PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

Samuel Rodrigues Costa

APLICAÇÃO DE TEXT MINING PARA ANÁLISE DA SATISFAÇÃO DOS USUÁRIOS COM SERVIÇOS BANCÁRIOS

Belo Horizonte
2020

Samuel Rodrigues Costa

APLICAÇÃO DE TEXT MINING PARA ANÁLISE DA SATISFAÇÃO DE USUÁRIOS COM SERVIÇOS BANCÁRIOS

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Belo Horizonte

2020

SUMÁRIO

1. Introdução

1.1. Contextualização

A experiência com serviços bancários está cada vez mais digital. Entender as necessidades e dores dos clientes é fundamental para garantir uma melhor satisfação.

A satisfação é geralmente acompanhada através de pesquisas NPS(Net Promoter Score), CSAT (Customer Satisfaction Score), CES(Customer Effort Score), CRC (Custo de retenção de clientes), taxa de cancelamento de produtos, etc. Outra forma de verificar isso é através da leitura dos comentários deixados por eles usuários nas lojas de aplicativos da Play Store e Apple Store.

Os *feedbacks* registrados podem ser úteis para determinar o quão satisfeito um cliente está com um produto ou serviço ou ainda se existe algum problema acontecendo que exige ação rápida. É possível ainda fazer diagnósticos sobre produtos, atendimento, aplicativo, experiência, etc.

1.2. O problema proposto

O processo de análise de comentários dos usuários pode ser cansativo, repetitivo, sem falar custoso. Ele geralmente consiste na análise de cada comentário, se tivermos poucos, essa prática já resolve. O problema se instala quando é necessário analisar uma massa de dados enorme e que só cresce mais a cada dia.

O objetivo desse trabalho é propor uma solução a esse problema de modo a diminuir o tempo de análise e permitir melhor gerenciamento da informação através da busca de padrões nas avaliações dos usuários aplicando técnicas de text mining. O processo de text mining tipicamente envolve o uso de técnicas de processamento de linguagem natural (NLP) para se extrair dados estruturados de uma narrativa estruturada.

Queremos mostrar que é possível entender melhor a satisfação e experiência do cliente tendo como ponto de partida apenas data e comentário.

Os dados analisados consistem de reviews presentes nas lojas de aplicativos do Google e da Apple referentes a bancos, corretoras e seguradoras em sua maioria concentrada entre os anos de 2019 e início de 2020.

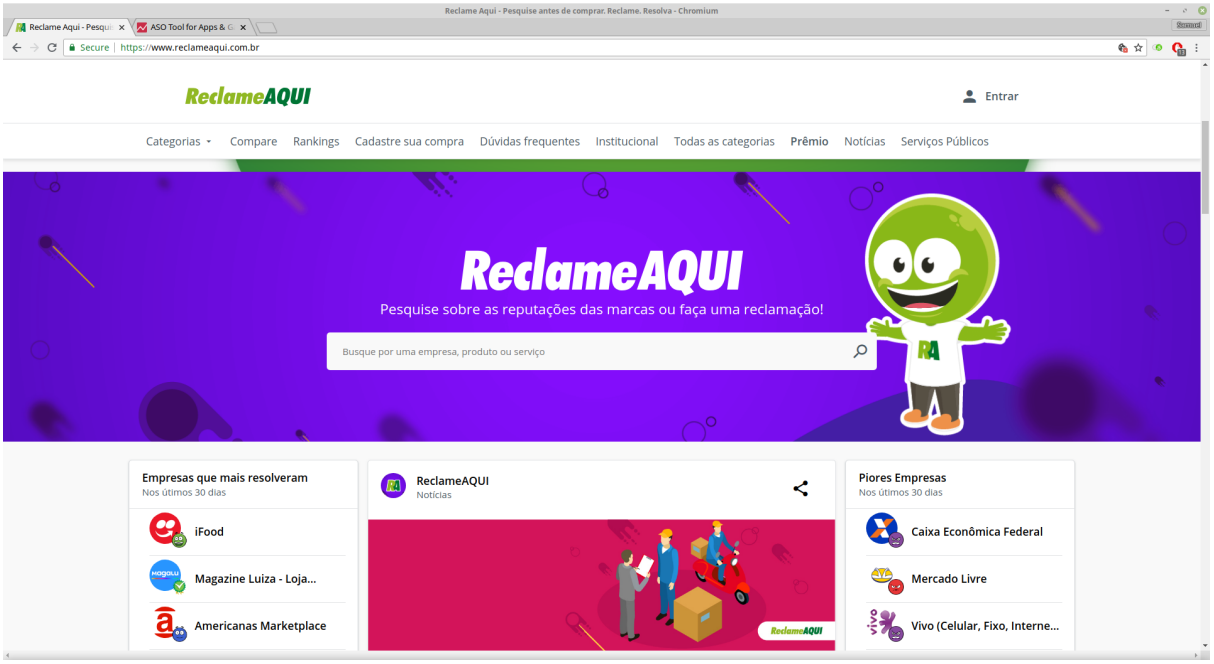
1.3 Estrutura do projeto

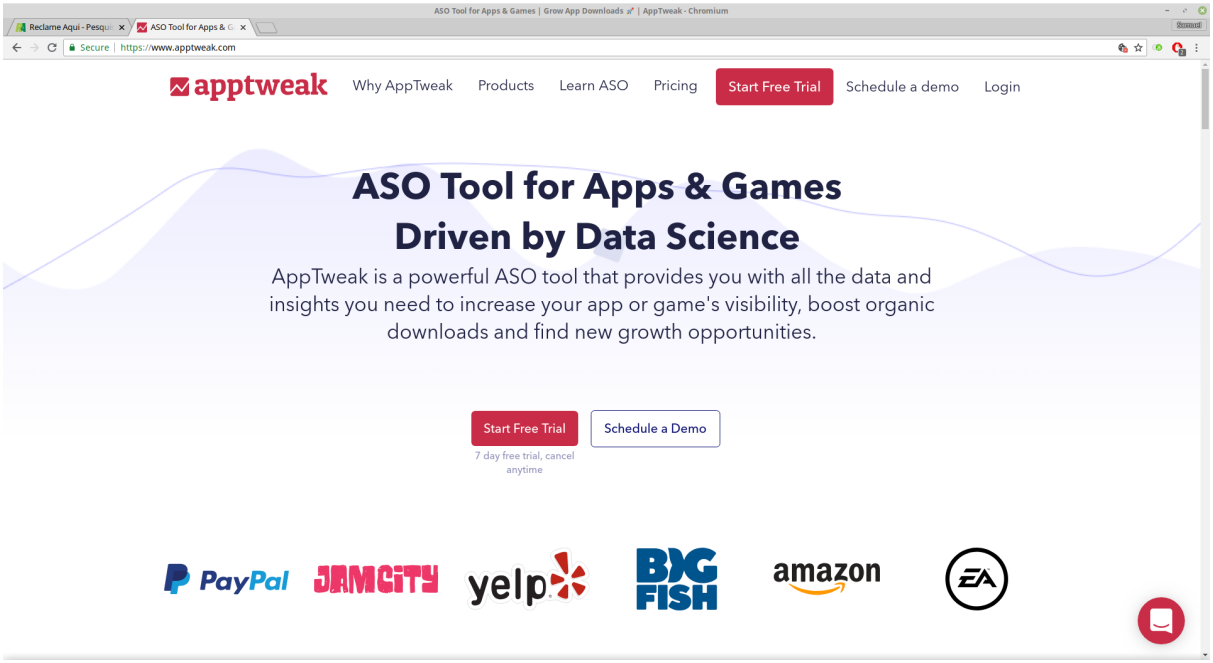
A disposição do projeto segue o seguinte padrão de pastas:

- *data_extraction*: centraliza todos os processos de extração de arquivos;
- *word2vec*: responsável pela criação de modelo word2vec;
- *data_analysis*: Armazena o notebook responsável pela análise exploratória;
- *sentiment_analysis*: cuida da criação de modelo de análise de sentimento;
- *data_visualization*: armazena todos os códigos e arquivos necessários para a vizualização das análises.

2. Coleta de Dados

Os dados foram obtidos através de requisições a API's disponibilizadas pelos sites https://www.reclameaqui.com.br/ e https://app.apptweak.com/.





A base do Reclame Aqui foi obtida em 07/2019. Precisamos destacar que esses comentários serão utilizados primariamente na construção do modelo word2vec uma vez que precisamos de um vocabulário grande para uma melhor generalização.

Os dados obtidos da API pública do site Apptweak são apenas uma amostra de todas as resenhas publicadas nas lojas de aplicativo. Para ter acesso a todas as reviews é necessário assinar o serviço provido pelo site ou realizar web scraping na página das lojas de aplicativos. A última atualização foi feita em 05/2020.

Os scripts de extração de dados estão localizados na pasta *data_extraction/* e todas as execuções de código desse capítulo são feitas a partir dela.

O download dos dados do site Reclame Aqui é realizada através do script *app_reclame_aqui.js* e os dados são salvos na pasta *csv/reclame_aqui/*.

Segue descrição das propriedades dos arquivos salvos:

Campo	Descrição	Tipo
id	id	String
data	Data da reclamação	String (Date)
empresa	Empresa	String
titulo	Título da reclamação	String

review	Texto da reclamação	String
reply1	Resposta 1 do cliente ou da empresa	String
reply2	Resposta 2 do cliente ou da empresa	String
reply3	Resposta 3 do cliente ou da empresa	String
reply4	Resposta 4 do cliente ou da empresa	String
reply5	Resposta 5 do cliente ou da empresa	String
interacoes	Número de respostas	Inteiro
categoria	Categoria da reclamação	String
tipo_problema	Tipo de problema	String
outros_problemas	Outro tipo de categoria	String
produto	Produto reclamado	String
estado	Estado	String
cidade	Cidade	String

As requisições à API do site Apptweak são feitas através do script *apptweak-selenium.py* e os dados das lojas do Google e IOS são salvos na pasta *csv/app/*.

As colunas do arquivo salvo mudam para cada empresa de acordo com a loja informada sendo algumas comuns aos dois, detalharemos apenas o que é comum.

Campo	Descrição	Tipo
application_id	Identificador do app na loja	String
date	Data da publicação da resenha mais recente	String(Date)
rating	Nota atribuída ao aplicativo	String
title	Título da resenha	String
review	Conteúdo da resenha	String
body_length	Tamanho da resenha	Inteiro
id	Id da resenha na loja	String
autor	Autor	String (dict)

3. Processamento/Tratamento de Dados

A fase de normalização dos textos é essencial para garantir uniformidade e expurgar construções gramaticais que nada agregam na elaboração dos modelos, permitindo assim que haja um aprendizado mais congruente.

O texto extraído precisa ser pré-processado a fim de remover caracteres especiais, quebras de linhas, acentos, números, endereço de sites e e-mails e alguns sinais de pontuações. Os números de 0 a 10, 100 e 1000 são substituídos por sua forma extensa, assim 0 é substituído por zero, 100 por cem e assim sucessivamente. Algumas palavras também são substituídas no processo por estarem com grafia incorreta ou são mudadas para uma forma mais inteligível, dessa forma “vc” se torna “voce” e “mt” em “muito”. Esse procedimento foi realizado através da aplicação de técnicas de busca usando regex com posterior aplicação das substituições. A lista completa está contida no arquivo *regex.py* na pasta *word2vec/*.

Outra parte importante no processo consiste na remoção de stopwords, termo em inglês para palavras que são bastante frequentes nos textos, mas que pouco agregam no sentido das sentenças. A lista utilizada inclui artigos, preposições, pronomes, verbos, adjetivos, nomes de pessoas, etc. Existem dois arquivos com stop words salvos na pasta *word2vec/input/*, o primeiro (*stop_words_.txt)* contém 4106 palavras em sua maioria nomes de pessoas, já o segundo (*stop_words.txt*) possui 6351 palavras e é será usado na construção de bigramas e no cálculo de frequência das palavras.

A limpeza dos dados ocorre em momentos diferentes. Para a construção do modelo word2vec é executado através do script *extract_reviews.py* da pasta *word2vec/* que extrai os comentários dos arquivos das pastas *data_extraction/csv/app* e *data_extraction/csv/reclame_aqui* e salva na pasta *word2vec/input/comentarios/*. Outro momento é antes a criação do modelo de análise de sentimento através do Jupyter notebook *sentiment_analysis.ipynb* presente na pasta *sentiment_analysis/*.

Após analisarmos os arquivos da base do Reclame Aqui pudemos verificar que dos mais de 1 milhão de registros não existem entradas duplicadas uma vez que os identificadores são únicos. Os campos reply1, reply2, reply3, reply4, reply5,

categoria, tipo_problema, outros_problemas e produto possuem uma quantidade relevante de registros nulos. No pré-processamento dos comentários o registros nulos são descartados.

1. Reclame Aqui

```
# Lista de arquivos
files = [f for f in glob('../data_extraction/csv/reclame_aqui/*.csv')]
## Tentar realizar o join desses arquivos pode resulter em problemas de memória
```

Registros vazios

```
df = check_null(files)
```

```
df.groupby('id').sum()
```

data	empresa	titulo	review	reply1	reply2	reply3	reply4	reply5	interacoes	categoria	tipo_problema	outros_problemas	produto	estado	cidade
0	0	1	0	82202	513632	925579	1019779	1049581	0	1067596	1067717	907770	1067741	5	5

```
total, duplicados = get_duplicates(files)
```

Quantidade de reclamações

```
total
```

```
1068254
```

Registros duplicados

```
duplidados
```

```
0
```

Os dados das lojas de aplicativos possuem mais de 600 mil registros. Existem 14.656 entradas duplicadas. As colunas version, tittle, review, sort_score, response possuem registros nulos. No pré-processamento dos comentários o registros nulos são descartados.

2. Lojas de aplicativos do Google e da Apple

```
# Arquivos ainda não pré-processados
files = [f for f in glob('../data_extraction/csv/app/*.csv')]
```

Registros vazios

```
df = check_null(files)
```

```
df.groupby('id').sum()
```

id	language	application_id	version	date	rating	title	review	body_length	author	sort_score	country	vote_count	vote_sum	is_edited	response
0	0.0	0	54054.0	0	0	540274	22	0	0	397029	0.0	0.0	0.0	0.0	48306.0

```
total, duplicados = get_duplicates(files)
```

Quantidade de reviews

```
total
```

```
639673
```

Registros duplicados

```
duplicados
```

```
14656
```

4. Análise e Exploração dos Dados

Análise disponível no Jupyter notebook *data_analysis.ipynb* da pasta *data_analysis/.*

4.1 Reclame Aqui

O processo de junção desses arquivos demanda bastante recursos do computador devido ao tamanho deles, decidimos então criar a função *describe* que após percorrer cada um disponibiliza o resultado num *Pandas dataframe*.

Banco do Brasil, Banco Itaú, Banco Bradesco, Banco Santander e Banco Cetelem se destacam como as cinco instituições mais reclamados dos dados analisados.

Top 10 empresas mais reclamadas

```
resumo['empresa'].head(10)
```

empresa	quantidade
Banco do Brasil	99992.0
Banco Itaú	99977.0
Banco Bradesco	99961.0
Banco Santander	99955.0
Banco Cetelem	85412.0
Caixa Econômica Federal	83908.0
Santander Cartões	71316.0
Cielo	62263.0
Bradesco Seguros	33804.0
Porto Seguro	32594.0

O título mais comum nas reclamações é cobrança indevida, mas os usuários tratam de portabilidade, cartão de crédito e juros abusivos.

Top 10 títulos das reclamações

```
resumo['titulo'].head(10)
```

titulo	quantidade
Cobranca Indevida	32545.0
Portabilidade	5145.0
Cartao De Credito	3022.0
Juros Abusivos	2649.0
Propaganda Enganosa	2635.0
Pessimo Atendimento	2485.0
Cancelamento	2242.0
Cobranca	2008.0
Descaso	1788.0
Falta De Respeito	1744.0

O campo categoria quase sempre não é preenchido quando o usuário abre uma reclamação, então a maioria dos registros não tem essa informação. O mesmo problema acontece com produto e tipo_problema.

Top 5 categorias mais reclamadas

```
resumo['categoria'].head(5)
```

categoria	quantidade
Não informada(o)	1067596.0
Cartões de Crédito	154.0
Bancos	129.0
Seguradoras	67.0
Não encontrei meu problema	56.0

Top 5 produdos mais reclamados

```
resumo['produto'].head(5)
```

produto	quantidade
Não informada(o)	1067741.0
Cartão de crédito	121.0
Outro Tipo de produto/Serviço	97.0
Problemas Gerais	46.0
Conta	42.0

Top 5 problemas mais relatados

```
resumo['tipo_problema'].head(5)
```

tipo_problema	quantidade
Não informada(o)	1067717.0
Outro problema	100.0
Cobrança indevida	70.0
Mau Atendimento	20.0
Cobrança indevida de tarifas	16.0

Outro fato interessante é que o Estado e a cidade de São Paulo são em disparado os locais que mais originam reclamações.

Top 5 registros por Estado

```
resumo['estado'].head(5)
```

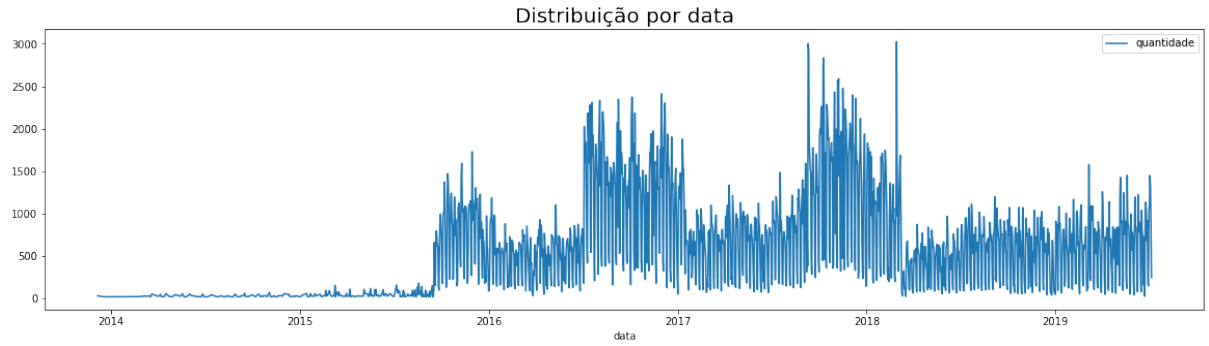
estado	quantidade
SP	459963.0
RJ	133699.0
MG	100362.0
PR	48963.0
BA	46359.0

Top 5 registros por cidade

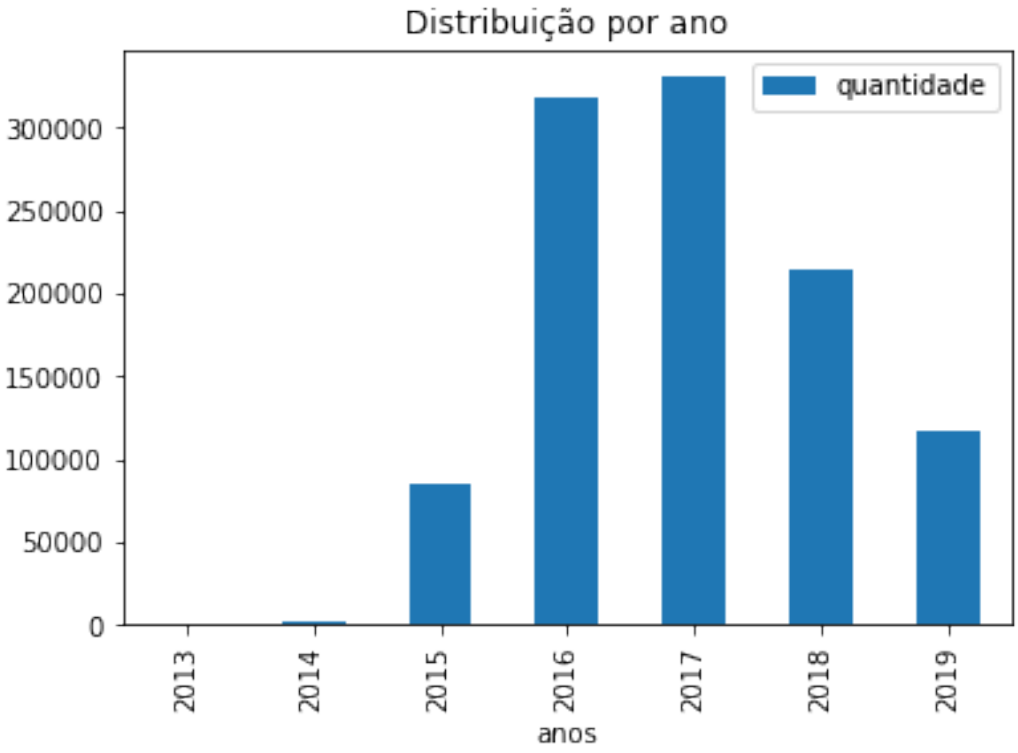
```
resumo['cidade'].head(5)
```

cidade	quantidade
São Paulo	207006.0
Rio de Janeiro	76980.0
Belo Horizonte	32008.0
Brasília	24774.0
Salvador	21299.0

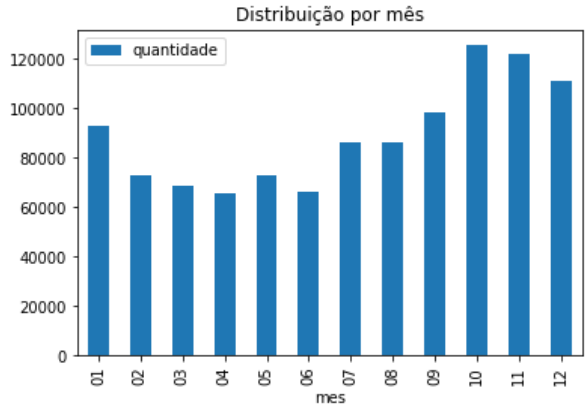
Traçamos uma linha temporal para compreender como os usuários se comportam ao longo do tempo.

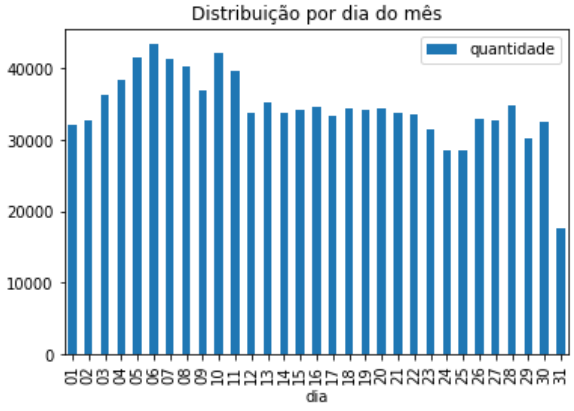


Podemos verificar que a quantidade de reclamações está diminuindo após pico em 2017. Precisamos destacar, no entanto que a série de 2019 não está completa.



Existe um certo aspecto de sazonalidade nas reclamações. Os usuários tendem a reclamar mais no último trimestre do ano e menos no segundo. O dia 31 costuma ser o dia com menos registros enquanto o dia 06 com mais.





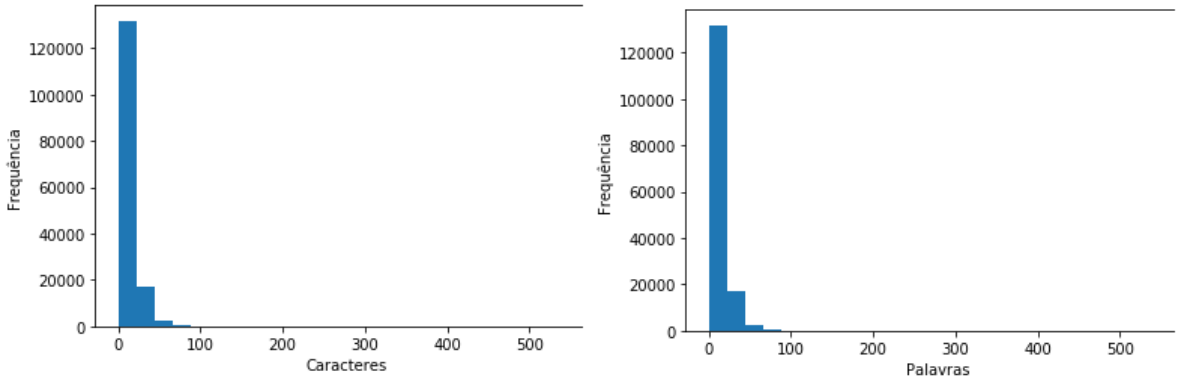
4.2 Google Play e Apple Store

Esse estudo será focado apenas nos comentários preprocessados que possui 152.259 registros.

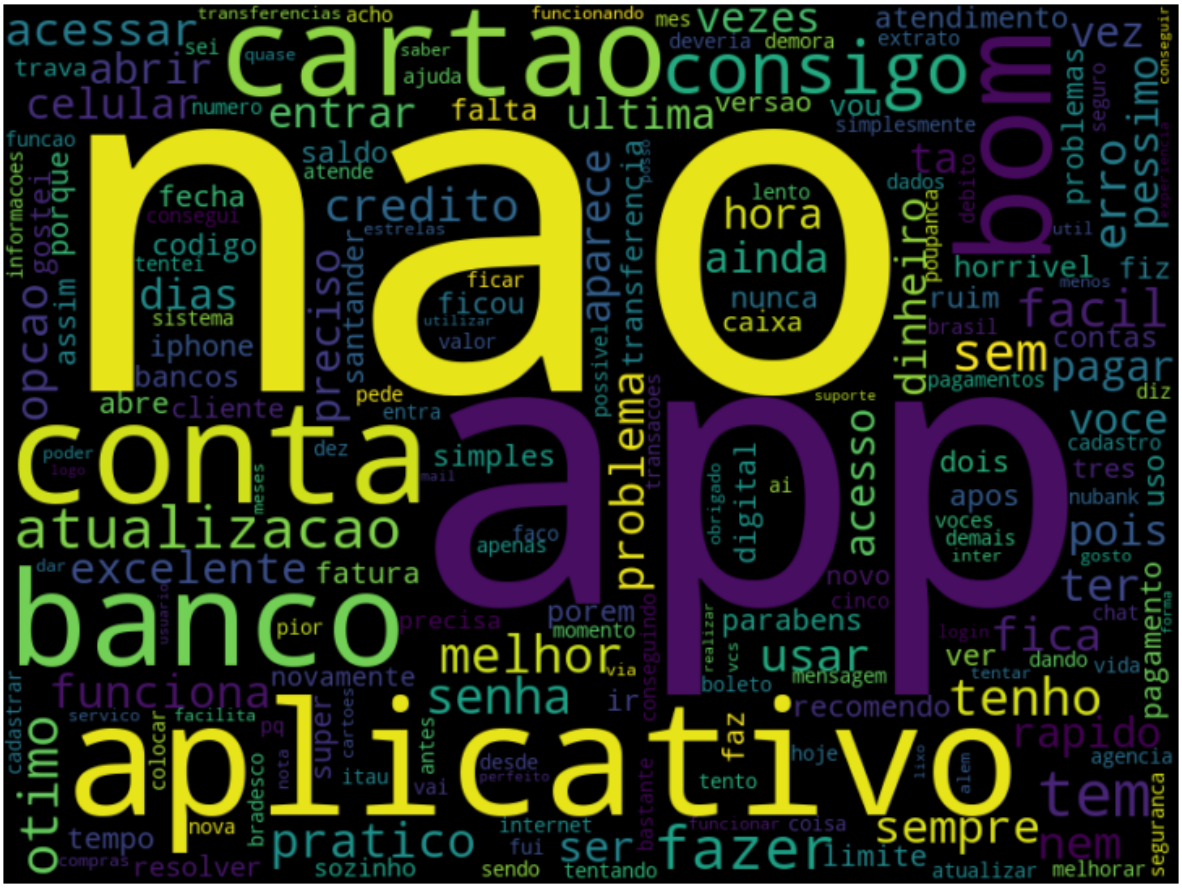
Após analisar os comentários verificamos que a média possui 110 caracteres e apenas 11 palavras. Após segmentar por quartis e criar histograma concluímos que nesse canal os usuários tendem a escrever pouco.

```
# Quantidade de caracteres
mean        110.141200
std         110.175405
min           5.000000
25%          38.000000
50%          75.000000
75%         145.000000
max        4868.000000
```

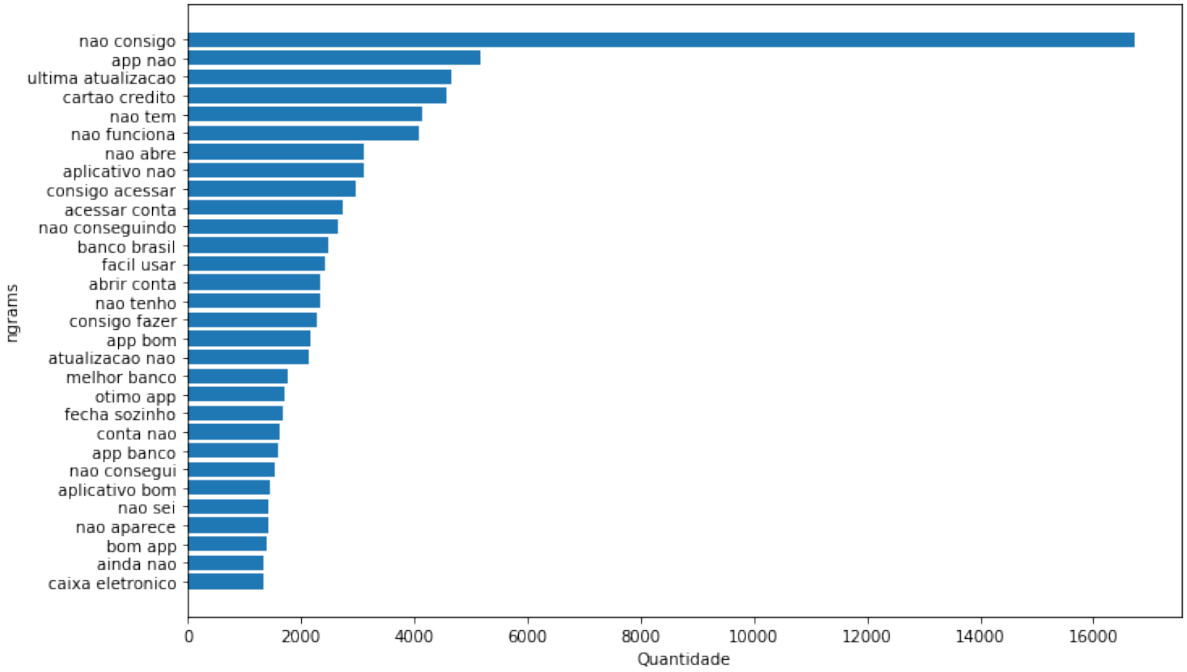
```
# Quantidade de palavras
mean         11.811052
std          11.668705
min           0.000000
25%           4.000000
50%           8.000000
75%          16.000000
max         556.000000
```

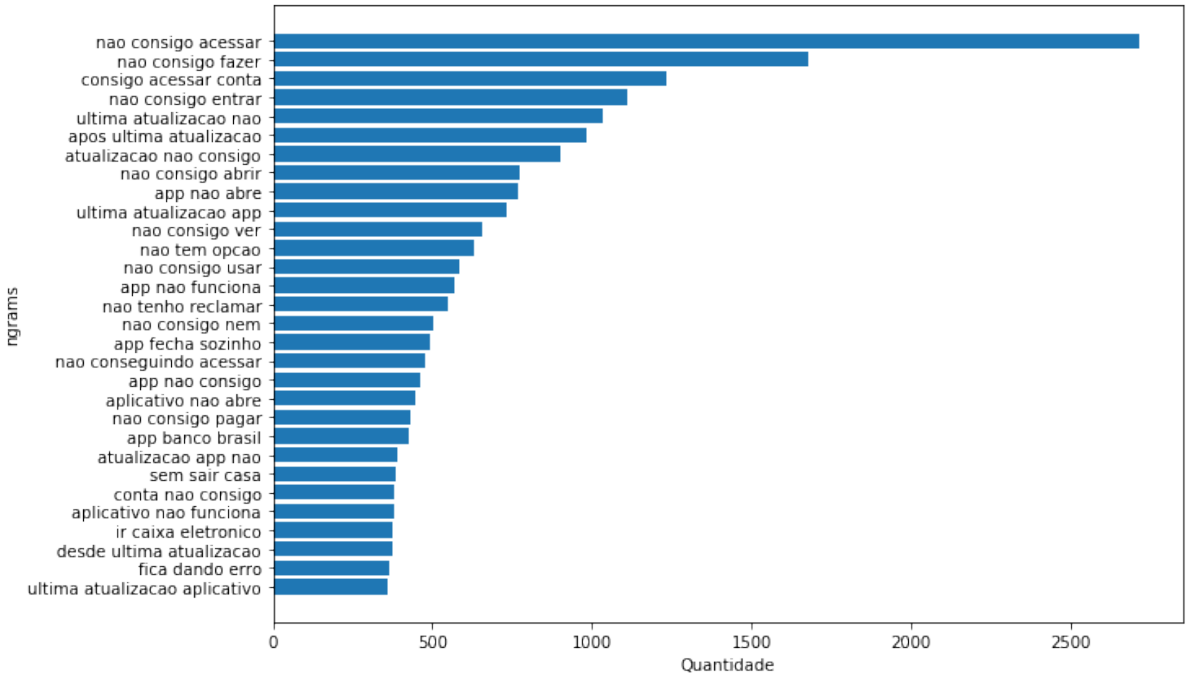


Depois de construir uma nuvem de palavras descobrimos que os assuntos mais comentados se referem a **aplicativo(app)**, **conta**, **cartão, banco** provavelmente com alguma conotação negativa devido à presença frequente da palavra **não**.



Uma importante forma de visualizar sobre o que os usuários estão falando é através da construção de n-grams. Descobrimos que é comum reclamarem que não estão conseguindo fazer alguma coisa ou que o aplicativo apresenta algum problema.


nao consigo
app nao
ultima atualizacao
cartao credito
nao tem
nao funciona
nao abre
aplicativo nao
consigo acessar
acessar conta
nao conseguindo
banco brasil
facil usar
abrir conta
nao tenho
consigo fazer
app bom
atualizacao nao
melhor banco
otimo app
fecha sozinho
conta nao
app banco
nao consegui
aplicativo bom
nao sei
nao aparece
bom app
ainda nao
caixa eletronico
ngrams
0
2000
4000
6000
8000
10000
12000
14000
16000
Quantidade


nao consigo acessar
nao consigo fazer
consigo acessar conta
nao consigo entrar
ultima atualizacao nao
apos ultima atualizacao
atualizacao nao consigo
nao consigo abrir
app nao abre
ultima atualizacao app
nao consigo ver
nao tem opcao
nao consigo usar
app nao funciona
nao tenho reclamar
nao consigo nem
app fecha sozinho
nao conseguindo acessar
app nao consigo
aplicativo nao abre
nao consigo pagar
app banco brasil
atualizacao app nao
sem sair casa
conta nao consigo
aplicativo nao funciona
ir caixa eletronico
desde ultima atualizacao
fica dando erro
ultima atualizacao aplicativo
ngrams
0
500
1000
1500
2000
2500
Quantidade

5. Criação de Modelos de Machine Learning

Para esse projeto foi criado um modelo word2vec usando a biblioteca gensim e o um de análise de sentimento usando XGBoost todos em Python.

Escolhemos o algoritmo word2vec por ser de fácil implementação e mais completo que o Bag of Words para representar sentenças sem falar na diminuição de dimensionalidade das features.

O XGBoost é um algoritmo de aprendizado supervisionado que implementa um processo de *Boosting* para gerar modelos precisos. Isso nos dá a vantagem de produzir modelos mais generalizáveis e menos propensos a *over fitting*.

5.1. Word2vec

Para construção do modelo word2vec o script *model_builder.py* da pasta *word2vec/* é acionado. Primeiro extraímos apenas os comentários dos arquivos salvos na pasta data_extraction/csv/reclame_aqui/ e data_extraction/csv/app/, depois preprocessamos e salvamos os arquivos na pasta *word2vec/input/comentarios/.* Esses arquivos serão percorridos para a construção do corpus, que contém 2.596.630 sentenças e de um vocabulário de 84.781 palavras distintas que serão usadas no cálculo dos vetores.

A estrutura principal do código é relativamente simples e composta pelos parâmetros abaixo:

- num_features – número de dimensões dos vetores de cada palavra;
- min_count – palavra precisa ter ocorrido pelo menos 5 vezes no corpus para fazer parte do vocabulário;
- window – janela deslizante de palavras;
- sample – diminui peso aleatoriamente de palavras muito frequentes;
- alpha – taxa de aprendizado inicial;
- total_examples – tamanho do vocabulário;
- epochs – número de iterações sobre o corpus.

```
#Setar parâmetros
num_features = 300
w2v_model = word2vec.Word2Vec(
  min_count=5,
  window=5,
  size=num_features,
```

```
  sample=0.0,
  alpha=0.025,
  workers=4
 )
 #Construção do vocabulário
 w2v_model.build_vocab(sentences)
 #Treinamento do modelo
 w2v_model.train(sentences, total_examples=w2v_model.corpus_count,
epochs=100)
 model_name = "models/w2v_model"+str(num_features)
```

Uma vez que o treinamento é finalizado extraímos apenas dos vetores das palavras e salvamos um modelo mais enxuto, em contrapartida ele não poderá ser usado para continuar o treinamento com outros textos:

```
 word_vectors = w2v_model.wv
 word_vectors.save(model_name + '.kv')
```

5.2. Análise de sentimento

Antes de iniciar o treinamento selecionamos uma amostra com mais de 11 mil registros da base das lojas de aplicativos e rotulamos os comentários como 1(positivo) e -1(negativo). Esse processo foi realizado manualmente tendo em vista falta de material em língua portuguesa.

O treinamento do modelo é feito através do Jupyter notebook *sentiment_analysis.ipynb* da pasta *sentiment_analysis/.*

Primeiramente carregamos em memória o modelo word2vec construído na etapa anterior que servirá de base para a criação de um modelo de análise de sentimento.

```
 w2v_path = '../word2vec/models/vectors.kv'
 w2v_model = word2vec.load(w2v_path)
```

Fazemos o *load* dos dados com as reviews rotuladas. Caso haja registros duplicados, eles serão eliminados e valores nulos serão tratados.

```
 pos = pd.read_csv("input/positivo.csv")
 neg = pd.read_csv("input/negativo.csv")
 sentiment = pd.concat([pos,neg]).drop_duplicates().reset_index(drop=True)
 sentiment = sentiment.fillna('')
```

A função *transform* é responsável por receber as sentenças e encadear o processo de normalização, adicionar a feature *bad* que ajuda encontrar as sentenças negativas e retornar os vetores que serão usados no treinamento.

```
trainVecs = transform(sentiment['review'], w2v_model)
data = pd.concat([sentiment.categoria, trainVecs], axis=1)
data = data.drop_duplicates()
```

O próximo passo é balancear as classes para garantir que nenhuma delas seja favorecida durante o aprendizado.

```
df_maior = data[data.categoria==-1]
df_menor = data[data.categoria==1]
df_ = resample(df_maior, replace=True, n_samples=df_menor.shape[0],
random_state=123)
data = pd.concat([df_menor, df_])
```

Os dados foram separados em 80% para treino e 20% para teste.

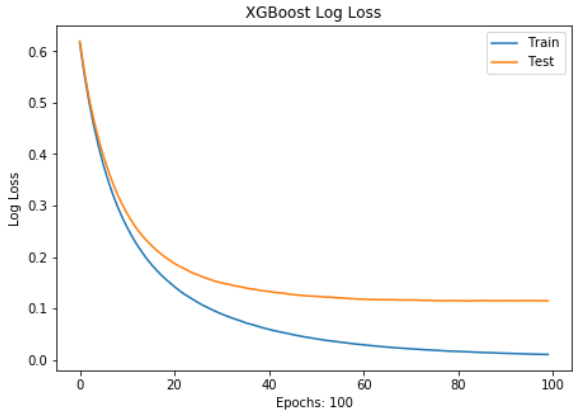
```
X_train, X_test, y_train, y_test = train_test_split(data.iloc[:, 1:], data.iloc[:, 0],
train_size=.8, stratify=data.iloc[:, 0], random_state=99)
```

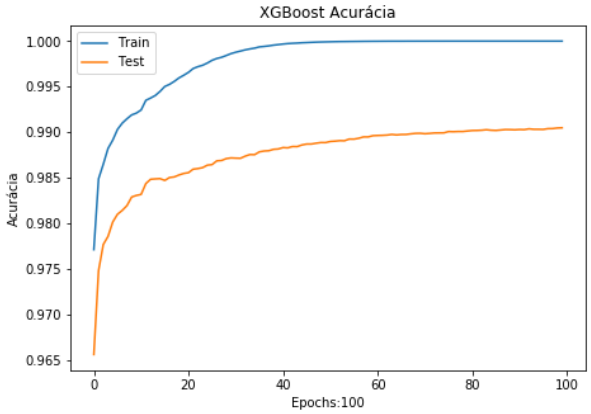
Iniciamos o algoritmo XGBoost no modo padrão configurando apenas a taxa de aprendizagem, os parâmetros de seleção aleatória dos dados de treinamento, quantas threads serão usadas(o que estiver disponível), o tipo do problema, que é de classificação binária e o tipo de booster:

```
xgb_model = xgb.XGBClassifier(
 learning_rate=0.1,
 nthread=-1,
 random_state=99,
 objective='binary:logistic',
booster='gbtree')
```

A acurácia de 96.19% observada é relativamente alta levando em consideração que iniciamos no modo default.

```
xgb_model = train(xgb_model, X_train, X_test, y_train, y_test)
eval_train(xgb_model)
```





Avaliar métricas

```
# Fazer previsões com dados-teste
y_pred = xgb_model.predict(X_test)
predictions = [value for value in y_pred]
accuracy = accuracy_score(y_test, predictions)
print("Acurácia: %.2f%%" % (accuracy * 100.0))
```

```
Acurácia: 96.19%
```

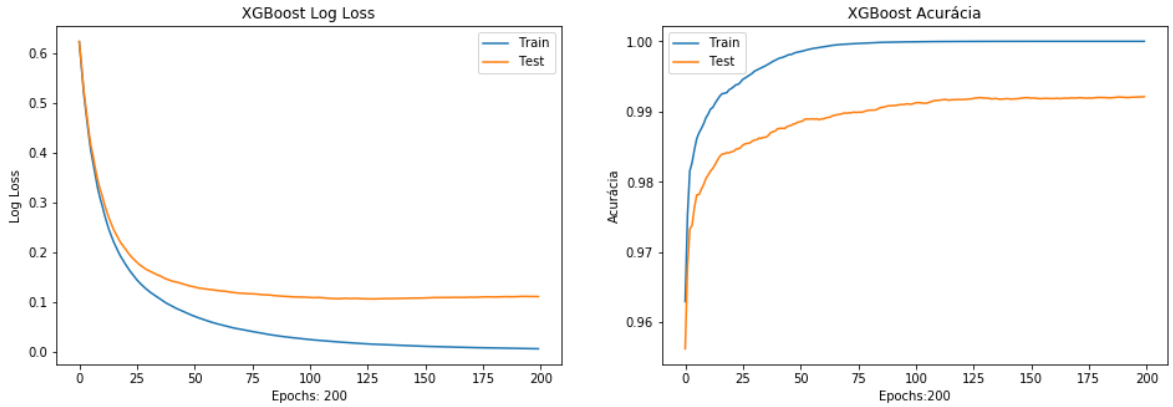
Embora o modelo já apresentasse resultados satisfatórios decidimos fazer o tuning dos demais parâmetros através de um processo de busca com validação cruzada. Esse processo é necessário para garantir não só um modelo otimizado, mas também generalizável.

```
scorers = {
 'accuracy_score': make_scorer(accuracy_score),
 'neg_log_loss': make_scorer(log_loss)
}
params = {
      'subsample': [i*.1 for i in range(5,10)],
      'colsample_bytree': [i*.1 for i in range(5,10)],
      'max_depth': [i for i in range(5,13,2)],
      'n_estimators': [i for i in range(100, 300, 50) ]
}
skf = StratifiedKFold(n_splits=3, shuffle = True)
grid = GridSearchCV(xgb_model,
 param_grid = params,
 scoring = scorers,
 n_jobs = -1,
 cv = skf.split(X_train, y_train),
```

```
    refit = 'accuracy_score',
    verbose=3,
    return_train_score=True)
grid.fit(X_train, y_train)
print('\n Best parameters:')
print(grid.best_params_)
xgb_model.set_params(**grid.best_params_)
```

De acordo com a imagem abaixo podemos ver que com o passar das épocas as métricas logloss e acurácia tendem para um valor mais próximo do ideal, a distância das métricas entre treino e teste também diminuíram e conseguimos melhorar a acurácia em 0.26 pontos percentuais.

```
xgb_model.set_params(**grid.best_params_)
xgb_model = train(xgb_model, X_train, X_test, y_train, y_test)
eval_train(xgb_model)
```



Avaliar métricas

```
# Fazer previsões com dados-teste
y_pred = xgb_model.predict(X_test)
predictions = [value for value in y_pred]
accuracy = accuracy_score(y_test, predictions)
print("Acurácia: %.2f%%" % (accuracy * 100.0))
```

```
Acurácia: 96.45%
```

```
#Matriz de confusão
conf = confusion_matrix(y_test, y_pred)
pd.DataFrame(conf, columns=xgb_model.classes_, index=xgb_model.classes_)
```

	-1	1
-1	919	25
1	42	903

Realizamos o teste com algumas sentenças novas:

```
sentences = [ "Não trava e ajuda vc evitar de ir ao banco",
              "Pode confiar",
              'Mui raramente não funciona. Melhor que o do Santander e da Caixa.',
              'Eita fila que demora!',
              'Minha senha de 6 dígitos está bloqueada',
              'Nota maxima para o app',
              'App muito bom, so que não',
              'mto dificil de ser utilizado, poderia ser mais prático como o da Caixa',
            'O banco não pensa no cliente',
             'nao, gosto muito do atendimento',
             'nao gosto muito do atendimento',
             ]
```

```
predict(pd.Series(sentences), w2v_model=w2v_model, xgb_model=xgb_model)
```

	texto	-1	1
0	Não trava e ajuda vc evitar de ir ao banco	3.83	96.17
1	Pode confiar	5.16	94.84
2	Mui raramente não funciona. Melhor que o do Sa...	4.46	95.54
3	Eita fila que demora!	99.65	0.35
4	Minha senha de 6 dígitos está bloqueada	95.52	4.48
5	Nota maxima para o app	0.42	99.58
6	App muito bom, so que não	99.76	0.24
7	mto dificil de ser utilizado, poderia ser mais...	95.52	4.48
8	O banco não pensa no cliente	78.25	21.75
9	nao, gosto muito do atendimento	12.01	87.99
10	nao gosto muito do atendimento	91.82	8.18

Por último salvamos o modelo.

filename = filename = 'models/xgb_model_sentiment-' + str(int(time.time())) + '.bin'

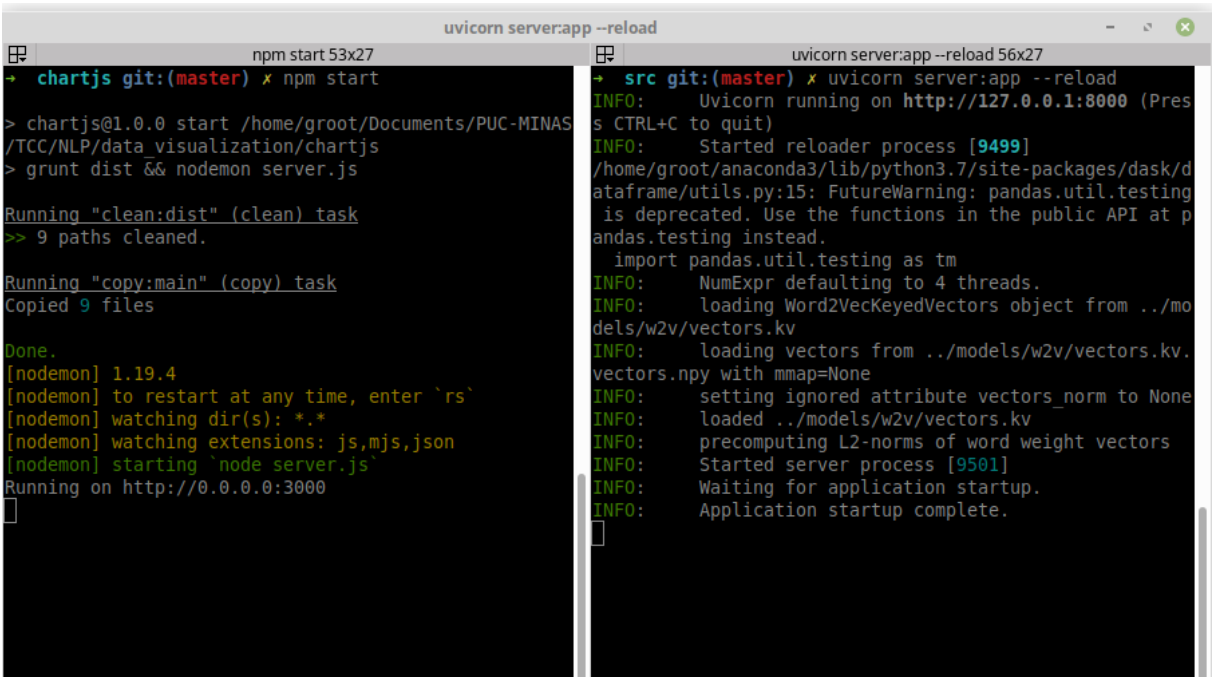
xgb_model.save_model(filename)

print(filename)

5.3. Construção de API e aplicação web

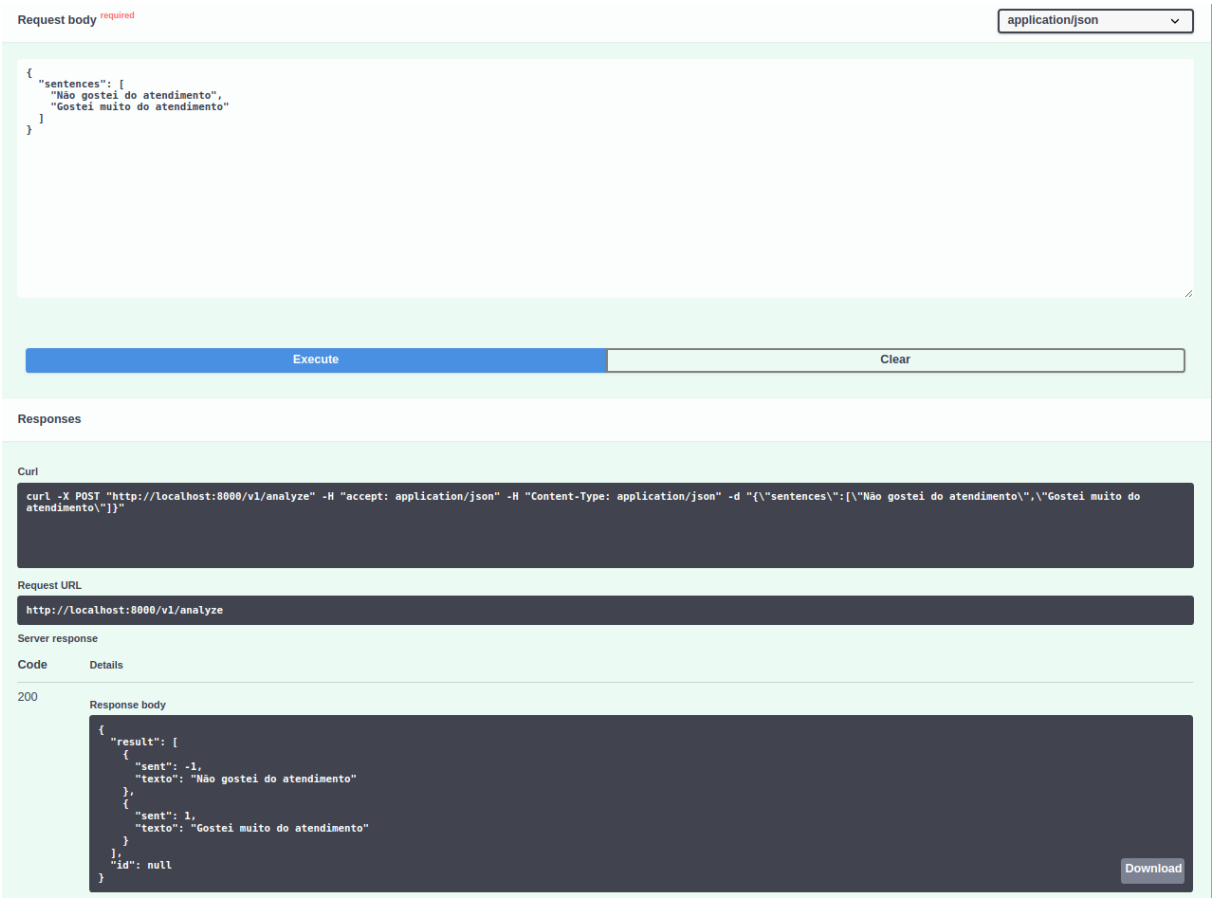
Para facilitar a apresentação dos dados construímos uma API com *FastApi* em Python para a análise de comentários e uma aplicação web com *Node.js, Express e Highcharts*.

As instruções de acionamento da API e da aplicação estão disponíveis nos arquivos *README.txt* das pastas *data_visualization/api_fastapi/* e *data_visualization/chartjs/.*



A tela inicial da API lista todos o endpoinsts disponíveis para utilização.

A página da documentação permite a realização de requisições diretamente, assim é possível simular a solicitação de uma análise e receber a resposta em tempo real.

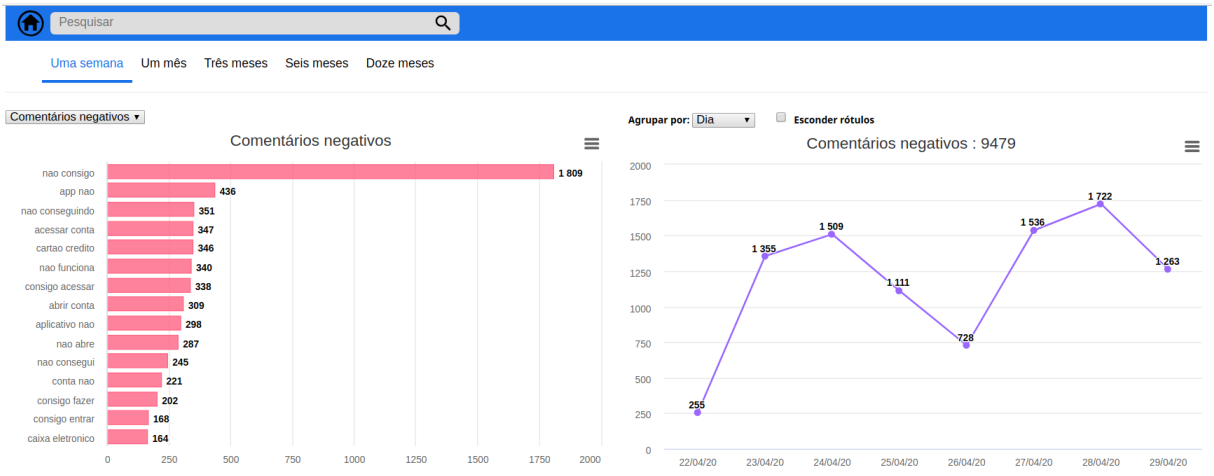

Request body required
application/json
{
"sentences": [
"Não gostei do atendimento",
"Gostei muito do atendimento"
]
}
Execute
Clear
Responses
Curl
curl -X POST "http://localhost:8000/v1/analyze" -H "accept: application/json" -H "Content-Type: application/json" -d "{\"sentences\":[\"Não gostei do atendimento\",\"Gostei muito do atendimento\"]}"
Request URL
http://localhost:8000/v1/analyze
Server response
Code
Details
200
Response body
{
"result": [
{
"sent": -1,
"texto": "Não gostei do atendimento"
},
{
"sent": 1,
"texto": "Gostei muito do atendimento"
}
],
"id": null
}
Download

6. Apresentação dos Resultados

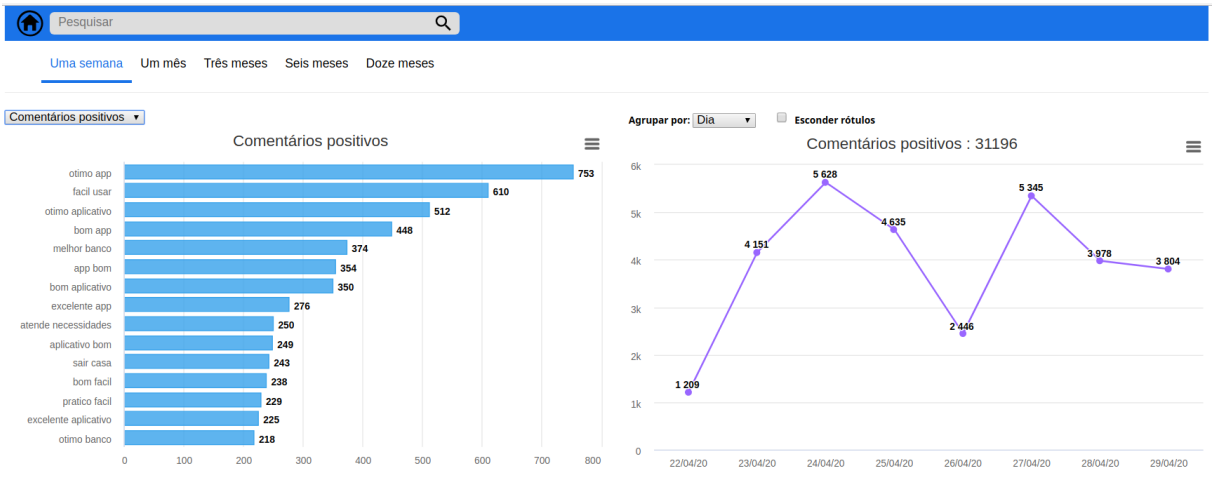
Os dados visualizados compreendem o período de 01/01/2019 a 30-04-2020, e é composto por 528.942 reviews de usuários da Play Store e da Apple Store.

A página inicial mostra uma visão geral dos comentários negativos dos últimos 7 dias. É possível trocar para a visão de comentários positivos.

O gráfico de barras exibe os 15 bigramas mais frequentes do período. Já o gráfico de linhas apresenta a série temporal da informação indicada no seu título. Essa série pode ser agrupada por dia, semana, mês ou ano.

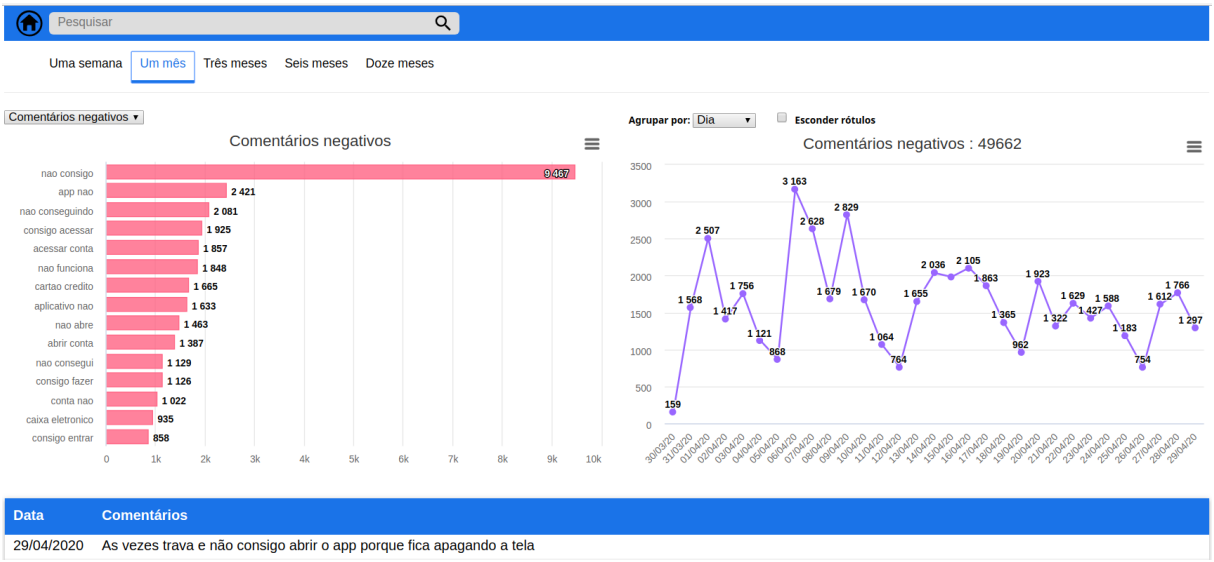


Data	Comentários
29/04/2020	É um bom app, porém, nunca consegui fazer um DOC ou TED, sempre fica fechando o app e pedindo feedback.
29/04/2020	As vezes trava e não consigo abrir o app porque fica apagando a tela

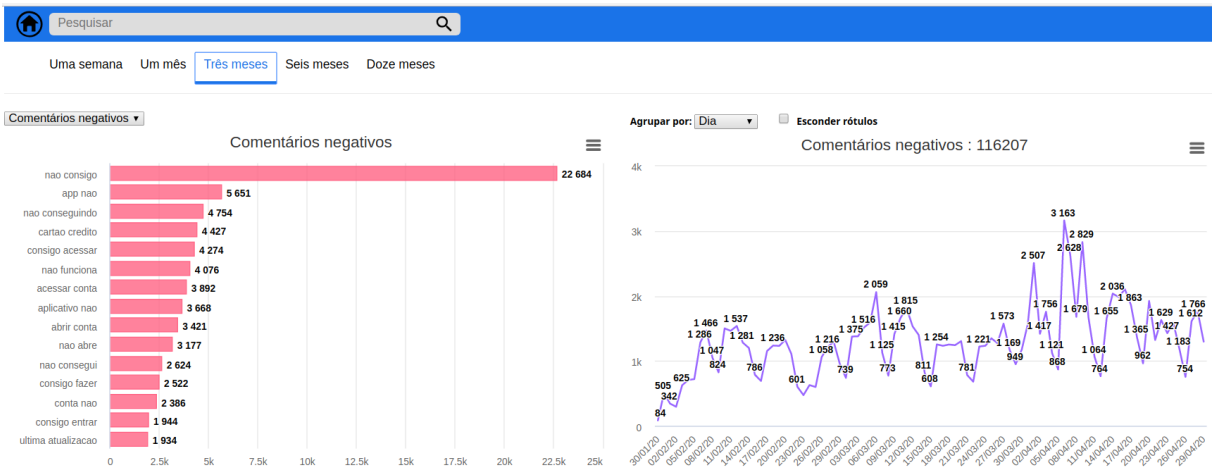


Data	Comentários
29/04/2020	Muito bom mesmo

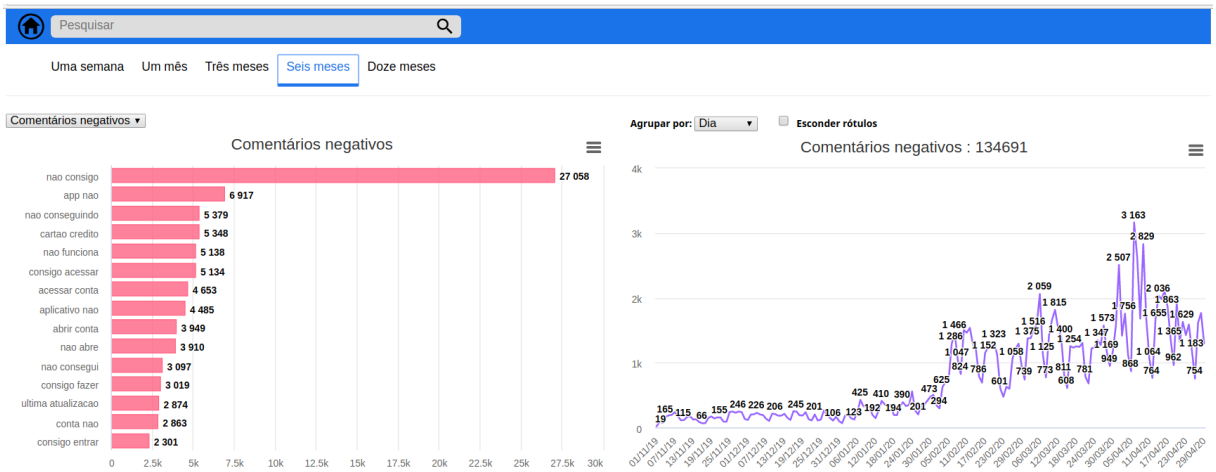
É possível navegar por períodos predeterminados. Isso permite que se identifique problemas e tendências. As visões se ajustam automaticamente aos novos dados.



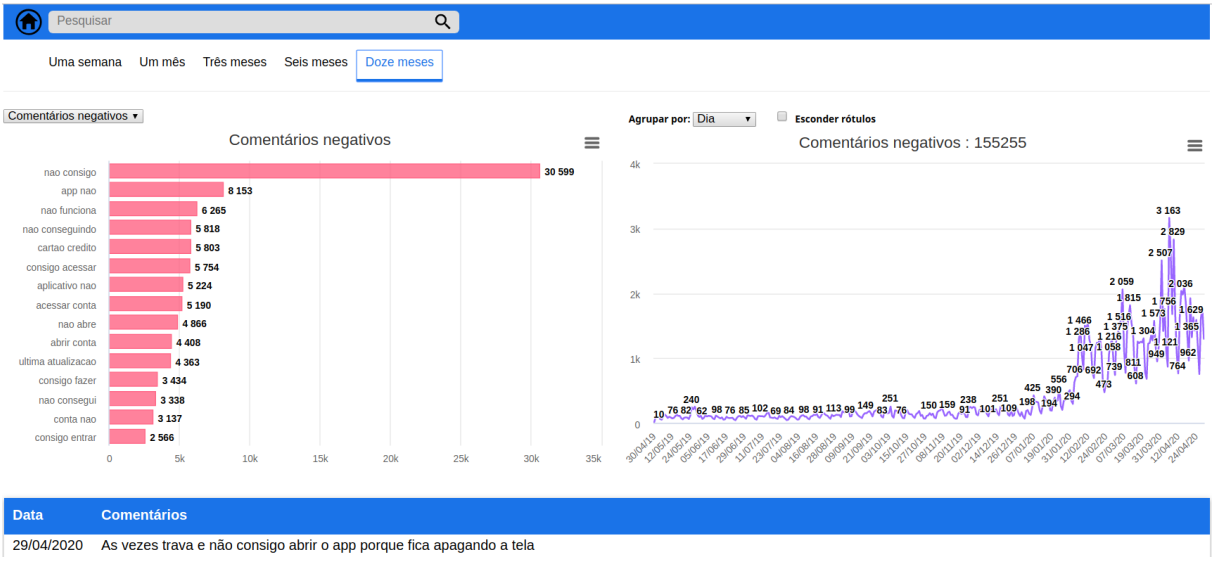
Data	Comentários
29/04/2020	As vezes trava e não consigo abrir o app porque fica apagando a tela



Data	Comentários
29/04/2020	As vezes trava e não consigo abrir o app porque fica apagando a tela

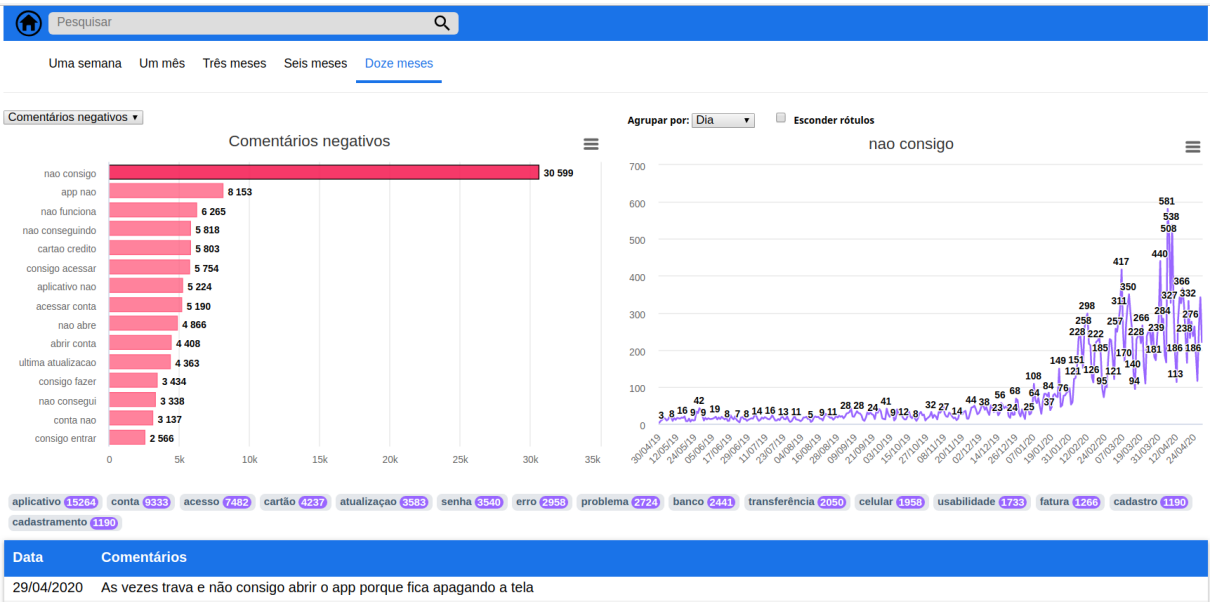


Data	Comentários
29/04/2020	As vezes trava e não consigo abrir o app porque fica apagando a tela



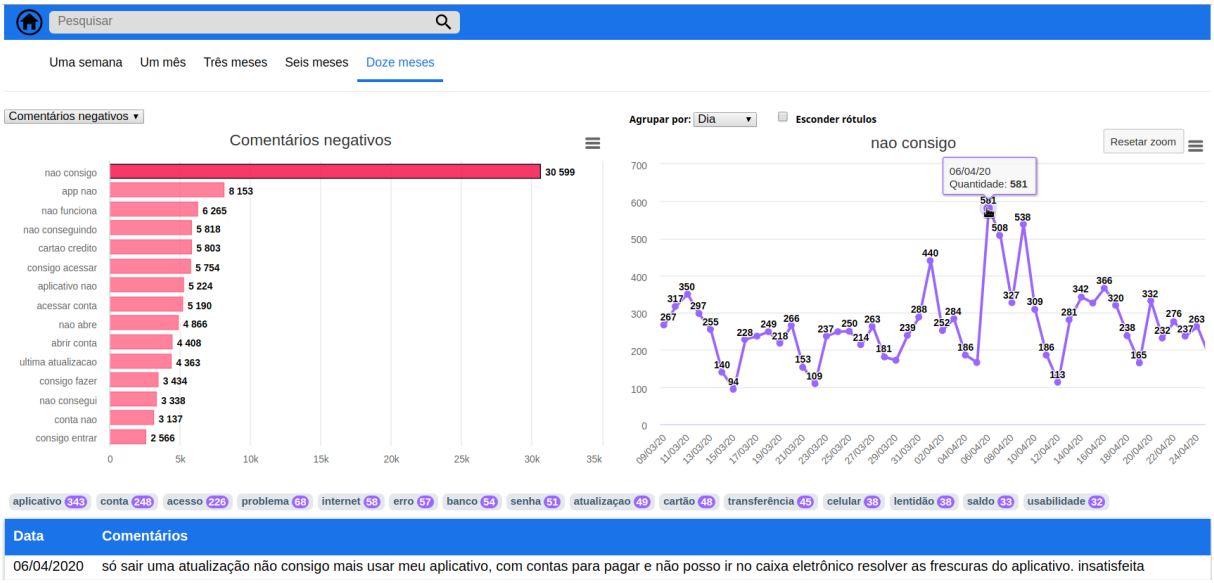
Data	Comentários
29/04/2020	As vezes trava e não consigo abrir o app porque fica apagando a tela

Os gráficos gerados são todos interativos e funcionam como controle para novas ações e filtros. Dessa forma podemos clicar sobre a barra relativa ao bigrama *não consigo* que vai gerar uma nova série temporal e mostrar as tags a ele associadas.



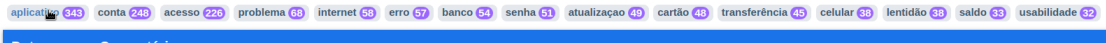
Data	Comentários
29/04/2020	As vezes trava e não consigo abrir o app porque fica apagando a tela

Podemos ainda clicar sobre o dia com mais menções a esse bigrama (06/04/2020) o que fará um novo filtro das tags para o período selecionado.



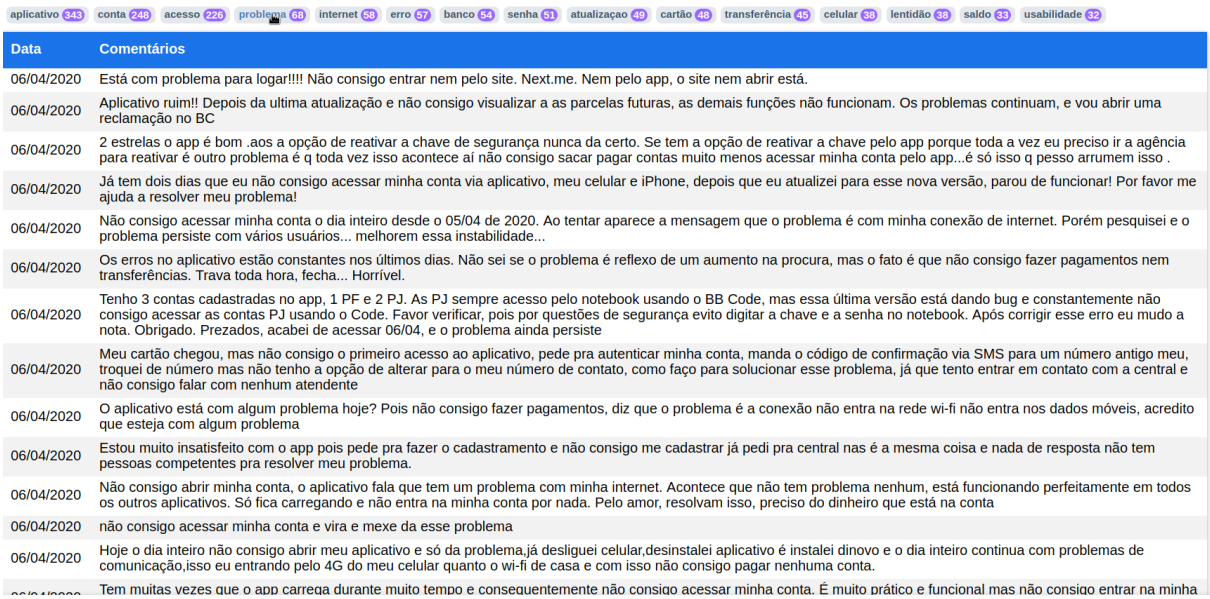
Data	Comentários
06/04/2020	só sair uma atualização não consigo mais usar meu aplicativo, com contas para pagar e não posso ir no caixa eletrônico resolver as frescuras do aplicativo. insatisfeita

Ao clicar sobre uma palavra-chave específica, os comentários a ela associados também mudarão na tabela. Para esse dia específico verificamos que existem 343 menções à palavra *aplicativo*. E ao clicarmos sobre ela é realizado mais um filtro sobre os comentários exibidos.



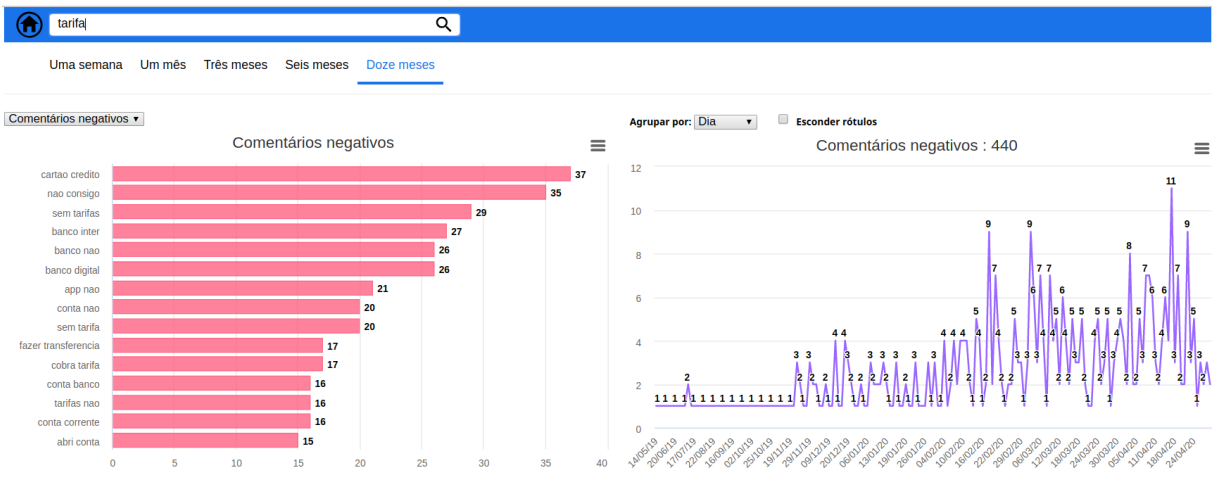
Data	Comentários
06/04/2020	só sair uma atualização não consigo mais usar meu aplicativo, com contas para pagar e não posso ir no caixa eletrônico resolver as frescuras do aplicativo. insatisfeita
06/04/2020	Eu não consigo baixa o app do banco inter, pq estou bloqueada, eu liguei para a central, me falaró que eu não posso ter mais acesso, ao app, eu vou fazer o que com o cartão? Eu não consigo ver o meu limite, o extrato, quanto eu vou gastar, quando vai vencer o cartão, desse jeito fica rui, eu tenho o cartão mais não vou utiliza, por causa disso.
06/04/2020	Eu tenho conta corrente no banco baixei o app mas não consigo fazer quase nada fala que tem que ir no caixa eletrônico do BB pra poder cadastrar o número para ter acesso aos serviços já fui 2 vezes e não consegui e muito complicado e não tinha ninguém no momento para auxiliar enfim não consigo usar só ver o saldo mesmo .
06/04/2020	Esse botão azul com o logo do facebook na barra inferior está dando MUITA AGONIA. Não consigo ficar com o aplicativo aberto por 1 minuto sequer. Desfaçam isso!!!
06/04/2020	Fazem 03 dias que busco código liberar o aplicativo SMS, não consigo e também sem apoio, preciso pagar minhas faturas.
06/04/2020	Eu não consigo acessá da erro na senha oque devo fazer depois avalio o App se Eu conseguir acessá e claro
06/04/2020	Eu ate gostava do app mas depois ficou péssimo, ja tem 2 meses que eu não consigo ver meu limite, e quando eu entro na opção ver limite diz que eu não cartão, sendo que eu tenho ,faço o pagamento todo mes certo e no mes passado eu fui fazer uma compra e não consegui
06/04/2020	Como mudar a senha desse App? Tô há dias tentando e ele simplesmente não abre e só trava, aparece senha bloqueada mas não consigo mais nem abrir ? Nem definir a senha. Consigo fazer nada. Alguém sabe o que fazer?
06/04/2020	Não consigo fazer nada com esse app
06/04/2020	Não atualiza saldo não consigo pagar cartão de crédito tá virado em nada o App está muito ruim
06/04/2020	Depois dessa última atualização não consigo mais abrir o app Está horrível
06/04/2020	Não consigo fazer nada pelo app..na hora de transferir ou pagar fatura o app para de funcionar... já atualizei e nada. Também já desinstalei e tive que instalar o itoken dnv e não adiantou nada. Vou ter que sair de casa para pagar minhas contas?
06/04/2020	Está com problema para logar!!!! Não consigo entrar nem pelo site. Next.me. Nem pelo app, o site nem abrir está.
06/04/2020	O App deve ser ótimo, mas não consigo baixar ele ! Minha internet não presta.
06/04/2020	Desde ontem não consigo acessar minha conta. Nem pelo app nem pelo computador! O que está acontecendo caixa?! Tenho pagamentos e transferências a fazer! Em época de corona vírus o internet banking é essencial. Porque vcs estão fazendo isso conosco?!
06/04/2020	Depois da atualização ficou muito ruim, eu não consigo redefinir a senha de acesso do aplicativo, só fica aparecendo dados inválidos
06/04/2020	Nao consigo fazer login nao conclui meu cadastro alguém me ajude preciso ativar o itoken e cadastrar meu cel no app tem pessoas que conheço deu certo mas tentei varias maneiras e nada ate agora 😬

Se selecionarmos outra tag, os comentários serão filtrados novamente.

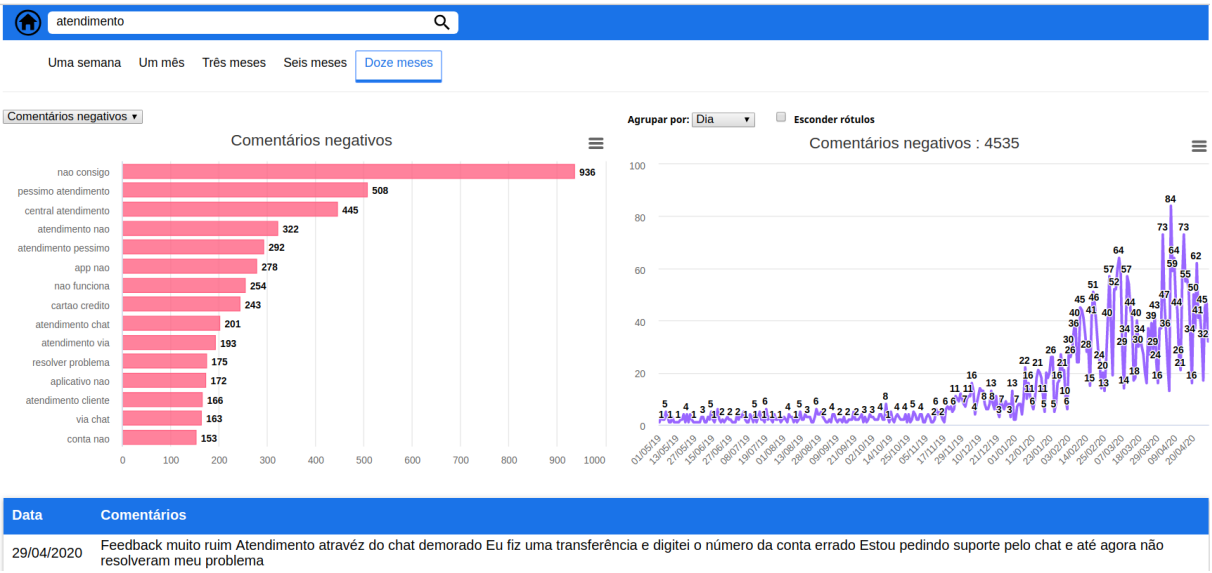


Data	Comentários
06/04/2020	Está com problema para logar!!!! Não consigo entrar nem pelo site. Next.me. Nem pelo app, o site nem abrir está.
06/04/2020	Aplicativo ruim!! Depois da ultima atualização e não consigo visualizar a as parcelas futuras, as demais funções não funcionam. Os problemas continuam, e vou abrir uma reclamação no BC
06/04/2020	2 estrelas o app é bom .aos a opção de reativar a chave de segurança nunca da certo. Se tem a opção de reativar a chave pelo app porque toda a vez eu preciso ir a agência para reativar é outro problema é q toda vez isso acontece aí não consigo sacar pagar contas muito menos acessar minha conta pelo app...é só isso q pesso arrumem isso .
06/04/2020	Já tem dois dias que eu não consigo acessar minha conta via aplicativo, meu celular e iPhone, depois que eu atualizei para esse nova versão, parou de funcionar! Por favor me ajuda a resolver meu problema!
06/04/2020	Não consigo acessar minha conta o dia inteiro desde o 05/04 de 2020. Ao tentar aparece a mensagem que o problema é com minha conexão de internet. Porém pesquisei e o problema persiste com vários usuários... melhorem essa instabilidade...
06/04/2020	Os erros no aplicativo estão constantes nos últimos dias. Não sei se o problema é reflexo de um aumento na procura, mas o fato é que não consigo fazer pagamentos nem transferências. Trava toda hora, fecha... Horrível.
06/04/2020	Tenho 3 contas cadastradas no app, 1 PF e 2 PJ. As PJ sempre acesso pelo notebook usando o BB Code, mas essa última versão está dando bug e constantemente não consigo acessar as contas PJ usando o Code. Favor verificar, pois por questões de segurança evito digitar a chave e a senha no notebook. Após corrigir esse erro eu mudo a nota. Obrigado. Prezados, acabei de acessar 06/04, e o problema ainda persiste
06/04/2020	Meu cartão chegou, mas não consigo o primeiro acesso ao aplicativo, pede pra autenticar minha conta, manda o código de confirmação via SMS para um número antigo meu, troquei de número mas não tenho a opção de alterar para o meu número de contato, como faço para solucionar esse problema, já que tento entrar em contato com a central e não consigo falar com nenhum atendente
06/04/2020	O aplicativo está com algum problema hoje? Pois não consigo fazer pagamentos, diz que o problema é a conexão não entra na rede wi-fi não entra nos dados móveis, acredito que esteja com algum problema
06/04/2020	Estou muito insatisfeito com o app pois pede pra fazer o cadastramento e não consigo me cadastrar já pedi pra central nas é a mesma coisa e nada de resposta não tem pessoas competentes pra resolver meu problema.
06/04/2020	Não consigo abrir minha conta, o aplicativo fala que tem um problema com minha internet. Acontece que não tem problema nenhum, está funcionando perfeitamente em todos os outros aplicativos. Só fica carregando e não entra na minha conta por nada. Pelo amor, resolvam isso, preciso do dinheiro que está na conta
06/04/2020	não consigo acessar minha conta e vira e mexe da esse problema
06/04/2020	Hoje o dia inteiro não consigo abrir meu aplicativo e só da problema,já desliguei celular,desinstalei aplicativo é instalei dinovo e o dia inteiro continua com problemas de comunicação,isso eu entrando pelo 4G do meu celular quanto o wi-fi de casa e com isso não consigo pagar nenhuma conta.
[illegible]	Tem muitas vezes que o app carrega durante muito tempo e consequentemente não consigo acessar minha conta. É muito prático e funcional mas não consigo entrar na minha

Outra funcionalidade importante na plataforma é a busca. Isso permite que as visões sejam construídas a partir de um assunto específico como, por exemplo, *tarifa* ou *atendimento*.



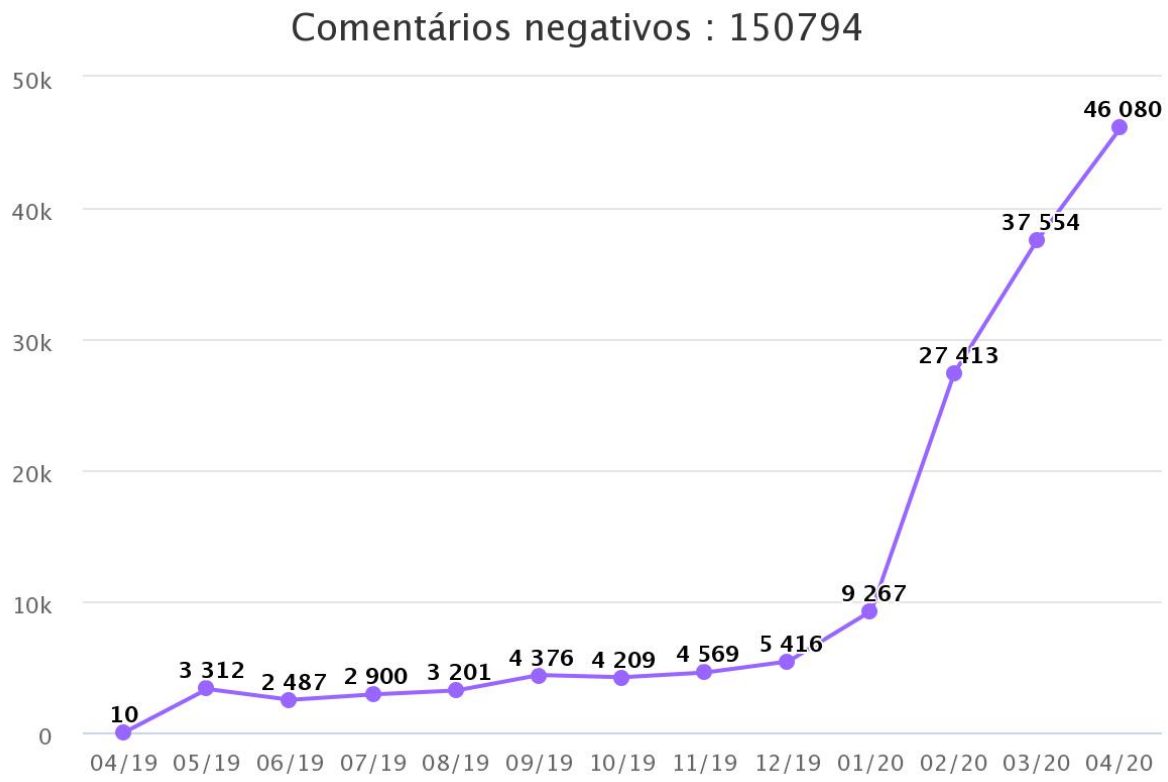
Data	Comentários
29/04/2020	Não gostei porque não consigo saber se meu cartão de crédito foi aprovado ou não é não consigo saber se essa conta que abro é corrente ou poupança e se tem tarifas não especifica nada muito ruim!!



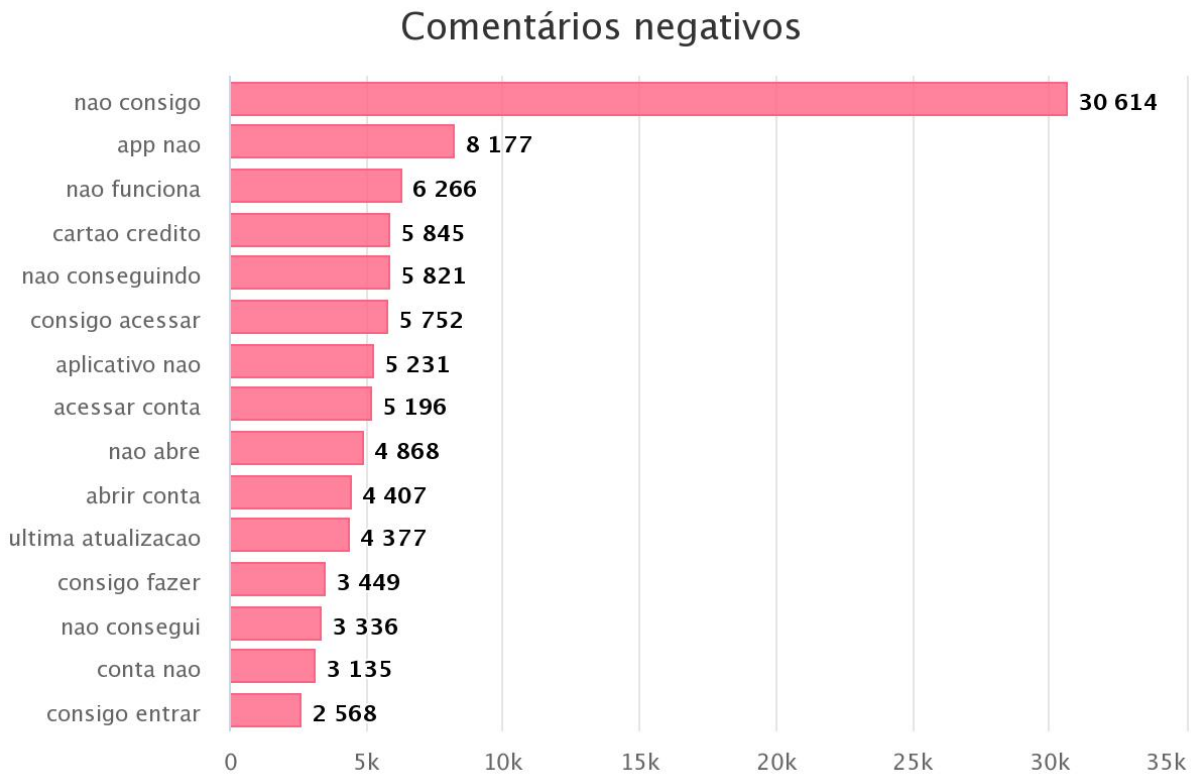
Data	Comentários
29/04/2020	Feedback muito ruim Atendimento atravéz do chat demorado Eu fiz uma transferência e digitei o número da conta errado Estou pedindo suporte pelo chat e até agora não resolveram meu problema

Após explorar essa plataforma de text mining pudemos observar que:

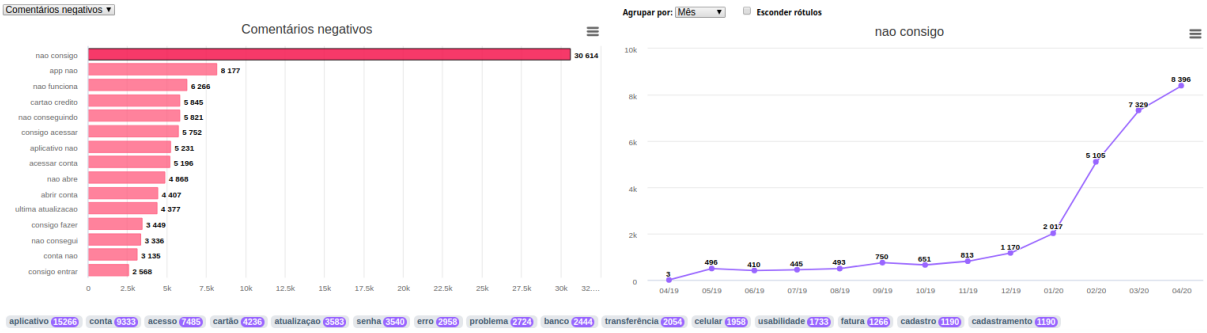
- A insatisfação tem aumentado nos últimos doze meses;



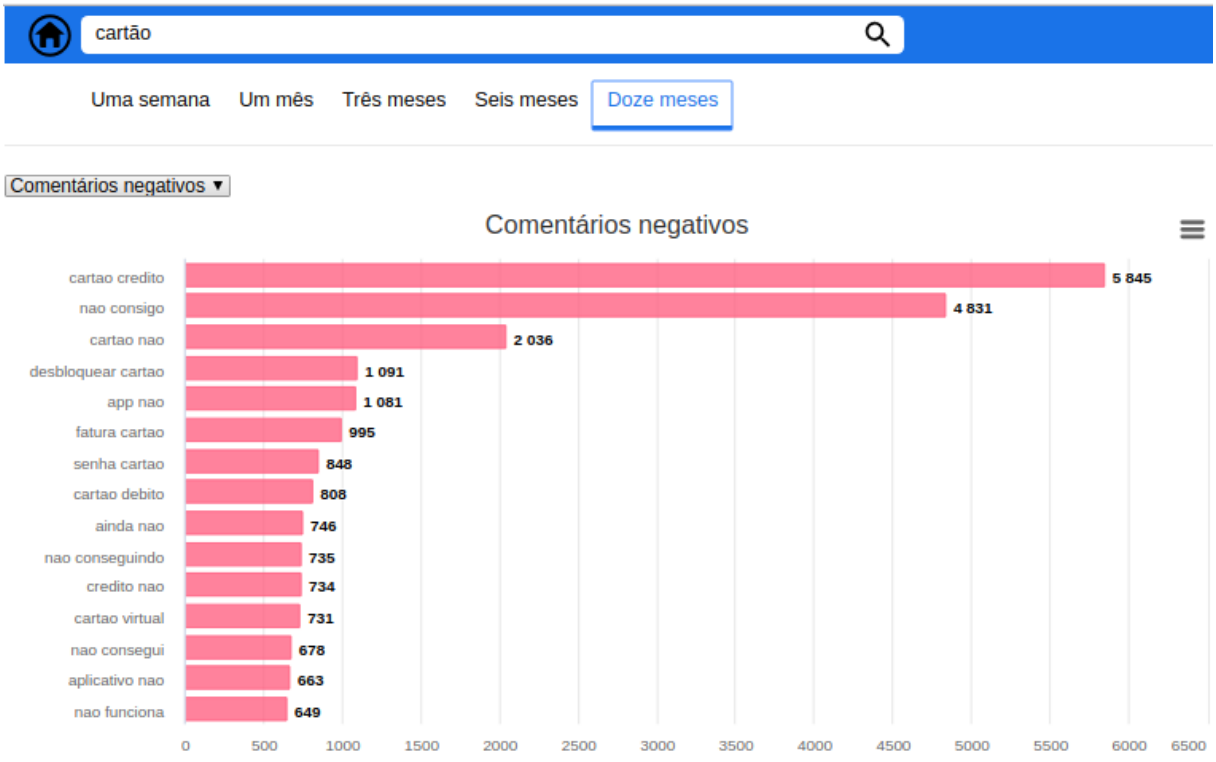
- Os usuários reclamam bastante de não conseguir fazer alguma coisa;
- O produto mais reclamado é o *cartão de crédito*;



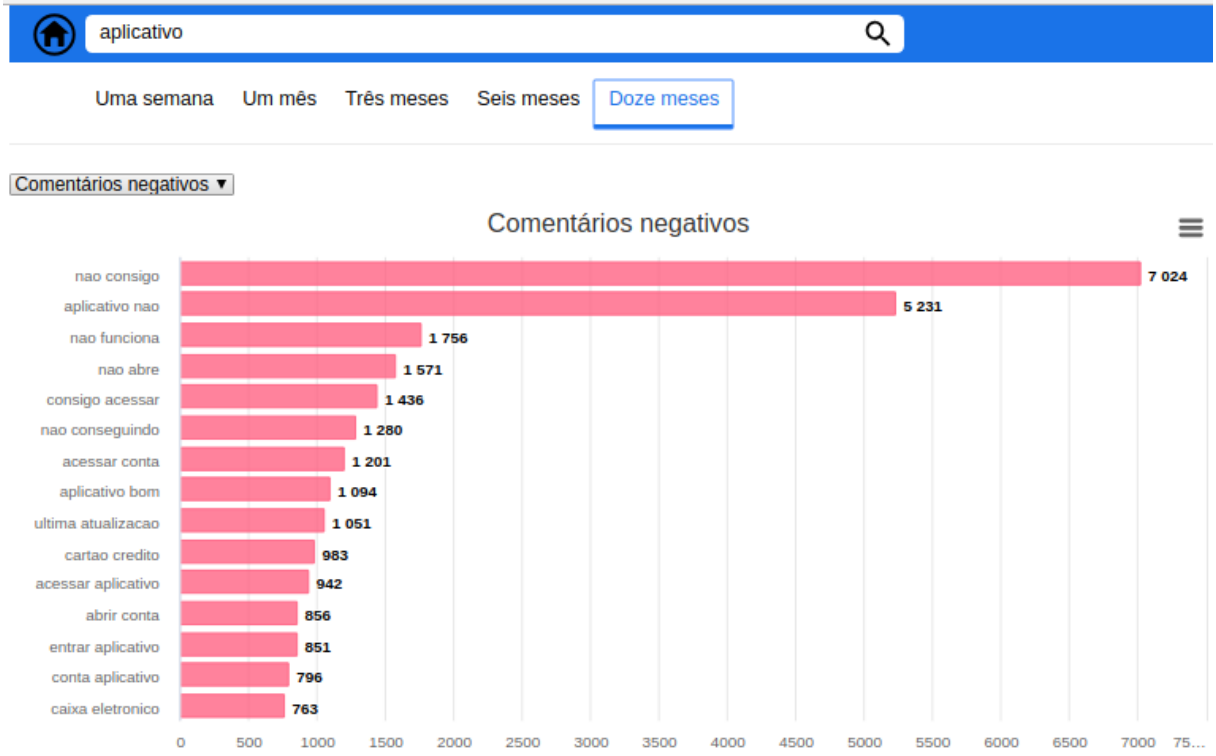
- As tags *aplicativo*, *acesso*, *senha*, *erro*, *problema e usabilidade* são frequentemente associadas ao bigrama *não consigo*;



- Ao pesquisar sobre *cartão* percebemos que é comum ter problemas relacionados a desbloqueio, fatura e senha.



• Ao pesquisar sobre *aplicativo* os bigramas *não consigo*, *não funciona e não abre* se destacam.



7. Conclusão

Pudemos verificar através desse trabalho que os clientes/usuários estão dispostos a informar as suas dores e experiência com o serviço prestado pela sua instituição financeira e isso tudo acontece de forma espontânea.

Percebemos que muita informação pode ser retirada de dados não estruturado, comentários, ao aplicarmos técnicas de text mining.

E por fim deixamos a recomendação mais importante que é ouvir o que o cliente tem a dizer, e se propor a resolver os problemas apontados da forma mais célere possível.

8. Links

Todos os scripts usados nesse projeto estão disponíveis no repositório https://github.com/src77/nlp-puc-minas.

Todos os arquivos de reviews extraídos do site Reclame Aqui e Apptweak, o modelo word2vec, o arquivo all_reviews.zip necessário para o funcionamento da aplicação web e o vídeo de apresentação estão disponíveis através do link https://1drv.ms/u/s!AuxviEfASIowhbVtOLmKYpHX1OSjUQ?e=9qLlfR.

REFERÊNCIAS

API Reference. **Express**, 2010. Disponível em: <https://expressjs.com/en/4x/api.html>. Acesso em: 01 de julho de 2019.

Documentation. **FastAPI**, 2018. Disponível em: <https://fastapi.tiangolo.com/>. Acesso em: 01 de julho de 2019.

JS API Reference. **Highcharts**, 2009. Disponível em: <https://api.highcharts.com/highcharts/>. Acesso em: 01 de julho de 2019.

KARANI, Dhruvil. Introduction to Word Embedding and Word2Vec. **Medium**, 2018. Disponível em: <https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa>. Acesso em: 20 de out. de 2019.

MORDE, Vishal. XGBoost Algorithm: Long May She Reign. **Medium**, 2019. Disponível em: <https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d>. Acesso em: 20 de out. de 2019.

Nodejs Documentation. **Nodejs**, 2009. Disponível em: <https://nodejs.org/en/docs/>. Acesso em: 01 de julho de 2019.

Pandas Documentation. **Pandas**, 2008. Disponível em: <https://pandas.pydata.org/docs/>. Acesso em: 01 de julho de 2019.

XGBoost Documentation. **XGBoost**, 2016. Disponível em: <https://xgboost.readthedocs.io/en/latest/>. Acesso em: 15 de jan. de 2020.