International Journal of Research in Engineering, Science and Management
Volume-1, Issue-8, August 2018
www.ijresm.com
ISSN (Online): 2581-5782

Sentiment Analysis

Sanjeev Maurya[1], Yagyanshi Anand[2], Ankita Dubey[3]

[1, 2, 3]*Student, Department of Computer Engineering, MGMCET, Navi Mumbai, India*

***Abstract*: In essence, Sentiment analysis is the process of determining the emotional tone behind the series of words, used to gain an understanding of the attitudes, opinions and emotions expressed within an online mention. With the help of sentiment analysis relevant recommendations in real time are made. Recommendation in real time requires the ability to correlate product, customer, inventory, supplier, and logistics and even social media sentiment data. This topic has the ability to instantly capture any new interest shown in customer's current visit or feedback. The application of this topic is to determine whether a piece of writing is positive, negative or neutral and based on the analysis the user is recommended the accurate result. This is the kind of insight one aims to find through market, online research.**

Keywords: Data Mining, Deep Learning, Decision Trees, Machine Learning, Naive Bayes, Natural Language Programming, Recurrent Neural Networks, Support Vector Machine, Sentiment Analysis.

I. Introduction

With the social media emerging, there is high availability of the information on Internet and the users have become prone to share their feelings on the internet regarding products, movies or any other experiences they want to share. For instance in Twitter or Facebook where the people share how they are feeling today or if it's a new car it is good or not. The ability to process this information has become important because, we can introduce a new product to the market and then wait for the feedback of the people on Internet, extract them in a useful form, and decide the future viability of this new product.

Most of this information on the internet isn't classified or rated in any kind of classification range that can be easily used and is hard to classify at massive scale manually or by using any other normal tool. For this reason, the development of tools that can learn to read texts and extract the feelings is important to the future.

The Natural language programming is a language where the combination of computer science, artificial intelligence and linguistics is used which helps the machines to analyse the sentiments of the people hidden in their language. In NLP there is field of Sentiment analysis that learns how to train the machine to process the text and classify them on the basis of the sentiments so that we can understand the data and use. The field of sentiment analysis uses different language processing algorithms for extracting the features, like words frequency,

Though there are number of ways to understand with the help of human mind but due to limited number of machines, we need to be careful with the type of texts that we want to process because it has a huge impact on the on the size of vocabulary that the algorithm needs to learn and the size of text that needs to be processed. For instance, Facebook and twitter are the sites where people post their feelings and sentiments in a short informal way where a single word could have different meanings. On the other hand if we consider a Movie review site like Imdb the reviews are large texts and have more formal language.

The different machine learning techniques can be used that have different ways to work and learns various methods to implement same data. These differences in the process of learning have impact on the performance of the software which is finally developed .In the document different machine learning methods that can be used in sentiment analysis is shown.

II. Machine Learning Algorithms

There are different machine learning that can be used with a good performance in sentiment analysis problem.

A. Naive Bayes

Naive Bayes is a classification algorithm for binary and multi-class classification problems .This methodology is easy to understand when it is described using binary or categorical input values. In this the probability of each feature contributes to the final probability. The calculation of the probabilities for each hypothesis is simplified to make their calculation traceable.

Predictions can be done by using the Bayes' Theorem.

$$MAP(\mathrm{h}) = \max(\mathrm{P}(\mathrm{d} \mid \mathrm{h}) \times \mathrm{P}(\mathrm{h})) \quad (1)$$

Where

- P (d|h) is the probability of data d given that the hypothesis h was true.
- P (h) is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h.
- P (d) is the probability of the data (regardless of the hypothesis).

B. Random Forest

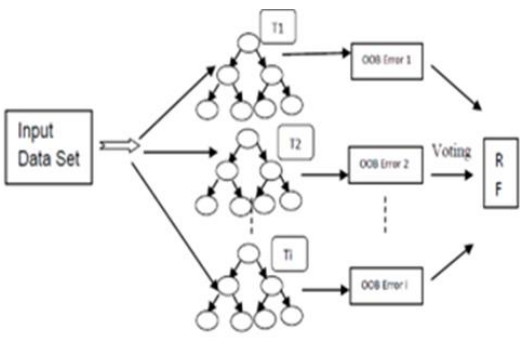

Fig. 1. Random forest algorithm

International Journal of Research in Engineering, Science and Management
Volume-1, Issue-8, August 2018
www.ijresm.com
ISSN (Online): 2581-5782

Random Forest (RF) is an algorithm that trains the decision trees. It is the most used algorithm, because of its simplicity and the fact that it can be used for both regression and classification. In short, RF builds multiple decision trees and merges them together to get a more accurate and stable prediction. The implementation provided by SciKit-Learn can be used for a project on sentiment analysis or the code for Random Forest can be built from scratch.

C. Support Vector Machine (SVM)

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

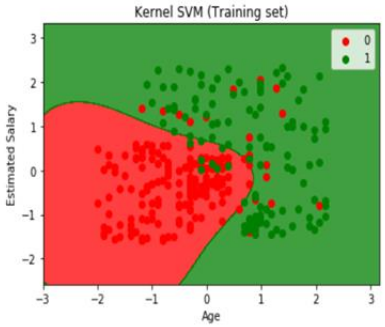

Fig. 2. Kernel SVM testing

D. Neural Networks

An Artificial Neural Network is an information processing paradigm inspired by the way biological nervous systems such as the brain process information. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. ANNs, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons. This is true of ANNs as well.

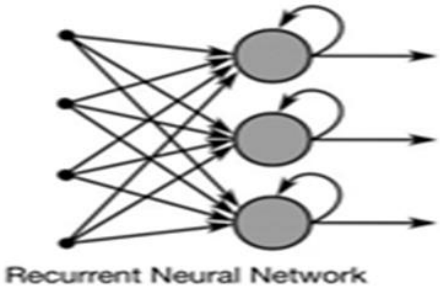

Fig. 3. Recurrent Neural Network (RNN)

The text classification problem can be viewed as a temporal distribution of features, characters or words. One can consider the use of RNN (Recurrent Neural Network) that is a special case of Neural Network that uses an internal memory on each neuron that represents the intermediate understanding between the features that can be accumulated or forgot by the neuron. For the implementation of RNN Tensor Flow and Long- Short Term Memory (LSTM) Cell is used.

III. OBJECTIVE

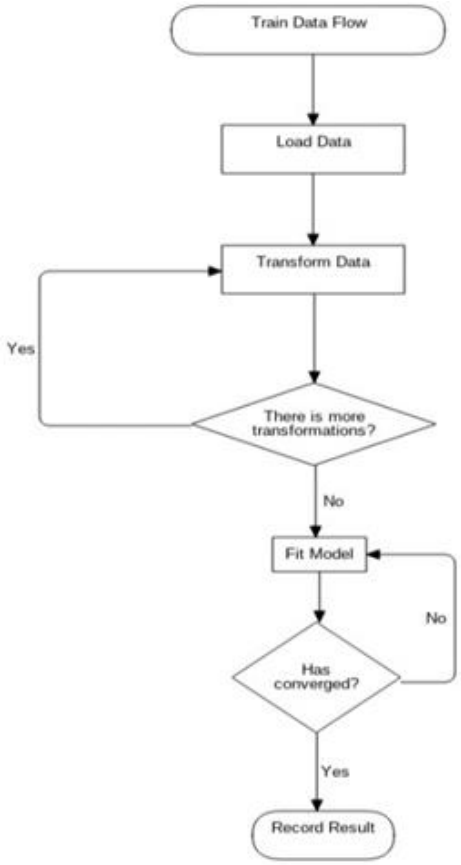

Fig. 4. Training process flow

There is a lot of important data present online that, actually is hard to use. Processing this data can give the power to predict future products, economy trends or social facts. The objective is to learn how to process this type of massive data using sentiment analysis and derive the results based on the analysis. Sentiment Analysis can be divided in four phases:

- *Data Mining:* Collecting data available Online. Data online is present in various form. For example, in review system stars can be used or on Facebook informal language is used for reviews. Thus, data should be mined properly.
- *Choosing an algorithm:* There are various algorithms of machine learning that can be used for sentiment analysis. But using the correct algorithm for the dataset is of major importance.
- *Training:* After the successful choice of algorithm, the

algorithm is applied on the machine and a supervised learning environment is created.

- *Testing:* Once the training of the machine is completed. The testing of machine is done to check whether it gives the correct output.

IV. Data Mining

This is the first step of sentiment analysis. As one should have a dataset prepared to start processing it.

Analyst can start with the type of data they want and create a data warehouse based on those specs. This data warehouse acts like the dictionary of words from where the user's opinions can be sorted and analyzed according to the dictionary designed. Regardless of how businesses and other entities organize their data, they use it to support management's decision-making processes.

Data mining programs analyze relationships and patterns in data based on what user's request. For example, data mining software can be used to create classes of information. To illustrate, imagine a restaurant wants to use data mining to determine when they should offer certain specials. It looks at the information it has collected and creates classes based on when customers visit and what they order.

V. Choosing an Algorithm

The most crucial part, while developing a project on sentiment analysis is to decide which algorithm should be applied so as to get the desired result. There are various machine learning algorithms that are used in the process of sentiment analysis as they are explained above the Naïve Bayes, Random Forest, Support Vector Machine and Neural Network.

Different results are obtained each of the following algorithm is applied on the same data set. Thus, determining the algorithm that best suits the project is important in sentiment analysis. This is usually done with the help of visual analysis of the dataset or by the result that one wants at the end.

VI. Training and Testing

Each machine learning method requires a time to develop and integrate inside the system. This process consists in three steps:

- *Develop the Model:* Integrate it in the system and concatenate the transformations each round.
- *Test and Optimization:* As I say in section 4 each machine learning method and text transformations pair must be tested that works correctly and optimize the hyper-parameters of the model to archive the maximum performance.
- *Collect Results:* After the optimization was done.

VII. Examples of Sentiment Analysis

A. Twitter Sentiment Analysis

A relative sentiment analysis score provides insight into the effectiveness of call center agents and customer support representatives and also serves as a useful measurement to gauge the overall opinion on a company's products or services. When sentiment analysis scores are compared across certain segments, companies can easily identify common pain points, areas for improvement in the delivery of customer support, and overall satisfaction between product lines or services.

By monitoring attitudes and opinions about products, services, or even customer support effectiveness continuously, brands are able to detect subtle shifts in opinions and adapt readily to meet the changing needs of their audience.

Sentiment analysis is used in variety of applications. It can be performed on twitter data to determine the opinion behind the feedback of the customer on certain products.

B. Sentiment Analysis used in Politics

Sentiment analysis has been used by political candidates and administration to monitor the overall opinions about the policy change and campaign announcements.

C. Stock Market Sentiment Analysis

It has been used for stock market analysis by checking the trend of the stock in last few years and recommending the customers to buy the stocks according to the analysis done.

D. Movie Reviews

In this example the movie reviews are calculated only on the basis of the starts/rating given by the users. These ratings thus can be used as the mode of classification. The classification is done in three categories:

1. Positive
2. Negative
3. Neutral

VIII. Conclusion

Sentiment analysis is a branch of machine and deep learning which is used in day to day life for the benefit of business, people, politics etc. Sentiment analysis can be further used for building a recommendation engine where the users are recommended with the most favorable product they are looking for.

References

[1] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Ga¨el Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In ECML PKDD Workshop: Languages for Data Mining and Machine Learning, pp. 108–122, 2013.

[2] P. Raghavan C.D. Manning and H. Schuetze. Introduction to information retrieval, 2008.

[3] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(12), 2009.

[4] Ozan Irsoy and Claire Cardie. Opinion mining with deep recurrent neural networks. In EMNLP, pages 720–728, 2014.

[5] Andrej Karpathy. The unreasonable effectiveness of recurrent neural networks, 2015.

[6] Edward Loper and Steven Bird. Nltk: The natural language toolkit. In Proceedings of the ACL-02 Workshop on Effective Tools and

Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.
[7] Tony Mullen and Nigel Collier. Sentiment analysis using support vector machines with diverse information sources. In EMNLP, volume 4, pages 412–418, 2004.
[8] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In LREc, volume 10, 2010.
[9] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up Sentiment classification using machine learning techniques. In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10, EMNLP '02, pages 79–86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.
[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay.Scikit learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825– 2830, 2011. O` scar Romero Llombart: Using Machine Learning Techniques for Sentiment Analysis
[11] Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani. Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold. 2013.
[12] J¨urgen Schmidhuber. Deep learning in neural networks: An overview. Neural networks, 61:85–117, 2015.
[13] Soroush Vosoughi, Helen Zhou, and Deb Roy. Enhanced twitter sentiment classification using contextual information.