Capstone Project-3

Predicting the News Popularity in Multiple Social Media Platforms

Content

- The Problem Statement
- The EDA
- Text Processing
- Preparing Dataset
- Model Evaluation
- Future Work

shutterstock.com • 642070846

Problem Statement

This is a large data set of news items and their respective social feedback on multiple platforms: Facebook, Google+ and LinkedIn.The collected data relates to a period of 8 months, between November 2015 and July 2016, accounting for about 100,000 news items on four different topics: Economy, Microsoft, Obama and Palestine.

Data Encapsulation

IDLink (numeric): Unique identifier of news items

Title (string): Title of the news item according to the official media sources

Headline (string): Headline of the news item according to the official media sources

Source (string): Original news outlet that published the news item

Topic (string): Query topic used to obtain the items in the official media sources

PublishDate (timestamp): Date and time of the news items' publication

SentimentTitle (numeric): Sentiment score of the text in the news items' title

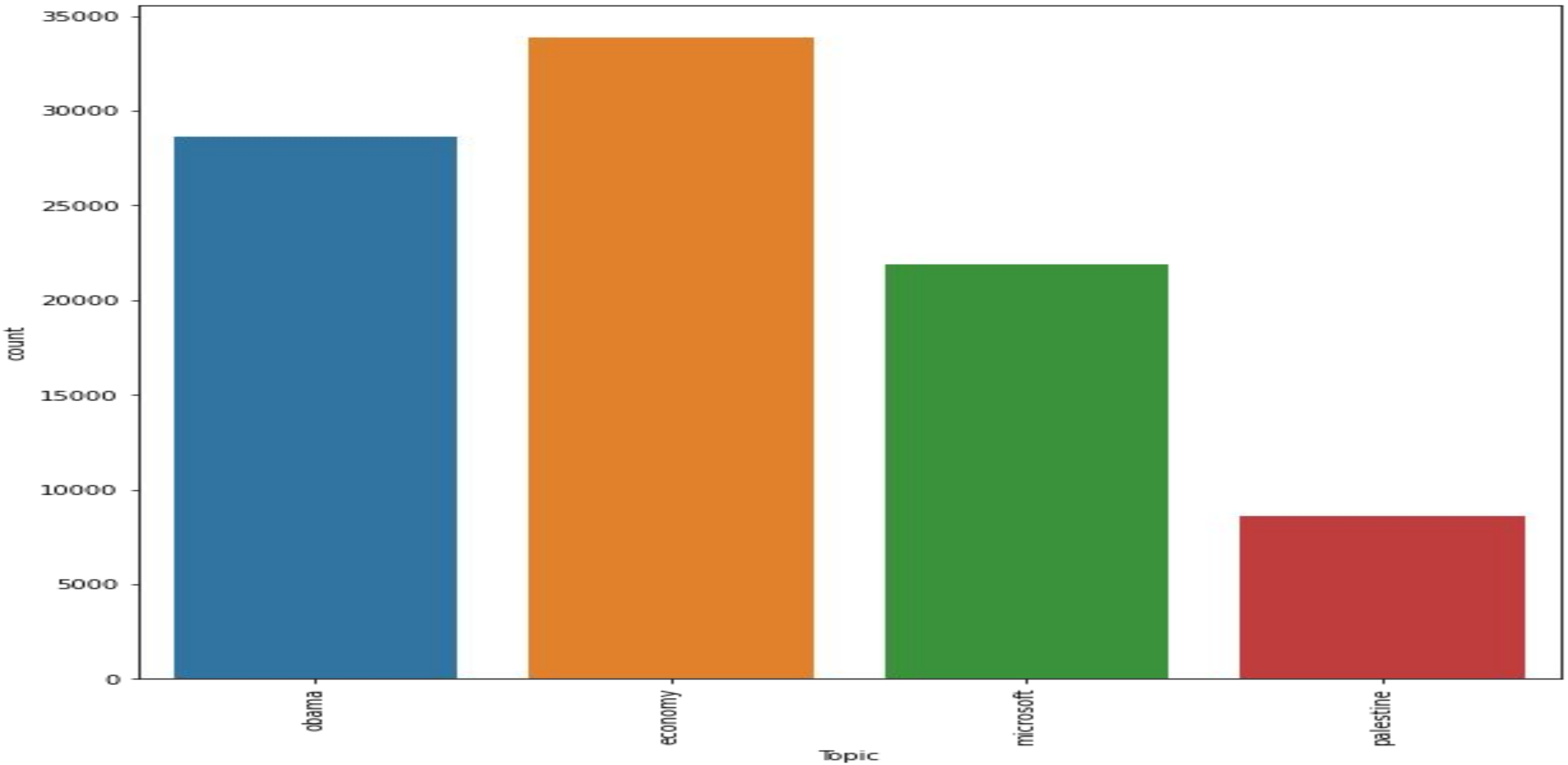
SentimentHeadline (numeric): Sentiment score of the text in the news items' headline

Facebook (numeric): Final value of the news items' popularity according to the social media source Facebook

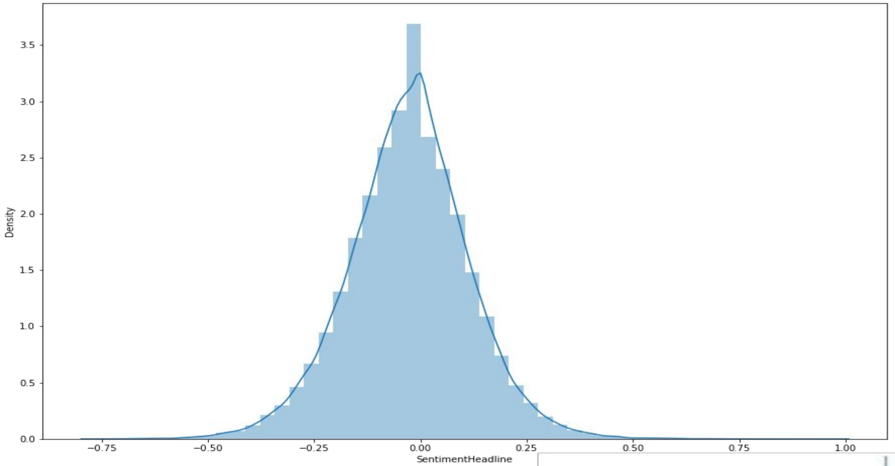
GooglePlus (numeric): Final value of the news items' popularity according to the social media source Google+

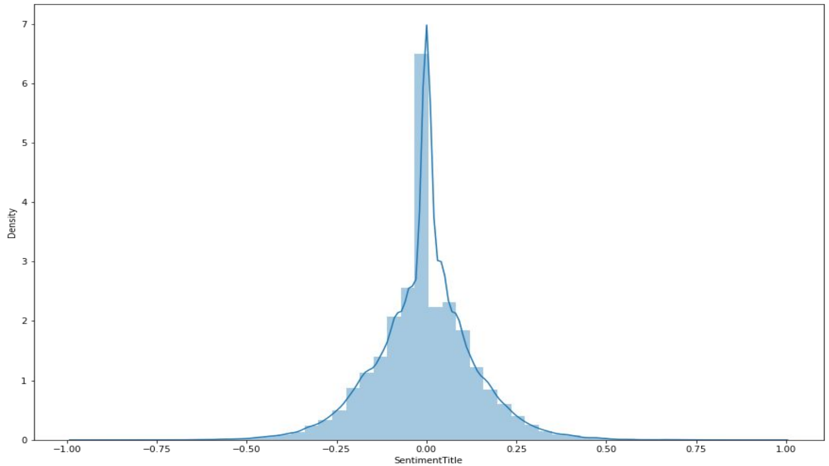
LinkedIn (numeric): Final value of the news items' popularity according to the social media source LinkedIn

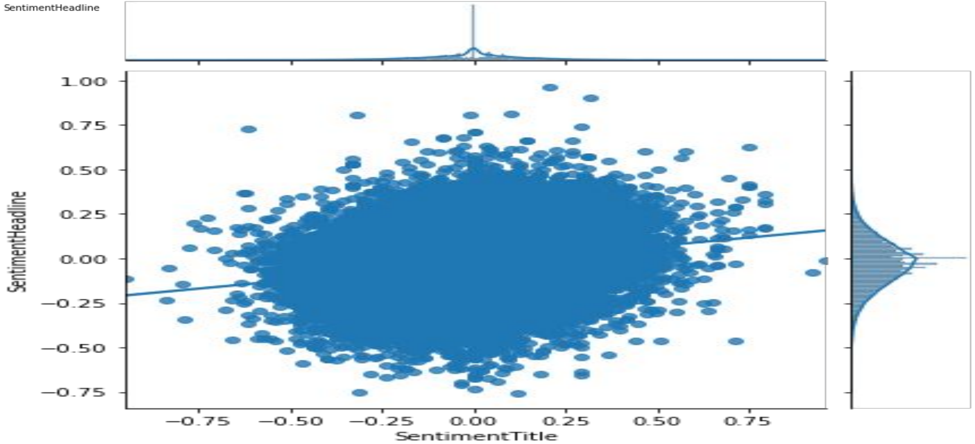
The EDA


35000
30000
25000
20000
15000
10000
5000
0
count
obama
economy
microsoft
palestine
Topic

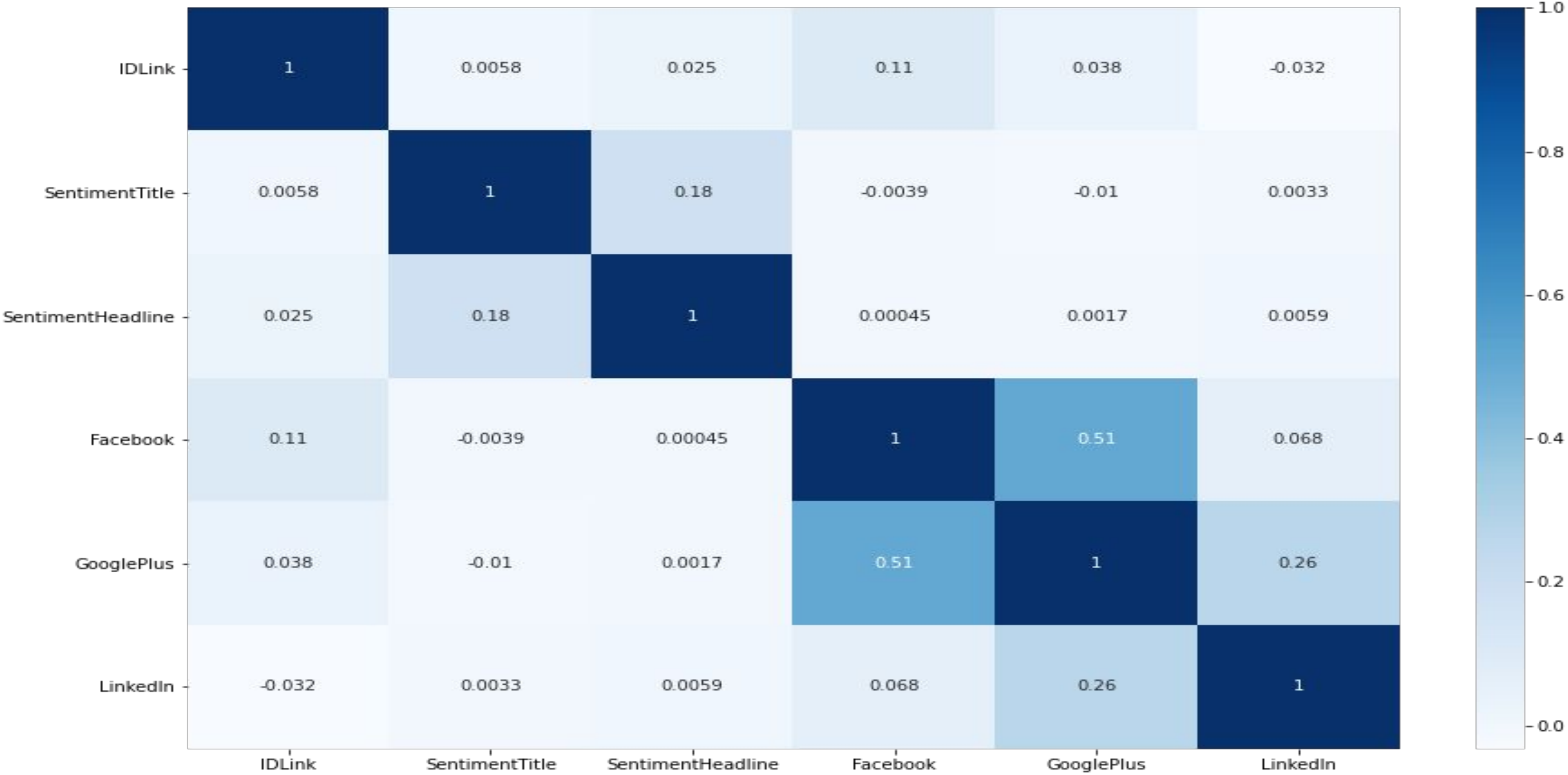
The EDA(Continue)





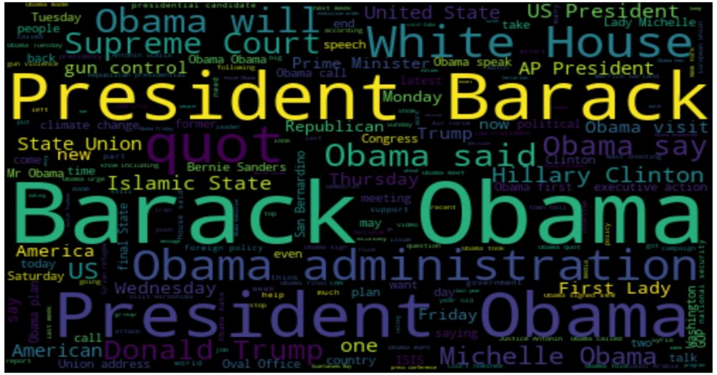


The EDA (Continue)



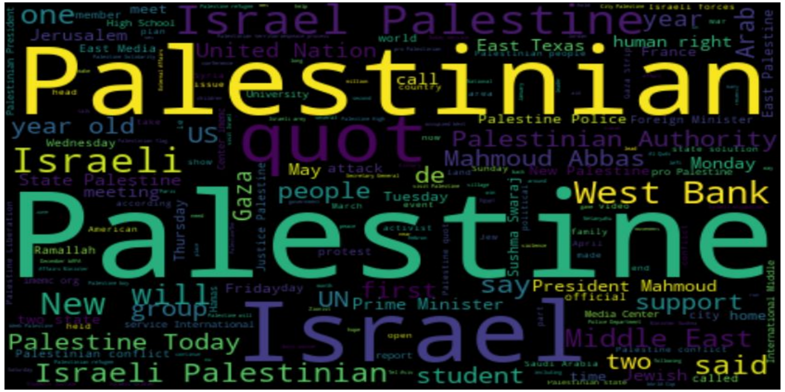
	IDLink	SentimentTitle	SentimentHeadline	Facebook	GooglePlus	LinkedIn
IDLink	1	0.0058	0.025	0.11	0.038	-0.032
SentimentTitle	0.0058	1	0.18	-0.0039	-0.01	0.0033
SentimentHeadline	0.025	0.18	1	0.00045	0.0017	0.0059
Facebook	0.11	-0.0039	0.00045	1	0.51	0.068
GooglePlus	0.038	-0.01	0.0017	0.51	1	0.26
LinkedIn	-0.032	0.0033	0.0059	0.068	0.26	1

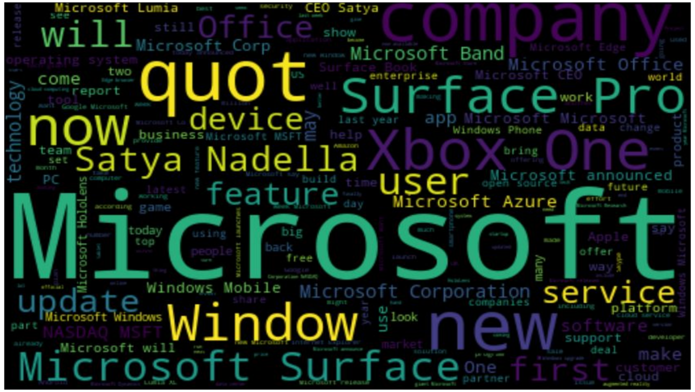
The EDA(Continue)



Obama

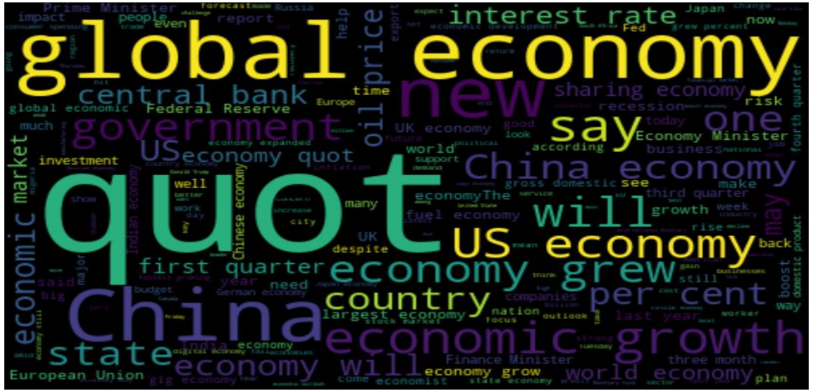
Palestine



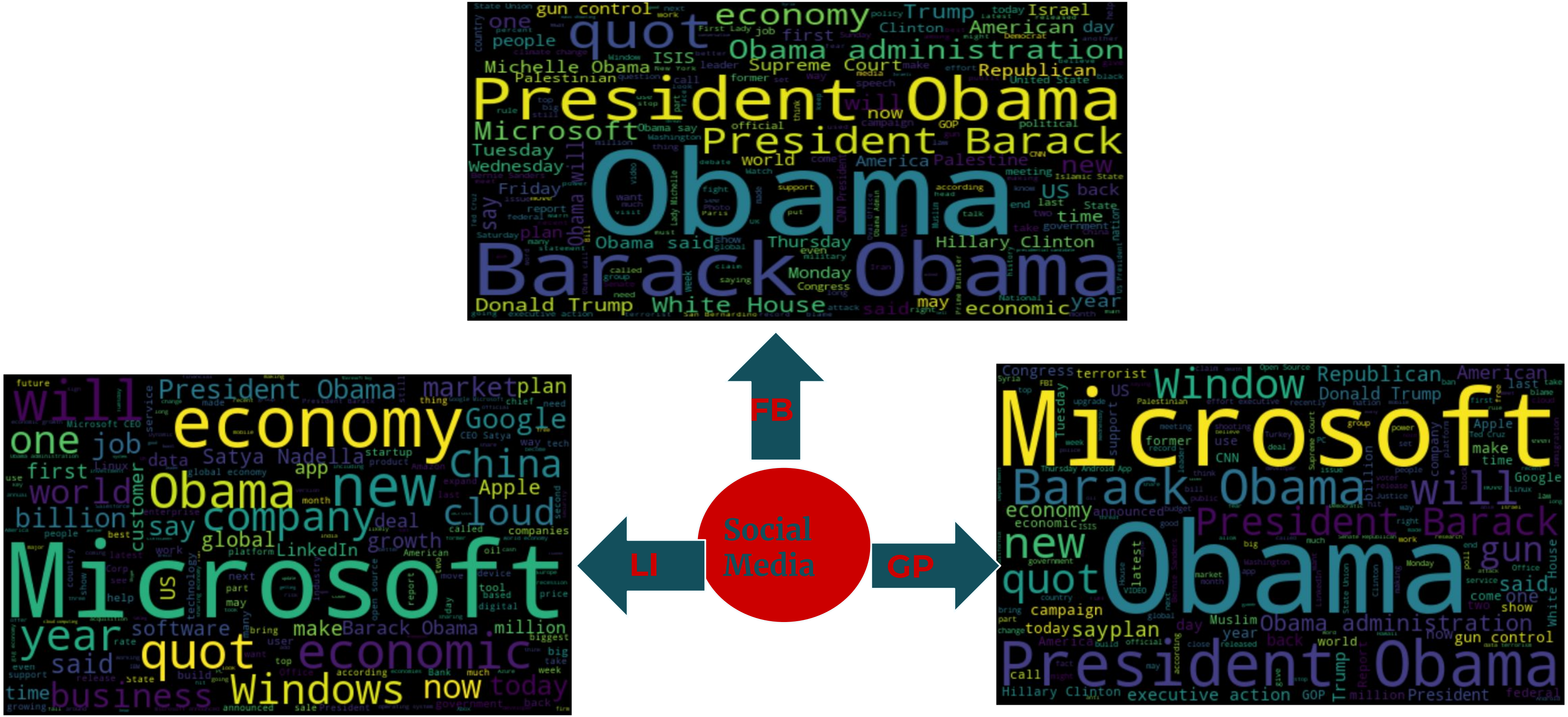


Microsoft

Economy



The EDA(Continue)



News With 100+ likes On Social Media

Text Processing

Tokenization.

Remove punctuations.

Remove stopwords.

Remove short words.

Lemmatization.

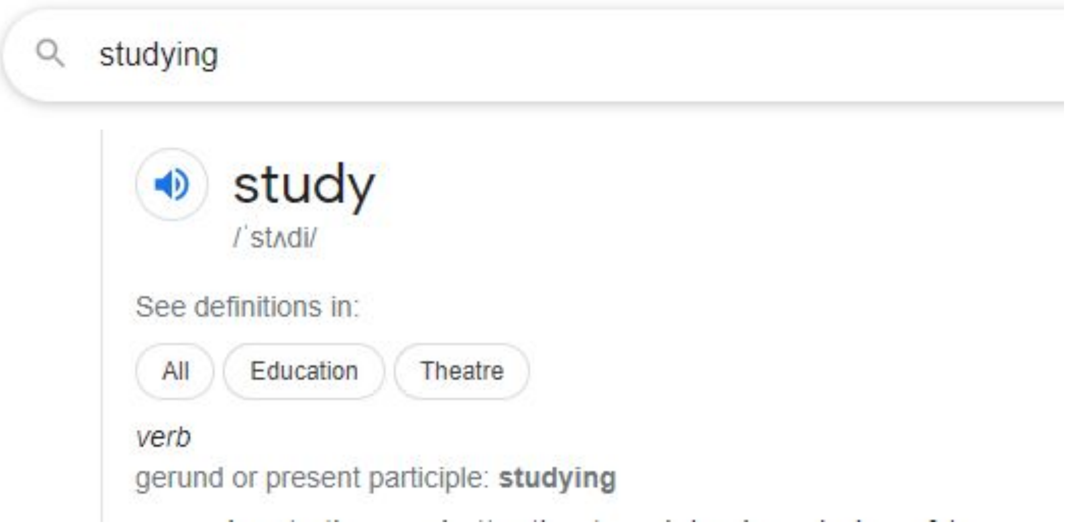
Stemming.

N-grams.

TF-IDF Vectorizer.

Text Processing(Continue)

The Lemmatization : Lemmatization (lemmatization) in linguistic is the process of grouping together the inflected forms of a word so they can be analysed as a single item, identified by the word's lemma, or dictionary form.



studying

study

/ˈstʌdi/

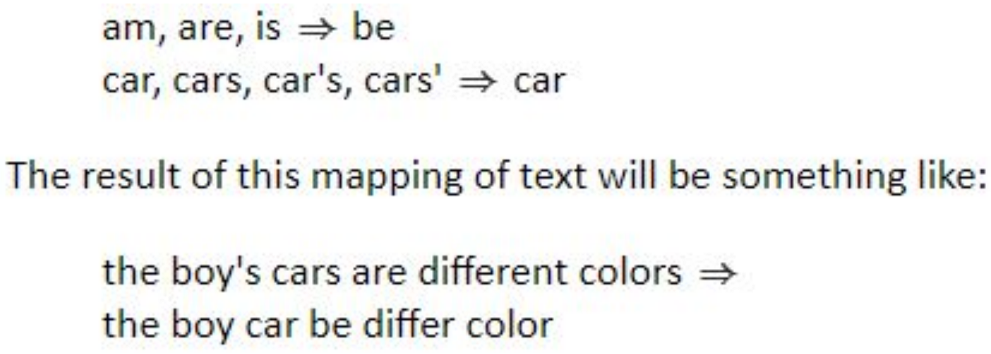
See definitions in:

verb

gerund or present participle: **studying**

Text Processing(Continue)

Stemming : Stemming is the process of producing morphological variants of a root/base word. Stemming programs are commonly referred to as stemming algorithms or stemmers. A stemming algorithm reduces the words “chocolates”, “chocolatey”, “choco” to the root word, “chocolate” and “retrieval”, “retrieved”, “retrieves” reduce to the stem “retrieve”.



am, are, is ⇒ be

car, cars, car's, cars' ⇒ car

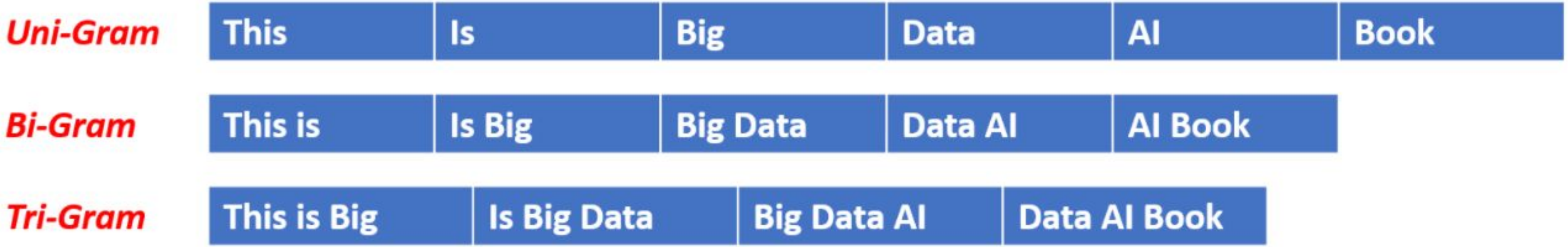
The result of this mapping of text will be something like:

the boy's cars are different colors ⇒

the boy car be differ color

Text Processing(Continue)

Ngram: An *n*-gram model is a type of probabilistic language model for predicting the next item in such a sequence



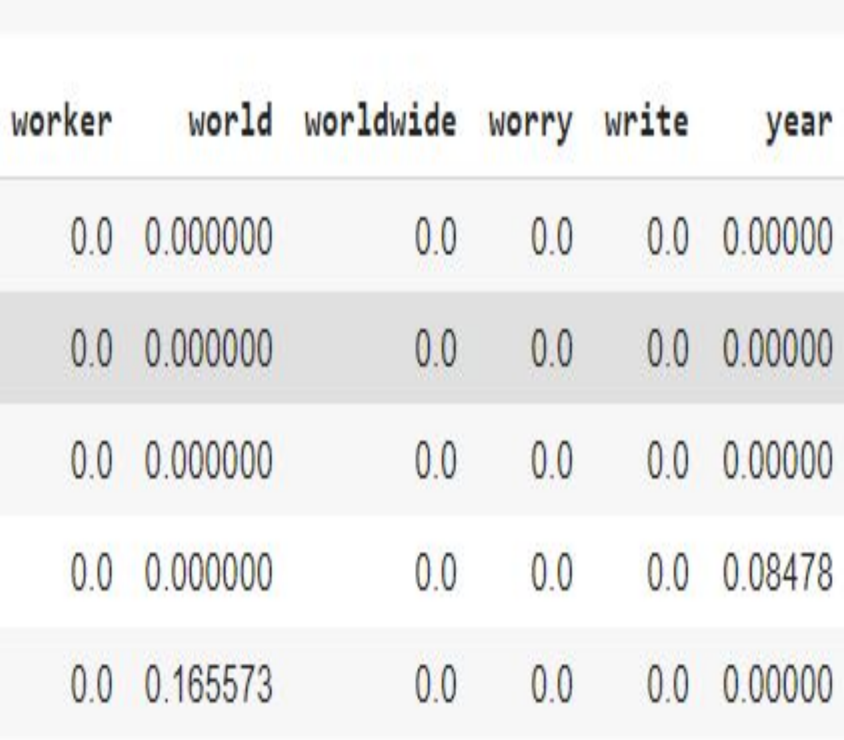
Uni-Gram	This	Is	Big	Data	AI	Book
Bi-Gram	This is	Is Big	Big Data	Data AI	AI Book	
Tri-Gram	This is Big	Is Big Data	Big Data AI	Data AI Book		

Text Processing(continued)

Max features:- Build a vocabulary that only consider the top max_features ordered by term frequency across the corpus.
Ie.If max_feature = n; It means that it is selecting the top n Feature on the basis of Tf-Idf value

Tf_idf We also experimented with changing min_df , and maxdf , which represent the minimum or maximum document frequencies of a word in order to be included as a feature. Empirically, we found that model performance was not very sensitive to the min_df and max_df parameters. In our analysis, we set min_df = 5, max_df = 80%.

Preparing Dataset



worker	world	worldwide	worry	write	year
0.0	0.000000	0.0	0.0	0.0	0.00000
0.0	0.000000	0.0	0.0	0.0	0.00000
0.0	0.000000	0.0	0.0	0.0	0.00000
0.0	0.000000	0.0	0.0	0.0	0.08478
0.0	0.165573	0.0	0.0	0.0	0.00000

Data Partition (80% -Train data 20%-Test Data)

Model Evaluation

RandomForestRegressor

bootstrap=True, ccp_alpha=0.0,
criterion='mse',
max_depth=11, max_features='auto',
max_leaf_nodes=None,max_samples=
None,
min_impurity_decrease=0.0,
min_impurity_split=None,
min_samples_leaf=60,
min_samples_split=50,
min_weight_fraction_leaf=0.0,
n_estimators=120, n_jobs=None,
oob_score=False,
random_state=None, verbose=0,
warm_start=False

XGBRegressor

base_score=0.5, booster='gbtree',
colsample_bylevel=1,
colsample_bynode=1,
colsample_bytree=1, gamma=0,
importance_type='gain',
learning_rate=0.1,
max_delta_step=0,max_dept=,
min_child_weight=1, missing=None,
n_estimators=100, n_jobs=1,
nthread=None,
objective='reg:linear',
random_state=0,
reg_alpha=0, reg_lambda=1,
scale_pos_weight=1, seed=None,
silent=None

Future Work

Complexity Measures

Natural Language Features We also added features for the counts of each part of speech using spaCy. We also added average sentence length, number of sentences, and average word length to our features array.

Embedding Matrix

The concept of an embedding matrix is an attempt to solve this relationship representation problem. To begin with, we pick a dimensionality of meaning — this can be somewhat arbitrary. Say we decide all meaning can map to some abstract space of three dimensions. Conceptually, that would mean that all words would exist as singular points in a 3D space and any word could be uniquely defined by their position within that space described by three numbers (x, y, z).

Conclusion

- We could conclude that if sentiment value is highly positive then the popularity of that new is high same goes for low sentiment value.

- If news contain certain kinds of words like growth, Fear, Danger, Government,business then our sentiment goes either positive or negative.

Challenges

- We have 3 dependent variables(Facebook, Google Plus,LinkedIn)
- We have a huge data set (it consists of 12 dataset which contain the information of likes of each topic on the given social media platform.)
- To implement TF-IDF vectorizer on corpus of having 100000 news document.
- Google Colab Crash problem.

Thank You