Technical Report: Final Project DS 5110: Introduction to Data Management and Processing

Team Members: Nicholas Nehemia,Rebecca Bronfeld
Khoury College of Computer Sciences
Data Science Program
nehemia.n@northeastern.edu, Bronfeld.r@northeastern.edu

November 25, 2024

Contents

1 Introduction

Provide a brief introduction to the project, including the background, objectives, and scope. In today's age, many individuals express the opinions of current events using various social media platforms. In our final project assignment, we will utilize social media, specifically Reddit data, to gather reactions of the iOS 17 and 18 upgrades, and conduct a sentiment and trend analysis based on sentiments found.

2 Literature Review

Summarize the existing research relevant to the project. Discuss the methodologies, findings, and gaps in the literature.
Remove symbols : https://saturncloud.io/blog/how-to-remove-special-characters-in-pandas-dataframe/: :text=Use

Source for how to scrape using PRAW:
`https://www.geeksforgeeks.org/scraping-reddit-using-python/`
PRAW Documentation : `https://praw.readthedocs.io/en/stable/`
Source for obtaining word counts : `https://www.earthdatascience.org/courses/use-data-open-source-python/intro-to-apis/calculate-tweet-word-frequencies-in-python/`

3 Methodology

Describe the methods and techniques used in the project. Include details about data collection, preprocessing, and analysis.
In this project , we will gather data using the PRAW python package, to access the reddit API to obtain comments, and then we will pass the data to the transformers package to label the classification of the data.

3.1 Data Collection

Explain how the data was collected, including the sources and tools used.
In order to collect the data, we will manually find the reddit post we would like to extract comments from , and we will add to a .CSV stored in our system. From there , the application will read into the CSV and extract the data for all links in the CSV.

3.2 Data Preprocessing

Describe the steps taken to clean and preprocess the data.
Our dataset contains 2 level of comments , the main comment and the 1st level replies to the 1st level comment. We only take the first reply higher level replies might go off topic. The notebook script will then find all parent comments ,and join the child comments , so we can have the parent and child comments in one row for easier identification and visualization purposes.
To generate the wordcloud , we also need to have a wordcount table. First , we will clean the data by removing the stopwords, and doing tokenization. We will then do a wordcount grouped by iOS version and classification label.

3.3 Analysis Techniques

Detail the analytical techniques and models used in the project. In this project , we use the transformers package to label the sentiment of all the comments.

4 Results

Present the results of the analysis. Use tables, figures, and charts to support the findings. Results: primarily negative reactions, but overall feeling is found to be indifferent after the initial review. Will display graphs here - bubble graph, heat map, and comparison bar chart. All screenshots from Tableau Dashboard.

5 Discussion

Interpret the results and discuss their implications. Compare the findings with the literature review and explain any discrepancies. (RB)

6 Conclusion

Summarize the key findings of the project. Discuss the limitations and suggest areas for future research.

7 References

References

A Appendix A: Code

Include any relevant code used in the project. For example:

```
def cleanCount(df,ios,sentiment):
    stop = stopwords.words('english')
    cleaned_stopwords=df[(df["IOS"]==17 )& (df["classification_result"
   ]==sentiment)]["unigrams"].apply(lambda x: [item for item in x if
   item not in stop])
    all_words_nsw = list(itertools.chain(*cleaned_stopwords))
    counts_nsw = collections.Counter(all_words_nsw)
    final_counts_nsw=counts_nsw.most_common()
    dfCounts=pd.DataFrame(final_counts_nsw,columns=["words","count"])
    dfCounts["ios"]=ios
    dfCounts["Sentiment"]=sentiment
    return dfCounts
```

Listing 1: Function to Remove stop words and generate word count

B Appendix B: Additional Figures

Include any additional figures or tables that support the analysis.