A Detail Report On Capstone Project: Intensity Analysis Using NLP And Python

Introduction:

The aim of this project is to develop an intelligent system to predict the intensity of emotions using Natural Language Processing (NLP). The model classifies text into three categories: happiness, angriness, and sadness. The project uses various NLP techniques, including text cleaning, feature transformation, and machine learning models..

The project will proceed through several key stages. It begins with data collection, where a wide variety of text data, each labeled with emotional intensity, will be gathered. Afterward, the data will undergo preprocessing, which includes cleaning, normalization, and transformation to ensure it is in the right format for further analysis. Feature engineering will then be carried out to extract meaningful features and identify the key factors that influence the emotional intensity in the text.

Next, we will select suitable machine learning models capable of classifying emotional intensity based on the processed data. Several algorithms will be tested and evaluated to determine which one performs best for this particular task. The chosen models will be trained on the cleaned and transformed data, aiming to improve their performance. Once the models are trained, their effectiveness will be assessed using relevant evaluation metrics. Hyperparameter tuning will be done to optimize the models and improve their prediction accuracy. Finally, the most successful model will be deployed in a live environment, where it will provide real-time predictions, ultimately contributing to enhanced customer satisfaction by enabling proactive improvements.

Creating an intelligent system that predicts emotion intensity in customer reviews can provide businesses with valuable insights into customer sentiment. This enables them to make informed, data-driven decisions that improve customer experiences and support overall business growth.

Data cleaning:

In this project, we utilized a combination of powerful tools and libraries, including NLTK, regex, and string, to preprocess and clean the collected textual data. NLTK (Natural Language Toolkit) was essential for performing key NLP tasks, such as tokenization, lemmatization, and removing stopwords. Regex (regular expressions) was crucial for tasks like pattern matching and manipulating the text, enabling us to efficiently remove punctuation, extract URLs, and handle special characters. By using these tools together, we were able to effectively clean, standardize, and transform the raw text into a format that was ready for training and analysis in machine learning models.

Analysing data:

In our project, we used word tokenization as a key part of the data preprocessing process. This technique involves breaking down the text into individual words or tokens, which allows us to analyze and process each word separately. We took advantage of NLTK's tokenization features to split the text into meaningful parts, considering punctuation, whitespace, and other linguistic elements. This approach helped us manage the data at a more detailed level, making it easier to perform further analysis and extract relevant features. Additionally, data visualization played a crucial role in exploring the frequency and distribution of words in the text reviews. By visualizing word frequencies through methods like bar charts, word clouds, and histograms, we were able to identify the most common words across the dataset. This exploration revealed key themes, sentiments, and topics within the reviews, offering valuable insights for later stages of analysis. Visualizing word frequencies also helped us uncover patterns or anomalies that could affect the emotional intensity in the text. In summary, word tokenization and data visualization were essential tools in our NLP workflow, providing a deeper understanding of the data and guiding our subsequent modeling decisions.

Model Selection and Training :

To begin with, we used Count Vectorizer to transform the text data into a matrix of token counts, capturing the frequency of each word in the corpus. This method helped us track how often individual words appeared without factoring in their relative importance or rarity. Additionally, we applied the TF-IDF technique, which takes into account not only word frequency but also the importance of each word across the entire dataset by reducing the weight of common words and increasing the weight of rarer ones. Next, we trained three different classification models using both the Count Vectorizer and TF-IDF representations of the data. The first model, Multinomial Naive Bayes, is a probabilistic approach often used for text classification, especially when dealing with sparse data like text. We also used Logistic Regression, a linear model that predicts probabilities through a logistic function and is particularly effective for binary classification tasks.

Finally, we applied Linear SVM, a robust algorithm for text classification, known for its ability to handle high-dimensional data and find the best hyperplane to separate different classes.

By testing various combinations of text representation methods and classification algorithms, our goal was to find the best approach for predicting emotion intensity in text reviews. We carefully evaluated and compared the performance of the models using key metrics like accuracy, precision, recall, and F1-score, with the aim of identifying the optimal combination that would deliver accurate and dependable predictions in practical, real-world situations.

Testing Analysis :

The **ML Algorithms** used for prediction are listed as follows:

Building models using different classifiers (Count vectorizer):

Model 1: **Multinomial Naive Bayes Classifier** - Accuracy **70%**

Model 2: **Linear SVM** - Accuracy **82%**

Model 3: **Logistic Regression** - Accuracy **90%**

Building models using different classifiers (TF-IDF vectorizer):

Model 1: **Multinomial Naive Bayes Classifier** - Accuracy **48%**

Model 2: **Linear SVM** - Accuracy **54%**

Model 3: **Logistic Regression** - Accuracy **48%**

During the testing phase of our project, we assessed the performance of various machine learning models and found that Logistic Regression, paired with Count Vectorization, achieved the highest accuracy of 90%. This result suggests that the logistic regression model, when trained on word counts, was particularly effective at predicting the intensity of emotions expressed in the reviews. The accuracy metric, which represents the proportion of correctly classified instances, proved to be a valuable measure of model performance. An accuracy of 86% indicates that the logistic regression model with count vectorization successfully identified key patterns in the text data, leading to accurate emotion intensity predictions. This outcome highlights the strong performance of logistic regression for text classification tasks and the value of count vectorization in transforming text data for analysis. By utilizing these techniques, our model shows great potential for accurately predicting emotion intensity in text reviews, providing businesses with actionable insights into customer sentiment and helping improve customer satisfaction.

Prediction of intensity from sentences

	input_text	predicted_emotion
0	I am so angry at you!!!!!	Anger
1	You ve hit a new low with a danger of blm fascist slogan please stop it before too late stop	Anger
2	I love my doggg	Happy
3	I think i’m gonna be sick :'â€‘(	Happy
4	I hate you so much	Happy
5	@TheTombert i was watching Harpers Island, lol... there was no vodka involved	Happy
6	sometimes i wish things could go back to the way they were the beginning of last summer	Sad
7	it's your 18th birthday finally!!! yippeeeee	Happy
8	oh no he is hospitalised!!!	Anger

Future Prospects :

The successful completion of the project paves the way for several exciting future opportunities and areas for growth:

Fine-tuning and Optimization: While an accuracy of 86% is impressive, there is always potential for further improvement. Future work could focus on refining model parameters, exploring alternative feature engineering methods, and testing advanced algorithms to boost predictive accuracy even more.

Integration with Feedback Systems: Incorporating the developed model into existing feedback systems or CRM platforms could offer businesses real-time insights into customer sentiment. This would allow for quicker issue resolution, more strategic decision-making, and personalized responses to customers based on their emotional feedback.

Sentiment Analysis in Multimodal Data: Expanding the project to process multimodal data—such as text, images, and audio—could offer a richer and more holistic view of customer sentiment. By applying advanced multimodal sentiment analysis techniques, businesses can tap into diverse forms of customer feedback, gaining deeper insights that support more informed decision-making.

Personalization and Recommendation Systems: By harnessing the model's predictive power, businesses can create personalized recommendation systems that cater to individual customer preferences and emotional responses. Analyzing both past behavior and real-time interactions, these systems could provide tailored product recommendations, content suggestions, and targeted marketing, ultimately improving customer engagement and satisfaction.

Deployment in Various Industries: The framework and methodology developed in this project can be adapted for use across numerous industries, including e-commerce, hospitality, healthcare, and finance. Customizing the model for specific sectors and use cases could present valuable opportunities to enhance customer experiences, streamline business processes, and gain a competitive edge in the market.

Ethical Considerations and Bias Mitigation: As with any AI system, it’s crucial to prioritize ethical considerations and reduce potential biases. Future work could focus on creating

strategies to ensure the model's predictions are fair, transparent, and accountable, helping to build trust and confidence among users and stakeholders.

In conclusion, the future of the project looks very promising, offering numerous opportunities to push the boundaries of customer sentiment analysis, improve decision-making processes, and ultimately enhance customer experiences across a wide range of industries and sectors.