Sentiment Analysis using Machine Learning

PROJECT SYNOPSIS
OF
MAJOR PROJECT

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE & ENGINEERING

SUBMITTED BY:

Group 55
Vishu Agarwal (2000290100192)
Ashmit Tayal (2000290100033)
Rajani Dixit (2000290100110)

TABLE OF CONTENTS

DECLARATION

We hereby declare that this submission is our work and that, to the best of us by another person or material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Name:

Vishu Agarwal (2000290100192)
Ashmit Tayal (2000290100033)
Rajani Dixit (2000290100110)

Date:

CERTIFICATE

This is to certify that the Project Report entitled “Sentiment Analysis using Machine Learning” which is submitted by Ashmit Tayal and Vishu Agarwal in partial fulfilment of the requirement for the award of degree B. Tech. in the Department of Computer Science & Engineering of Dr. A.P.J. Abdul Kalam Technical University, Lucknow is a record of the candidates own work carried out by them under my supervision. The matter embodied in this report is original and has not been submitted for the award of any other degree

Date:

Mentor:
Ms. Bharti Chugh
(Group- 55)

ACKNOWLEDGEMENT

It gives us great pleasure to present the synopsis of the B.Tech Major Project undertaken during B.Tech, Third Year. We owe a special debt of gratitude to Dr. Vineet Sharma, Head of the Department of Computer Science & Engineering, KIET Group of Institutions, Delhi-NCR, Ghaziabad, for his constant support and guidance throughout our work. His sincerity, thoroughness, and perseverance have been a constant source of inspiration for us. It is only his/her cognizant efforts that our endeavours have seen the light of day.

We also do not like to miss the opportunity to acknowledge the contribution of all the faculty members of the department for their kind assistance and cooperation during the development of our project.
Last but not the least, we acknowledge our friends for their contribution to the completion of the project.

Ashmit Tayal (2000290100033)
Vishu Agarwal (2000290100192)
Rajani Dixit (2000290100110)

ABSTRACT

Data Analysis has become an important part of our life, as it helps us to draw useful information, decisionmaking conclusions on any particular raw data. One component of digital data analysis is sentiment analysis. Sentiment analysis evaluates the emotional tone of textual data and classifies it as neutral, negative, or positive. In business, this analysis is frequently utilized for market social media monitoring, market research, consumer feedback and many more. In order to automate the procedure, machine learning models that recognize patterns in labeled data are frequently used. The objective is to derive meaningful insights from textual data so that decisions can be made with knowledge, effectively. In order to provide important insights into public opinion and behavior, this study uses machine learning techniques to investigate the sentiments conveyed in Twitter data. A wide range of Twitter posts, including ones about trending hashtags, events, and subjects, are gathered as part of the process.

Sentiment analysis is done using machine learning algorithms. Sentiment analysis and examples of human language are used to train machine learning models. Following this training, machine learning is used by sentiment analysis software to evaluate and rank human language according to its previous training.

The current study proposes a technique that best analyzes the sentiments of the text used in social media is gradient boosting classifier. With this method, an additive model can be built in a step-by-step, forward fashion and any differentiable loss function can be optimized. There are n-class regression trees that are fitted on the negative gradient of the loss function, which could be the multiclass

log loss or the binary loss. This process is repeated in each step. In the particular case of binary classification, just one regression tree is induced.

The computational results demonstrates that gradient boosting approach with an accuracy of 93.62% which outperforms is preferable to other classifiers.

INTRODUCTION

An intriguing area of machine learning, sentiment analysis extracts and understands textual emotions. Sentiment analysis analyzes human language using machine learning. Sentiment analysis uses contextual meaning to estimate a brand's social sentiment and if its product will sell in the market. The dataset is used to train a variety of ML models, including Naïve bayes, K-nearest neighbour classifier, Support vector machines, etc to find patterns and correlations between words and sentiments. The performance level is analysed with the metrics in mind like accuracy, F1 score, precision. The study tackles issues including managing sarcasm, slang, and context-specific terms in an effort to better grasp public emotion on Twitter. Market research, brand management, political analysis, and other fields can all benefit from the sentiment analysis's results. They offer businesses and scholars insightful information gleaned from the enormous volume of viewpoints expressed on the Twitter network.

Potential uses of this research include real-time public sentiment monitoring, which enables proactive response and decision-making tactics across a range of industries. Agarwal et al. (2011) examines a well-known microblog called Twitter and develops models to categorize "tweets" into sentiment categories that are good, negative, and neutral. Sentiment Analysis is needed to store data efficiently and cheaply. You can solve all real-time scenarios with sentiment analysis. Businesses use sentiment analysis to analyze customer opinions, build brand reputation, create better goods, and personalize content. It supports market research, policymaking, and automation. Text sentiment is measured using metrics to determine positive, negative, or neutral. Today, the internet offers several means to convey feelings.

Machines are trained with text examples of emotions to identify sentiment without human input. To put it briefly, computers can learn new tasks using machine learning without the need for explicit programming. Sentiment analysis models may learn context, misapplied words and sarcasm beyond definitions. As many as methods and complicated system of equation commands and train machines to carry out sentiment analysis. However, combined they can produce great effects.

Natural language and sentiment are employed in machine learning model training. Sentiment analysis software uses machine learning to score human language after this training. Among the algorithms that are utilized the most are Linear Regression, Support vector machines (SVM), Naïve Bayes, and many more. Every new model proceeds in the direction of least prediction error within the space of possible predictions for every training example. Wang et al. (2022) states that opinion mining and sentiment analysis are two exciting new areas of research that are used to find out what people think and feel about particular subjects. As a result, opinion mining and sentiment analysis are frequently used synonymously to convey the same concept. Chalathom et al. (2015) illustrates that the content of tweets has emerged as a hot research topic concerning the positive or negative polarity of sentiment. Our work with sentiment analysis of tweet contexts demonstrates that group learning, which is made up of the majority vote of the Stacking, support vector machine, SentiStrength, and naïve bayes, can enhance and be more successful in achieving accuracy performance. Xia et al. (2011) tries to illustrate why popular text classification techniques like maximum entropy, support vector machines and naïve bayes are necessary. Saif et al. (2012) states the characteristic short length and asymmetrical structure of microblogs like Twitter present various additional obstacles for sentiment analysis.

The research literature on sentiment analysis on microblogs identifies two primary study avenues. Gautam et al. (2014) encompasses the examination of the content on the Internet spanning numerous domains that are seeing a rapid increase in both quantity and number as websites are devoted to particular product categories and have expertise gathering customer reviews from other websites, including Amazon, twitter, etc. Naalakandan et al. (2020) states that the data sentiment analysis is a categorization method based on finding if a review is good, negative, or neutral based on the given opinion. Finding the review writer's position at the document, sentence, or aspect levels is SA's main goal.

METHODOLOGY

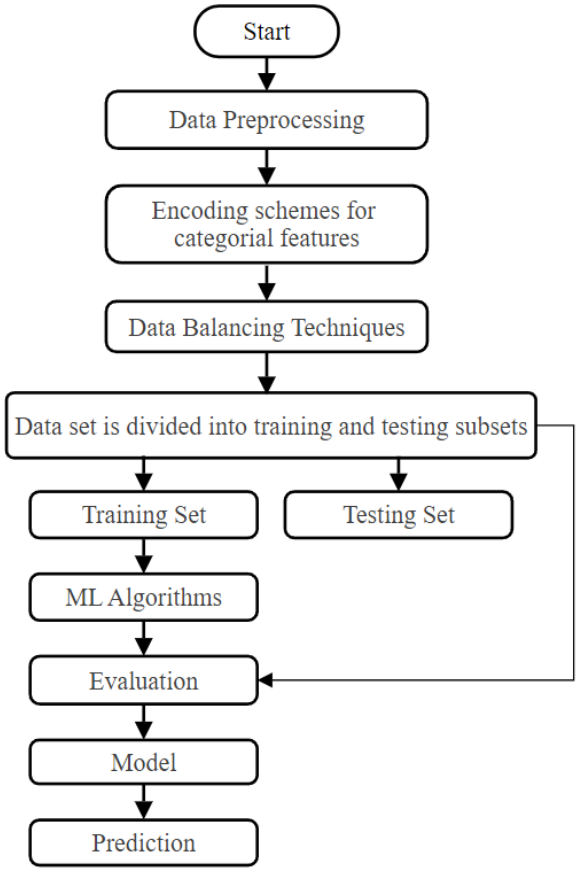

Figure: Flow Chart of the proposed model

Step I: Twitter dataset has been pre-processed and then encoded.
Step II: Generate models for an unbalanced dataset.
Step III: Select the top models for the evaluation of the results.
Step VI: Perform the Re-sampling techniques to balance the dataset.
Step V: Optimization of the model by applying hyperparameters.
Step VI: Prediction based on test data.

CONCLUSION

Improving the accuracy of the model that has been suggested is the primary goal of the research that has been proposed. Consequently, in order to accomplish this goal, the presence of a balanced dataset is one of the primary prerequisites. This is because an imbalanced dataset can result in bias towards classes that have a greater number of samples. Every classifier has its unique set of benefits, which are determined by the kind of data that is used for training. It is seen that the results of the Gradient boosting classifier were superior to those of the other classifiers. The purpose of this endeavour is to employ a variety of optimization strategies aimed at selecting the most advantageous features, in conjunction with resampling techniques, in order to ensure that classifiers produce satisfactory outcomes. The first step is to select the five most accurate classifiers and nominate them for further consideration. Each model is assessed according to its level of accuracy. In conclusion, the Gradient boosting classifier beats the other classifiers in terms of accuracy, surpassing prior efforts that were performed with the same dataset by a margin of 93.62%. The work that will be done in the future can involve other datasets to work on.

REFERENCES

1. Agarwal, Apoorv, et al. "Sentiment analysis of twitter data." Proceedings of the workshop on language in social media (LSM 2011). 2011.
2. Wang, Yili, et al. "Sentiment analysis of Twitter data." Applied Sciences 12.22 (2022): 11775.
3. Chalothom, Tawunrat, and Jeremy Ellman. "Simple approaches of sentiment analysis via ensemble learning." information science and applications. Springer Berlin Heidelberg, 2015.
4. Xia, Rui, Chengqing Zong, and Shoushan Li. "Ensemble of feature sets and classification algorithms for sentiment classification." Information sciences 181.6 (2011): 1138-1152.
5. Saif, Hassan, Yulan He, and Harith Alani. "Semantic sentiment analysis of twitter." The Semantic Web–ISWC 2012: 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I 11. Springer Berlin Heidelberg, 2012.
6. Gautam, Geetika, and Divakar Yadav. "Sentiment analysis of twitter data using machine learning approaches and semantic analysis." 2014 Seventh international conference on contemporary computing (IC3). IEEE, 2014.
7. Neelakandan, S., and D. Paulraj. "A gradient boosted decision treebased sentiment classification of twitter data." International Journal of Wavelets, Multiresolution and Information Processing 18.04 (2020): 2050027.