Sentiment Analysis on Customer Support dataset

HUSSAIN SHAHBAZ KHAWAJA[1]

[1] *Submitted to solvpath® as test evaluation for the Data Science position.*
[*] *hskhawaja@live.com*

Compiled May 16, 2023

Customer support tweets often contain valuable insights into customer satisfaction. In this report, we describe our sentiment analysis process on a dataset of customer support tweets. Our analysis includes details on data, preprocessing, model architecture, and visualized results.

1. INTRODUCTION

This report presents the results of a sentiment analysis project conducted on a customer support dataset.The dataset includes various columns, including the tweet text, tweet ID, date, and user information.

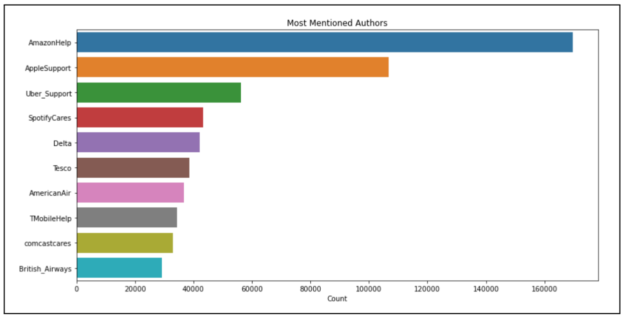
The project involved several tasks, including data ingestion, preprocessing, visualization, model selection, and analysis of sentiment trends over time. The insights gained from this analysis can help businesses better understand customer sentiment, improve customer support strategies, and enhance overall customer satisfaction.

2. EXPLORATORY DATA ANALYSIS

In this section, we perform exploratory data analysis on the customer support dataset to gain insights into the data. We start by identifying the top mentioned authors in the dataset and then analyze the language used by customers using the langdetect package.

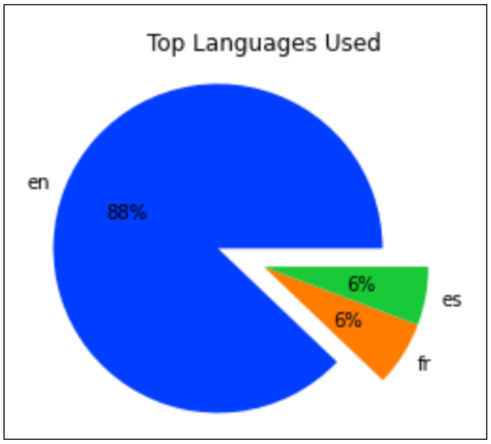
A. Top Mentioned Authors

We identified the top mentioned authors by counting the frequency of user mentions in the tweet data. By aggregating this information, we obtained a list of the most frequently mentioned authors, indicating their significance in the customer support interactions.



B. Most Used Language

Next, we created a subset of the data that only includes tweets mentioning the top author i.e. AmazonHelp. This subset allows us to focus on the interactions related to this particular author and delve deeper into the language used by customers in these interactions.



3. SENTIMENT ANALYSIS

In this section, we conduct sentiment analysis on the customer support dataset. From this point onwards, to speedup the process, we only focus on the top language i.e. English tweets. Since the dataset is unlabelled, we utilize a pre-trained RoBERTa-base model, which has been trained on a large corpus of approximately 124 million tweets collected from January 2018 to December 2021. This pre-trained model has been further fine-tuned specifically for sentiment analysis using the TweetEval benchmark.

A. Model Architecture

The RoBERTa-base model is based on the Transformer architecture, which is a type of deep learning model that has achieved remarkable success in various natural language processing tasks. The Transformer architecture is known for its ability to capture contextual dependencies and effectively model long-range dependencies in text data.

The RoBERTa-base model consists of a stack of transformer layers, including self-attention mechanisms, which allow the model to attend to different parts of the input sequence while encoding contextual information. The model also incorporates positional encoding to account for the order of words in the input text.

Following table presents sentiment labels of a few random samples:

Text	Sentiment
@AmazonHelp It's not working I told u.what is ur problem I share screenshot of my order chk it and revert	negative
@AmazonHelp The delivery date given was today 10th November	neutral
@AmazonHelp Can you pls help me on this	neutral
@AmazonHelp Seriously, I don't have time. I worked for whole day and now it's time for some sleep.	negative
@AmazonHelp Thank you I'll check the link right now	positive

B. Evaluation Mechanism

Since the dataset does not include labeled sentiment information, evaluating the assigned sentiment labels without any test data can be challenging. In such cases, here is an alternative evaluation mechanism for assessing the sentiment analysis task without relying on test data or manual annotations:

1. Utilize sentiment lexicons or dictionaries, which contain pre-defined sentiment polarity (positive, negative, or neutral) for words.

2. Assign sentiment labels to the words in the dataset based on the sentiment lexicon.

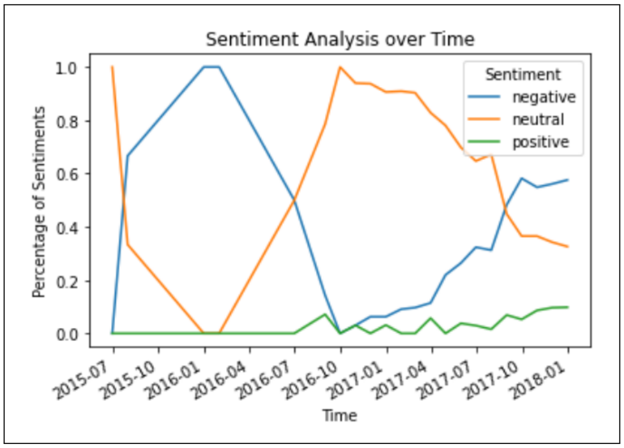
3. Calculate the sentiment distribution in the dataset by counting the number of positive, negative, and neutral words.

4. Analyze the sentiment distribution and compare it with expectations based on domain knowledge or general assumptions. This comparison can provide insights into the sentiment capture capabilities of the model.

This evaluation mechanism provide a simpler approach to assess the sentiment analysis task without requiring additional labeled test data.

4. BONUS: TEMPORAL ANALYSIS

In this section, we analyze the sentiments of the text over time in the customer support dataset to identify any trends or patterns. By examining the temporal aspect of sentiment, we can gain insights into how customer sentiments have evolved or fluctuated over different time periods.



5. CONCLUSION

In this project, we performed sentiment analysis on a customer support dataset consisting of over 2 million tweets.

During the exploratory data analysis, we identified the top mentioned authors in the dataset. We then focused on the tweets from the top author to analyze the language used by customers. By leveraging the langdetect package, we determined the predominant languages present in the dataset, providing valuable insights for further analysis and understanding customer demographics.

For sentiment analysis, we employed a pre-trained RoBERTa-based model because of its effectiveness in capturing the sentiment nuances present in Twitter data.

We also conducted an analysis of the sentiments of text over time. By grouping tweets into time intervals and calculating sentiment statistics within each interval, we identified trends and patterns in customer sentiments.

The findings from this sentiment analysis can guide businesses in making data-driven decisions to enhance customer satisfaction, tailor support approaches, and address issues promptly. The insights gained from analyzing customer sentiment over time can help businesses adapt their strategies, identify opportunities for targeted interventions, and ultimately foster stronger customer relationships.

ACKNOWLEDGMENTS

Thanks to HuggingFace for providing high-performing open source models.