App Dev (BIG DATA) Hackathon

Niharika
NIT Durgapur
Email: niharikacpl@gmail.com

Contents

Introduction

This project is built using Koalas which is pandas API on Apache Spark. Pandas is a very user-friendly language and has emerged as the most widely used tool for data scientists but it is limited to small datasets. Therefore we need PySpark for handling large datasets.

PySpark is a great language to work with big data however converting code from pandas to PySpark is quite confusing and difficult. Koalas serve the purpose to overcome this difficulty.

Koalas has APIs similar to pandas which is built on top of PySpark which allows the users to take the benefits of PySpark without explicitly learning any new framework.

Case Description

We are trying to solve the problem of social bookmarking by capturing customer interests and targetting them for driving business.

ABC Bank is a FinTech bank that aims to enhance the customer experience and would like to use/analyse the public data available on various social bookmarking sites (like twitter, reddit etc.) for this purpose. Som of the important datasets they would like to analyse to find out how user sentiment is trending on social media are customer comments about (not just limited to!)

- Their payment experience using credit cards e.g. were they able to make the payment using the credit card, overall availability of payment services.
- Mobile based payment solutions across various platforms e.g. What do their customers think about these solutions, the ease/speed of transactions, success rate.
- Special offers/discounts, cashbacks, loyalty points on the credit cards

Solution Approach

Sentiment Analysis

We have used NLTK for sentiment analysis. One can also create his/her own sentiment analysis ML model using Spacy or deep learning.

Finding most common words

Finding most common words allows us to create a relationship with the sentiments and word co-occurance.

Understanding @askamex vs. customers tweets

Classifying tweets based on user names helps us understand the difference between sentiment of tweet.

Visualization of final results

Visualization is the best tool to understand the final outcome easily.

Technology Used to create the project

We have used the following technologies to make the project:

1. **Twint** : To extract data from twitter in large scale
2. **PySpark** : For big data processing of data and data analysis
3. **Koalas** : Its a wrapper on top of pyspark allowing pandas syntax to work with pyspark
4. **Matplotlib and Seaborn** : We used matplotllib and seaborn for plotting plots
5. **Jupyter notebook** : it allows interactive visualization and programming for working on data science projects
6. **Google Colab** : It is provided by Google for working with big datasets allowing high and free RAM and GPU
7. **NLTK** : It is a python package for natural language processing and sentiment analysis

Most of the packages are made using python. The complete data analysis is done on around 0.5 million tweets that was scraped from twitter.

Data Mining

The data has been gathered using **Twint** which is a Twitter scraping tool written in Python that allows for scraping Tweets from Twitter profiles without using Twitter's API.

All the tweets have been gathered from @A**skAmex** which is **American Express** twitter team for identifying customers problems. We targetted only a specific company for our tweets as we did not wanted to bias our data by using another company's data for analysis. We also targetted specifc keywords like payment, service, offers, etc. to create categories for data analysis.

Note : Similar can be implemented for other credit companies also. We have only used tweets for our project, similar can be re-implemented for Reddit comments as well.

The twint project : https://github.com/twintproject/twint

How to run twint to scrape tweets and save to file

>>> twint -s "@askamex" --since "2014" -o file.json --json

Tweet Pre-processing

Text pre-processing is an important part for data analysis when it comes to sentiment analysis.

We defined our tweet pre-processing in the following steps:

1. removing *@mentions* and urls from tweets
2. assigning sentiments to each tweets using NLTK
3. removing most occured words (stop words)
4. adding category tags to specially targetting words to scrape tweet from twitter

And finally we have kept only the most important words for further analysis.

Sentiment analysis & Word Vocabulary

Sentiment analysis is a good tool to summarize the overall sentiment of the customers. Then the company can target the customers having negative sentiment/ feedback to improve their customer experience.

Word vocabulary can help understanding the trend of sentiments and different kinds of categories of feedback the customers have and which is the most basic kind of occuring requests/ problems the customer faces. The company can then automate that specific problem and focus on the other more priority issues.

Most of the tweets were found to be about American Express cards. Other than those, we were able to see quite a good difference between words linked to Positive and Negative sentiments.

We divided the occurance of words as per Overall , Positive and Negative sentiment and ranked them on the basis of high frequency in tweets.

Most Occured Words

The word frequency and most occured words were found to be changing with respect to change in sentiment in Overall tweets.

We can see that words like 'please' have low ranks (high occurace in negative tweets) for negative tweets.

word	count	Rank positive	Rank negative
card	26926	1	1
please	6305	3	27
information	4554	5	19
product	2796	14	99
account	2780	22	5
thanks	2569	15	194
assist	2098	23	109
pls	1687	34	12

Understanding word occurace and sentiment changes

We also noticed some specifc credict card names in the tweets that mentions that the customers of those cards have very positive/ negative feedback for that card. This helps with product related sentiment understanding.

We created two different vocabularies.

1. Overall/ positive/ negative vocabularies for all tweet frequencies
2. Overall/ positive/ negative vocabulary of only customer tweets to understand customer feedback (tweets by @askamex were filtered and removed)

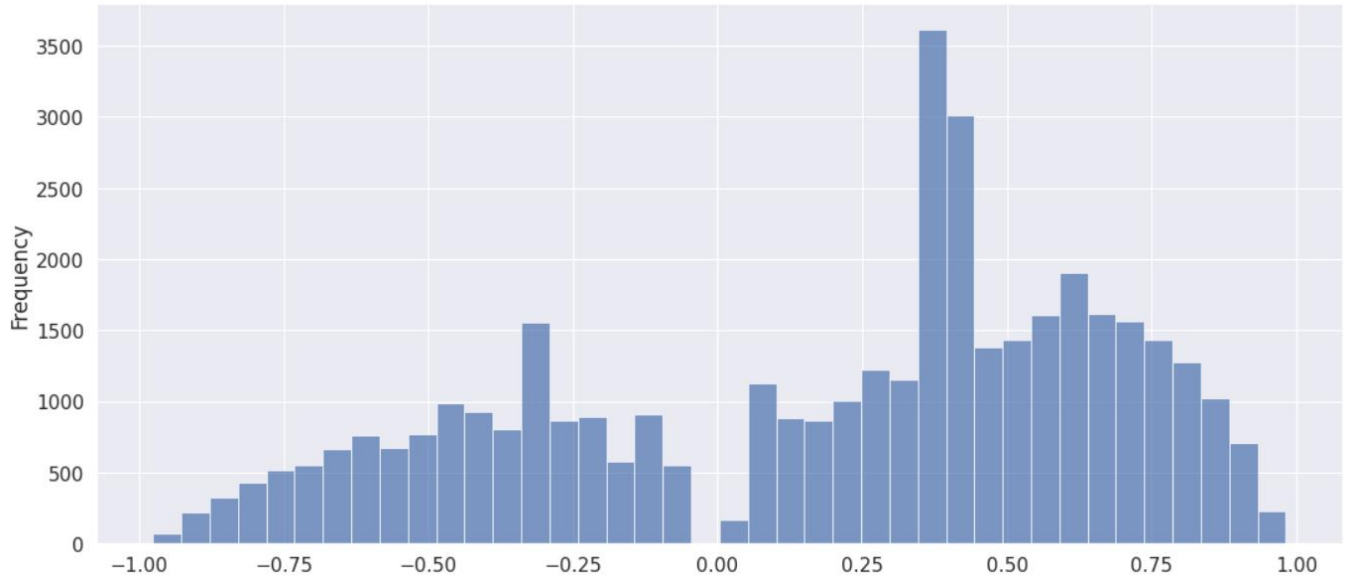
The customer specific tweets vocabulary showed us the trend of sentiment in word ranking for postive/ negative sentiment tweets. For detailed analysis refer to Appendix at the end of this presentation.

Note : For complete vocabulary refer to **frequency.xlsx** excel sheet or the link below to download the file.

https://drive.google.com/drive/folders/11175RZ3wraUAxJq4Jb6MLojvceL4RrY8?usp=sharing

Overall sentiment

The overall sentiment analysis showed that most of the tweets were positive sentiment driven. Below is the histogram to show the sentiment distribution.

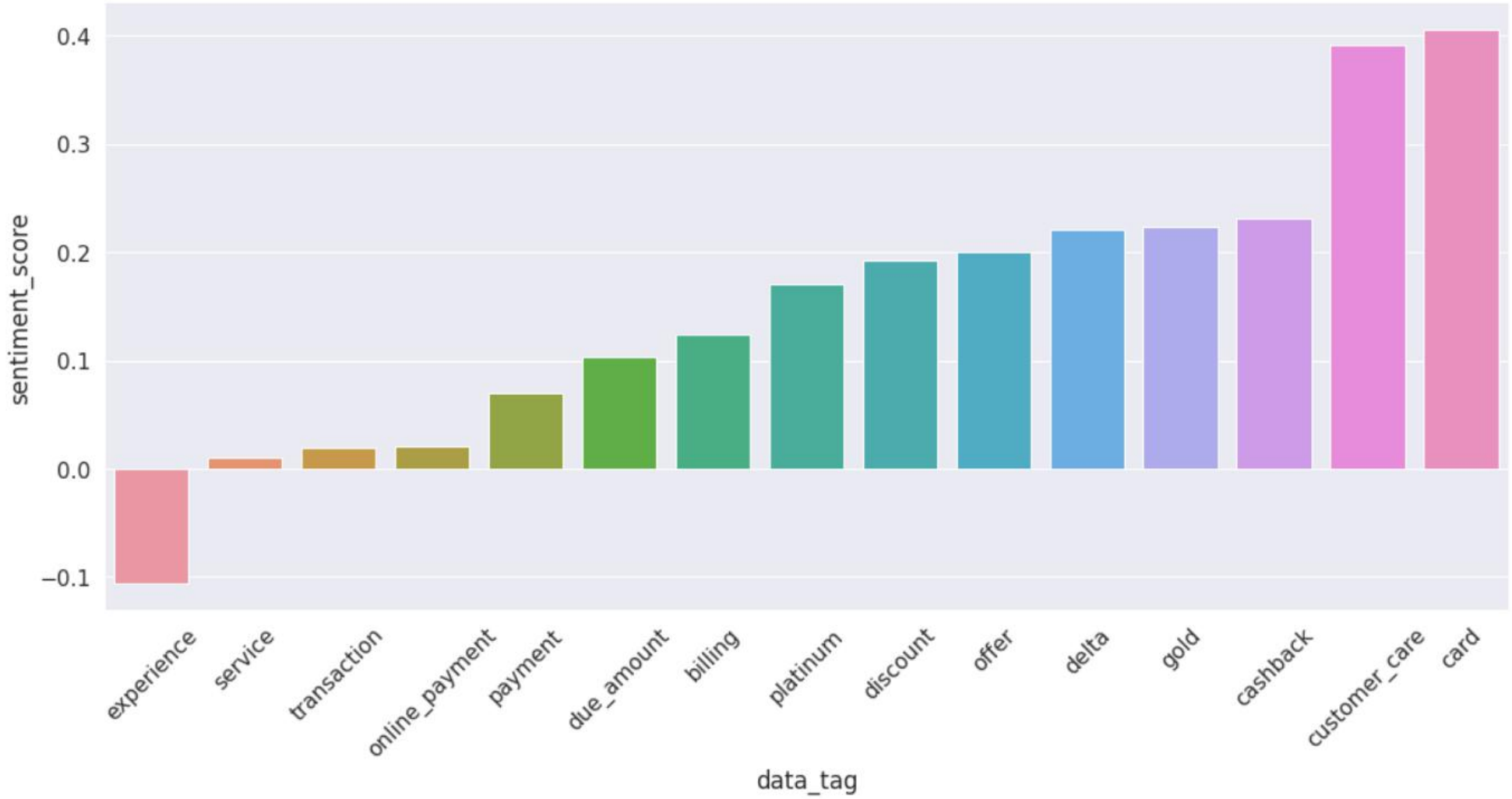


increasing Positive Sentiment --->

Positive tweets (1)	27264
Negative tweets (0)	13068
Neutral tweets (-1)	13909

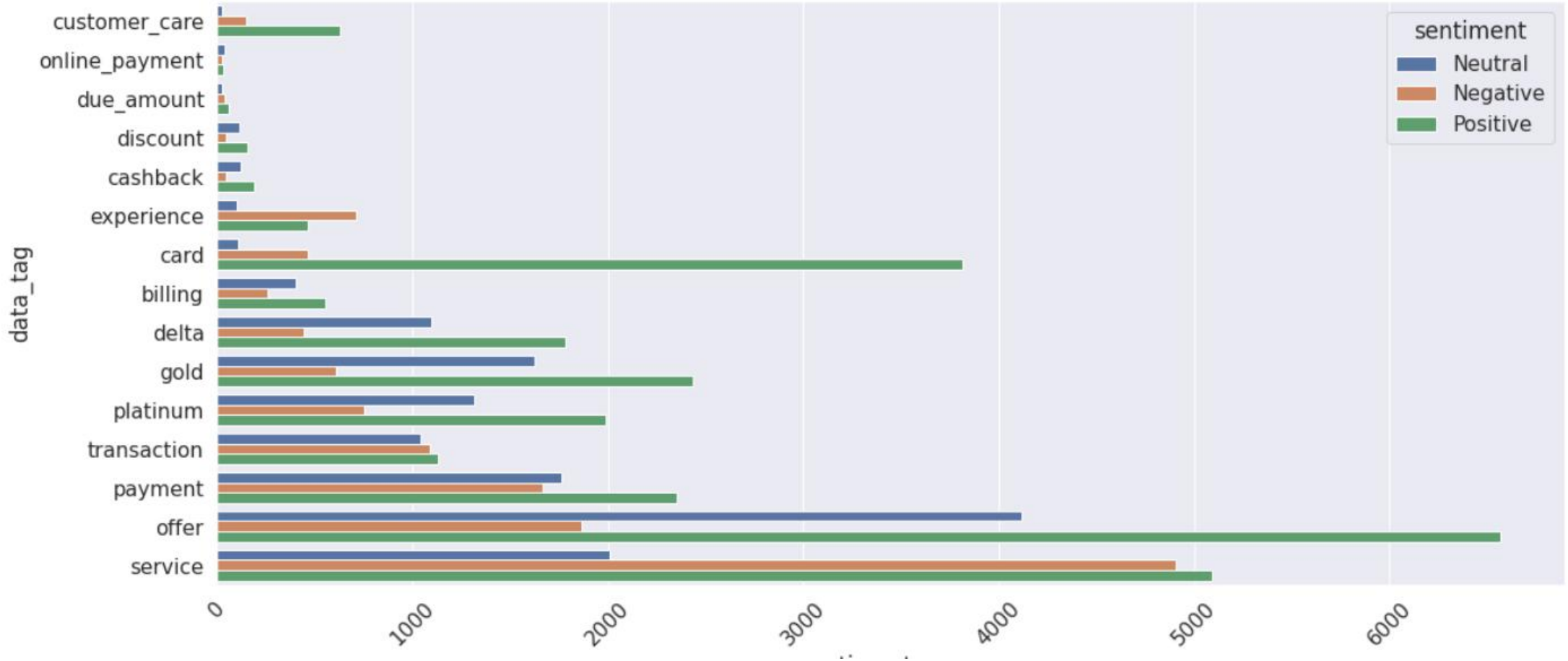
Overall sentiment (contd.)

AmEx has very high positive sentiment for customer care and cashback and premium cards.



Overall sentiment (contd.)

We have low number of tweets to support bad sentiment score for online payment.



Solution

data_tag	sentiment_score
experience	-0.105401
service	0.010705
transaction	0.019615
online_payment	0.020506
payment	0.070087
due_amount	0.102684
billing	0.12394
platinum	0.170705
discount	0.192802
offer	0.199884
delta	0.220983
gold	0.222846
cashback	0.231105
customer_care	0.390924
card	0.405224

- On analyzing the user experience we find that the **major issues faced by the users are regarding the payments being made via credit cards** i.e, platinum, gold and delta cards as these are the most frequent words used by the customers.
- The word “experience” and “transaction” has a very low mean of sentiment score therefore it can be **improved by providing more assistance in using credit cards and transaction activities**.
- Another major issue faced by the customers is during online payments. There are certain offers given for online purchasing which attracts users to do cashless transactions. Customers were happy about customer care offers and cashbacks.
- However we have very small dataset for online transaction tweets so the sentiment score can be **biased**.

Solution (contd.)

The gold and delta cards were found to have more sentiment scores than platinum. We have the following conclusions to support the sentiment scores

- The sentiment score could be driven by high fees for platinum cards
- From the dataset we were able to identify that the negative sentiment for platinum cards is driven by service, payment, account and transaction related tweets.

word	count
service	5739
payment	2169
account	1427
transaction	1296
call	957
experience	934
pay	914
credit	901

Video for the project

https://drive.google.com/drive/folders/11175RZ3wraUAxJq4Jb6MLojvceL4RrY8?usp=sharing

Challenges while creating the project

1. Twitter does not allow high data scraping using its basic twitter API and requires premium paid accounts.
2. Scraping and data pre processing requires a lot of CPU resources and time.
3. Sentiment analysis does not provide good results when not trained on original dataset.
4. Understanding and using Big Data can be complex for some people without proper education/ training

Conclusion

Data analysis of tweets provides the company with a lot of details about the needs and sentiment of the products they are providing and the customer service. The feedback in these cases are almost sudden which allows a company to immediately respond by improving their service and other market approach.

Big data solves this problem very easily by processing large amount of datasets and customer responses. There are many companies that already leverage the big data tools for processing customer feedback and drive customer satisfaction or for targetting new customers through sales.

Appendix

word	count (Overall)	Rank Overall (pos+neg)	Rank positive	Rank negative
credit	1347	2	2	8
account	1226	3	4	2
platinum	1138	4	3	4
gold	671	5	5	-
service	620	6	12	3
call	526	9	11	9
money	508	10	-	5
delta	489	11	7	-
my	457	12	18	10
how	443	13	17	16
points	439	14	10	-
pay	436	15	13	6
cards	384	16	16	-
trying	380	18	19	-
bonus	371	19	6	-
email	348	20	-	-