WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Sentiment Analysis Using Machine Learning

Akshat Jain[1], Rohit Bankar[2], Ankit Khalane[3],Yogita Ganage[4]

[123]Research Scholar, Department of Information Technology, Rajiv Gandhi Institute of Technology, Mumbai
[4]Professor, Department of Information Technology, Rajiv Gandhi Institute of Technology, Mumbai

A B S T R A C T

Text analytics has become a valuable technique for extracting information on human emotional state as the number of user-generated content on the web has expanded dramatically in recent years. The purpose of this project is to analyse text to determine what the user was attempting to say and whether their thoughts on it were favourable or negative, as well as to detect hate speech and extract problematic events using a machine learning approach. Because not every user types in the same way and different parts of the world attach different meanings with the same terms, the complexity of natural language constructions makes this work extremely difficult. To deal with this complexity, we conduct extensive experiments with several machine learning architectures to learn semantic word embeddings. With their ability to capture both syntactic and semantic aspects of text, machine learning algorithms have emerged as a viable option for fulfilling these goals. It detects the underlying attitude automatically, allowing the machine to interpret the meaning of a sentence in the same way that people do. The goal of this research is to show how Natural Language Processing can be combined with machine learning to create a model that can read text like a human.

Keywords: Opinion Mining, Sentiment Analysis,

1. INTRODUCTION

Text analysis is the process of analysing text material and determining the true meaning of that text using natural language processing. Text analysis can help a company understand its consumers' feelings by analysing what they're saying, how they're saying it, and what they mean. Analysing this information can help them better build their business and improve their goods. It can also be beneficial for review writers or rookie reporters to obtain real-time public opinion on a topic via a popular social media site in order to write a more complete article.

As with many other fields, advances in machine learning have brought text analysis into the foreground of cutting-edge algorithms allowing us to accurately interpret text. With the boom in number of social media platform users across the globe the amount of user opinion available online has increased. There is a opportunity to turn these opinions into useful information for business of almost every sector like news channels, manufacturers, retailers, reviewers, etc. People share their opinion on platforms like Facebook, Twitter, IMDB, Reddit and many more platforms, they can be positive or negative and our goal here is to interpret, and analysis such data of thousands of users to get a shared common opinion. .Humans learn to communicate with one another from the moment they are born. They learn different languages, and as long as they know a given language, they can read and understand any material written in that language, but robots are not the same. As a result, the initial stage in our endeavour was to train a machine to read.

Machine learning is a branch of artificial intelligence that enables computers to learn to accurately predict outcomes without being explicitly taught to do so. In our scenario, we want the machine to figure out what a tweet's meaning is. This can be accomplished by training a machine learning model to understand and process human language.For this reason, we had to build a model capable of understanding human language to some extent.

* *Corresponding author.* Tel.: +0-000-000-0000 ; fax: +0-000-000-0000.
E-mail address: author@institute.xxx

2. LITERATURE REVIEW

Natural language processing is still in its infancy, and there has been little progress until recently. With the availability of new and advanced machine learning and deep learning algorithms, it has become a very popular issue, and all enterprises and businesses are attempting to capitalise on this model. There has been an explosion of user-generated material as a result of the rise of social networking sites. Because of its unique short and straightforward method of speech, microblogging sites have millions of users sharing their opinions every day.They propose to test a paradigm for extracting sentiment from Twitter, a popular real-time microblogging service where users submit real-time comments and views on "anything." They present a hybrid technique to determining the semantic orientation of opinion terms in tweets in this study, which employs both corpus-based and dictionary-based methodologies.

Existing Research Papers

They consider the problem of classifying documents by overall sentiment, e.g., determining whether a review is positive or negative. Using movie reviews as data, they found that standard machine learning techniques definitively outperform human-produced baselines. However, the three machine learning methods they employed (Naive Bayes, maximum entropy classification, and support vector machines) did not perform as well on sentiment classification as on traditional topic-based categorization.

They also pointed out that sarcasm and irony are very difficult to identify [1]

This paper was focused on sentiment analysis of movie review.

They found out that the results produced via machine learning techniques are quite good in comparison to the human generated baselines. In terms of relative performance SVM tends to produce better results than Naïve Bayes, but the results produced were still relatively poor with accuracy ranging from 50% to 69 % [2]

This paper found some very interesting results on product reviews. Some reviewers use terms that have negative connotations, but then write an equivocating final sentence explaining that overall, they were satisfied with the product. This confused the model whether the review was actually positive or not. It is easy for humans to understand this but for a machine to be able to interpret this they would need relatively large dataset. [3]

This paper had a unique way of data collection. They used the 21-million-word 1987 wall street journal corpus with parts of speech tag [4]

The main difference in this paper was the use of different deep learning features like Word2Vec,Doc2Paragragh and Word Embedding to apply to deep learning algorithms like RNN and CNN to get better results. Even when compared to SVMs the results produced were about 10% to 15% better.The only problem was the dataset used here was relatively small. Only 300 training and 150 testing data was used. So the results produced may have been inaccurate. [5]

In this paper a supervised learning model was used to predict on a large amount of product review dataset which was unlabeled. They used Linear SVM, Random Forest, Naïve Bayes, Logistic regression, Decision tree and Stochastic gradient descent.

The results achieved were significantly better with an F1 score of over 90% on SVM. The trick used was the use of combination of Bag of words, TF-IDF (each word is given a score or weight of its positiveness or negative-ness) along with the Machine learning model and to average the obtained result to get a better overall rating [6]

In this paper the distribution of data largely affected the results. Getting equal number of samples for every state in real world was difficult. Which resulted in the inaccurate predictions for states with small size of data. The problem here was mainly with the real-life test data and not in the way the model actually worked. [7]

3. PROPOSED SYSTEM

To achieve our final goal of text analysis, we use a combination of strategies in this model. The procedure's steps are outlined below.

1. **Data Retrieval**: Public data is mined utilising the many APIs for data extraction that are currently available. We have chosen to use the API because of the convenience of data extraction. Data would be selected based on a few chosen keywords relevant to the subject of our concern, for example, product reviews.

2. **Pre-processing**: In this stage, the data is pre-processed to remove identifying information such as the user's name, the message's timestamps, and embedded links and videos. Such information is mainly unimportant and may cause our system to produce erroneous results.

3. **Data Correction**: Because data is prepared for human consumption, it is prone to slang, misspellings, and other extraneous information. As a result, we correct the misspellings in the sentences and look for words from normal English that are closely related to the slang in question to replace the slang in the sentences. This technique is important so that slang phrases can be considered as part of the emotion portrayed, as slang can be used to express a wide range of sentiments, frequently with more emotional effect.

4. **Predicting the output:** We would predict whether the given data is of positive or negative genre

Natural Language Processing

Natural Language Processing, or NLP for short, is the automated manipulation of natural language by software, such as speech and text.

The way we, as humans, communicate is referred to as natural language. Speech and text, to be specific. Text is all around us, whether it's on social media, in emails, SMS, or on websites. This may be simple for people to read and comprehend, but for a machine, it's just 1s and 0s. NLP is required for a machine to comprehend text in the same way that a human does.

NLP is ideal for analysing this type of information. Machine learning and text analysis are used frequently to enhance an application's utility. A brief list of application areas follows:

1. Searching: This locates certain text elements. It might be as simple as looking for a name in a document, or it could entail using synonyms and alternate spelling/misspelling to locate entries that are similar to the original search phrase.
2. NER (Named Entity Recognition): This entails reading text and extracting names of places, people, and things. This is usually used in conjunction with other NLP activities, such as query processing.
3. Information classification: This is an important task that involves taking textual material and organising it into categories that represent the document's content. You've definitely come across a number of websites that organise data based on your requirements and list categories on the left-hand.

Text analysis with machine learning

Machine learning-based systems can anticipate the future based on what they've learned from previous observations. Multiple instances of texts, as well as the predicted predictions (tags) for each, must be supplied into these systems. This is referred to as training data. The better your final predictions are, the more consistent and accurate your training data is.
When you train a machine learning-based classifier, the training data must be converted into vectors, which a machine can interpret (i.e., lists of numbers which encode information). The system can extract relevant features (pieces of information) using vectors, which will aid it in learning from existing data and making predictions about future texts.

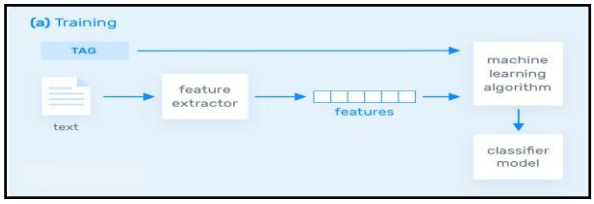

Fig. 1- Training model

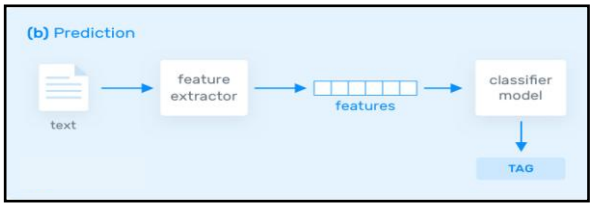

Fig. 2- Prediction model

Logistic Regression:

Logistic regression is a supervised learning technique as well. A categorical dependent variable's output is predicted using logistic regression. As a result, the result must be a discrete or categorical
value. It can be Yes or No, 0 or 1, true or false, and so on, but instead of giving exact values like 0 and 1, it delivers probabilistic values that are somewhere between 0 and 1. Among the models we evaluated on the dataset, this had the highest accuracyused to assess how well the model has learned. When we try to test the model on this testing data, all we give it is plain text to forecast. The plain text labels are then used to cross-check if the model has made its predictions.

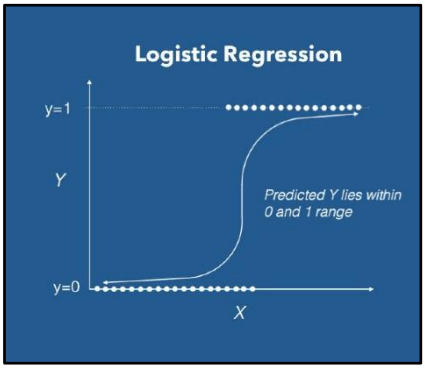

Fig.3- Logistic Regression

4. Method of Validation

After the model has predicted the value for test data, it is critical to validate the prediction and determine the model's accuracy. The training data is frequently divided into two portions for this
purpose. We separated the complete training data into training and testing data in a 70:30 ratio. We utilise 70% of the data to train the model, while the remaining 30% isThe trained model will transform unseen text into a vector, extract its relevant features, and make a prediction:

$$Accuracy = \frac{True\ Positive}{True\ positive + True\ Negative}$$

Fig.4- F1 Score Formula

5. Result

In this section, we'll compare various machine learning techniques. Based on our observations of training and testing data, we discovered that the logistic regression algorithm outperforms all other models in terms of accuracy. The models we tested were Linear regression, XGBoost classifier, Random Forest, Support Vector Classifier). On both training and test data, logistic regression provided a superior accuracy matrix as well as a higher accuracy score. As a result, the main reason we use logistic regression for sentiment analysis classification is because of high accuracy and better result
Dataset used for this project consisted of 75,250 samples. It included datasets from various different sources like amazon review, twitter sentiment, IMDB review, Yelp review. This variety of dataset ensured that the training data included all kinds of things people may talk about, since users on twitter can tweet about anything that one may not even expect. Training data and test data was split on a 0.7 to 0.3 ratio which meant that 52,675 sample data points were used for training while testing sample size was 22,575.

Algorithm	Correctly classified	Incorrectly Classified
Logistic Regression	96.32%	3.68%
XGB	95.55%	4.45%
Decision Tree	93.24%	6.76%
SVC	95.21%	4.79%

Fig.5- Model Accuracy Comparison

Then we calculated F1 score of the same trained models to compare their performance.

Algorithm	F1 Score	True positive	True Negative
Logistic Regression	59.3	7385	140
XGB	57.4	7285	240
Decision Tree	53.7	7137	388
SVC	49.8	7335	190

Fig.6 – Model F1 Score

6.Conclusion

We have discussed the text processing and sentiment analysis using NLP & Machine learning algorithm in detail. Sentiment analysis is a machine learning technique that allows companies to automatically understand text data, such as tweets, emails, support tickets, product reviews, and survey responses automatically. Along with that it helps in knowing the sentiment of the text. As we know data has become a very important part of our life, so our intention of making this project is to make use of machine learning to process this raw text data into useful information so that companies and business can profit from them and help them improve their company model. After searching current issues with their keywords, we got our output accurately according to the situation's sentiments.

References

[1]Kumar, Akshi, and Sebastian, Teja, Mary. (2017). “Sentiment analysis. A perspective on its past present and future.” International Journal of Intelligent Systems and Applications, 4 (10): 1-14.
[2]B. Pang, L. Lee, and S. Vaithyanathan,“Thumbs up?:Sentiment classificationusingmachinelearning techniques,”inProc.ACL02Conf.Empirical Methods Natural Lang. Process., 2016, pp. 79–86.
[3]K. Dave, S. Lawrence, and D. M. Pennock, “Mining the peanut gallery: opinion extraction and semantic classification of product reviews,” in Proc. 12th Int. Conf. World Wide Web, New York: ACM, 2003, pp. 519–528.
[4]V. Hatzivassiloglou and K. R. McKeown, “Predicting the semantic orientation of adjectives,” in Proc. 8th Conf. Eur. Chap. Assoc. Comput.Linguist., Morristown, NJ: Assoc. Comput. Linguist, 2017, pp.174–181.
[5]A. Esuli and F. Sebastiani, “Determining the semantic orientation of terms through gloss classification,” in Proc. 14th ACM Int. Conf. Inf. Knowl.Manage., 2017, pp. 617–624.
[6]R. Bandana, "Sentiment Analysis of Movie Reviews Using Heterogeneous Features," 2018 2nd International Conference on Electronics,MaterialsEngineering&Nano-Technology (IEMENTech), 2018, pp. 1-4, doi: 10.1109/IEMENTECH.2018.8465346.
[7]T. U. Haque, N. N. Saber and F. M. Shah, "Sentiment analysis on large scale Amazon product reviews," 2018 IEEE International Conference on Innovative Research and Development (ICIRD), 2018, pp. 1-6, doi: 10.1109/ICIRD.2018.8376299.
[8]F. Nausheen and S. H. Begum, "Sentiment analysis to predict election results using Python," 2018 2nd International Conference on Inventive Systems and Control (ICISC), 2018, pp. 1259-1262, doi: 10.1109/ICISC.2018.8399007.