Sentiment Analysis Through Natural Language Processing

FYP 24090 Detailed Project Plan

Student	Supervisor
YUANTO Salvatore Patrick	Chan Tsz Hong Hubert
KASMAN Winson Sutanto	

Table of Content

1. Background

In today's digital age, understanding the nuances of consumer sentiment is a crucial part of business strategy. The rise of social media, online reviews, and forums has created an explosion of unstructured data that holds valuable insights into customer opinions, preferences, and behaviors. As of July 2024, there are a staggering total of 5.45 billion internet users worldwide, amounting to 67.1% of the global population [2]. Furthermore, it is estimated that around 5% to 10% of consumers leave online reviews [3]. This makes the internet a valuable source of information for businesses to learn more about customer sentiment.

Given the vast amount of data available on the internet, it is not feasible to analyze their overall sentiment manually. Sentiment Analysis, a subfield of Natural Language Processing (NLP) also known as opinion mining, offers a solution to this problem. Through leveraging machine learning algorithms and lexical knowledge, overall sentiment from data can be extracted automatically.

The importance of sentiment analysis in a business setting cannot be overlooked. By analyzing customer sentiment, businesses can identify areas for improvement, measure brand reputation, inform product development, enhance customer experience [1] and gain a competitive edge._For instance, sentiment analysis can help pinpoint specific products, services, or features that require improvement to meet customer expectations. Additionally, tracking changes in public perception and sentiment towards the brand over time can provide valuable insights into the effectiveness of marketing strategies and customer engagement initiatives. For this reason, there is a demand for sentiment analysis tools.

2. Objective

This project aims to develop a robust sentiment analysis model that can help businesses tap into the collective voice of their customers. This project involves creating a user-friendly dashboard that provides personalized insights into customer sentiment, enabling companies to make data-driven decisions to drive growth and improvement. By leveraging modern NLP techniques and machine learning algorithms, providing a scalable and accurate sentiment analysis solution that meets the needs of businesses across various industries.

Ultimately, this project seeks to bridge the gap between customer feedback and business strategy, empowering companies to respond appropriately to customer concerns, improve overall customer satisfaction, and stay ahead of the competition.

3. Methodology

3.1 Data Collection

The data for this research will be sourced from various online platforms, including social media, search engines, articles, and web pages. The data collection process involved the use of Application Programming Interfaces (APIs) and web scraping techniques.

Specifically, the social media platforms to be used in this study are X and Reddit. X's API will be used to collect posts, while Reddit's API will be used to collect comments and posts from relevant subreddits. For articles or forums, search engine results obtained from Custom Search API from Google will be used. Online articles and web pages were crawled using web scraping techniques to collect relevant data.

The data collection process will be limited to the past 10 years, and the data collected contains some form of human sentiment. The data will be stored in a .csv or .xlsx file.

3.2 Data Preprocessing

The preprocessing of data is required to ensure it is in a suitable format for analysis. This process includes tokenization, stopword removal, stemming or lemmatization, removing special characters and punctuation, and handling missing values.

Tokenization involves breaking down text into individual words or tokens. Stopword removal involves removing common words like "the," "and," etc. that do not add much value to the sentiment analysis. Stemming or lemmatization involves reducing words to their base form as a form of dimensionality reduction. Removing special characters and punctuation marks is also required to improve the accuracy of the sentiment analysis model. Handling missing values involves replacing missing values with suitable alternatives to prevent bias in the analysis.

3.3 Sentiment Analysis Model

The preprocessed data will be analyzed using various NLP techniques, including text classification, sentiment intensity analysis, and aspect-based sentiment analysis.

Text classification involves classifying text as positive, negative, or neutral using transformer models such as BERT, GPT-3, etc. Sentiment intensity analysis involves analyzing the intensity of sentiment in text using techniques like sentiment lexicons and rule-based approaches. Aspect-based sentiment analysis involves identifying specific aspects or features of a product or service that are being praised or criticized.

This project will test large language models (LLMs) like GPT-3.5, which has an accuracy of approximately 88%[5], as well as explore other models such as BERT, RoBERTa, and GPT-4 for sentiment analysis. The final model will be selected based on a price-to-performance metric, taking into account project budget constraints.

3.4 Evaluation of Sentiment Analysis model

The performance of the sentiment analysis model will be evaluated using metrics like accuracy, precision, recall, and F1-score.

Accuracy involved measuring the proportion of correctly classified instances. Precision involved measuring the proportion of true positives among all positive predictions. Recall involved measuring the proportion of true positives among all actual positive instances. F1-score involves measuring the harmonic mean of precision and recall.

3.5 Dashboard Development

The deliverables of this project include a user-friendly dashboard which allows users to input keywords and visualize the statistics and results of the sentiment analysis, providing a comprehensive overview of customer opinions and sentiment towards the keywords inputted which could be a particular product, service, or brand. Data visualization involved visualizing sentiment analysis results using charts, graphs, and heatmaps. Filtering and sorting involved allowing users to filter and sort data based on specific criteria like date, product, or sentiment.

The interactive dashboard will be built using React.JS and includes features such as data visualization, filtering and sorting. Users can also add their own products to get a personalized analysis report. The backend server will be built using the Django framework. The database management system used for this project is MongoDB.

4. Schedules and Milestones

Date	**Objective**	**Status**
Oct 1, 2024	Deliverable 1: • Detailed project plan • Set up project web page	completed
Nov 1, 2024	Data Collection	In progress
Nov 15, 2024	Data Preprocessing	Not started yet
Jan 13-17, 2025	First presentation	Not started yet
Jan 26, 2025	Deliverables 2 • Preliminary implementation • Detailed interim report	Not started yet
Mar 21, 2025	Dashboard Development	Not started yet
Apr 21, 2025	Deliverables 3 • Finalized tested implementation • Final report	Not started yet
Apr 22-26, 2025	Final presentation	Not started yet
Apr 30, 2025	Project exhibition • 3-min video	Not started yet

5. References

[1] Han T, Liu C, Yang W, Jiang D (2019) A novel adversarial learning framework in deep convolutional neural network for intelligent diagnosis of mechanical faults. Knowl-Based Syst 165:474–487

[2] https://www.statista.com/statistics/617136/digital-population-worldwide/

[3] https://luisazhou.com/blog/online-review-statistics/#

[4] Shaha Al-Otaibi, Allulo Alnassar, Asma Alshahrani, Amany Al-Mubarak, Sara Albugami, Nada Almutiri and Aisha Albugami, “Customer Satisfaction Measurement using Sentiment Analysis” International Journal of Advanced Computer Science and Applications(ijacsa), 9(2), 2018. http://dx.doi.org/10.14569/IJACSA.2018.090216

[5] Krugmann, J.O., Hartmann, J. Sentiment Analysis in the Age of Generative AI. *Cust. Need. and Solut.* 11, 3 (2024). https://doi.org/10.1007/s40547-024-00143-4