MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE GWALIOR
(A Govt. Aided UGC Autonomous Institute Affiliated to RGPV, Bhopal)
NAAC Accredited with A++ Grade

Project Report

on

Sentiment Analysis Using Python

Submitted By:

JAYESH KUSHWAH

0901AD211022

Faculty Mentor:

Mr. Arun Kumar

CENTRE FOR ARTIFICIAL INTELLIGENCE
MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE
GWALIOR - 474005 (MP) est. 1957

JULY-DEC. 2023

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE GWALIOR

(A Govt. Aided UGC Autonomous Institute Affiliated to RGPV, Bhopal)

NAAC Accredited with A++ Grade

CERTIFICATE

This is certified that **Jayesh Kushwah** (0901AD211022) has submitted the project report titled **Sentiment Analysis Using Python** under the mentorship of **Mr. Arun Kumar**, in partial fulfilment of the requirement for the award of degree of Bachelor of Technology in **Artificial Intelligence and Data Science** from Madhav Institute of Technology and Science, Gwalior.

23 Nov 2023

Mr. Arun Kumar
Faculty Mentor
Assistant Professor
Centre for Artificial Intelligence

23/11/23

Dr. R. R. Singh
Coordinator
Centre for Artificial Intelligence

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE GWALIOR

(A Govt. Aided UGC Autonomous Institute Affiliated to RGPV, Bhopal)

NAAC Accredited with A++ Grade

DECLARATION

I hereby declare that the work being presented in this project report, for the partial fulfilment of requirement for the award of the degree of Bachelor of Technology in **Artificial Intelligence and Data Science** at Madhav Institute of Technology & Science, Gwalior is an authenticated and original record of my work under the mentorship of **Mr. Arun Kumar**, **Assistant Professor**, Centre for Artificial Intelligence

I declare that I have not submitted the matter embodied in this report for the award of any degree or diploma anywhere else.

Jayesh Kushwah
0901AD211022
III Year,
Centre for Artificial Intelligence

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE GWALIOR

(A Govt. Aided UGC Autonomous Institute Affiliated to RGPV, Bhopal)

NAAC Accredited with A++ Grade

ACKNOWLEDGEMENT

The full semester project has proved to be pivotal to my career. I am thankful to my institute, **Madhav Institute of Technology and Science** to allow me to continue my disciplinary/interdisciplinary project as a curriculum requirement, under the provisions of the Flexible Curriculum Scheme (based on the AICTE Model Curriculum 2018), approved by the Academic Council of the institute. I extend my gratitude to the Director of the institute, **Dr. R. K. Pandit** and Dean Academics, **Dr. Manjaree Pandit** for this.

I would sincerely like to thank my department, **Centre for Artificial Intelligence,** for allowing me to explore this project. I humbly thank **Dr. R. R. Singh**, Coordinator, Centre for Artificial Intelligence, for his continued support during the course of this engagement, which eased the process and formalities involved.

I am sincerely thankful to my faculty mentors. I am grateful to the guidance of **Mr. Arun Kumar**, Assistant Professor, Centre for Artificial Intelligence, for his continued support and guidance throughout the project. I am also very thankful to the faculty and staff of the department.

Jayesh Kushwah
0901AD211022
III Year,
Centre for Artificial Intelligence

ABSTRACT

This project explored the domain of logic using the Natural Language Toolkit (NLTK) library for Amazon product reviews. Understanding people's needs in the dynamic environment of e-commerce is important for businesses. The NLTK library is known for its powerful language tools and forms the basis of our analysis. The main goal is to develop a predictive model that can identify Amazon product reviews as Positive, Negative, or neutral.

This work includes data collection, pre-processing, pattern removal and model training using machine learning algorithms such as Naive Bayes and Support Vector Machines. Address challenges inherent in emotional analysis, such as sarcasm and context-sensitive thinking. The effectiveness of the NLTK library in solving these problems is measured by metrics such as accuracy, precision, recall, and F1 score.

Analysis results show the customer's opinion about the product, give a good idea for the company to improve business strategy, strengthen production and increase customer satisfaction. This project simply demonstrates the use of NLTK in sentiment analysis and demonstrates its ability to extract meaningful content from large datasets of unstructured data in the summary of Amazon product reviews.

Keyword: Natural Language Toolkit (NLTK), Amazon Product Reviews, E-commerce, Predictive Model, NLTK Library, Data Collection, Data Preprocessing, Machine Learning Algorithms, Naive Bayes, Support Vector Machines, Emotional Analysis, Sarcasm, Context-sensitive Thinking, Effectiveness, Metrics, Accuracy, Precision, Recall, F1 Score, Business Strategy, Production, Customer Satisfaction, Sentiment Analysis, Unstructured Data.

सार:

इस प्रोजेक्ट ने अमेज़ॅन उत्पाद समीक्षाओं के लिए नेचुरल लैंग्वेज टूलकिट (एनएलटीके) लाइब्रेरी का उपयोग करके तर्क के क्षेत्र का पता लगाया। ई-कॉमर्स के गतिशील वातावरण में लोगों की ज़रूरतों को समझना व्यवसायों के लिए महत्वपूर्ण है। एनएलटीके लाइब्रेरी अपने शक्तिशाली भाषा उपकरणों के लिए जानी जाती है और यह हमारे विश्लेषण का आधार बनती है। मुख्य लक्ष्य एक पूर्वानुमानित मॉडल विकसित करना है जो अमेज़ॅन उत्पाद समीक्षाओं को सकारात्मक, नकारात्मक या तटस्थ के रूप में पहचान सके।

इस कार्य में नाइव बेयस और सपोर्ट वेक्टर मशीन जैसे मशीन लर्निंग एल्गोरिदम का उपयोग करके डेटा संग्रह, प्री-प्रोसेसिंग, पैटर्न हटाना और मॉडल प्रशिक्षण शामिल है। भावनात्मक विश्लेषण में निहित चुनौतियों का समाधान करें, जैसे व्यंग्य और संदर्भ-संवेदनशील सोच। इन समस्याओं को हल करने में एनएलटीके लाइब्रेरी की प्रभावशीलता को सटीकता, सटीकता, रिकॉल और एफ। स्कोर जैसे मेट्रिक्स द्वारा मापा जाता है।

विश्लेषण के परिणाम उत्पाद के बारे में ग्राहक की राय दिखाते हैं, कंपनी को व्यावसायिक रणनीति में सुधार करने, उत्पादन को मजबूत करने और ग्राहक संतुष्टि बढ़ाने के लिए एक अच्छा विचार देते हैं। यह परियोजना केवल भावना विश्लेषण में एनएलटीके के उपयोग को प्रदर्शित करती है और अमेज़ॅन उत्पाद समीक्षाओं के सारांश में असंरचित डेटा के बड़े डेटासेट से सार्थक सामग्री निकालने की इसकी क्षमता को प्रदर्शित करती है।

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1: INTRODUCTION

In the dynamic world of e-commerce, understanding customer needs is critical to business success. This project uses the Natural Language Toolkit (NLTK) library, a powerful tool in natural language processing, to understand the sentiment of Amazon product reviews. Our main goal is to build a model to determine sentiment (positive, negative, or neutral) in user reviews. We solve problems like sarcasm and context-aware thinking through data collection, pre-processing, and Lexical based Analysis such as the NLTK Vader sentiment analyzer, involves using a set of predefined rules and heuristics to determine the sentiment of a piece of text. The performance of the NLTK library is measured by metrics such as accuracy, precision, recall, and F1 score. The results not only reveal customers' opinions, but also provide valuable information for companies to adjust their strategies, improve production and increase customer satisfaction in the highly competitive e-commerce environment. This project shows NLTK in action for sentiment analysis, providing a valuable tool for businesses looking to leverage the power of customer feedback.

1.1 Objective:

The objective of this project is to implement sentiment analysis on Amazon product reviews using the Natural Language Toolkit (NLTK) library. Specifically tailored to the dynamic e-commerce landscape, the project seeks to create a predictive model for the categorization of reviews into positive, negative, or neutral sentiments. Leveraging the capabilities of the NLTK library, the project encompasses tasks such as data collection, pre-processing, and Lexical based Analysis such as the NLTK Vader sentiment analyzer, involves using a set of predefined rules and heuristics to determine the sentiment of a piece of text. Addressing challenges inherent in sentiment analysis, including sarcasm, the project assesses the NLTK library's effectiveness using metrics such as accuracy. Ultimately, the project aims to empower businesses with a tool for extracting actionable insights from Amazon reviews, facilitating informed decision-making in the competitive marketplace.

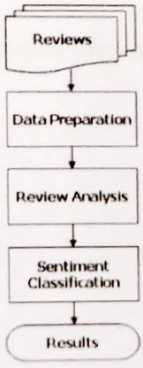

Figure 1.1 Flow chart of basic model

1.2 System Requirement:

Operating System:

Compatible with Windows, macOS, and Linux.

Python:

Requires Python 3.x.

Memory (RAM):

A minimum of 4 GB RAM is recommended.

Storage:

Adequate space for the dataset, NLTK library, and related libraries/models.

Processor:

A standard processor is sufficient, but a multi-core processor is beneficial for faster processing.

Internet Connection:

Necessary for downloading NLTK resources during setup.

Development Environment:

Use a code editor or IDE like Jupyter Notebook, PyCharm, or VSCode.

NLTK Library:

Install NLTK using the command: pip install nltk.

Dependencies:

Ensure all necessary dependencies, including tokenizers and corpora, are installed.

Chapter 2: Literature Survey

A literature survey on sentiment analysis using the NLTK library for Amazon product reviews reveals a growing body of research in this domain. Numerous scholarly articles and research papers emphasize the importance of sentiment analysis in deciphering customer opinions, particularly within e-commerce contexts. Common themes include the application of natural language processing (NLP) tools, machine learning algorithms, and sentiment lexicons. The literature underscores the challenges inherent in sentiment analysis, such as dealing with sarcasm and context-dependent sentiments, aligning with the objectives of this project. Additionally, there is considerable attention given to evaluating the effectiveness of the NLTK library in sentiment analysis, especially in the context of product reviews, highlighting its practical applicability. This literature survey serves as a foundation for the project, guiding the selection of methodologies and offering insights into gaps or areas requiring further exploration. The existing research findings inform the approach taken in developing a sentiment analysis model using the NLTK library for Amazon product reviews.

Chapter 3: Preliminary Design:

The preliminary design of a sentiment analysis system using the NLTK library for Amazon product reviews involves outlining the key components and steps required for the project. Here is a brief overview of the preliminary design:

3.1 Data Collection:

Gather a dataset of Amazon product reviews, ensuring it covers a diverse range of products and sentiments.

	reviewText	Positive	sentiment
0	one best apps acording bunch people agree bomb...	1	1
1	pretty good version game free . lot different ...	1	1
2	really cool game . bunch level find golden egg...	1	1
3	silly game frustrating , lot fun definitely re...	1	1
4	terrific game pad . hr fun . grandkids love	1	1
...	...	...	...
19995	app fricken stupid.it froze kindle wont allow ...	0	0
19996	please add ! ! ! ! ! need neighbor ! ginger101...	1	1
19997	love ! game . awesome . wish free stuff house ...	1	1
19998	love love love app side fashion story fight wo...	1	1
19999	game rip . list thing make better & bull ; fir...	0	1

20000 rows × 3 columns

Figure 3.1 Data Collection

3.2 Data Preprocessing:

Clean the data by handling missing values, removing irrelevant information, and standardizing the text. Tokenize the reviews into individual words.

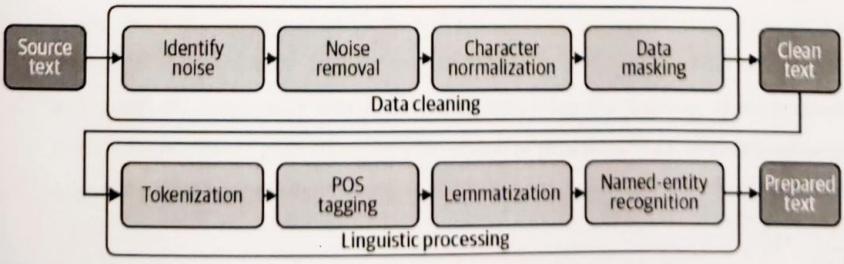

Figure 3.2 Data Preprocessing

3.3 NLTK Setup:

Install and set up the NLTK library, including downloading necessary resources such as tokenizers and sentiment lexicons.

To install NLTK library:

```
pip install nltk
```

Figure 3.3.1 Install NLTK

To download all resources:

```
import nltk

nltk.download('all')
```

Figure 3.3.2 Download Resources

3.4 Feature Extraction:

Extract relevant features from the preprocessed data. This could involve using techniques like bag-of-words.

3.4.1 Bag of word model:

The bag-of-words model is a fundamental technique in natural language processing (NLP) for converting text data into numerical features. In this model, each document is treated as an unordered set of words, and each word becomes a distinct feature in the resulting vector representation. The value assigned to each feature corresponds to the frequency of the respective word within the text.

This approach is essential in NLP as it enables the application of machine learning algorithms, which typically operate on numerical input. The conversion of text data into numerical features facilitates tasks such as text classification and sentiment analysis using machine learning models.

In the following section, an example will demonstrate the application of the NLTK Vader model for sentiment analysis on the Amazon customer dataset. Although the NLTK Vader API directly accepts text input, it's noteworthy that in scenarios involving the training of supervised machine learning models for sentiment prediction, the conversion of processed text into a bag-of-words model becomes a crucial preprocessing step.

	about	bird	heard	is	the	word	you
About the bird, the bird, bird bird bird	1	5	0	0	2	0	0
You heard about the bird	1	1	1	0	1	0	1
The bird is the word	0	1	0	1	2	1	0

Figure 3.4.1 Bag of Words Model

3.5 Model Evaluation:

Evaluate the model's performance using metrics like accuracy, precision, recall, and F1 score. Adjust the model parameters if needed.

3.6 Challenges Addressing:

Implement strategies to address challenges in sentiment analysis, such as handling sarcasm and context-sensitive sentiments.

Dynamic Language Evolution:

Challenge: Language evolves, and sentiment expressions may change over time.

Addressing Strategy: Regularly update sentiment lexicons and models to adapt to evolving language trends. Monitor changes in sentiment expressions over time.

Handling Multilingual Text:

Challenge: Sentiment analysis across multiple languages introduces complexities due to linguistic differences.

Addressing Strategy: Utilize multilingual sentiment analysis tools and models. Translate text to a common language before analysis or employ language-specific sentiment models.

Contextual Sentiments:

Challenge: Sentiments often depend on context, making it challenging to accurately interpret without contextual understanding.

Addressing Strategy: Utilize context-aware sentiment analysis techniques, considering the broader context in which the text is situated. This may involve analyzing surrounding sentences or paragraphs.

Sarcasm and Irony:

Challenge: Textual expressions of sarcasm or irony may convey sentiments opposite to the literal meaning.

Addressing Strategy: Incorporate contextual analysis to identify cues indicating sarcasm. Advanced machine learning models, such as deep learning approaches, can enhance sarcasm detection.

Chapter 4: Final Analysis and Design:

To conduct sentiment analysis using NLTK in Python, it is essential to preprocess the text data through procedures like tokenization, stop-word removal, and either stemming or lemmatization. Following the preprocessing step, the text is then fed into the Vader sentiment analyzer to assess its sentiment, determining whether it is positive or negative.

4.1 Import libraries and load dataset:

```
# import libraries
import pandas as pd

import nltk

from nltk.sentiment.vader import SentimentIntensityAnalyzer

from nltk.corpus import stopwords

from nltk.tokenize import word_tokenize

from nltk.stem import WordNetLemmatizer


# download nltk corpus (first time only)
#import nltk

#nltk.download('all')



# Load the amazon review dataset

df = pd.read_csv('https://raw.githubusercontent.com/pycaret/pycaret/master/datasets/amazon.csv')

df
```

Figure 4.1 Import libraries and load dataset

Output of figure 4.1:

	reviewText	Positive
0	This is a one of the best apps acording to a b...	1
1	This is a pretty good version of the game for ...	1
2	this is a really cool game. there are a bunch ...	1
3	This is a silly game and can be frustrating. b...	1
4	This is a terrific game on any pad. Hrs of fun...	1
...	...	...
19995	this app is fricken stupid.it froze on the kin...	0
19996	Please add me!!!!! I need neighbors! Ginger101...	1
19997	love it! this game. is awesome. wish it had m...	1
19998	I love love love this app on my side of fashio...	1
19999	This game is a rip off. Here is a list of thin...	0

20000 rows × 2 columns

Figure 4.1.1 Output

4.2 Preprocess Text:

Let's create a function preprocess_text in which we first tokenize the documents using word_tokenize function from NLTK, then we remove step words using stepwords module from NLTK and finally, we lemmatize the filtered_tokens using WordNetLemmatizer from NLTK.

```
# create preprocess_text function
def preprocess_text(text):

    # Tokenize the text

    tokens = word_tokenize(text.lower())



    # Remove stop words

    filtered_tokens = [token for token in tokens if token not in stopwords.words('english')]



    # Lemmatize the tokens

    lemmatizer = WordNetLemmatizer()

    lemmatized_tokens = [lemmatizer.lemmatize(token) for token in filtered_tokens]



    # Join the tokens back into a string

    processed_text = ' '.join(lemmatized_tokens)

    return processed_text

# apply the function df

df['reviewText'] = df['reviewText'].apply(preprocess_text)
df
```

Figure 4.2 Preprocess Text

Output of figure 4.2:

	reviewText	Positive
0	one best apps acording bunch people agree bomb...	1
1	pretty good version game free . lot different ...	1
2	really cool game . bunch level find golden egg ...	1
3	silly game frustrating . lot fun definitely re...	1
4	terrific game pad . hr fun . grandkids love	1
...	...	...
19995	app fricken stupid it froze kindle wont allow ...	0
19996	please add ! ! ! ! need neighbor ! ginger101...	1
19997	love ! game . awesome . wish free stuff house ...	1
19998	love love love app side fashion story fight wo...	1
19999	game rip . list thing make better & bull ; fir...	0

20000 rows × 2 columns

Figure 4.2.1 Output

4.3 NLTK Sentiment Analyzer:

- Initialize Sentiment Intensity Analyzer from nltk.sentiment.vader.
- Create a function get_sentiment(text) using polarity_scores method to obtain sentiment scores.
- Check if the positive score is greater than 0; assign sentiment score 1 for positive, 0 otherwise.
- Apply get_sentiment function to 'reviewText' column in the DataFrame (df) using the apply method.
- Create a new column 'sentiment' to store sentiment scores for each review.
- Display the updated DataFrame (df).

```
# initialize NLTK sentiment analyzer

analyzer = SentimentIntensityAnalyzer()


# create get_sentiment function

def get_sentiment(text):

    scores = analyzer.polarity_scores(text)

    sentiment = 1 if scores['pos'] > 0 else 0

    return sentiment



# apply get_sentiment function

df['sentiment'] = df['reviewText'].apply(get_sentiment)

df
```

Figure 4.3 Initialize NLTK sentiment analyzer

Output of figure 4.3:

	reviewText	Positive	sentiment
0	one best apps acording bunch people agree bomb...	1	1
1	pretty good version game free . lot different ...	1	1
2	really cool game . bunch level find golden egg...	1	1
3	silly game frustrating , lot fun definitely re...	1	1
4	terrific game pad . hr fun . grandkids love	1	1
...	...	...	...
19995	app fricken stupid it froze kindle wont allow ...	0	0
19996	please add ! ! ! ! ! need neighbor ! ginger101 ...	1	1
19997	love ! game . awesome . wish free stuff house ...	1	1
19998	love love love app side fashion story fight wo...	1	1
19999	game rip . list thing make better & bull ; fir...	0	1

20000 rows × 3 columns

Figure 4.3.1 Output

4.4 Evaluate the performance:

The NLTK sentiment analyzer returns a score between -1 and +1. We have used a cut-off threshold of 0 in the get_sentiment function above. Anything above 0 is classified as 1 (meaning positive). Since we have actual labels, we can evaluate the performance of this method by building a confusion matrix.

```
from sklearn.metrics import confusion_matrix

print(confusion_matrix(df['Positive'], df['sentiment']))
```

```
[[ 1131  3636]
 [  576 14657]]
```

Figure 4.4.1 Confusion Matrix

We can also check the classification report:

```
from sklearn.metrics import classification_report

print(classification_report(df['Positive'], df['sentiment']))
```

```
              precision    recall  f1-score   support

           0       0.66      0.24      0.35      4767
           1       0.80      0.96      0.87     15233

    accuracy                           0.79     20000
   macro avg       0.73      0.60      0.61     20000
weighted avg       0.77      0.79      0.75     20000
```

Figure 4.4.2 Classification Report

4.5 Applications:

1.Business Strategy Optimization:

Provide businesses with insights into customer sentiments to refine marketing strategies, enhance products, and stay competitive in the market.

2.Customer Feedback Management:

Enhance customer service by promptly addressing negative sentiments and identifying areas for improvement based on customer feedback.

3.Competitor Analysis:

Analyze sentiment scores in comparison to competitors to pinpoint strengths and weaknesses, allowing for strategic positioning in the market.

4.Brand Reputation Monitoring:

Monitor and manage brand reputation through the analysis of sentiments associated with the brand in online reviews and social media.

5.E-commerce Platform Enhancement:

Improve the user experience on e-commerce platforms by addressing common issues highlighted in product reviews, ultimately boosting overall customer satisfaction.

4.6 Conclusion:

In conclusion, the sentiment analysis project employing a lexical-based approach, specifically utilizing the Natural Language Toolkit (NLTK), has provided valuable insights into the sentiments expressed in Amazon product reviews. The project encompassed various stages, including data collection, preprocessing, and the application of NLTK's tools such as tokenization, stop-word removal, and lemmatization.

The lexical-based approach, exemplified by NLTK's Vader sentiment analyzer, demonstrated its effectiveness in discerning sentiments, offering a nuanced understanding of the polarity of the reviews. This approach proves particularly useful in scenarios where context and nuances play a crucial role in sentiment interpretation.

While NLTK presents a robust framework for lexical-based sentiment analysis, it is essential to acknowledge its limitations, such as challenges in handling sarcasm, context-dependent sentiments, and subjectivity. Addressing these challenges necessitates a combination of advanced techniques and domain-specific adjustments.

Ultimately, the lexical-based sentiment analysis using NLTK provides a foundational understanding of customer sentiments in the context of Amazon product reviews. This approach equips businesses with a valuable tool for informed decision-making, enabling them to enhance products, refine marketing strategies, and ultimately improve customer satisfaction. As the field of natural language processing continues to evolve, further exploration of advanced techniques and integration with machine learning models holds the potential to enhance the accuracy and applicability of sentiment analysis in diverse real-world scenarios.

References

1. https://raw.githubusercontent.com/pycaret/pycaret/master/datasets/amazon.csv
2. http://www.nltk.org/book/
3. https://www.geeksforgeeks.org/
4. https://www.nltk.org/