EDA and Twitter Sentimental Analysis

Kenee Patel
Department of Computer Science and Engineering, Devang Patel Institute of Advance Technology and Research (DEPSTAR), Charotar University of Science and Technology (CHARUSAT), Anand, India.
patelkenee2804@gmail.com

Parth Patel
Department of Computer Science and Engineering, Devang Patel Institute of Advance Technology and Research (DEPSTAR), Charotar University of Science and Technology (CHARUSAT), Anand, India.
parth.642001@gmail.com

Shruti Patel
Department of Computer Science and Engineering, Devang Patel Institute of Advance Technology and Research (DEPSTAR), Charotar University of Science and Technology (CHARUSAT), Anand, India.
19dcs102@charusat.edu.in

Krishna Patel
Department of Computer Science and Engineering, Devang Patel Institute of Advance Technology and Research (DEPSTAR), Charotar University of Science and Technology (CHARUSAT), Anand, India.
krishnapatel.ce@charusat.ac.in

Parth Goel
Department of Computer Science and Engineering, Devang Patel Institute of Advance Technology and Research (DEPSTAR), Charotar University of Science and Technology (CHARUSAT), Anand, India.
er.parthgoel@gmail.com

Amit Ganatra
Department of Computer Engineering, Devang Patel Institute of Advance Technology and Research (DEPSTAR), Faculty of Technology and Engineering (FTE), Charotar University of Science and Technology (CHARUSAT) , Anand, India.
amitganatra.ce@charusat.ac.in

***Abstract*— In today's scenario, where we are generating data at an unprecedented rate and we are living in an era where future oil of the world is going to be the data. We are witnessing a paradigm shift where more and more companies and organizations are relying on the data driven decision making. We this, comes the advancement and dominance of the fields of data analytics and data science. These two domains are heavily data dependent and thus the importance of the data is increased. EDA stands for Exploratory Data Analysis [1] which is an important is an important step in any Data Analysis or Data Science project. EDA is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset.[1] EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better. With this, the need of sentiment analysis is also coming into demand as the use of social media is firing to a record. The project consists of 2 subprojects which will be developed in phased manner. The first one is about creating an EDA app which will give a data analyst a generalized information about the data set, which will be helpful in the decision making and second one is about developing a dashboard for twitter sentiment analysis.**

Keywords—EDA, Sentiment analysis, data analytics, machine learning, streamlit, naïve bayes classifier

I. Introduction

The overall project is divided into two phases.
The two phases are as follows
1. Development of exploratory data analysis application
2. Development of Twitter sentiment analysis dashboard
So talking about the phase one of the project the project is all about development and analysis of the exploratory data analysis application which is also commonly known as EDA app.

Exploratory Data Analysis, or EDA, is an important step in any Data Analysis or Data Science project. EDA is the pro cess of investigating the dataset to discover patterns, and a nomalies (outliers), and form hypotheses based on our und erstanding of the dataset. EDA involves generating summ ary statistics for numerical data in the dataset and creating various graphical representations to understand the data bet ter. In this article, we will understand EDA with the help of an example dataset.

The second phase of project that is the development and th e analysis of Twitter sentiment analysis dashboard. This is regarding the analysis of sentiments or which is commonly called as sentiment analysis. this kind of analysis is very famous and very common and in demand nowadays because lots and lots of companies are trying to influence their customers with the help of social media. social media is a place millions of users swipe and oh content switching now and then 24 by 7. Sentiment analysis, also referred to as opinion mining, is an approach to natural language processing that identifies the emotional tone behind a body of text. This is a popular way for organizations to determine and categorize opinions about a product, service, or idea. It involves the use of data mining, machine learning (ML) and artificial intelligence (AI) to mine text for sentiment and subjective information

II. EDA

A. Explanation

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. [2]

It is a good practice to understand the data first and try to gather as many insights from it. EDA is all about making sense of data in hand, before processing them with different approaches to make use of it. [2]

Imagine your group of friends decides to watch a movie you haven’t heard of. There is absolutely no debate about that, it will lead to a state where you find yourself puzzled with lot of questions which needs to be answered in order to make a decision. Being a good chieftain the first question you would ask, what is the cast and crew of the movie? As a regular practice ,you would also watch the trailer of the movie on YouTube. Furthermore, you’d find out ratings and reviews the movie has received from the audience.

B. *Advantages*

It helps us understand the given dataset and helps clean up the given dataset. It gives you a clear picture of the features and the relationships between them. It Provides guidelines for essential variables leaving behind or removing non-essential variables.[1] It is especially helpful in handling missing values or human error. Morever it can also be useful in identifying outliers. The EDA process would be maximizing insights of a dataset.

C. *Process*

The process in formal EDA generally consists of:

- Renaming of columns in the dataset
- Dropping the duplicate rows
- Dropping the missing or null values
- Uni-variate Analysis, Bi-variate Analysis and Multi-variate Analysis
- Finding Data Types of Column(s)
- Dropping irrelevent Columns
- Detecting Outliers
- Missing values treatment
- Variable transformation
- Feature Creation.

III. SENTIMENTAL ANALYSIS

Sentiment analysis (or opinion mining) is a natural language processing (NLP) technique used to determine whether data is positive or negative. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs. [3]

A. *Types*

Sentiment analysis focuses on the polarity of a text (positive, negative or in some cases even neutral) but it can also goes beyond polarity to detect specific feelings and emotions (angry, happy, sad, etc), urgency (urgent or not urgent) and even intentions (interested or not interested).

Depending on how you want to interpret customer feedback and queries, you can define and tailor your categories to meet your sentiment analysis needs. In the meantime, here are some of the most popular types of sentiment analysis:

- Graded Sentiment Analysis

  If polarity precision is important to your business, you might consider expanding your polarity categories to include different levels of positive and negative: [3]

  - Very positive
  - Positive
  - Neutral
  - Negative
  - Very negative

  This is usually referred to as graded or fine-grained sentiment analysis, and could be used to interpret 5-star ratings in a review, for example:

  - Very Positive = 5 stars
  - Very Negative = 1 star

- Emotion detection

  Emotion detection sentiment analysis allows you to go beyond polarity to detect emotions, like happiness, frustration, anger, and sadness.

  Many emotion detection systems use lexicons (i.e. lists of words and the emotions they convey) or complex machine learning algorithms.

  One of the downsides of using lexicons is that people express emotions in different ways. Some words that typically express anger, like bad or kill (e.g. your product is so bad or your customer support is killing me) might also express happiness (e.g. this is bad ass or you are killing it).

- Aspect-based Sentiment Analysis

  Usually, when analyzing sentiments of texts you’ll want to know which particular aspects or features people are mentioning in a positive, neutral, or negative way.

  That's where aspect-based sentiment analysis can help, for example in this product review: "The battery life of this camera is too short", an aspect-based classifier would be able to determine that the sentence expresses a negative opinion about the battery life of the product in question.

- Multilingual sentiment analysis

  Multilingual sentiment analysis can be difficult. It involves a lot of preprocessing and resources. It involves things as sentiment lexicons or translated corpora or noise detection algorithms.

B. *Why we need Sentimental analysis*

Since humans express their thoughts and feelings more openly than ever before, sentiment analysis is fast becoming an essential tool to monitor and understand sentiment in all types of data.

Automatically analyzing customer feedback, such as opinions in survey responses and social media conversations, allows brands to learn what makes customers happy or frustrated, so that they can tailor products and services to meet their customers’ needs.

For example, using sentiment analysis to automatically analyze 4,000+ open-ended responses in your customer satisfaction surveys could help you discover why customers are happy or unhappy at each stage of the customer journey.

Maybe you want to track brand sentiment so you can detect disgruntled customers immediately and respond as soon as possible. Maybe you want to compare sentiment from one quarter to the next to see if you need to take action. Then you could dig deeper into your qualitative data to see why sentiment is falling or rising.

C. *Pros of Sentimental analysis*

- Sorting Data at Scale:
  Can you imagine manually sorting through thousands of tweets, customer support conversations, or surveys? There's just too much business data to process manually. Sentiment analysis helps businesses process huge amounts of unstructured data in an efficient and cost-effective way.[3]
- Real-Time Analysis:
  Sentiment analysis can identify critical issues in real-time, for example is a PR crisis on social media escalating? Is an angry customer about to churn? Sentiment analysis models can help you immediately identify these kinds of situations, so you can take action right away.
- Consistent criteria:
  It's estimated that people only agree around 60-65% of the time when determining the sentiment of a particular text. Tagging text by sentiment is highly subjective, influenced by personal experiences, thoughts, and beliefs.

  By using a centralized sentiment analysis system, companies can apply the same criteria to all of their data, helping them improve accuracy and gain better insights.
  The applications of sentiment analysis are endless. So, to help you understand how sentiment analysis could benefit your business, let's take a look at some examples of texts that you could analyze using sentiment analysis.

IV. ML MODEL

The Nave Bayes method is a supervised learning algorithm for addressing classification issues that is based on the Bayes theorem. [4] It is mostly utilised in text classification tasks that require a large training dataset. The Nave Bayes Classifier is a simple and effective classification method that aids in the development of fast machine learning models capable of making quick predictions. It's a probabilistic classifier, which means it makes predictions based on an object's probability.
Nave Bayes is a fast and simple machine learning algorithm for predicting a class of datasets. It's suitable for both binary and multi-class classifications.
In comparison to the other Algorithms, it performs better in Multi-class predictions.It is the most often used method for text classification.

V. PROJECT OBJECTIVE

Main objective behind the development of this project was to facilitate and is of working for the Aspiring and current data analysts and data scientists. We know that in each and every project of data science and Data Analytics EDA is an important part. In order to decide whether which data set is correct for the project and which data set is not correct for the project each and every analyst and data scientist has to code extra 60 to 120 lines extra for each and every data set. So we have tried to automate this process. The exploratory data analysis application that we have built tries to automate the process and also aims at saving time of the analyst.
The second phase of our project was to develop Twitter sentiment analysis dashboard. We need to clarify that dashboard means feeding on live data. Now this tool will help sentiment analysts to analyse the sentiment for the current trend of emotions of the people towards certain tweets. Twitter is a Battleground of tech companies, its Battleground of the world where people battle with hashtags and their words. In order to determine the sentiment of an overall population towards a particular tweet many companies and Organisation have started building up sentiment analysis teams which performs analysis on the sentiments of people on the social media. This is particularly helpful for the company in order to develop their future publicity campaign in order to make changes in the environment so that the productivity is increased and also their profit is increase which is the ultimate goal of each and every organisation.

VI. IMPLEMENTATION

We used python as our language to develop machine learning model. We have used naïve bayes classifier as our model. We also used flask framework, vue js framework and also the supporting libraries and dependencies in order to develop the project. For the EDA app, we have used python and streamlit to complete the development.

Talking about the phase-1, where we worked on the EDA app, we started with the development and completed the phase with 102 lines of code, which was further optimized to 52 lines.

Talking about the phase-2, where we tried to developed the twitter sentiment analysis dashboard to simplify the sentiment analysis dashboard, by providing the user to visualize the statistics of the tweets in an interactive and lucid manner, in such a manner that they can easily get the insights out of it.

VII. TOOLS AND TECHNOLOGIES USED.

- Python (Version = 3.10.0)

Python is a high-level, general-purpose programming language. We used python to develop our machine learning model. [5] We made use of many libraries also.

- Node js

Node.js is an open-source, cross-platform, back-end JavaScript runtime environment that runs on the V8 engine and executes JavaScript code outside a web browser. [6] We used node js in our backend development.

- Vue js

Vue.js is an open-source model–view–viewmodel front end JavaScript framework for building user interfaces and single-page applications. [7] We used Vue js in our frontend development.

- VS Code

Visual Studio Code, also commonly referred to as VS Code, is a source-code editor made by Microsoft for Windows, Linux and macOS. Features include support for debugging, syntax highlighting, intelligent code completion, snippets, code refactoring, and embedded Git. [8] We used vs code as our primary IDE.

- Jupyter Notebook

Project Jupyter is a project and community whose goal is to "develop open-source software, open-standards, and services for interactive computing across dozens of programming languages". [9] We used jupyter notebook for developing our machine learning model.

- Nltk

The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing for English written in the Python programming language. [10]

- Venv

The module used to create and manage virtual environments is called venv [11]

- Bootstrap

Bootstrap is a free and open-source CSS framework directed at responsive, mobile-first front-end web development. It contains HTML, CSS and JavaScript-based design templates for typography, forms, buttons, navigation, and other interface components.[12]

VIII. FACTS OF PROJECT

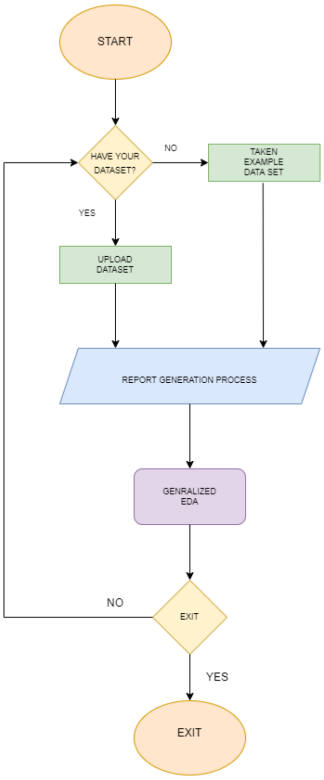
A. Limitations

- Gives general EDA not Specific EDA
- Dirty Data might affect the EDA
- Dashboard usage limited to twitter's scope
- Flask
- Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. We used flask as our framework for backend development.
- Incorrectly Targeted Sentiment.
- Inability to perform well in different domains, inadequate accuracy and performance.
- Insufficient labeled data, incapability to deal with complex sentences that require more than sentiment words and simple analyzing

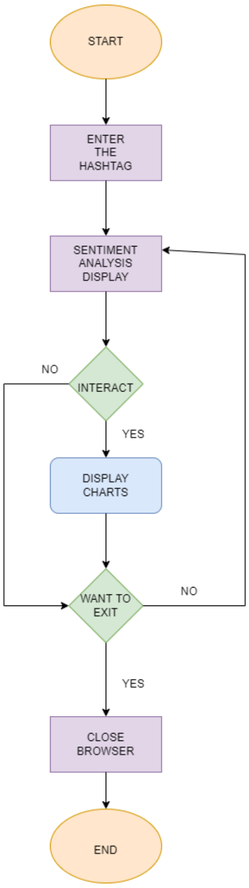
B. Expected Outcomes

- This project will save initial cost and time taken by data analyst to finalize the data set or to judge whether the data set is appropriate or not.
- This project will give a glimpse of data and it’s underlying impurities.
- The Project provides a detailed information about the data and their types
- The project provides correlation between the data from the data set.
- The project provides analysis dashboard for the tweets.
- The project provides a platform to interact with the sentiment analysis.

IX. FLOW CHART



EDA FLOW CHART



Sentiment app flowchart

X. CONCLUSION

Exploratory Data Analysis, or EDA, is an important step in any Data Analysis or Data Science project. EDA is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset. Sentiment analysis, also referred to as opinion mining, is an approach to natural language processing that identifies the emotional tone behind a body of text. This is a popular way for organizations to determine and categorize opinions about a product, service, or idea. It involves the use of data mining, machine learning (ML) and artificial intelligence (AI) to mine text for sentiment and subjective information. we have completed the project work using software engineering and system analysis and design approach. This project is completed with the primary functionalities. Due to lack of skilled knowledge and time constraints, the project cannot be fully completed so far. we have created an experience which performs the basic functionalities but yet there is a scope of improvement and advancement which can be taken care of in the near future.

ACKNOWLEDGMENT

We hereby avail this opportunity to express our gratitude to number of people who extended their valuable time, full support and cooperation in developing the project. Every work that one completes successfully stands on the constant encouragement, good will and support of the people around. I hereby avail this opportunity to express my gratitude to number of people who extended their valuable time, full support and cooperation in developing the project We would like to sincerely thank Dr. Amit Ganatra, Dr. Parth Goel for their expertise and guidance they have provided during our research. We would also like to sincerely appreciate the efforts of Prof. Krishna Patel, who since beginning is guiding us throughout the project and has patiently listened to our queries and offered the solutions to our problems. It is because of them, that we were prompted to do hard work, adopting new technologies. They altogether provided us favorable

environment, and without them it would not have been possible to achieve my goal.

REFERENCES

[1]https://www.analyticsvidhya.com/blog/2021/05/exploratory-data-analysis-eda-a-step-by-step-guide/

[2]https://towardsdatascience.com/exploratory-data-analysisdcb5e7189c4e#:~:text=Exploratory%20Data%20Analysis%20(EDA)%20refers,summary%20statistics%20and%20graphical%20representations.

[3]https://monkeylearn.com/sentiment-analysis/#:~:text=Sentiment%20analysis%20(or%20opinion%20mining, feedback%2C%20and%20understand%20customer%20needs.

[4] https://www.javatpoint.com/machine-learning-naive-bayes-classifier

[5] https://www.python.org/doc/essays/blurb/

[6] https://en.wikipedia.org/wiki/Node.js

[7] https://en.wikipedia.org/wiki/Vue.js

[8] https://en.wikipedia.org/wiki/Visual_Studio_Code

[9] https://en.wikipedia.org/wiki/Project_Jupyter

[10]https://en.wikipedia.org/wiki/Natural_Language_Toolkit

[11] https://docs.python.org/3/library/venv.html

[12] https://getbootstrap.com/

[13]https://www.journaldev.com/53190/exploratory-data-analysis-python#:~:text=Exploratory%20Data%20Analysis%20%E2%80%93%20EDA,or%20through%20some%20python%20functions