Sentiment Analysis and Natural Language Processing for Marketing using Python

Mohamed Affan M[1], Mohammed Saqui T[2] , Dr. Khader Babu SK[3]
Department of Mathematics, School of Advance Sciences
Vellore Institute of Technology, Vellore –632014, Tamil Nadu, India.
E-mail: mohamedaffan.m2022@vitstudent.ac.in, mohammedsaqui.T2022@vitstudent.ac.in, debaroti.das@vit.ac.in

***Abstract*— Sentiment analysis and Natural Language Processing (NLP) are becoming increasingly important in the field of marketing and customer feedback analysis. This paper presents a case study of a newly launched startup that is introducing its first video game to the market. The startup wants to analyze the sentiments of its potential customers regarding its video game. This paper discusses the various techniques used for sentiment analysis and NLP using Python. In order to understand what makes a video game appealing to gamers and worth purchasing, it is necessary to gain a deeper understanding of the language used in their feedback and reviews. This involves analyzing the linguistic features of their statements, such as the words and phrases they use to describe the game and their overall sentiment towards it. In order to carry out this task, we will employ different NLP methods because it is very time efficient. The sentiment analysis techniques used to analyze reviews with the dictionary-based sentiment analysis tools, which are part of NLTK, a natural language toolkit, used in Python. Test our algorithm to see if it worked well and performing the Data evaluation with scikit-learn in Python. A visual representation of preferred and non-preferred words related to video games. The results of the study show that sentiment analysis and NLP can provide valuable insights into customer feedback. These insights can be used to improve the marketing and overall success of the newly launched video game.**

***Keywords*— Sentiment analysis, Natural Language Processing, Natural Language Toolkit, Support Vector Machine, Opinion lexicon, Data visualization etc.**

I. INTRODUCTION

Nowadays, Due to the high quality, quick logistics system, and substantial discounts offered by online retailers, online shopping is now more popular than ever. It also makes shopping wonderfully comfortable. Because of this, user reviews and comments are crucial sources of data used by businesses to enhance their goods and determine what consumers think of them. One of the more common approaches is to have researchers conduct surveys to gather opinions and reviews. The client is now adopting social media, which is an unorganised form to get opinion and reviews, due to advancements in technology and the computer world. Social media opinions can be grouped in order to distinguish between all different kinds of favourable, unfavourable, and neutral reviews of the content that was submitted. The Sentiment Analysis NLP for Marketing project is a machine learning-based approach for analyzing customer feedback. This project aims to help businesses understand customer sentiment towards their products or services by automatically classifying feedback as positive, negative, or neutral. This analysis can be used to improve product offerings, customer service, and brand reputation.The project is built using Natural Language Processing (NLP) techniques, which enable the algorithm to understand the context of a customer's statement, and accurately classify their sentiment. The project's output is a sentiment analysis report, which includes various metrics such as accuracy, precision, re-call, and F1 Score. The report also includes visualizations such as bar graphs to help businesses understand the most common positive and negative sentiments expressed by their customers. Overall, the Sentiment Analysis NLP for marketing project is a valuable tool for businesses looking to gain insights into customer sentiment towards their products or services. Businesses can save time and money by automating the sentiment analysis process while getting insightful data that can guide their decision-making.

II. LITERATURE REVIEW

In a study paper [1], the association between the words was not taken into account while using the Bow (Bag of words) method for sentiment analysis. The sentiment of each word in the sentences was independently determined in order to get the sentiment for the entire sentence. Values were then compiled using a grouping technique. You can utilise the features-driven opinion summarising technique. Each product has a specific feature with its associated attributes, and each product class has a general feature. Next, polarity is assigned to each feature with the aid of Sequential Minimal Optimisation and Support Vector Machines. The goal is to determine the most substantial, according to a study article [3] that summarised the prior work. Learn how to get reliable results that are relevant to the techniques being utilised and the difficulties in sentiment analysis.Another study by Liu (2012) explores the use of sentiment analysis for brand reputation management. The study discusses the challenges associated with analyzing sentiment towards brands and proposes various solutions, including the use of NLP techniques and machine learning algorithms. The Sentiment Analysis NLP for Marketing project can be seen as an implementation of some of the proposed solutions, as it uses NLP techniques to preprocess the data and machine learning algorithms to classify customer feedback[8].In a similar study, Jansen et al. (2009) discuss the importance of sentiment analysis in social media monitoring and highlight the need for automated sentiment analysis tools. The Sentiment Analysis NLP for Marketing project provides an example of such a tool, as it aims to automate the sentiment analysis process to provide valuable insights for decision-making[9].Overall, the Sentiment Analysis NLP for Marketing project is a valuable example of how machine learning-based approaches and NLP techniques can be used to automate sentiment analysis and gain insights into customer sentiment towards products or services. These insights can be used to improve product offerings, customer

service, and brand reputation, making this project a useful tool for businesses looking to gain a competitive edge in their respective markets.

III. METHODOLOGY USED

The e-commerce site Amazon offers a lot of reviews, as it is one of the popular shopping site. It is recommended that labeled data be used for supervised learning models. Amazon data, however, is unlabeled, and it must be converted to labeled data before performing the analysis. This diagram illustrates the complete workflow for sentiment analysis. The Amazon data we obtained is in JSON format as follows: "reviewTime": "10 17, 2015", "reviewerID": "A1HP7NMA4N", "asin": "07026657", "reviewerName": "Ambra075", "reviewText": "This game is a bit hard to get the hang of, but it's great.", "summary": "but it's great.", "unixReviewTime": 14040000

A. Data Collection

Amazon.com product reviews are used for the analysis in this paper. Seven data points are available in every review: reviewer's ID, reviewer's name, product's ID, product rating, timestamp, effectiveness, and review content. Ratings can only be discrete (for example. 2 stars), not continuous (for example. 2.5 stars), and range from 0 to 5. Using Amazon.com dataset of Video game product reviews, training model is created. In order to set the levels, we separated the ratings according to one to five stars. The obtained results distribution is shown in Fig. 1.

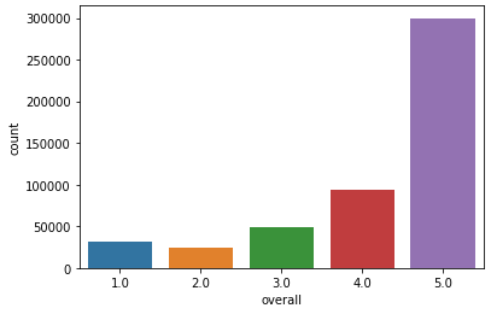

Fig. 1 Histogram of Ratings

B. Data Pre-Processing

a) Tokenization: Tokenization is the process of breaking up a sentence into symbols, keywords, and phrases. Some characters are removed in tokenization, such as exclamation marks and semicolons.

b) Removal of Stopwords: Stopword in text mining are those parts of a sentence that are not required in any segment and are typically eliminated to improve the analysis's efficiency. The formats of stop words vary depending on the language and country.

c) Parts Of Speech tagging/labelling: The technique of POS tagging involves identifying the word's part of speech, it contains noun, pronoun, verb, adjective and their subcategories.

d) Stemming : The process of integrating a word's modified versions into its typical meaning is known as stemming. This method of information retrieval (IR) from the text of the document based on the statements is employed in text processing.

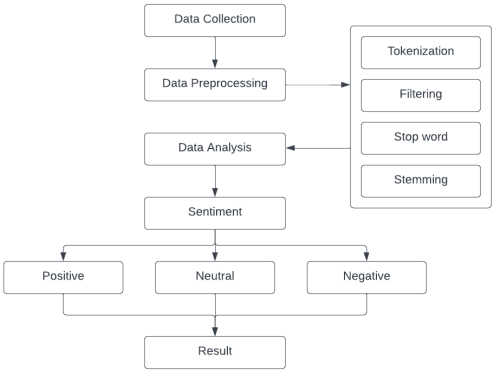

Fig. 2 Work flow

C. Text Sentiment Extarction and Parts of Speech Tagging

The review was initially transformed into tokens into distinct English words and The remaining words participate in POS tagging after STOP words is removed, filtered, and stemmed such as "in", "is", "are" ,"but," etc.. The POS tagging technique in NLP is well-known and useful for identifying the function of words in sentence and was created to arrange words according to their parts of speech The POS tagger is extremely helpful for opinion mining for the following two primary reasons: Nouns and pronouns typically do not express opinions about the goods, and they can be readily removed by utilising a POS tagger. A POS tagger additionally helps in identifying words used in various parts of speech.

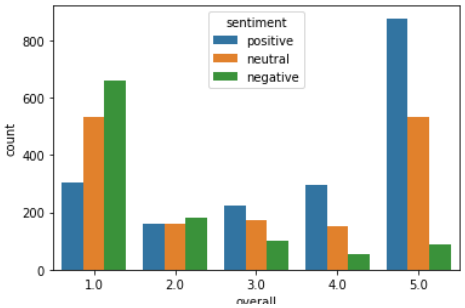

Fig. 3 Sentiment count

IV. IMPLEMENTATION

A. Classification Classifier

The process of identifying and categorising a particular opinion depending on orientation (positive, negative, or neutral) is known as the sentiment classification, also known as the polarity categorization. Support Vector Machine (SVM) is the categorization model that was selected. Scikit-learn [11], an open source machine learning toolkit used in the Python programming language for data analysis and data mining, is the software library used for this work.

B. Sentiment Analysis using Unsupervised Lexicon-Based Models

The NLTK (Natural Language Toolkit) library in Python provides a pre-built Opinion Lexicon that can be used for sentiment analysis tasks. The Opinion Lexicon contains lists of positive and negative words, along with their associated polarity. It's important to note that lexicon-based approaches, like the Opinion Lexicon, have certain limitations. They may

struggle with sarcasm, negation, context-dependent sentiment, and new or domain-specific words that are not present in the lexicon. Despite its limitations, the Opinion Lexicon can be a useful resource for basic sentiment analysis tasks or as a starting point for developing more sophisticated sentiment analysis systems.

V. RESULT

We developed a model using the machine learning classification approach SVM to determine if customer reviews were positive or negative, and we calculated the training model's accuracy using the findings of Recall, Precision, F1. When we tested our model, we got the best accuracy. The confusion matrix we found is as follows:
for the unsupervised lexicon model, We retrieve a small corpus of test data.This data is used in the model. In confusion matrix the predicted label shows the predicted

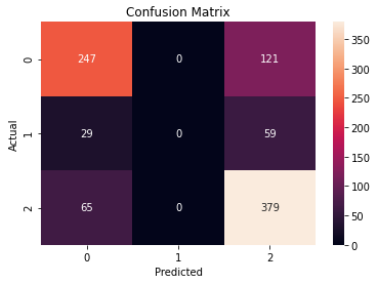

Fig.4 Confusion Matrix of SVM model

sentiment and the Actual label has test sentiment. We evaluate model performance and accuracy using these labels. Model performance using precision score, re-call, f1-Score, accuracy for both classes. The confusion matrix we found is

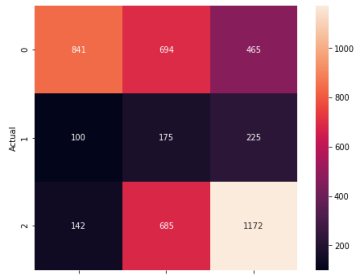

Fig.5Confusion matrix of Opinion Lexicon

as shown in fig.6:
Comparing the Results of SVM and Opinion Lexicon.
On comparing both supervised and unsupervised Lexicon models, The SVM model outperforms the Opinion lexicon model with the highest Positive percentage 85.3% accuracy

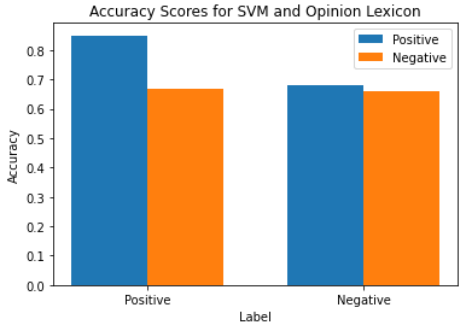

Fig.6Accuracy of model

and negative percent 67.1%, which is depicted in the graph in Fig. 6.

Table.1 shows the Accuracy, precision, recall and F1 score of both supervised and unsupervised model and we see SVM performs better then NLTK opinion Lexicon in both positive reviews and negative reviews.

SVM using TF-IDF				
	Accuracy	Precision	Recall	F1
Positive	0.853	1.0	0.853	0.921
Negative	0.671	1.0	0.671	0.803
NLTK Opinion Lexion				
	Accuracy	Precision	Recall	F1
Positive	0.681	0.749	0.593	0.609
Negative	0.663	0.771	0.402	0.52

Table.1 Comparative result of SVM and Opinion Lexicon model

VI. CONCLUSION

The Amazon.com reviews that were used for the sentiment analysis were successful.In this study, we use both supervised and unsupervised models to analyse sentiment. In the classification model for SVM, was employed, and for the unsupervised lexicon model we used the opinion lexicon. The accuracy of the Opinion lexical model was found to be 68.1 percent. The best supervised learning model is the SVM model on TF-IDF features, which has an accuracy rate of 85.3 percent. By comparing the models from both supervised and unsupervised models, we may get the conclusion that traditional supervised models perform better than the lexical model.

VII. REFERENCES

1. P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in Proceedings of the 40th annual meeting of the association for computational linguistics, Dec. 2002, pp. 417-424. https://arxiv.org/abs/cs/0212032.
2. D. M. E.-D. M. Hussein, "A survey on sentiment analysis challenges," J.King Saud Univ. - Eng. Sci., vol. 30, pp. 330-338, Oct. 2018,https://doi.org/10.1016/j.jksues.2016.04.002
3. Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology, 60(11), 2169-2188.
4. Utz, S., Kerkhof, P., & Bos, J. V. (2012). Consumers rule: How consumer reviews influence perceived trustworthiness of online stores. Electronic Commerce Research and Applications, 11(1), 49-58.doi:10.1016/j.elerap.2011.07.010
5. Tan, L. K., Na, J., Theng, Y., & Chang, K. (2011). Sentence-Level Sentiment Polarity Classification Using a Linguistic Approach. Digital Libraries: For Cultural Heritage, Knowledge Dissemination, and Future Creation Lecture Notes in Computer Science, 77-87. doi:10.1007/978-3-642-24826-9 13
6. Krikorian, Raffi. (VP, Platform Engineering, Twitter Inc.). ”New Tweets per second record, and how!” Twitter Official Blog. August 16, 2013.
7. Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

8. Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology, 60(11), 2169-2188.
9. [5] Mukherjee, S., & Bhattacharyya, P. (2012, March). Feature specific sentiment analysis for product reviews. In *International conference on intelligent text processing and computational linguistics* (pp. 475-487). Springer, Berlin, Heidelberg.
10. [6] Rashid, A., & Huang, C. Y. (2021). Sentiment Analysis on Consumer Reviews of Amazon Products. *International Journal of Computer Theory and Engineering*, *13*(2).