Section A-Research paper

AN APPROACH ON BIG DATA FOR SENTIMENTAL ANALYSIS APPLICATIONS

Mrs.Smitha Nayak MCA
Department of Computing
Muscat college
Sultanate of Oman
smitha@muscatcollege.edu.om

Dr.JanakiSivakumar
Associate professor,Department of Computing and IT,
Global College of engineering and Technology,Sultanate of Oman
janaki.s@gcet.edu.om

Abstract— Increasingly, buyers turn to the Internet for reviews of goods and services. It is quite difficult for an application to keep track of all the data that is available on the web. They also come in a wide variety of forms as well as a fast-moving nature. As a result, an efficient method for classifying and analysing online reviews in the context of large data is required. Opinion mining, or sentiment analysis, is a term used to describe the process of identifying and evaluating such aggregate online data. Using both information retrieval and computational linguistic tools, sentiment analysis is a particularly hard and promising science that deals with a source's reviews. Sentiment analysis and machine learning algorithms for sentiment categorization are discussed in this paper, as well as issues in opinion mining for massive data.

Keywords— Analysis of Big Data, Sentiment Mining, Text Classification, and Web Mining

I. INTRODUCTION

Social Media, such as blogs, forums, wikis, review sites, social networks, tweets, and so on, allow people to share their knowledge, experiences, and ideas with the world. As a result, in Web 2.0, individuals are communicating and influencing others' social, political, and economic behaviour in new and innovative ways. People will be given a voice on a global scale via the usage of read-write Web and user-generated content as part of the Web 2.0, which promises to improve human cooperation. There are many different ways to express one's feelings regarding an entity or an element of an entity. An opinion is simply a good or negative sentiment, perspective, attitude, emotion, or evaluation from an opinion holder at any given moment.

Product/service, event, person, organisation or subject may all have aspects (features/properties) that reflect both components and attributes of the entity. Companies, legislators, service providers, social psychologists, academics, and other players must assess the growth of user-generated ideas in order to make better decision choices. Though the phrase "sentiment analysis" was coined recently, the field of sentiment/opinion research has existed for some time before to that. Literature in this field covered a wide range of subjects from business to computer and social studies to management because of the importance of sentiment analysis to society as a whole for tasks such as subjective expressions (sentiment of a word), subjective sentences (sentiment of a phrase), and topics.

Sentiment data analysis is a multi-step procedure including five separate stages. Here are the steps:

User-generated content (UGC) on blogs, forums, and social networks is the initial source of data for sentiment analyses. The information is disjointed and conveyed in a variety of ways, including via the use of slang, multiple vocabulary, and the context in which it was written. It is almost hard to do manual analysis. For this reason, text mining techniques such as natural language processing and text analytics are used.

Text preparation: the process of preparing the data for analysis by removing unnecessary information. Excessive non-textual information is removed when sentiment identification is performed on the extracts. Objective communication (facts, information) is deleted while subjective expressions (opinions, beliefs, and perspectives) are maintained;

Sentiment classification: here, subjective statements are categorised into positive or negative; good or poor; like or dislike; or many points; presentation of output: this is the major goal in sentiment analysis, which is to turn unstructured text into useful data. Analytical data are presented graphically in the form of pie, bar, and line

graphs. A sentiment time line may be constructed using the selected value (frequency, percentages, and averages) and presented visually through time.

II. LITERATURE REVIEW

G.Vinodhini, R.M.Chandrasekaran (2012) There have been new openings for data analytics to imagine significant insights from unstructured data because of developments in IoT technology as well as widespread appreciation and adoption of social media tools and apps. OMSA, an opinion mining and sentiment analysis technique, has found use in the age of big data as a valuable tool for classifying public opinion and gauging how it is feeling. Furthermore, OMSA has evolved over the years to include a wide range of procedures that may be used to a variety of data sets and experimental conditions. This work includes a full systematic literature review, which seeks to examine both the technical aspects of OMSA (techniques and kinds) and the non-technical aspects in the form of application fields. Furthermore, this article addressed both technical and non-technical elements of OMSA, highlighting both technical and non-technical issues related to the use of the approach. As a future research path, these challenges are laid forth here.

DikshaSahni, Gaurav Aggarwal (2015) We're in the midst of the "big data" age right now. Users generate enormous amounts of text data via a variety of means, including social media sites, e-commerce sites, and many kinds of scientific investigations. With this "Text data," companies may have a better understanding of how the public perceives their brand and use that information to guide future business choices. As a result, businesses must use emotional Big Data from social media to generate forecasts. It is necessary to use open-source big data tools and machine learning algorithms to handle massive amounts of text data in real time. To this end, we developed a machine learning algorithm-based system for analysing sentiment in large datasets. It is easier, faster, and more scalable to use Nave Bayes and SVM classification algorithms for text analysis on Apache Spark datasets than previous methods. Accuracy was also used as a measure of the algorithms' performance. Experimental findings show that the Algorithms are quite successful at managing large sentiment datasets, as shown by the results. It will be more beneficial for businesses, governments, and people to increase their worth.

Charu C. Aggarwal (2015) The article provides an overview of the various methods and tools for sentiment analysis. The study begins with an introduction and then goes on to classify I methods based on features/methodologies and benefits/limitations, and (ii) tools based on the various sentiment analysis techniques employed. The study also discusses sentiment analysis's several application domains, including business, politics, public affairs, and finance.

Rizvaan Irfan, Christine K. King (2004) As the internet's reach expands, an increasing number of individuals are taking to blogs, tweets, forums, social networking sites, and customer review sites to air their views. A favourable, negative, or neutral attitude is connected with each of these opinions. But the issue is that there is just too much information to process.. It is possible to measure feelings that are captured in digital form using supervised machine learning or lexical-based techniques. In addition to studying the effects of current events on social media, sentiment analysis has also been used to the study of consumer attitudes about various goods and services. Sentiment analysis applications and the issues they face are discussed in this chapter. Sentiment analysis researchers who read this chapter will have a comprehensive understanding of how to conduct their work in this area.

III. METHODOLOGY

It is common practice to use descriptive examples as input for supervised learning since the intended output is already known. Predictive analytics is mostly employed in applications where previous data is used to forecast future data.

Classifiers for Supervised Learning: Decision Trees

A simple and widely used classification process is the decision tree classifier. Output from classifying problems is the primary goal of this classifier. Each time a response is received, the Decision Tree Classifier generates a new set of connected questions. This process continues until a conclusion is reached and recorded. Inverse document frequency and the relevance of the term identified were used to extract review characteristics from IMDb. Principal Component Analysis (PCA) as well as CART were used to choose attributes depending on how important the work was in terms of a text as a whole. According to the LVQ method, 75 percent of the time, the correct categorization was made. It is suggested to use data mining tools to investigate emotional disparities in teenage age and the underlying reason for these variances. Using a decision tree and classifying emotions, a variety of emotional states may be documented. Decision trees may also generate if-then rules.

Classifiers that are linear

Support vector machines and neural networks are utilised in linear classifiers.

Hyper planes are created in an N-dimensional space by a support vector machine and then utilised for a number of tasks, including classification and regression. There are many artificial neurons that make up a neural network, which is a computational model used in machine learning.

A Classifier Based on Rules

A collection of rules is used to represent all of the data in a rule-based classifier. DNF (Disjunctive Normal Form) is used to represent a feature set, while a condition on the feature set is used to represent a class.

Classifier Based on Probability

In the probabilistic classifier, Nave Bayes, Bayesian network and maximum entropy are all included. This is the simplest and most often used classifier, Nave Bayes. It determines a class's posterior probability based on the word distribution in a record. The Naive Bayes classifier's primary goal is attribute independence. Predicting that all of the qualities are completely interdependent is one of the assumptions used in this process. The conditional exponential classifier, commonly known as the Maximum Entropy Classifier, uses encoding to convert labelled attribute sets into vectors. This encoded vector may be used to calculate weights for each attribute.

IV. RESULTS

Figures demonstrate the results of both text analysis machine learning methods on the amazon cell labeled dataset. Training models are tested by comparing their findings to those of a test dataset after they have been run on the training data. As a result, the Nave Bayes model generates the probabilities and predictions from the data. The LinearSVM model, on the other hand, just generates predictions and can be readily compared to the labels' base values.

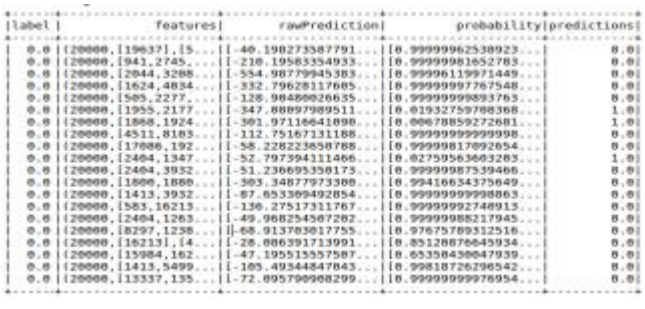

Figure1: Sentiment Analysis Results by using Naïve Bayes

V. MODEL EVALUATIONS

Depending on the kind of data being analysed, the evaluation may either be a Regression Evaluator for regression issues, or a Multiclass Classification Evaluator for problems involving many different types of classifications of data. Measuring a project's success or failure is usually dependent on the machine learning job itself. Linear regression, classification, clustering, and recommendation all employ different metrics. There's a Multiclass Classification Evaluator in here, and an Evaluator is used to assess a model's prediction ability or efficacy. Evaluate accepts a Data Frame as an input and returns a scalar value as output. Accuracy is a straightforward measure for evaluating models. The proportion of labels successfully predicted by a model is used as an assessment statistic for various algorithms. If a model successfully predicts the labels for 85 out of 100 observations in a test dataset, for example, its accuracy is 85%.

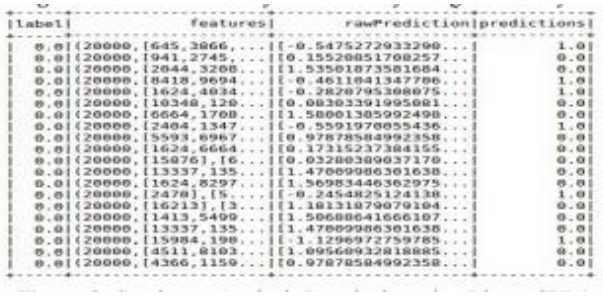

Figure2: Sentiment Analysis Results by using Linear SVM

To see how the model performs on both the training and test datasets, retrieve the predictions for each observation in both datasets. Second, assess the models based on the test datasets' correctness. K-fold Cross Validation with k=5 Iterations was also performed to increase the accuracy assessment findings. This dataset's average accuracy assessment metrics show that linear SVM outperforms a Nave Bayes model when applied to the test data. The findings are shown in figure.

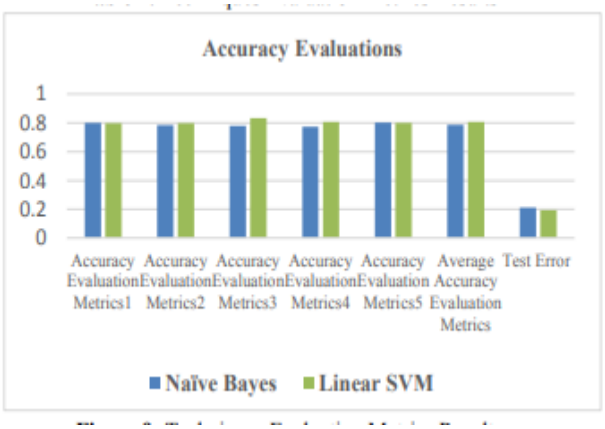

Figure3: Techniques Evaluation Metrics Results

Because the accuracy score will be between 0 and 1, and high Accuracy score is better, it means that has less error. All the above work is implemented by using supervised machine leaning algorithms and big data machine learning libraries.

VI. CONCLUSION

More promising approaches exist for sentiment analysis, which is a technically demanding task that will only grow in importance as the number of internet users who make purchases and voice their thoughts grows. Product makers and consumers alike may benefit greatly from summarizing customer feedback, which has a broad range of uses. As more and more people use the internet to communicate, massive amounts of data are being created on a daily basis. As a result, effective sentiment analysis requires a distributed parallel computing environment.

Reference:

[1] G.Vinodhini, R.M.Chandrasekaran, "Sentiment Analysis And Opinion Mining: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering, June 2012

[2] Jiawei Han, Micheline Kamber and Jian Pei, "Data mining Concepts and Techniques", Third Edition, Morgan Kaufmann Series in Data management Systems

[3] DikshaSahni, Gaurav Aggarwal, "Recognizing Emotions and Sentiments in Text: A Survey ", International Journal ELSEVIER, 2013 of Advanced Research in Computer Science and Software Engineering , 2015

[4] Charu C. Aggarwal, "Data Mining: The Textbook", Springer, 2015

[5] Rizvaan Irfan, Christine K. King ,"A Survey on Text Mining in Social Networks", The Knowledge Engineering Review, 2004

[6] A.Jeyapriya, C.S.KanimozhiSelvi,"Extracting Aspects And Mining Opinions In Product Reviews Using Supervised Learning Algorithm", IEEE, 2015

[7] JeevanandamJotheeswaran, Dr. Y. S. Kumaraswamy, "Opinion Mining Using Decision Tree Based Feature Selection Through Manhattan Hierarchical Cluster Measure", Journal of Theoretical and Applied Information Technology, 2013

[8] BlessySelvam1 , S.Abirami2, "A Survey On Opinion Mining Framework", International Journal of Advanced Research in Computer and Communication Engineering, 2013

[9] Kai Gao, Hua Xu, JiushuoWanga, "A Rule-Based Approach To Emotion Cause Detection For Chinese Micro-Blogs", ELSEVIER, 2015

[10] ChetashriBhadane,HardiDalal, HeenalDoshi, "Sentiment Analysis: Measuring Opinions", Science Direct, 2015

[11] Weiyuan Li, Hua Xu, "Text-based emotion classification using emotion cause extraction",

[12] Richa Sharma, Shweta Nigam, Rekha Jain, "Polarity Detection at Sentence Level", International Journal of Computer Applications, Volume 86- No 11, 2014

[13] Vikrant Yadav. 2016. thecerealkiller at SemEval-2016 Task 4: Deep learning based system for classifying sentiment of tweets on two point scale. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), San Diego, US.

[14] V. Sahayak, V. Shete, and A. Pathan, "Sentiment Analysis on Twitter Data", IJIRAE, 2015.

[15] Yunxiao Zhou, Zhihua Zhang, and Man Lan. 2016. ECNU at SemEval-2016 Task 4: An empirical investigation of traditional NLP features and word embedding features for sentence-level and topic-level sentiment analysis in Twitter. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), San Diego, US.