IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114131

Sentimental Analysis On Twitter Data For Product Evaluation

S.Rathinakumar, Dr.D.Shanmuga Priyaa,

[1]M.Sc Computer Science, Department of Computer Science, Karpagam Academy of Higher Education

[2]Professor, Department of Computer Science, Karpagam Academy of Higher Education

Abstract: As more people are expressing their views and opinions on various social websites there has been a surge of data generated by the users, these websites have people sharing their thoughts daily because of a short and simple form of expression. We can consider such type of data as a resource and performance sentiment analysis on data of various products and services to make better data-driven decisions. This paper highlights the use of sentiment analysis along with the type of data that is being analyzed, the complex process involved in analyzing the data, the different approaches that can be used, and an experimental observation using the Machine Learning approaches.

Keyterms: Machine Learning, sentiment analysis, Naive Bayes Classifier.

I. INTRODUCTION

Sentiment Analysis with the help of text mining technique for identifying the sentiments of the human form product or service, also sometimes referred to as opinion mining involves creating a system to collect reviews about a product and categorize them as positive, negative, and neutral. It helps market researchers evaluate the performance of the product, and estimate the success of the product. All the categories of reviews for a product/service can be analyzed.

For example, a review for a digital camera can be positive based on picture quality but can be negative depending on how heavy it is. This kind of analyzed information in a systematic way helps the market researchers with a clear picture of public sentiments for the product, similar to surveys as the data is created by customers.

Sentiment analysis is an algorithmic process where a word can be analysis a positive and negative situation. For example the word "long". If a customer said a laptop's battery life was long, that would be a positive opinion. If the customer specifies that the laptop's boot time was long, that would be a negative sentiment. As a result, we need to create a new system to analyze opinions for each type of product/service experience. Also, people can show the contradiction of opinions in their statements.

1.1 Machine Learning

Machine learning (ML) is a category of algorithm that allows software applications to become more accurate in predicting outcomes without being explicit programming. The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available.

The processes involved in machine learning are similar to that of data mining and predictive modeling. Both require searching through data to look for patterns and adjusting program actions accordingly. Many people are familiar with machine learning from shopping on the internet and being served ads related to their purchases. This happens because recommendation engines use machine learning to personalize online ad delivery in almost real-time. Beyond personalized marketing, other common machine learning use cases include fraud detection, spam filtering, and network security. 1.2 Types of Machine Learning

Machine learning is sub-categorized into three types:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Supervised Learning

Supervised Learning is where you can consider the learning is guided by a teacher. We have a dataset that acts as a teacher and its role is to train the model or the machine. Once the model gets trained it can start making a prediction or decision when new data is given to it.

Unsupervised Learning

The model learns through observation and finds the structures in the data. Once the model is given a dataset, it automatically finds the patterns and relationships in the dataset by creating clusters in it. What it cannot do this add

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114131

labels to the cluster, like it cannot say this is a group of apples or mangoes, but it will separate all the apples from mangoes.

Reinforcement Learning

An agent can interact with the environment and find out what is the best outcome. It follows the concept of the hit and trial methods. The agent is gained or suffered from a point for a correct or a wrong answer, and based on the positive reward points gained the model trains itself. And again once trained it gets ready to predict the new data presented to it.

1.3 Types of machine learning algorithms

There are nearly limitless uses of machine learning, there is no shortage of machine learning algorithms. They range from the fairly simple to the highly complex. Here are a few of the most commonly used models:

This class of machine learning algorithms involves identifying correlations generally between two variables and using that correlation to make predictions about future data points.

- Decision trees. These models use observations about certain actions and identify an optimal path to arriving at the desired outcome.
- K-means clustering. This model groups a specified number of data points into a specific number of groups based on like characteristics.
- Neural networks. These deep learning models utilized large amounts of training data to identify correlations between many variables to learn to process incoming data in the future.
- Reinforcement learning. This area of deep learning involves models iterating over many attempts to complete a process. Steps that produce favorable outcomes are rewarded and steps that produce undesired outcomes are penalized until the algorithm learns the optimal process.

1.4 Datamining And Techniques

Data mining is a powerful technology with great potential to help companies focus on the important information in the data they have collected about the behaviors of customers and potential customers. It discovers information within the data that queries and reports can't be expressed effectively reveal.

Classification:

This analysis is used to retrieve important and relevant information about data, and metadata. This data mining method helps to classify data into different classes.

Clustering:

Clustering analysis is a data mining technique to identify data that are like each other. This process helps to understand the differences and similarities between the data.

Regression:

Regression analysis is the data mining method for identifying and analyzing the relationship between variables. It is used to identify the likelihood of a specific variable, given the presence of other variables.

Association Rules:

This data mining technique helps to find the association between two or more Items. It discovers a hidden pattern in the data set.

Outer detection:

This type of data mining technique refers to the observation of data items in the dataset which do not match an expected pattern or expected behaviors.

This technique can be used in a variety of domains, such as intrusion, detection, fraud or fault detection, etc. Outer detection is also called Outlier Analysis or Outlier mining.

Sequential Patterns:

This data mining technique helps to discover or identify similar patterns or trends in transaction data for a certain period.

Prediction:

The prediction has used a combination of the other data mining techniques like trends, sequential patterns, clustering, classification, etc. It analyzes past events or instances in the right sequence for predicting a future event.

In previous work, we used Support Vector Machine techniques for classification. SVMs are used in many applications, among these applications are classifying reviews according to their quality. Have used two multiclass SVM-based approaches: One-versus-All SVM and Single-Machine Multiclass SVM to categorize reviews.

They used a method for evaluating the quality of the information in product reviews considering it a classification problem. They also adopted an information quality (IQ) framework to find an information-oriented feature set. They worked on digital cameras and MP3 reviews. However, there is no fundamental difference between document and sentence level classifications because sentences are just short documents. Classifying text at the document level or the sentence level does not provide the necessary detail needed opinions on all aspects of the entity which is needed in many applications, to obtain these details; we need to go to the aspect level.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114131

To overcome the above drawbacks, this proposed work highlights the usefulness of sentiment analysis along with the type of data that is being analyzed, the complex process involved in analyzing the data, and the different approaches that can be used. Once the data is cleansed it's ready for classification, into positive, negative, and neutral tweets. There are various approaches to sentiment analysis like Machine Learning, Lexicon-based, and Hybridapproachesh. Also, there are some other approaches like Natural Language Processing and Nero Linguistic Programming. Machine Learning involves a training dataset and a testing dataset, where we used the training data and train the classifier using one of many algorithms like Naïve Bayes classification.

II. REVIEW OF LITERATURE

A Naïve Bayesian Classifier for Educational Qualification [1] Manual classification of the individuals into different categories based on their educational qualifications is a tedious task and it may vary respectively to the considered scenario. This paper proposes a classification methodology utilizing the benchmark Naïve Bayesian classification algorithm for the classification of persons into different classes based on several attributes representing their educational qualifications. The experimental results are appreciable indicating that the proposed classification method can be a promising one and can be applied elsewhere.

A Survey On Sentiment Analysis And Opinion Mining [2] With www expanding its reach to anything and everything related to our daily lives, people are becoming more and more vocal to express their views or ideas on online portals, blogs, etc. So there are a million reviews for a product. As a result, it becomes difficult to track the opinions of customers. Sentiment analysis finds the subjective information from the source data by using natural language processing. There are many techniques available to classify the polarity of opinions.

A Survey on Sentiment Analysis of (Product) Reviews [3] With the help of wireless technology, the internet becomes a valuable place for online learning, exchanging ideas, and reviews for a product or service. Reviews on the internet could be millions for a product or service which makes it difficult to track and understand customer opinions. Sentiment analysis is an emerging area of research to extract the subjective information in source materials by applying Natural Language Processing, Computational Linguistics, and text analytics and classify the polarity of the opinion stated.

Analysis of Polarity Information in Medical Text [4] Knowing the polarity of clinical outcomes is important in answering questions posed by clinicians during inpatient treatment. We treat the analysis of this information as a classification problem. Natural language processing and machine learning techniques are applied to detect four possibilities in medical text: no outcome, positive outcome, negative outcome, and neutral outcome.

A supervised learning method is used to perform the classification at the sentence level. Five feature sets are constructed: UNIGRAMS, BIGRAMS, CHANGE PHRASES, NEGATIONS, and CATEGORIES. The performance of different combinations of feature sets is compared.

Determining the Polarity and Source of Opinions Expressed in Political Debates [5] In this work we investigate different approaches we developed to classify opinion and discover opinion sources from text, using effect, opinion, and attitude lexicon. We apply these approaches to the discussion topics contained in a corpus of American Congressional speech data. We propose three approaches to classifying opinion at the speech segment level, firstly using similarity measures to the effect, opinion, and attitude lexicon, secondly dependency analysis, and thirdly SVM machine learning.

III. PROPOSED SYSTEM

This paper highlight the usefulness of sentiment analysis along with the type of data that is being analyzed, the complex process involved in analyzing the data, and the different approaches that can be used. Once the data is clean it's ready for classification, into positive, negative, and neutral tweets. There are various approaches to sentiment analysis like Machine Learning, Lexicon-based, and Hybrid approaches. Also, there are some other approaches like Natural Language Processing and Nero Linguistic Programming. Machine Learning involves training dataset and testing dataset, where we used the training data and train the classifier using one of many algorithms like Naive Bayes classification.

Sentiment analysis is the main aim to generate meaningful information from raw data. After the analysis is complete, we can perform visualization to create bar graphs, Time series, and pie charts. Bar graphs can be used to measure the sentiment of the tweets as positive, negative, and neutral. Time Series can be used to measure the likes, retweets, and average length. The Sentiment Classifier is created by using the movie Review dataset where the reviews are marked positive and negative. The feature extraction is done by positive and negative reviews and the data is trained using a Naive Bayes Classifier.

Naive Bayes Classifier - Naïve Bayes classification algorithm is a basic classification algorithm that assumes that the classification of entities based on their attributes and attributes are independent of each other without any correlation between them. It is "Naive" because the probabilistic calculation for each hypothesis is simplified by calculating the

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114131

value of each attribute independently without considering the conditional dependencies between the various attributes. Let us consider Hypothesis (hypo) may be a class that can be assigned to a data instance (data).

Proposed Methodologies

3.1 Importing the packages - At the start of the machine learning implementation using python, first have to import the packages or libraries which are required for implementing the machine learning models and processing the input data. In this way, Pandas is a package that is familiar with reading and processing the input text file or CSV file, NumPy is a package that contains numerical value handling and array handling. scikit-learn is a package for model training, data splitting, and classification process. These libraries or packages are imported using the terminal in the Anaconda Environment.

3.2 Generate Twitter Account Credentials - If you don't have a Twitter account, you have to create a Twitter account on the Twitter official app or website. After registration and login into the Twitter user page, choose the options for the developer control access, which is used to generate the Twitter API credentials. With the credentials, can access the tweets on Twitter, using python code. For accessing the Twitter API, four essentials are needed, that are mentioned in the following,

1. consumer key
2. consumer secret key
3. access token key
4. access token secret key

3.3 Pre-processing of data - Data pre-processing is the preliminary step that should be required in all data science and machine learning processes.

It processes the raw information into the model readable format. The input content retrieved from real-time is always in improper formats, namely, text may contain symbols, numbers, not consistent and not a complete data, or may contain some flaws. Pandas are used to read the data, after reading the data, have to pre-process or normalize the data, which includes

- Convert the input data to case requirement as uppercase to lowercase or vice versa
- Remove symbols and punctuations
- White space will get removed
- New line will get removed
- Stopwords Removal

This section is the next preprocessing step, which is the removal of the stopwords. The stopwords mean the commonly used English words or supportive words, the name is, was, the, in, on, as, etc. Usually, the words given above do not have any meaning, so these are not required for machine learning models, waste of time processing the stopwords, and for this reason, have to remove the stopwords from the input data. Natural Language Toolkit (NLTK) is used for removing the stopwords in the input data. It contains many libraries for natural language processing.

3.4 Tokenization - Tokenization means separating the input data into the individual words. The input data maybe a paragraph or a sentence; it should be a set of more words. The input of the tokenization is the set of words, and the result is an individual word as separate. The result may be in the form of an array.

3.5 Sentiment Classification using Naïve Bayes Classification - Naïve Bayes algorithm is a supervised learning algorithm, which is based on the Bayes theorem and is used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

It is a probabilistic classifier, which means it predicts based on the probability of an object. Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on conditional probability.

The formula for Bayes' theorem is given as:

$$P(A|B)= \frac{P(B|A)P(A)}{P(B)}$$

Where,

- P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.
- P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.114131

- P(A) is Prior Probability: Probability of hypothesis before observing the evidence.
- P(B) is Marginal Probability: Probability of Evidence.

IV. CONCLUSION

The increase of various social platforms of Twitter where people can use short messages to express their views and opinions helps us to create technologies that can help in analyzing sentiments. This paper used the Naïve Bayes Algorithm to train the Movie Review Dataset, and also uses using the Text Blob package in python to calculate the sentiments of the tweets. Along with the sentiments of tweets ts we are also able to extract various characteristics of the tweets e.g. Likes, Retweets. The most liked tweet and the number of times Retweeted. The classification accuracy can be improved by using better models which can be trained using a larger dataset. The process thus defined is exploratory and we can improve it by using better approaches and algorithms.

REFERENCES

[1]. Niu, Y., Zhu, X., Li, J., Hirst, G. 2005: Analysis of polarity information in medical text. In Proceedings of the American Medical Informatics Association 2005 Annual Symposium

[2]. AlessiaD"Andrea, Fernando Ferri, PatriziaGrifoni, TizianaGuzzo: Approaches, Tools, and Applications for Sentiment Analysis Implementation

[3]. Balahur, A., Kozareva, Z., Montoyo, A. 2009: Determining the polarity and source of opinions expressed in political debates.

[4]. Medhat, W., Hassan, A., Korashy, H. 2014. Sentiment analysis algorithms and applications: A survey, Ain Shams Eng.

[5]. Apoorv Agarwal BoyiXie Ilia Vovsha Owen Rambow Rebecca Passonneau: Sentiment Analysis of Twitter Data

[6]. S. Karthika* and N. Sairam :A Naïve Bayesian Classifier for Educational Qualification.

[7]. Jebaseeli, A. N., &Kirubakaran, E. 2012. A survey on sentiment analysis of (product) reviews. International Journal of Computer Applications, 47(11).

[8]. Kaur, A., & Gupta, V. 2013. A survey on sentiment analysis and opinion mining techniques. Journal of Emerging Technologies inWeb Intelligence, 5(4), 367-371.