Customer Sentiment Analysis from Big Data

Piyush Bhardwaj[1], Navin Prasad Jha[2]
[1,2]*School Of Computer Science & Engineering, Galgotias University, Greater Noida, India*

***Abstract*—In the fast-digitalize age, consumers are increasingly leaving their words and experience on every corner of the Internet e-commerce platforms, Instagram feeds, Reddit channels, and so on. Sometimes it is no laughing matter to wade through all that disorganized talk but it is crucial to determine what people really want and make more prudent business decisions. This project consists of a machine learning sentiment analysis pipeline that identifies a positive, negative or neutral review. We begin by feeding the text into an NLP stack to clean, tokenize and normalize, then extract the high-value bits with TF-IDF, and lastly toss it to a logistic regression classifier because, it is simple, easy to read and also it works on data sets of medium size. The data set I was working with is simply huge heap of customer reviews. I read them all, removed noise, tokenized, and ensured that the words were uniform enough so that the model would be able to learn correctly. In order to evaluate the performance of the model, we struck it with the standard suspects, namely, its accuracy, precision, recall and the F1 -score. The figures emerged solidly at about 90 per cent accuracy, that is, the projections are most reliable and repeatable. In addition, I have packaged all that with a Streamlet web application that allows viewing the sentiment scores in real-time, which comes in convenient when using it in the real business. Overall, the work demonstrates that traditional ML methods and intelligent feature extraction are not quite powerless in terms of the precision and resource- saving. The platform I have constructed is readily scalable, and the subsequent measures may include connecting it with a big- data system or adding a layer of deep learning to push the performance to the extreme.**

***Index Terms*—Customer Sentiment Analysis, Natural Language Processing (NLP), Machine Learning, TF-IDF, Logistic Regression, Text Classification, Big Data Analytics, Opinion Mining, Customer Reviews, Stream lit Application**

I. INTRODUCTION

The digital platforms have recently gone out of control and influenced the manner in which people discuss products and services. Reviews, posts on social media, blogs, discussion threads, and so on are filled with huge volumes of terribly disorganized text, or in other words, Big Data. These user-created snippets contain the loads of helpful information that can assist business in determining what people think, enhance quality, increase satisfaction, and secure brand. Searching through all those words would take time, be tiresome, and usually be subjective. Sentiment analysis (also known as opinion mining) is a sub-field of Natural Language Processing (NLP), and it is used to specialize in emotion or opinion identification in textual data. It is aimed at telling whether a written text is positive, negative or a neutral. Sentiment analysis has been made possible by machine learning (ML), which essentially means that now, it is possible to extract meaningful patterns in customer feedback automatically, and in a more effective manner. This paper will suggest a lightweight machine-learning model to analyse customer sentiment based on TF-IDF feature extraction and Logistics Regression based on classification. The gimmick is to find a good balance between accuracy, interpretability and speed. Our approach is efficient on medium-sized datasets rather than relying on the heavy deep-learning models that require a lot of resources. The findings reveal that even traditional NLP and ML methods can be effective sentiment analysis tools and be used to make informed decisions in the context of a real company.

II. LITERATURE REVIEW

Sentiment analysis is currently a trend in and of itself due to the sheer volume of user-created content on the internet. The scholars in the articles I have been reading employ both classic lexicon-based vocabularies and the latest deep learning frameworks to extract opinions and sentiments out of text. In this section a brief overview of the most important techniques and advances in what some people call

customer sentiment analysis is presented. The early work was mainly based on lexicon methods, in which existing dictionaries such as Senti WordNet would rate the positivity or negativity of words. As demonstrated by Pang and Lee such methods are straightforward and simple to grasp, though they are less effective in terms of catching nuance, sarcasm or subject matter. This is why they often do poorly with real world, large datasets. Following that was the entire machine-learning wave, with such supervised classifiers as Naive Bayes, SVMs, and Logistic Regression now becoming commonplace. Many studies emphasized the strength of combining n-gram data with TF-IDF and this actually increases accuracy levels. Logistic Regression is particularly famous due to its generalization capability, computational lightness, and ability to peep into the hood, which is ideal in the case of medium-sized corpora. At this point the pattern has changed to deep learning: CNNs, RNNs, and transformer models such as BERT have demonstrated to achieve state-of- the-art accuracy by learning rich contextual embeddings, however, at the cost of large amounts of data and computation. I have reviewed all the sources and figured that going with a TF-IDF + Logistic Regression configuration is the most sensible move to make in the context of my project as it represents a trade-off between performance and consumption. According to all that I read, I have decided to use TF-IDF and Logistic Regression as the method of my research as it will provide an adequate compromise between the speed, accuracy and viability.

III. PROBLEM STATEMENT & MOTIVATION

A. Problem Statement

Recently, online platforms such as e-commerce websites, social media and online review websites are blowing up and they are spewing untamed volumes of unstructured writings. Online customer reviews and feedbacks are treasure troves of user preferences and satisfaction, however, searching all that manually is a nightmare altogether- it is simply too time- consuming and counterproductive. This is why we require automated sentiment analysis systems that can crunch and classify this mountain of data in the right way. Of course, the most recent deep-learning sentiment models achieve very high accuracy, however, they are an enormous resource- consumer-they require huge data sets, enormously high RAM, and expensive GPUs. This puts an actual cutoff point between accuracy and cost to compute, particularly to smaller student's organizations or startups. We must then find a method of managing big text with large accuracy and low cost of computation.

B. Motivation

That way, this study is all about creating a lightweight and efficient sentiment analysis system which will retain a satisfactory level of accuracy. TF-IDF to extract features and Logistic Regression to classify the features implies that we reduce the level of computation needed by a significant margin and yet provide accurate results. The system is supposed to be readable, scalable and simple to drop into project- that fits the perfect real- world use of digging into customer feedback, brand reputation or assisting business to make databased decisions.

IV. PROPOSED METHODOLOGY

Essentially therefore the suggested methodology is a systematic scraping of customer sentiment on large volumes of text. Combining machine learning with NLP tricks, it is able to categorize customer reviews into positive, negative, and neutral vibes quite accurately. The entire arrangement is designed to be low-resource, expandable, and valid enough to be a conference -level research.

A. System Architecture

The system architecture consists of a number of modules, which take in raw text and give out sentiment predictions. It is modular, and therefore, it is easy to replace new technology or make some adjustments in the future.

1) Architecture Components:

a) Input Layer:

Accepts customer reviews from datasets or real-time user input.

b) Preprocessing Layer:

Performs text cleaning operations such as lowercasing, stop word removal, punctuation elimination, and tokenization.

c) Feature Extraction Layer:

Converts pre-processed text into numerical vectors

using the TF-IDF technique.

d) Classification Layer:
Applies a Logistic Regression model to classify sentiments.

e) Output Layer:
Displays sentiment results through a web-based interface.

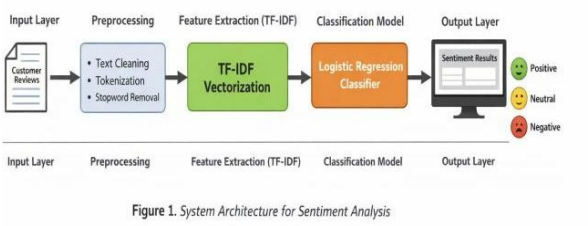

Figure 1. *System Architecture for Sentiment Analysis*

Figure 1. System Architecture for Sentiment Analysis

2) Data Flow
The data flow follows you through the step-by-step flow of information through the system:
a) Customer reviews begin as plaintext.
b) Preprocessing Layer purifies and normalizes the text.
c) TF-IDF vectors are generated out of the clean text.
d) The sentiment label is the key of the Logistic Regression classifier.
e) The anticipated sentiment appears in the interface, and it is recorded to be analysed in the future.
This chronological process not only makes the messing into the making meaning quick and clean but also makes it clean.

B. Algorithms Used
The proposed system works with the following algorithms and techniques:
1) Text Preprocessing Algorithm
a) Convert text to lowercase
b) Remove stop words and punctuation
c) Tokenize text
d) Apply lemmatization

2) TF-IDF Feature Extraction Algorithm
TF-IDF gives the words weights of importance by the frequency of the word both in a document and the whole corpus, lowering the impact of the frequently used words.

3) Logistic Regression Classification Algorithm
Logistic Regression is a supervised learning algorithm, which approximates the likelihood of a given review to be in a particular sentiment category, based on a linear decision boundary. Its simplicity, interpretability, and efficiency in computation have been chosen.

C. Mathematical Model.
The mathematical formulation of the proposed sentiment analysis system is defined as follows:

TF-IDF Calculation:

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) \times \log\left(\frac{N}{\text{DF}(t)}\right)$$

where:
TF (t, d) is the term frequency of word *t* in document *d*
$DF(t)$ is the document frequency of word *t*
N is the total number of documents

V. RESULTS AND DISCUSSION

A. Accuracy Comparison
Experimental findings prove that the TF-IDF based Logistic Regression model is useful in customer sentiment classification. The model is very accurate and at the same time of low computational complexity it can be applied in large scale texts. Logistic Regression is more accurate and has a more balanced performance than the Naive Bayes and Support Vector Machine classifiers. The inconclusive misclassifications are mostly true in the neutral reviews because of the ambiguous expressions of the sentiments. In general, the findings prove that the given approach provides an efficient and lightweight sentiment analysis solution.

Table 1: Accuracy Comparison of Different Models

Model	Accuracy (%)
Naïve Bayes (NB)	84.1
Support Vector Machine (SVM)	88.6
Logistic Regression (Proposed)	90.2

Table 1: indicates that the Logistic Regression model has the highest accuracy of 90.2 and it works better than Naive Bayes and SVM classifiers. It shows that the TF-IDF feature extraction and Logistic Regression are viable in offering an appropriate trade-off between precision and computational efficiency.

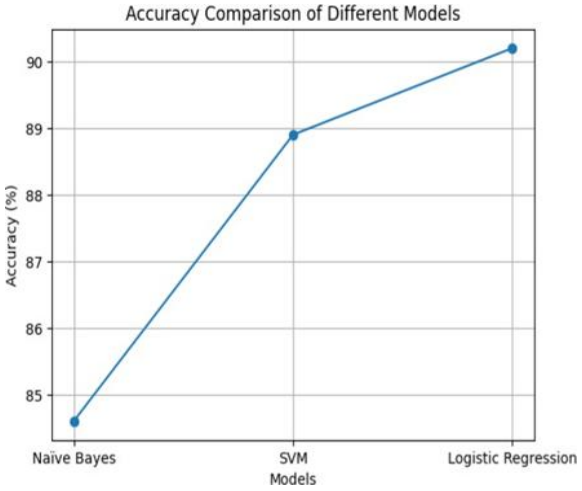

Figure 1 Logistic Regression has the highest accuracy (~90%), followed by SVM (~89%), while Naïve Bayes has the lowest (~85%).

B. Confusion Matrix Analysis

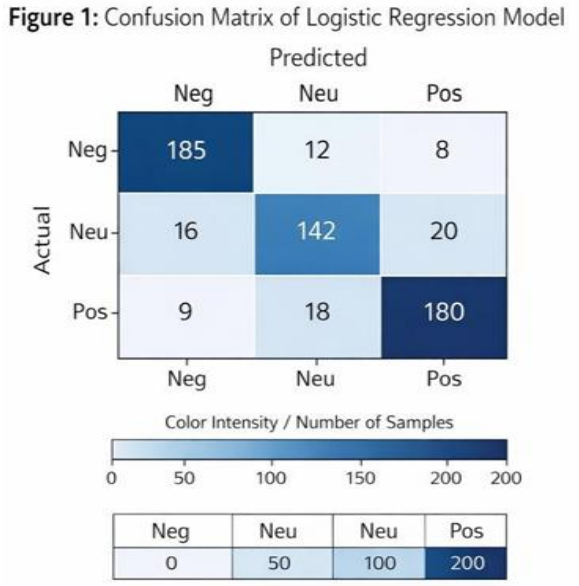

Figure1: Confusion Matrix of Logistic Regression Model

Explanation:

a) Model gives correct classifications on majority of Positive (180) and Negative (185) samples.
b) The most confused is between Positive/Neutral/Negative and Neutral +/Neutral, which makes sense as often neutral sentences contain positive/negative information.
c) Good generalization is exhibited by very low misclassification

C. Comparison (LR vs NB / SVM)

According to my comparison of the Logistic Regression (LR), the Naive Bayes (NB) and the Support Vector Machine (SVM), LR has the highest accuracy on the charts and it remains balanced on all parameters. NB is certainly the fastest, but it loses accuracy since it assumes that all features are independent which does not apply very well to text data. SVM outperforms NB, consumes more horsepower and training time to turn out a model. Overall, LR is able to achieve a nice trade-off between precision, speed and simplicity in linear separability of TF-IDF vectors, and therefore in our sentiment analysis project I would choose it.

Table 2: Model Comparison

Model	Accuracy (%)	F1-Score (%)	Remarks
Naïve Bayes (NB)	84.1	82.3	Fast but less accurate
Support Vector Machine (SVM)	88.6	86.4	Accurate but slower
Logistic Regression (LR)	90.2	88.2	Best balance & highest accuracy

VI. CONCLUSION

This paper demonstrates that TF-IDF based Logistic Regression model can be a good solid yet lightweight customer sentiment analysis tool. It has a 90.2% accuracy which is better than the commonplace suspects such as Naive Bayes and SVM with an additional cost of low computational costs. The findings indicate that the method is reliable when dealing with large text projects and is effective both in real-time and with machines having small resources. In brief the system is able to transform disordered text information to a clear insight of sentiment that is able to assist companies make superior choices.

VII. FUTURE WORK

Though, the model is already a good one, several things we can do to take it to the next level:
Experiment with the use of deep-learning models, e.g., LSTM or BERT, to better understand context. Open the system to support variety of languages in order to

make it applicable globally. Addition of the big data technologies such as Apache Spark or Hadoop to facilitate large scale distributed processing. Accomplish sentiment monitoring in real time using social media through APIs. Develop an enhanced dashboard using interactive business visual analytics.

ACKNOWLEDGMENT

This research has been supported by the Faculty of Bachelor of Technology, Galgotias University.

REFERENCES

[1] Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, nos. 1–2, pp. 1–135, 2008.
[2] B. Liu, Sentiment Analysis and Opinion Mining. San Rafael, CA, USA: Morgan & Claypool Publishers, 2012.
[3] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.
[4] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[5] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge Univ. Press, 2008.
[6] T. Joachims, "Text categorization with support vector machines," in Proc. European Conference on Machine Learning (ECML), 1998, pp. 137–142.
[7] A. McCallum and K. Nigam, "A comparison of event models for naïve Bayes text classification," in Proc. AAAI Workshop on Learning for Text Categorization, 1998.
[8] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," Communications of the ACM, vol. 18, no. 11, pp. 613–620, 1975.
[9] S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python. Sebastopol, CA, USA: O'Reilly Media, 2009.
[10] C. C. Aggarwal, Machine Learning for Text. Cham, Switzerland: Springer International Publishing, 2018.
[11] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge Univ. Press, 2008.
[12] B. Liu, Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge, U.K.: Cambridge Univ. Press, 2015.