Customer Support on Twitter

By : Tawaah Alhaddad

INTRODUCTION

In this project, the study will focus on modern customer support practices on one of the most popular social media platforms 'Twitter' for Over 3 million tweets and replies from the biggest brands and companies its impact. This will give a wide view to enhance my job as customer service representative in a mall which customer support on twitter is new filed to my company.

Original Dataset

```
full_df = pd.read_csv("twcs.csv", nrows=10000)
df = full_df[["text"]]
df["text"] = df["text"].astype(str)
full_df.head()
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
Requirement already satisfied: pyspellchecker in /usr/local/lib/python3.7/dist-packages (0.6.2)
```

	tweet_id	author_id	inbound	created_at	text	response_tweet_id	in_response_to_tweet_id
0	1	sprintcare	False	Tue Oct 31 22:10:47 +0000 2017	@115712 I understand. I would like to assist y...	2	3.0
1	2	115712	True	Tue Oct 31 22:11:45 +0000 2017	@sprintcare and how do you propose we do that	NaN	1.0
2	3	115712	True	Tue Oct 31 22:08:27 +0000 2017	@sprintcare I have sent several private messag...	1	4.0
3	4	sprintcare	False	Tue Oct 31 21:54:49 +0000 2017	@115712 Please send us a Private Message so th...	3	5.0
4	5	115712	True	Tue Oct 31 21:49:35 +0000 2017	@sprintcare I did.	4	6.0

DATA PREPARING

Some of the common text preprocessing / cleaning steps are:

- Lower casing
- Removal of Punctuations
- Removal of Stopwords
- Removal of Frequent words
- Removal of Rare words
- Conversion of emoticons to words
- Conversion of emojis to words
- Removal of URLs
- Chat words conversion
- Spelling correction

DATASET LINK

https://www.kaggle.com/thoughtvector/customer-support-on-twitter/code?datasetId=4133&sortBy=voteCount

Data after processing

```
full_df = pd.read_csv("twcs.csv", nrows=10000)
df = full_df[["text"]]
df["text"] = df["text"].astype(str)
full_df.head()
```

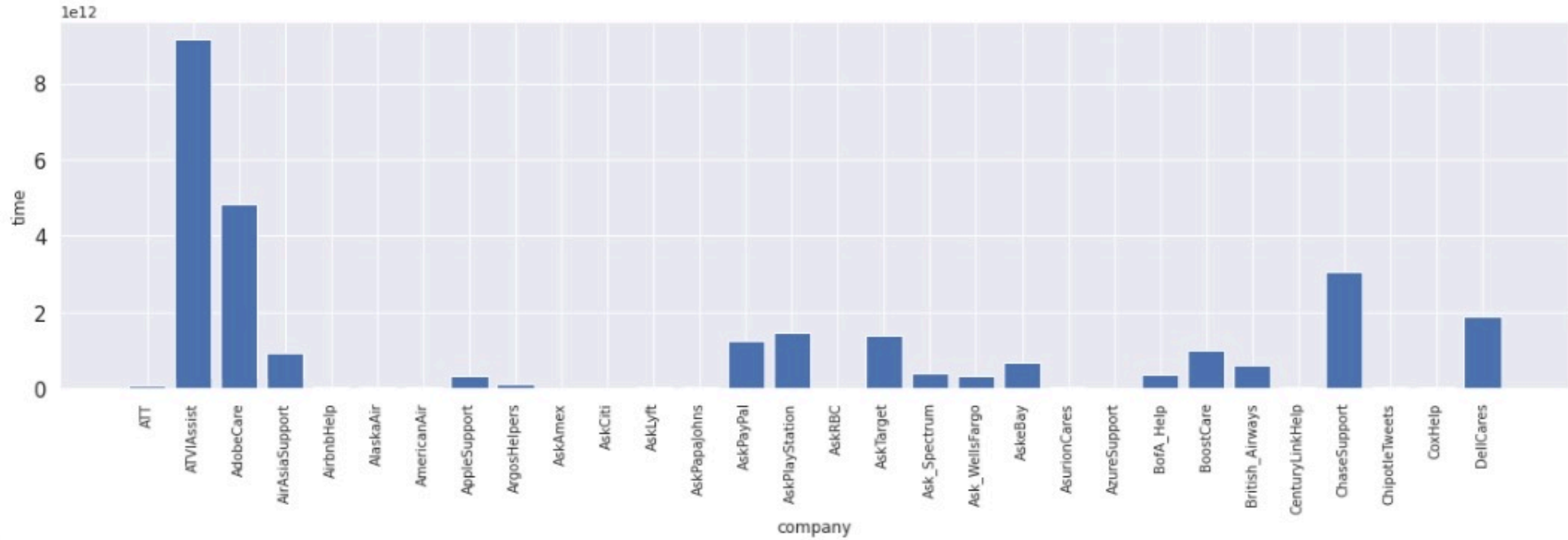
```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
Requirement already satisfied: pyspellchecker in /usr/local/lib/python3.7/dist-packages (0.6.2)
```

	tweet_id	author_id	inbound	created_at	text	response_tweet_id	in_response_to_tweet_id
0	1	sprintcare	False	Tue Oct 31 22:10:47 +0000 2017	@115712 I understand. I would like to assist y...	2	3.0
1	2	115712	True	Tue Oct 31 22:11:45 +0000 2017	@sprintcare and how do you propose we do that	NaN	1.0
2	3	115712	True	Tue Oct 31 22:08:27 +0000 2017	@sprintcare I have sent several private messag...	1	4.0
3	4	sprintcare	False	Tue Oct 31 21:54:49 +0000 2017	@115712 Please send us a Private Message so th...	3	5.0
4	5	115712	True	Tue Oct 31 21:49:35 +0000 2017	@sprintcare I did.	4	6.0

Answering

EDA Questions

01 The max time of delay for each company



EDA Questions

02 The time difference between the customer tweet and the company respond.

```
query_date = pd.to_datetime(Q_inrespond['created_at_x'])
respond_date = pd.to_datetime(Q_inrespond['created_at_y'])


Q_inrespond['DifferenceofTime']= pd.to_timedelta(respond_date - query_date)/ 60

Q_inrespond['DifferenceofTime']
```

```
0       0 days 00:00:01.233333333
1          0 days 00:00:03.200000
2          0 days 00:00:06.600000
3       0 days 00:00:04.033333333
4          0 days 00:00:05.250000
                    ...
2795       0 days 00:00:04.100000
2796       0 days 00:04:43.500000
2797    0 days 00:00:09.066666666
2799       0 days 00:08:25.200000
2800    0 days 00:01:10.266666666
Name: DifferenceofTime, Length: 2327, dtype: timedelta64[ns]
```

Answering

EDA Questions

03 Finding the subjectivity and polarity analysis for tweets.

	author_id_x	created_at_x	text_x	author_id_y	created_at_y	text_y	Sujectivity	Polarity	Sentiment	Analysis
0	115712	Tue Oct 31 21:45:10 +0000 2017	sprintcare worst customer service	sprintcare	Tue Oct 31 21:46:24 +0000 2017	send private message gain details account	1.000000	-1.000000	0	Negative
1	115713	Tue Oct 31 19:56:01 +0000 2017	y'all lie "great" connection bars lte still wo...	sprintcare	Tue Oct 31 19:59:13 +0000 2017	h wed definitely like work long experiencing i...	0.750000	0.800000	1	Positive
2	115715	Tue Oct 31 22:03:34 +0000 2017	whenever contact customer support tell shortco...	sprintcare	Tue Oct 31 22:10:10 +0000 2017	send private message send link access account fr	0.000000	0.000000	1	Natural
3	115716	Tue Oct 31 22:01:35 +0000 2017	actually thats broken link sent incorrect info...	Ask_Spectrum	Tue Oct 31 22:05:37 +0000 2017	information pertaining account assumption corr...	0.250000	-0.200000	0	Negative
4	115717	Tue Oct 31 22:06:54 +0000 2017	yo askspectrum customer service reps super nic...	Ask_Spectrum	Tue Oct 31 22:12:09 +0000 2017	hello apologies frustrations inconvenience i'd...	0.666667	0.333333	1	Positive
...	...	...	...	...	...	...	...	...	...	...
1329	117287	Wed Nov 22 08:12:16 +0000 2017	lagos nigeria week tmobile data work nigeria a...	TMobileHelp	Wed Nov 22 11:13:26 +0000 2017	knowing kind coverage expect traveling key wan...	0.566667	-0.066667	0	Negative
1330	117288	Wed Nov 22 06:59:32 +0000 2017	heres experience customer years amp bought pix...	TMobileHelp	Wed Nov 22 11:06:58 +0000 2017	limited time offer ended highlighted specifies...	0.527273	0.418182	1	Positive
1331	117289	Mon Oct 30 02:15:59 +0000 2017	dear new update sucks it's insanely hard navig...	hulu_support	Tue Oct 31 23:43:06 +0000 2017	isnt way roll back working making changes navi...	0.324053	-0.113826	0	Negative
1332	117290	Mon Oct 30 02:13:57 +0000 2017	guys maybe stuff longest time captions lasts o...	hulu_support	Tue Oct 31 23:41:54 +0000 2017	uh oh captions stay device happening specific ...	0.000000	0.000000	1	Natural
1333	117291	Mon Oct 30 01:57:36 +0000 2017	platform available specifically argentina	hulu_support	Tue Oct 31 23:40:03 +0000 2017	hulu available right well sure share interest ...	0.400000	0.400000	1	Positive

1104 rows x 10 columns

Imbalanced classes

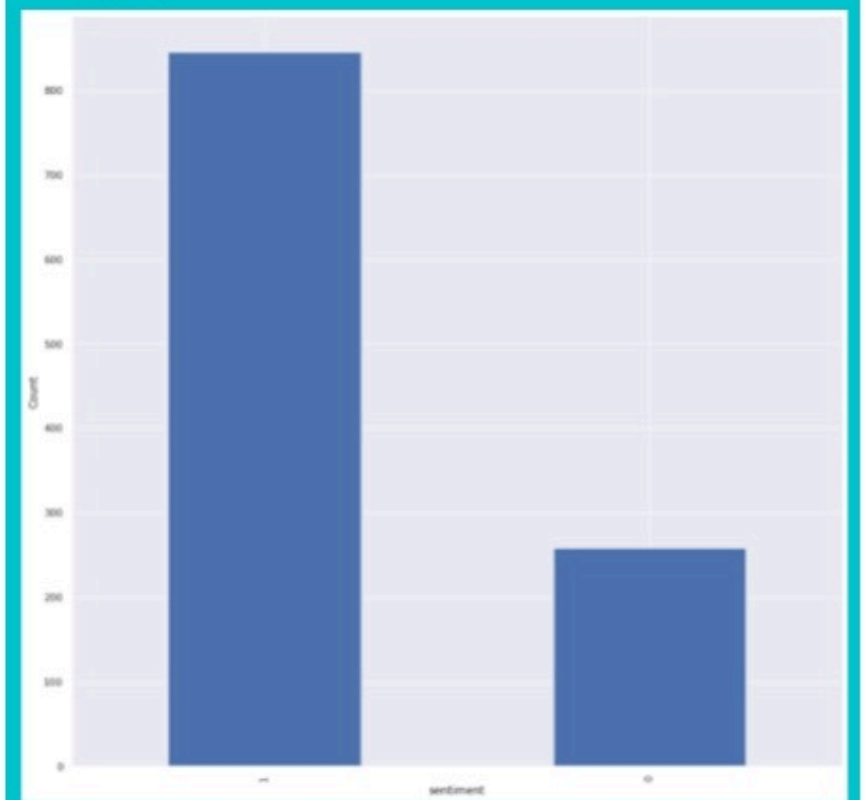
Discovering imbalance in dataset will be dealt with in two ways

#1

```
[34] #discovering imbalance in dataset
     Q_inrespond['Sentiment'].value_counts()

     1    846
     0    258
     Name: Sentiment, dtype: int64
```

#2



Algorithms

- Due the huge size of the dataset over 3M and the limited resources, I had to reduce the size of observation to 5000 and restricted the features to 1200.
- I have used only the tweets and sentiment columns.
- I have used logistic regression and Naive Bayes.
- I have used CountVectorizer and TF-IDF as word embedding and Create a logistic regression model to compare their performances. And run the experiment on imbalanced and balanced classes using SMOTE classes.

Result

I chose Logistic Regression with TFIDF balanced

	LR_CV	LR1-TFIDF	LRCV-balanced_class_smote	LRTFIDF-balanced_class_smote	Bayes_balanced_smote
Accuracy	0.810	0.760	0.679	0.819	0.756
Precision	0.805	0.761	0.829	0.843	0.838
Recall	0.988	0.994	0.725	0.934	0.838
F1 Score	0.887	0.862	0.774	0.886	0.838

THANK YOU!

Tawaah Alhaddad

tawaah@outlook.com