International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 05 | May 2022 www.irjet.net p-ISSN: 2395-0072

Review on Sentiment Analysis on Customer Reviews

Rohan Shiveshwarkar[1], Om Shende[2], Soudagar Londhe[3], Siddhesh Ramane[4], Prof. Prajakta A Khadkikar[5]

[1,2,3,4] *Student of Pune Institute of Computer Technology, Pune*
[5] *Assistant Professor at Pune Institute of Computer Technology, Pune*

Abstract - *Feedbacks and Reviews carry a lot of weightage when it comes to E-commerce websites/stores. Reviews have time and again proven to be a key to the decision-making process. They are utterly helpful in improving their product and services. These reviews come from the end-users and the greater the product or system, more the volume of the reviews. Practically it is not possible, or rather time and resource consuming to read each and every review and perform an analysis. For this purpose, it is very helpful to introduce NLP and Machine Learning algorithms to easily analyze the negative reviews from customers and display graphical representations for the users to derive actionable business-critical insights.*

Key Words: **Sentiment Analysis, Tokenization, Stemming, Feedback, Reviews, Natural Language Processing, Lemmatization, Machine Learning, Bernoulli's Naive Bayes, Logistic Regression.**

1. INTRODUCTION

Sentiments are nothing but thoughts or judgments prompted by feelings. Sentiment analysis is recognizing these judgments and emotions present in the text and conveying them to the end-user. Sentiment Analysis is used to determine whether the text in focus (here – feedbacks/customer reviews) is positive or negative. We can perform such analysis using Natural Language Processing and Machine Learning. In short, Sentiment Analysis consists of characterizing the reviews received. Tweets are frequently valuable in producing an immense measure of opinion information for determining the core issues faced by the end-users/customers of a product of an e-commerce store. Reviews are generally made of noisy, incomplete, grammatically incorrect, not well-shaped words and/or irregular expressions. Due to this, we have to perform several tasks on the text in order to extract the features and distinguish them based on labels. On the text, we have to perform a series of pre-processing tasks which include removing/replacing unnecessary words which helps in reducing the noise and unwanted data. Once the data is pre-processed and cleaned, vectorizing of data needs to be done. Vectorization transforms data into a numerical form which later on serves as an input to our ML Classification model. The ML model then categorizes the data and provides the output. This output is then used for analysis and business insights. These insights are helpful for recognizing the key problems in features of the business' product/services and improve their said product/services as per their needs. This technique has proven to sustain such e-commerce stores and businesses and customize the services as per the customers' needs and suggestions. Businesses can collect such feedback and suggestions from their own website, forums, blogs, and social media platforms and watch out for their brand reputation. Furthermore, they can roll out polls and survey forms to receive responses regarding their brand.

The benefits of Sentiment Analysis to benefit the company and its products and services are:

- Track client sentiments about their products.
- Analyse the positive/negative impact of products and services on the customers
- Find out about the issues faced by the customers which were not considered before
- Improve the products by working on the suggestions received.
- Track the response from the customers when an upgrade is rolled out in the product
- Find out your target audience and consider all aspects to comfort each type of audience

One of the major sources for acquiring such data is Twitter, among other social media platforms. Twitter has more than 303 million clients using its platform per month and over 500 million tweets are posted on a daily basis. With such a huge volume of tweets and users posting everything about everything - this social media platform has eventually become a hotspot for almost all companies and organizations to produce a strong impact when it comes to the reputation of their brand, products, and services. Since Twitter is a platform where opinions and reviews reach millions, there are many tweets about brands and companies which can be analyzed to benefit the said company. Opinion investigation offers these companies and organizations the upper hand in screening different web-based entertainment locales progressively.

1.1 Dataset

The dataset being used is the sentiment140 dataset. This dataset has 1,600,000 tweets fetched from Twitter-API. The sentiment values are ranged from 0 to 4(where 0 is negative and 4 is positive) which are used to represent sentiment.

This dataset has the following six fields:
sentiment: which shows polarity of tweets (0 = negative, 4 = positive)

ids: The id of the tweet
date: the date of the tweet
flag: The query. If there is no query, then the value is NO_QUERY.
user: the user that tweeted
text: the text of the tweet
Here we only require fields 'sentiment' and 'text', so we will remove other fields.
Again, we're transforming sentiment values (0 to 4) to (0 to 1). (0 = Negative, 1 = Positive)

1.2 Literature Survey

Review Paper on Sentiment Analysis Strategies for Social Media. - This Review paper published by Shaikh Sayema Anwer and V.S. Karwande focuses on sentence compression using NLP without any loss of important data and uses Naive Bayes classifier to categorize the sentiments of data.

Sentiment Analysis of Twitter Data: A Survey of Techniques- Vishal A. Kharde and S.S. Sonawane provided a survey and study over some existing techniques used for opinion mining which includes machine learning and lexicon-based approaches, while finally getting the most accurate models as SVM and Naive Bayes Classifier.

2. ALGORITHMS

2.1 Bernoulli Naïve Bayes

The Bernoulli Naive Bayes is a type of Naive Bayes that is particularly beneficial to use in a binary distribution of data points: especially when the output label is either available or missing to us and helpful to be employed when the dataset is in a parallel dispersion where the result data points are either missing or available. The primary benefit of this particular algorithm is that the features it takes as input are accepted only in the form of binary values. For example:

- 0 or 1
- True or False
- Left or Right
- Present or Absent
- Yes or No

Benefits of using this algorithm for binary classification:

- It is exceptionally quick compared to other classification algorithms.
- Sometimes machine learning algorithms don't function well if the dataset is small and less data formatting, however, this isn't accurate in this case because it gives more precise outcomes compared with other classification algorithms in the case of a small dataset.
- It's quick and can also handle and deal with irrelevant features.

2.2 Linear Support Vector Machine

SVM (support vector machine) is a supervised machine learning model that involves the classification of data into 2 different categories. After training over some input data, for new data, it outputs the category where this new data falls in. The basic functioning of the Support Vector machines can be best understood with the help of a basic model. Consider that we have two tag labels: blue and red. Our information data consists of 2 features: x and y. Our aim is to procure a classifier that, for a pair of (x, y) coordinates, organizes and produces outputs based on the assumption that it is blue or red, below you can see that the plot is previously been labeled training data on a plane.

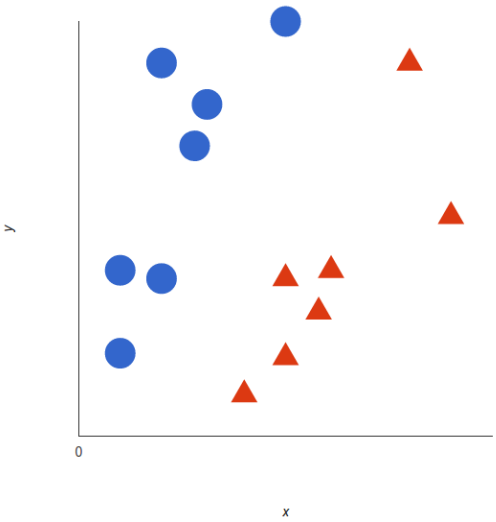


Fig 1. Initial data points plot

A support vector gives an output which is a hyperplane studying all input data points which as we can see represented in 2-dimensional form gives a simple line that separates the two tags of a particular aspect. This line is called a decision boundary. This then classifies the data into different categories based on which side data points are present from the hyperplane.

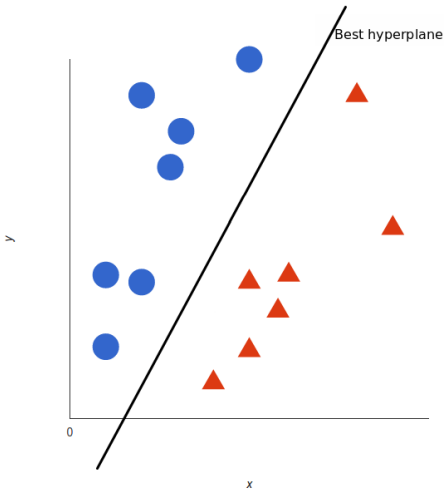


Fig 2. 2D hyperplane representation

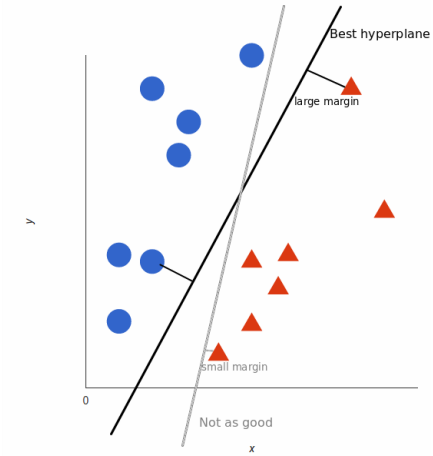


Fig 3. Two different hyperplane representations

2.3 Logistic Regression

Logistic Regression (LR) is a classification technique used in machine learning. To model the dependent variable, it uses a logistic function. The dependent variable used can be divided dichotomous in nature, i.e., there could only be present two possible classes section (e.g.: either the prediction could be cancer is malignant or not). This technique is more often used when we deal with binary data.

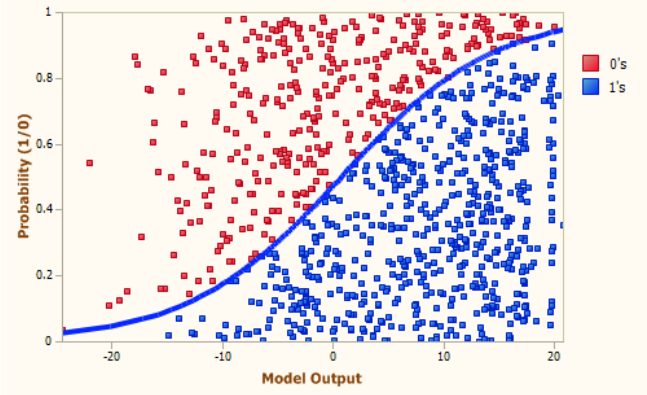


Fig 4 . Data points using Logistic Regression

2.3.1 Types of Logistic Regression

Logistic Regression is generally used for predicting binary data-target variables, it can be further classified into three different types below:

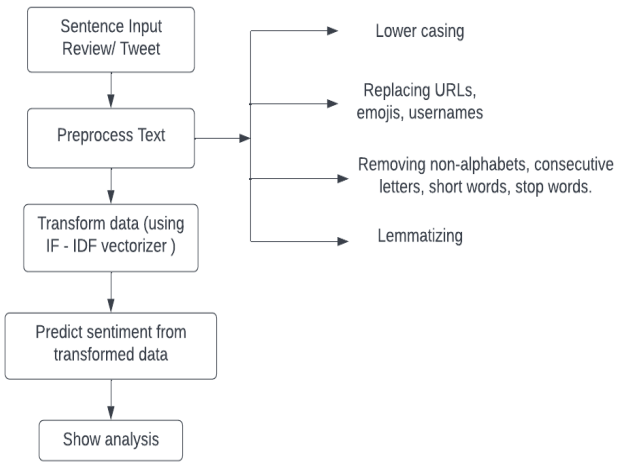
a) Binomial: Where the target data variable has only two possible classification types. e.g.: Predicting a mail as spam or not, Output is Yes or No, 1 or 0.

b) Multinomial: Where the target data variable has at least three possible types, which might not have any quantitative importance. e.g.: Predicting illness.

c)Ordinal: Where the target data variables have an ordered number of categories. e.g.: Movie ratings from 1 to 10.

In logistic regression, the sigmoid function maps and combines the predicted data values to the data probabilistic function. This sigmoid function maps and combines any real significant data value into another value in the range between 0 to 1 only. This function has a positive subordinate at each fixed point and exactly one inflection data point.

3. System Architecture



NLP (Natural Language Processing) is used for text pre-processing which is then forward to Machine Learning algorithms.

3.1 Text Preprocessing

Text pre-processing helps to transform text data into a more digestible form so that machine learning algorithms can perform better. Following are Pre-processing Steps:

1.Lower casing: Each text is converted into lowercase.
2.Replacing URLs: Links starting with 'http' or 'https' or 'www' are replaced by 'URL' keyword.
3.Replacing Emojis: Replace emojis by using a pre-defined dictionary containing emojis with meaning. E.g., ':)' to EMOJI smile.
4. Replacing Usernames: Replacing @Usernames with word USER e.g., @Kaggle to USER.
5. Removing Non- Alphabets: Replacing characters except Digits and Alphabets with a space.
6. Removing Consecutive letters: 3 or more consecutive letters are replaced by 2 letters. E.g., 'Heyyyy' to 'Heyy'.
7. Removing Short Words: Words with lengths less than 2 are removed.
8. Removing Stop words: Stop words are English words which does not add much meaning to a sentence. These can be simply removed from the text without any loss of weight carrying words. E.g., 'the', 'he', 'have'.
9. Stemming is the method where words are reduced to their core stem. E.g., 'Writing' to 'Write'
10. Lemmatization is the method where a word is converted to its root/base form. E.g., 'Better' to 'Good'.

3.2 Transform data

Word Clouds are plotted on positive and negative reviews so as to understand the weightage of particular words in the respective sentiment reviews.TF-IDF is a vectorizer that converts a collection of raw text into a numerical form with TF-IDF features.

4. Conclusion

The proposed system employs Natural Language Processing and different machine learning classification algorithms to derive an analysis on customer feedback and reviews. This method has recently gained popularity and is being employed by many e-commerce companies to uplevel their game. The said system analyzes the feedback and suggestions received and determines whether one is positive or negative. Among the three algorithms - Bernoulli's Naive Bayes, Linear Support Vector Classifier, and Logistic Regression Classifier - the highest accuracy was given by Logistic Regression.

REFERENCES

[1] Che, Wanxiang, Yanyan Zhao, Honglei Guo, Zhong Su, and Ting Liu. "Sentence Compression for Aspect-Based Sentiment Analysis." Audio, Speech, and Language Processing, IEEE/ACM Transactions on 23, no. 12 (2015): 2111-2124

[2] David Zajic1, Bonnie J. Dorr1, Jimmy Lin1, Richard Schwartz. "Multi-Candidate Reduction: Sentence Compression as a Tool for Document Summarization Tasks." University of Maryland College Park, Maryland, USA, 2BBN Technologies 9861 Broken Land Parkway Columbia, MD 21046.

[3] Trevor Cohn and Mirella Lapata. "Sentence Compression Beyond Word Deletion". Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 137–144 Manchester, August 2008.

[4] . Seyed Hamid Ghorashi, Roliana Ibrahim, Shirin Noekhah, and Niloufar Salehi Dastjerdi. "A Frequent Pattern Mining Algorithm for Feature Extraction of Customer Reviews" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 1, July 2012 ISSN (Online): 1694-0814.

[5] Dipanjan Das Andre, F.T. Martins. "A Survey on Automatic Text Summarization" Language Technologies Institute Carnegie Mellon University, November 21, 2007.

[6] Kapil Thadani and Kathleen McKeown. "Sentence Compression with Joint Structural Inference".Department of Computer Science Columbia University New York, NY 10025, USA

[7] LuWang,Hema Raghavan,Vittorio Castelli ."A Sentence Compression Based Framework to Query-Focused Multi-Document Summarization".Cornell University, Ithaca, NY 14853, USA ,T. J. Watson Research Center, Yorktown Heights, NY 10598, USA.

[8] Ghorashi, Seyed Hamid, Roliana Ibrahim, Shirin Noekhah, and Niloufar Salehi Dastjerdi. "A frequent pattern mining algorithm for feature extraction of customer reviews." In IJCSI International Journal of Computer Science Issues. 2012

[9] Jin, Jian, Ping Ji, and Ying Liu. "Translating online customer opinions into engineering characteristics in QFD: A probabilistic language analysis approach." Engineering Applications of Artificial Intelligence 41 (2015): 115-127.

[10] Kim, Soo-Min, and Eduard Hovy. "Automatic identification of pro and con reasons in online reviews." In Proceedings of the COLING/ACL on Main conference poster sessions, pp. 483-490. Association for Computational Linguistics, 2006.

[11] Go, Alec & Bhayani, Richa & Huang, Lei. (2009). Twitter sentiment classification using distant supervision. Processing. 150.

BIOGRAPHIES

Prof Prajakta A Khadkikar
Assistant Professor at Pune Institute of Computer Technology

Rohan Shiveshwarkar
Student at Pune Institute of Computer Technology

Om Shende
Student at Pune Institute of Computer Technology

Soudagar Londhe
Student at Pune Institute of Computer Technology

Siddhesh Ramane
Student at Pune Institute of Computer Technology