ISSN 2319 - 6629
Volume 8, No.3, April - May 2019
International Journal of Wireless Communications and Networking Technologies
Available Online at http://warse.org/IJWCNT/static/pdf/file/ijwcnt11832019.pdf
https://doi.org/10.30534/ijwcnt/2019/11832019

An Empirical Comparative Study on Methodologies of Sentimental Analysis

Aksa Mariam George[1], Joel Abraham Chekkulathu[2], Neena Joseph[3]
[1] APJ Abdul Kalam Technological University,India, aksamariyam2020@gmail.com
[2] APJ Abdul Kalam Technological University, India, joelabraham098@gmail.com
[3] APJ Abdul Kalam Technological University, India, neena.joseph@mangalam.in

ABSTRACT

This decade has a rapid growth over the field of data extractions, retrievals, and mining. Sentiment Analysis (SA) is the process of computationally deriving meaning from human-generated textual information. Textual classifications and processing approaches vary as per the methodologies found out till date. The limitations of methodology along with the results of how this can be addressed are an essential insight. Understanding and using the correct method is the important aspect behind the improved and accurate results

Key words : Sentiment Analysis, opinion, data mining

1. INTRODUCTION

SA leverages the use of various algorithms to make out the emotion (sentiment) from a subjective text. This is done by breaking down the textual data into components that can be fed into algorithms that weigh various aspects of specific constructs [1]. SA has found its way into many applications ranging from product review analysis to predicting the results of elections. There are many algorithms used for the same. Most of them widely used for these purposes only perform the basic binary analysis. This comes with a cost of reduced accuracy when dealing with sentences or words with multiple implied meanings [2]. This is particularly prevalent in opinions and comments where SA is usually performed.
SA can be performed on various levels. In each level, the actual meaning is finer grained and deals with varied complexities [2].
Surveys have been done covering all aspects of various methods and are referenced where stated**.**

2. USECASES

SA is done on textual data in the form of comments, movie review, product review, customer feedback etc. All these data are generated by people from all around the world and have a great amount of diversity in them. Contexts, figures of speech, slang, colloquial usages, internet slang convey meanings that are not in the conventional sense. Platforms like social media and shopping sites are prime use cases for SA to be conducted.

2.1 Social media

Twitter is the most used social media for sharing information and updates in the form of text. Various other social media platforms analysis of tweets are done in order to understand the general sentiment about a topic. This insight is valuable as it can be used for opinion analysis in matters that affect the general public[3]. A possible application is to analyze comments on various social media about governmental policies and how people respond to it.

2.2 Marketing

Businesses that have an online presence can monitor how their product is reviewed by users by accessing shopping website comments related to that particular topic. This approach is far better than traditional surveys and polls as there are no constraints of choice. Analysis can be performed on very large datasets and at the fraction of the time. A variety of information and insights can be derived from this collected data which can then be used to improve the experience and suggestions provided for the consumer. This has enabled businesses to employ strategies that best suit the audience.

2.3 Movie reviews

Movies and TV shows are all the hype nowadays. People are divided into multiple opinions. The general sentiment is a crucial insight as directors and telemarketers can captivate the audience by analyzing the response generated for the respective genre. This advantage can mean the success of their work compared to its competitors.

3. APPROACHES

There are many algorithms that can be used to perform SA. Based on the type of data, algorithms used can affect the efficiency and results of the analysis. For example, some algorithms rated on the basis of training done on specific data say social media comments may not produce data as accurate if done on movie reviews. This problem arises as different algorithms classify in different ways, for example, a deep learning algorithm might rely on the training it received. The type of data affects the decision the algorithm takes. Whereas a lexicon based algorithm might rely on the database of lexicons that might not be updated. This causes a problem as social media is flooded with internet slang and emoticons that were popularised fairly recently. These are used as modifiers for a lot of expressions and hence convey greater sentiment than conventional words. Machine learning examples are subject to bias while lexicon based ones simply might not contain certain words if not updated properly. The most widely used approaches are discussed below.

3.1 RANDOM FOREST

Random forest algorithm takes the decision tree concept further by producing a large number of decision tree. The scheme behind the algorithm is, first of all, a random sample of the data is taken out and then identifies a key set of features to grow the created decision tree [4]. These trees then will have their out of bag error determined and then the collection of decision trees are compared to find the join set of variables that produce the strongest classification mode. There is no need for feature normalization. Individual decision trees can be trained in parallel. Random forests are widely used. They reduce over fitting. They’re not easily interpretable. They’re not a state-of-the-art algorithm. The usage of random trees leads to over fitting if it's precisely smaller than the required and the only solution is the usage of an optimal number of decision trees.

3.2 THE NAIVE BAYES CLASSIFIER

The Naive Bayesian classifier is based on the Bayes theorem with the concept of independence assumption between predictions. The Bayesian model is easy to build with no iterative or complicated parameters which makes it useful, particularly for big data’s or huge datasets.

Abide its simplicity, the naive Bayes classifier always do surprisingly well and widely used because it’s often a well performed and sophisticated classifier. As Naive Bayes is super fast, it can be used for making predictions in real time [5]. This algorithm can predict the posterior probability of multiple classes of the target variable. Some of the applications of naive Bayes are. The main use of naive Bayes is in the text classification field it is only because of their multiclass problems and independence rule and also they are holding the higher success rate as compared to other algorithms. so it is in spam filtering, sentimental analysis. To predict and filter the unseen pieces of information Naive Bayes classifier is used with collaborative filtering for the accuracy in the recommendation system. The first disadvantage is that the Naive Bayes classifier makes a very strong assumption on the shape of your data distribution, i.e. any two features are independent given the output class. Due to this, the result can be (potentially) very bad - hence, a “naive” classifier. This is not as terrible as people generally think, because the NB classifier can be optimal even if the assumption is violated (see the seminal paper by Domingo’s and Pazzani [5], and its results can be good even in the case of sub-optimality.

3.3 SUPPORT VECTOR MACHINE

An SVM is a discriminative classifier.The algorithm always gives an output in hyperplane which categorizes new examples. In a 2D Space, a hyperplane is a line dividing a plane into two parts wherein each lay in either side.These models are of supervised learning models [6]. Uses are of classification and regression analysis. In this model only a labeled mechanism is possible .if data is unlabelled then supervised learning is not possible. It is one of the most widely used clustering algorithms in industrial application.

SVM works by determining boundaries between elements that belong to a group or those that do not belong to a group [7]. These elements are called vectors. By using this we can match the words accordingly and classify sentiment.

3.4 VADER (Valence Aware Dictionary and Sentiment Reasoner)

It is a rule-based model for sentiment analysis [9] that relies on a database of lexicons (dictionary of words with associated sentiment information). The values associated might be very subjective. To counter this, the creators of VADER sentiment analysis enlisted not just one, but a number of human raters and averaged their ratings for each word. This relies on the concept of the wisdom of the crowdie collective opinion is oftentimes more trustworthy than individual opinion.

VADER is superior to other methods of SA when it comes to analysis of social media comments because of its rich set of lexicons and weight for punctuations, emoticons, and capitalization of words [8]. For example: “The canyon was so deep it was scary to look into.”Would be rated lower than say, “The canyon was SO DEEP!! It was TERRIFYING to look at?!”As the sentence has more positive elements, the value of positivity increases.

These elements are far more likely to be missed by traditional SA tools and algorithms that are not adjusted to take these into consideration. VADER is also made open source so that it can be used for free [9].

3.5 BOOSTED TREE CLASSIFIER

A combined approach over a decision tree and a boosting tree are exactly what a boosting tree classifier is. The structured design of the boosted tree classifier aims at the reduction of the complexity probed over, by the supervised learning. Initially, weighted trees are developed followed by its combination into a single unit of a predictive model [10]. In short, it results in the execution of two algorithms. The main advantage it points out is all about its fast training algorithm without compromising the criteria accuracy. It can handle different types of data types and missing data management. But it holds a major disadvantage of computing conditional class probabilities

4. INFERENCE

Table 1: Comparison of discussed approaches

Methodology	Performance	Time Requirement	Result Accuracy
SVM	Excellent	Good	Best
VADER	Excellent	High	High
Naive Bayes	Better	Low accuracy	Variable
Boosted Tree	Moderate	High	Incremental
Random Forest	Excellent	Recruit learning with every novel dataset	Incremental

5. CONCLUSION

As a result of comparing various approaches, we can see that VADER better understands sentiment from text by assigning weights to different parameters. This might not perform well with new buzzwords etc. if the lexicon is not updated frequently. SVM is only as good as the data it is trained using or how it is optimized. Each model is suitable in their respects as discussed. This is because of how each approach varies. So one should choose a model based on requirement. There is no one best approach towards sentiment analysis. We can compound each method to get a better overall result by analyzing and comparing

REFERENCES

1. Megha Joshi. **A Survey on Sentiment Analysis,** *International Journal of Computer Applications Vala*(0975 – 8887) Volume 163 – No 6, pp. 1-3,april 2017.
   https://doi.org/10.5120/ijca2017913552
2. https://www.lexalytics.com/technology/sentiment-analysis
3. Ms. K. Nirmala Devi, Ms. K. Mouthami, Dr. V. MuraliBhaskaran**. Sentiment Analysis and Classification Based on Textual Reviews**',2012.
4. Ajay Kumar Mishra, Bikram Kesari Ratha **Study of Random Tree and Random Forest Data Mining Algorithms for Microarray Data Analysis** ISSN(Online) Volume -3, Issue -4, 2016
5. Sona ThaheriI, Musa Mammadov. **Learning The Naive Bayes Classifier using Optimisation methods Int. J. Appl. Math. Comput.** Sci., 2013, Vol. 23, No. 4, 787–795 2013.
   https://doi.org/10.2478/amcs-2013-0059
6. www.sciencedirect.com/topics/neuroscience/support-vector-machines
7. www.medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72
8. C.J. Hutto Eric Gilbert. **VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media** Text Georgia Institute of Technology, Atlanta, GA 30032
9. https://www.kdnuggets.com/2018/08/emotion-sentiment-analysis-practitioners-guide-nlp-5.html
10. Amit Gupte. **Comparative Study of Classification Algorithms used in Sentiment Analysis** / *(IJCSIT) International Journal of Computer Science and Information Technologies*, Vol. 5 (5) , 2014, 6261-6264