Group 1

PREPARED BY:

Jose German, Mohammed Araf, Abdela Mossa, Luiz Tavares

CSML-1000

York University

Group Project

PREPARED FOR:

Hashmat Rohian

Instructor

CSML-1000

York University

Table of Contents

Executive Summary

In today's digital world, social media is crucial for businesses to understand customer feedback. Our airline client wants to use social media analytics to make customers happier. We suggest using a special computer program to analyze tweets about the airline. This will help the airline respond better to customers and make their flights better.

Our idea can really improve customer happiness in the airline industry. By understanding tweets better, the airline can fix problems faster and make customers happier. This plan not only helps right away but also sets the airline up for success in the future. We believe this will make customers more loyal, improve the airline's reputation, and make it stand out from the competition.

As we go ahead with this plan, we promise to deliver real results for our client and their customers. Together, we're on a mission to make customers happier and keep improving in the airline industry.

Introduction

In today's fast-paced digital age, social media platforms have become invaluable sources of real-time feedback and insights for businesses across various industries. Recognizing the significance of this trend, our client, an esteemed airline company, seeks to leverage the power of social media analytics to enhance customer satisfaction and improve overall service quality.

This document outlines the business problem faced by our client and proposes a solution centered around the development and implementation of a semi-supervised machine learning model. Specifically, our client aims to analyze Twitter messages pertaining to their airline services, classify them into positive and negative sentiments, and extract actionable insights to drive strategic decision-making.

By harnessing the capabilities of machine learning algorithms, our client endeavors to gain deeper insights into customer sentiments expressed on Twitter, thereby enabling them to proactively address issues, optimize service delivery, and ultimately elevate the overall customer experience.

This document presents a comprehensive overview of the business problem, the proposed solution, key objectives, and the anticipated benefits of implementing the suggested approach. Additionally, it outlines the scope, methodology, and timeline for the project, laying the foundation for a successful collaboration between our client and our team.

Through this initiative, our client seeks to reaffirm their commitment to customer-centricity, innovation, and continuous improvement, underscoring their dedication to delivering exceptional service and fostering lasting customer relationships in the dynamic aviation industry.

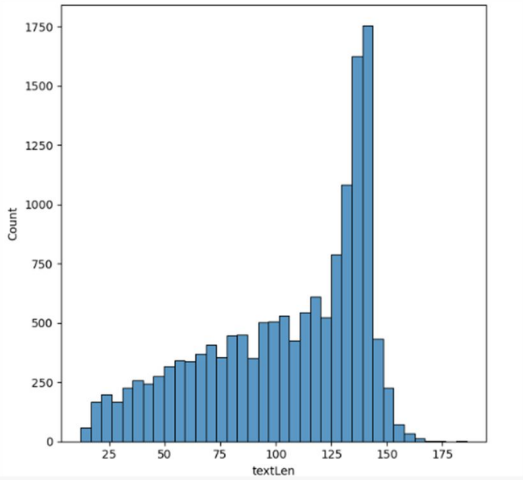
Exploratory Data Analysis:

Data Elements:

We found the original dataset has 14640 records (rows) and 15 columns.
We have decided to use only a handful of columns for the Sentiment analysis. The list of columns used are as follows:

1. 'tweet_id',
2. 'airline_sentiment',
3. 'text'

Calculated Fields:
We have calculated the length of each tweet to understand the average length or size of the texts within the tweets to be analyzed.We found the smallest tweet in this dataset is about 12 characters long and the largest tweet is about 186 characters long. The average length of the tweet is 104(approx) characters.

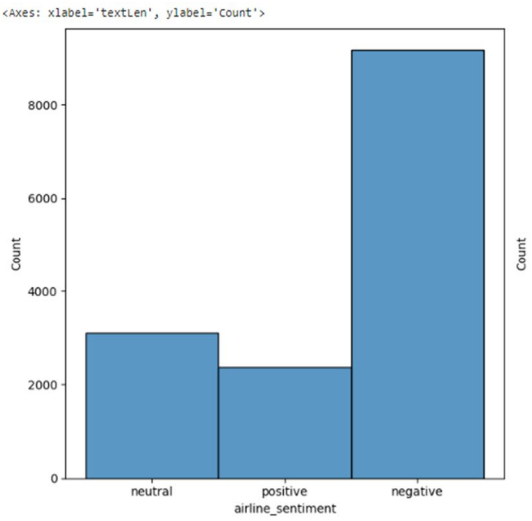


Checking Imbalancing:
We have calculated the number of tweets for each of the sentiment categories (“Positive”, “Negative” and “Neutral”) and found there are much more negative tweets (~63%) present in the dataset compared to other two categories (Neutral is 21% and Positive is 16%) . Hence, it is recommended to apply Oversampling techniques to balance the test data to avoid any bias in the model to be developed.

-

```
dfTweetSubset.groupby('airline_sentiment').count()/14640
```

airline_sentiment	tweet_id	text	textLen
negative	0.626913	0.626913	0.626913
neutral	0.211680	0.211680	0.211680
positive	0.161407	0.161407	0.161407



Duplicate Removal:

We found there are some (127) duplicate tweets present within the dataset. We have removed such tweets as part of data preprocessing.

Word Cloud: We have used the “Word Cloud” library to identify the most frequently used words in Positive, Negative and Neutral tweet categories.

Text Cleaning Steps: Following cleaning steps were performed:

(a) Remove username, http or https:// and any non alphanumeric characters
(b) Transforming the text into lower case
(c) Remove the punctuations
(d) Remove unicodes
(e) Removing the stop words excluding "not"

Modeling and Training

The development of our airline sentiment analysis classification model was carried out using the conventional approach of natural language processing (NLP) text classification in Python. Subsequently, we built and investigated ten distinct machine learning models utilizing the Bag of Words Term Frequency (BoW-TF) embedded text data. The table below shows the results of various classification algorithms both before and after the implementation of hyperparameter tuning, following the utilization of BoW-TF as the embedding technique.

Our findings show varying performance metrics across algorithms, substantial enhancements in performance in almost all of the evaluation metrics , highlighting the significance of tuning in ML model development for sentiment analysis tasks. As shown in the table, it is the Logistic Regression model that outperformed others in terms of AUC and Accuracy metrics, while Random Forest excelled in Recall. Additionally, utilizing SMOTE sampling effectively balanced the initially unbalanced dataset, resulting in enhanced performance across various evaluation metrics. Thus, addressing class imbalance also proved crucial in achieving reliable and accurate sentiment analysis predictions.

Classification Algorithms	Evaluation Metrics							
	AUC		ACC		PREC		RECALL	
Random Forest	82.1	82.24	70.36	70.67	56.64	56.91	**86.66**	**87.15**
Ada Boost	83.14	83.31	74.73	75.05	62.82	63.41	78.58	77.85
Gradient Boost	82.71	83.67	74.23	75.68	62.27	64.42	77.97	77.36
Naive Bayes	78.71	78.84	70.99	71.27	57.83	58.13	81.4	81.4
Logistic Regression	**83.4**	**83.4**	**75.5**	75.59	64.08	**64.45**	77.72	77.97
SVM	82.52	77.25	75.77	73.63	**64.86**	63.58	76.13	68.18
KNN	76.1	77.78	64.16	71.04	51.12	59.54	83.48	69.16
Decision Tree	72.67	75.5	69.95	71.04	57.12	58.63	77.11	75.28
LDA	82.79	82.86	74.18	73.86	61.86	61.54	79.8	79.31

App Development

Application development/deployment: https://huggingface.co/

Our team decided to utilize Hugging Face which is a machine learning (ML) and data science platform that is community based. It facilitates building, training, and deploying machine learning models. Hugging Face provides the necessary infrastructure to demonstrate, run and deploy ML and AI in live applications.

Hugging Face is known for its Python based Transformers library, which simplifies the process of training ML models. This library allows developers to take advantage of hosted pre-trained base models in their workflows and create ML pipelines. Although our team carried out the traditional methods of processing data, training and testing our model, for the application our team took a slightly different approach when building our airline twitter sentiment analysis model. This method allowed the team to use a base model which is then fine tuned using our dataset. Below I will show some of the key code that is executed to train our model.

These types of projects required GPU to speed up training:

```
import torch
torch.cuda.is_available()
```

Here is the code used to load our hosted dataset:

```
# Load data
from datasets import load_dataset
tweet = load_dataset("jos-ger/tweet-sentiment-airlines")
```

Define DistilBERT as our tokenizer:

```
from transformers import AutoTokenizer
tokenizer = AutoTokenizer.from_pretrained("distilbert-base-uncased")
```

Define DistilBERT as our base model:

```
from transformers import AutoModelForSequenceClassification
model = AutoModelForSequenceClassification.from_pretrained("distilbert-base-uncased", num_labels=2)
```

Define a new trainer with all objects we constructed. The repo_name is the name we are giving to the model to save into our hosted environment:

```
from transformers import TrainingArguments, Trainer

repo_name = "finetuning-sentiment-model-tweet-sentiment-3000-samples"

training_args = TrainingArguments(
    output_dir=repo_name,
    learning_rate=2e-5,
    per_device_train_batch_size=16,
    per_device_eval_batch_size=16,
    num_train_epochs=2,
    weight_decay=0.01,
    save_strategy="epoch",
    push_to_hub=True,
)

trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=tokenized_train,
    eval_dataset=tokenized_test,
    tokenizer=tokenizer,
    data_collator=data_collator,
    compute_metrics=compute_metrics,
)
```

Finally we train our model:

```
trainer.train()
```

The dataset URL hosted on Hugging Face:

https://huggingface.co/datasets/jos-ger/tweet-sentiment-airlines

The application URL hosted on Hugging Face:

https://huggingface.co/spaces/jos-ger/sentimentanalysis

Conclusion

In conclusion, our proposed semi-supervised machine learning solution holds tremendous potential to revolutionize customer satisfaction in the aviation industry. By effectively analyzing Twitter messages and categorizing them into positive and negative sentiments, our solution empowers our airline client to proactively engage with their customers, address concerns promptly, and enhance overall travel experiences.

Through this collaborative endeavor, we have not only addressed the immediate need for improved customer service but also paved the way for long-term success in the dynamic aviation landscape. With our solution in place, our client can expect increased customer loyalty, improved brand reputation, and a competitive edge in the market.

As we move forward with implementation, we remain committed to delivering tangible results and driving positive outcomes for our client and their valued customers. Together, we embark on a journey toward elevated customer satisfaction and continued innovation in the aviation industry.