widely used in social applications. Some of the applications are currently in place include monitoring V-Lance by detecting tweet violence polarity, prediction of electoral results and public awareness, determination of location satisfaction and recommendation, and monitoring and monitoring the views of students in education.

2. Medical and Health applications

By applying medical data to sentiment analysis can help to identify, predict and track patient satisfaction levels by tweet rates, depression rates, and healthy and unhealthy areas. It can help in the prediction of suicide rate.

3. Industrial applications

The brand monitor, stock market predictions, boxing results according to user tweets and measuring user satisfaction levels are the applications of sentiment analysis in the industrial sector.

IV. PROPOSED WORK

In the previously cited research paper, major is work done in fuzzy clustering, multi-direction cluttering by employing simple human feelings, opinions, and emotions. In this section, the paper discusses the proposed structure on sentiment analysis by using product reviews posted on online portals like amazon. Figure 1 illustrates the process and working of the proposed structure for feeling identification. Firstly the raw data which is the data set which comprises of product negative and positive reviews manually written and posted by multiple users on online web portals. That data set is categorized into two equal portions namely train data and test data. Both the data portions under do preprocess stage where pre-defined logical and analytical operations are carried out on the separated data. Preprocessing step refines the separated data by removing the noises embedded in that data. Preprocessing is done by creating a vocabulary of acronyms and abbreviations with their expansion in the English language.

Frequently used acronyms are monitored and collected in a batch. Such as "lol: laughing out loud". Some product reviews also contain a long arrangement of recurring characters, such as "gooooooooood, loveelly". Words like this are transformed into "good", "lovely".

Stemming and lemmatization process is employed to recognize a common word for words such as "accepted", "considered". These words are deliberated and reduced to "accept", "consider".

During the preprocessing step we eliminate the special characters like "*", "/", as they do not reflect any sentiment, feeling or emotion in text with an exception "!" because "!" reflects an emotion, sentiment.

Next step is feature extraction, which is implemented by unigram with POS tag and natural language toolkit which is widely used for computational linguistic and comprise of a set of libraries and algorithm written in Python programming language. With the application of the unigram model using natural language toolkit, each word in the dataset is labeled with POS tag. Some selected tags are already defined in the algorithm.

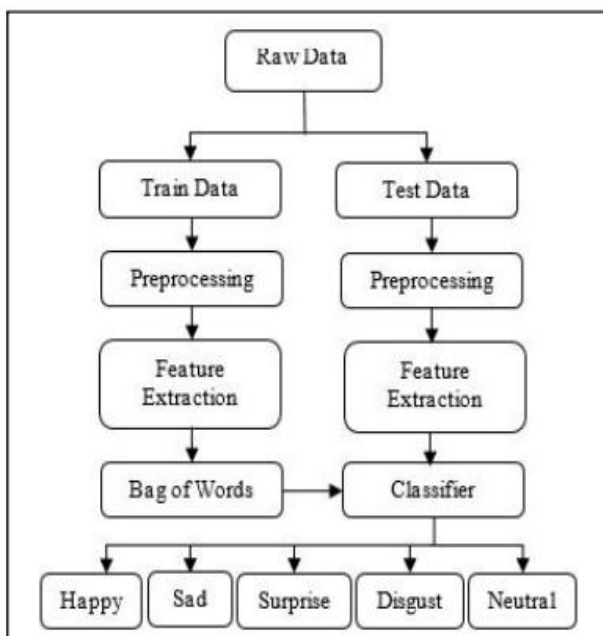


Figure 1 Proposed workflow

The extracted feature is represented by a bag of words module, in natural language toolkit document is transmuted to an unordered array of words which is further used for categorizing texts by scanning their intensity present in the document. The intensity is used as a feature for training the algorithm.

The tagged data is now classified by Multinomial Naïve Bayes, which is a simple model for performing text classification. It scans the intensity data present in the document and for refined data to understand the data Laplace algorithm is also employed.

$$P(x_i|c_j) = \frac{\text{count}(x_i, c_j)}{\sum_{v \in V} \text{count}(x, c_j)}$$

Wherein, v is the vocabulary.

V. RESULT

Sentiment Analysis analyses how users upload their views to the micro-blogging platforms, forums and electronic enterprise on their product, service, event, person or idea. It scans how texts can be studied. The most common use of feeling analysis is to classify a text into a class. Sentiment classification depending on the data set and reason may be binary (positive or negative) or a multi-class (3 or more) problem. Depending on the perspective between researchers and stakeholders, you may also find some similar or totally different views about the link between emotional detection and analysis. But regardless of the outcome or approach, they all adopt the same techniques. Many people use well-known techniques such as graphs, histograms and confusion matrices for visualizing Sentiment Analysis. Visualization approaches such as word cloud, cross-cutting maps, and Sparkline plots are a result of several current data domains and tasks.

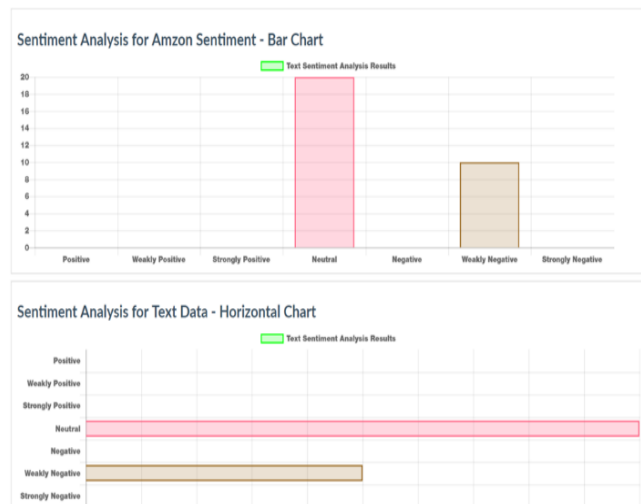


Figure 2. Amazon Reviews

Sr.No.	Topic Name	Total Tweets	Analysis Date	Action
1	First Topic Testing	Tweets	4 April,2019	View Sentiments Delete
2	Happy	Tweets	4 April,2019	View Sentiments Delete
3	Amzon Sentiment	Tweets	4 April,2019	View Sentiments Delete
4	Amazon 2	Tweets	4 April,2019	View Sentiments Delete

Figure 3 Database of sentiment analysis

Comparison with other paper's result is also being carried out. The comparison table is stated below:

Table I shows the comparison of the accuracy of the proposed structure with the framework presented by Agarwal et al. [1].

Table 1 Comparison with structure presented by Agarwal et al [1]

Feature Extraction	Accuracy of our system		Accuracy of Agarwal et al. [1]	
	4-way classification	5-way classification	2-way classification	3-way classification
Unigram	81%	66%	71.35%	56.58%

V. CONCLUSION

Sentiment analysis or opinion mining is a field of computational study that analyzes the mankind feelings, attitudes or emotions for certain entities. This paper addresses the basic issue of sentiments analysis. The data used for this study are chosen from Amazon.com's online product reviews. The new approach is proposed to represent the textual review of a product in the form of visual representation by open-source software. To increase the effectiveness of the analysis, the classification of the text based on natural language processing is employed. To improve analytical speed and reliability, the map reduction environment is used. We have focused on general sentiment analysis in this research. There is potential for work with the partially known context in the field of sentiment analysis. We have noted some examples, that the users of websites generally use keyword types which can be split into a number of different classes, namely politicians, celebrities, products/brands, sports/sportspeople, and media/film/ music. So, we can test and compare the results obtained by using general feelings analysis in tweets belonging to only one of the tweets (i.e. training information wouldn't be general but specific to one of the categories).

REFERENCES

1. Agarwal, Apoorv, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. "Sentiment analysis of twitter data." in Proceedings of the workshop on languages in social media, pp. 30-38, 2011.
2. Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." CS224N Project Report, 2009.
3. Bakhtawar Seerat, Farouque Azam, "Opinion Mining: Issues and Challenges (A Survey)", International Journal of Computer Applications, Vol49 No 9 July 2012Pg No 42-51.
4. Meena Rambocas, Joao Gama, "Marketing Research: The Role of Sentiment Analysis".
5. G. Vinodhini et al, "Sentiment Analysis and Opinion Mining: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering , Vol 2, Issue 6, June 2012.
6. Bluesy Selvam, A. Abirami, "A Survey on Opinion Mining Framework", International Journal of Advanced Research in Computer and Communication Engineering, Vol 2, Issue 9, Sep 2013Pg No 3544-3549.
7. Rudy Prabowo, Mike Thelwell, "Sentiment Analysis: A Combined Approach"
8. Dr. Ritu Sindhu, Ravendra Ratan Singh Jandail, Rakesh Ranjan Kumar, "A Novel Approach for Sentiment Analysis and Opinion Mining", International Journal of Emerging Technology and Advanced Engineering (Vol 4, Issue 4, April 2014).
9. Hussein, Doaa Mohey El-Din Mohamed. "A survey on sentiment analysis challenges." Journal of King Saud University-Engineering Sciences 30, no. 4 (2018): 330-338.
10. Lin, Bin, et al. "Sentiment analysis for software engineering: How far can we go?." 2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE). IEEE, 2018.
11. Ireland, Robert, and Ang Liu. "Application of data analytics for product design: Sentiment analysis of online product reviews." CIRP Journal of Manufacturing Science and Technology 23 (2018): 128-144.
12. Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." Foundations and Trends® in Information Retrieval 2.1–2 (2008): 1-135.
13. Fang, Xing, and Justin Zhan. "Sentiment analysis using product review data." Journal of Big Data 2.1 (2015): 5.
14. Suresh, P., and S. MuthuKumaran. "Sentiment Analysis of Product Reviews Using LDA Method based on Customer Text Content." International Journal of Computer Science Engineering & Technology 5.12 (2015).
15. Anchal Kathuria, Dr. Saurav Upadhyay, "A Novel Review of Various Sentimental Analysis Techniques" International Journal of Computer Science and Mobile Computing IJCSMC, Vol. 6, Issue. 4, April 2017, pg.17 – 22