GENERAL ASSEMBLY

Misk Academy
أكاديمية مسك
Empower The Next Generation

DSI Capstone Presentation

A topic-based sentiment analysis

Lujain Felemban

AGENDA

- PROBLEM STATEMENT
- DATA SCIENCE PROBLEM
- DATA COLLECTION AND PRE-PROCESSING
- FEATURE EXTRACTION
- MODELLING
- RESULTS
- CONCLUSION


WHY!

Target Account

- Analyze the Sentiments of ALL of target account followers
- Predict Approval (Perceived Sentiment) Distribution towards a new tweet to be written by target account

DS Work Flow

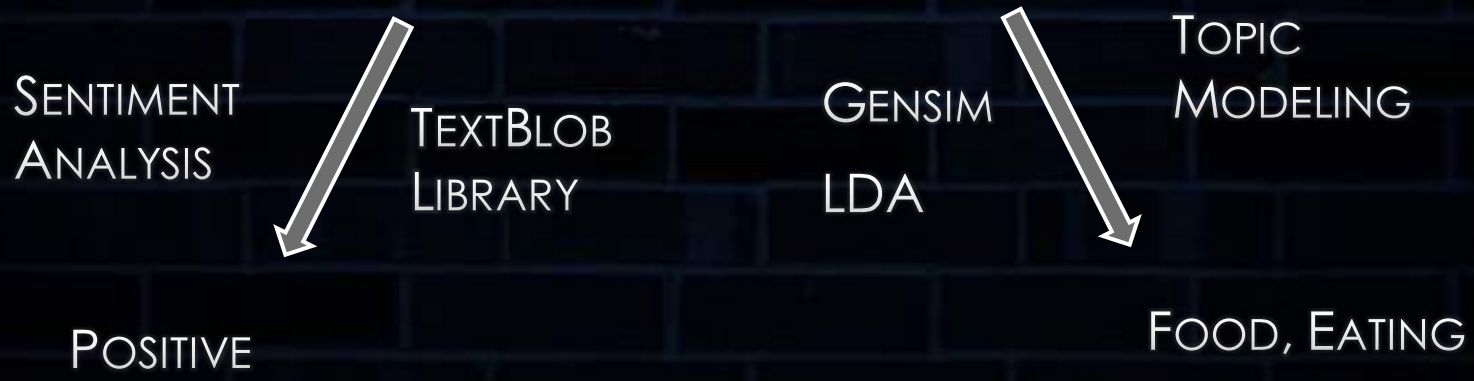
- Choose Account (PIF)
- Fetch all account followers username (40K followers)
- Check dominant Language (only 700 appear to be EN)
- Fetch all en Tweets (2000 max, 100 min per user)

date	time	username	tweet	mentions	hashtags
2018-09-27	06:47:22		sure lucy couldnt miss seeing speaking line lo...	['lucylucyprior', 'xrailgroup', 'transcityrail']	NaN
2018-09-03	18:48:27		darbaarbyabdul hi guys closed trying phone guy...	['darbaarbyabdul']	NaN
2018-08-13	04:15:16		british airways safety lucky serve tea	['johnmsv', 'emirates', 'british_airways']	NaN
2018-07-11	19:30:04		ive entered autotradergoals chance win free car	NaN	['#autotradergoals']
2018-07-02	10:45:58		hi becci sent dm ongoing issue thank	['landrover_uk']	NaN

- ~300, 000 observations

FEATURE EXTRACTION!

'I LIKE TO EAT SUSHI'



Topic modeling gensim LDA

Import data:
from .csv into pandas dataframe

Clean data (text analysis):
remove ascii, unicode, stopwords and remove affixes from words.

Construct a document-term matrix (DTM) :
from gensim import corpora, models

Create dictionary:
split sentences into tokens, assigning a unique integer id to each unique token while also collecting word counts and relevant statistics.

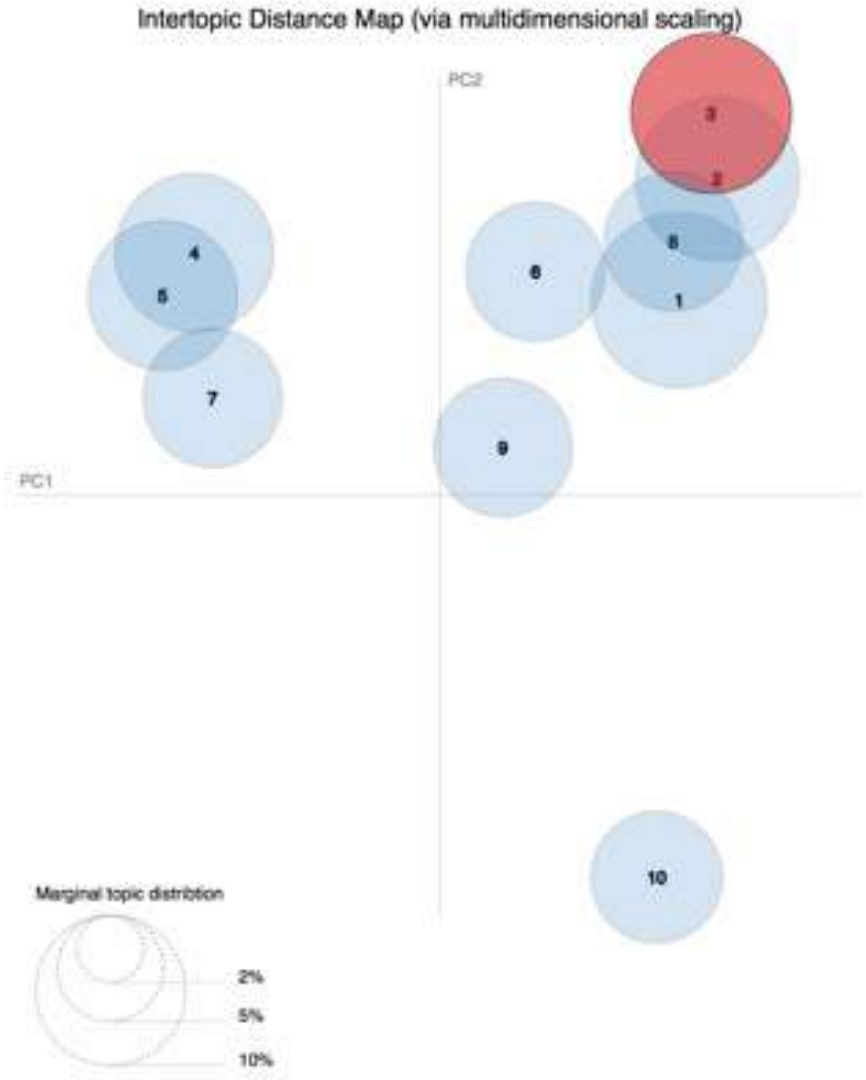
Create corpus:
doc2bow() function converts dictionary into a bag-of-words. The result, corpus, is a list of vectors equal to the number of documents. In each document vector is a series of tuples.

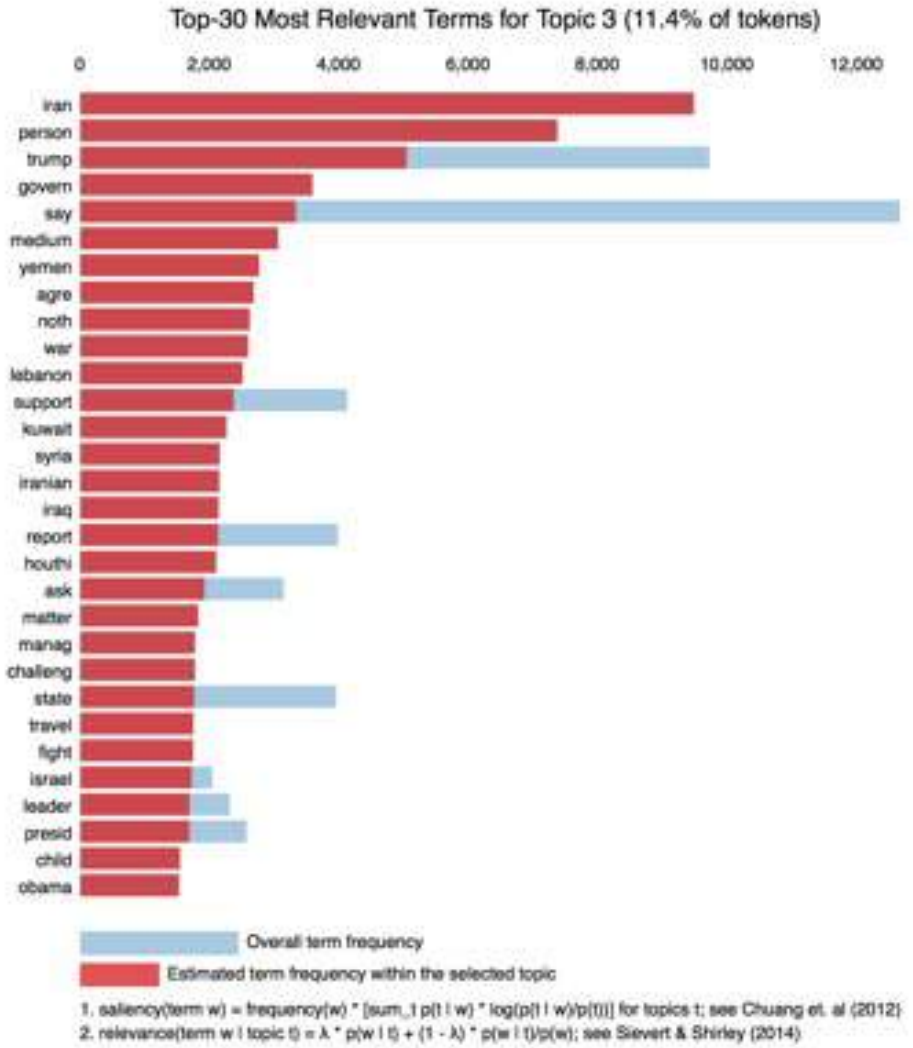
Decide number of clusters, passes and alpha:
Apply the LDA model

Examine the results:
each topic contains bags of words with multinomial distribution, analyze these set and assign meaningful labels to topic.

New document to be classified:
Clean data (text analysis), pass to LDA model, sort results based on topic probability and choose one with maximum probability or using another distance calculations techniques.

Topic Modeling of All Fetched Tweets





FEATURE EXTRACTION!

'I LIKE TO EAT SUSHI'

SENTIMENT ANALYSIS → TEXTBLOB LIBRARY → POSITIVE

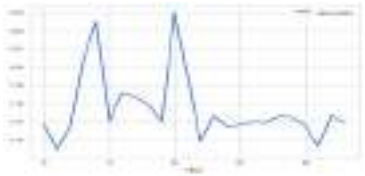
LDA → TOPIC MODELING → FOOD, EATING

0	1	2	3	4	5	6	7	8	9	DayOfWeek	Month	Hour	sentiment
0.010058	0.010058	0.228744	0.487173	0.113148	0.010058	0.010058	0.010058	0.010060	0.110585	3	9	6	0.500000
0.000000	0.129514	0.303850	0.000000	0.000000	0.000000	0.000000	0.000000	0.508295	0.000000	0	9	18	-0.085185
0.157130	0.157147	0.014292	0.014292	0.157099	0.157097	0.014292	0.300065	0.014292	0.014292	0	8	4	0.333333
0.157129	0.157156	0.014287	0.014288	0.299989	0.014289	0.157153	0.157131	0.014290	0.014287	2	7	19	0.600000
0.020004	0.219996	0.020004	0.419946	0.020008	0.020004	0.020004	0.020004	0.020007	0.220022	0	7	10	0.000000

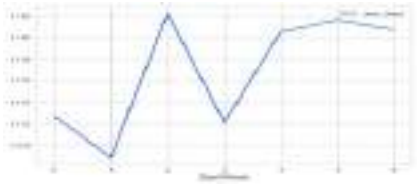
EDA

Sentiment EDA of tweets of EN followers of the Public Investment Fund

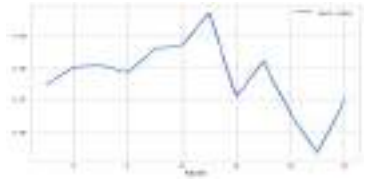
By Hour



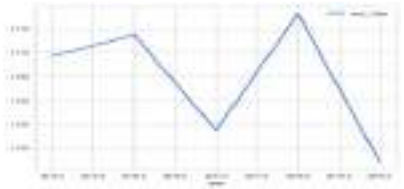
By Day



By Month

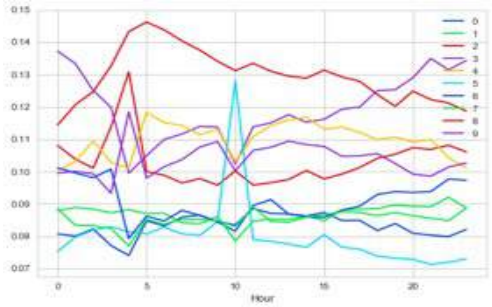


By Year

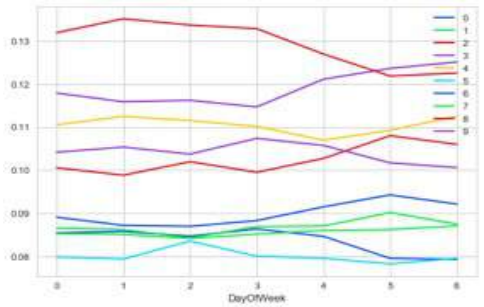


Average Sentiments of All Tweets Grouped by Time

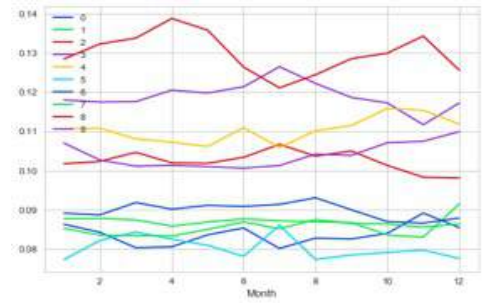
By Hour



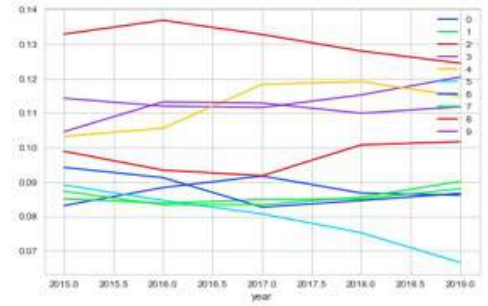
By Day



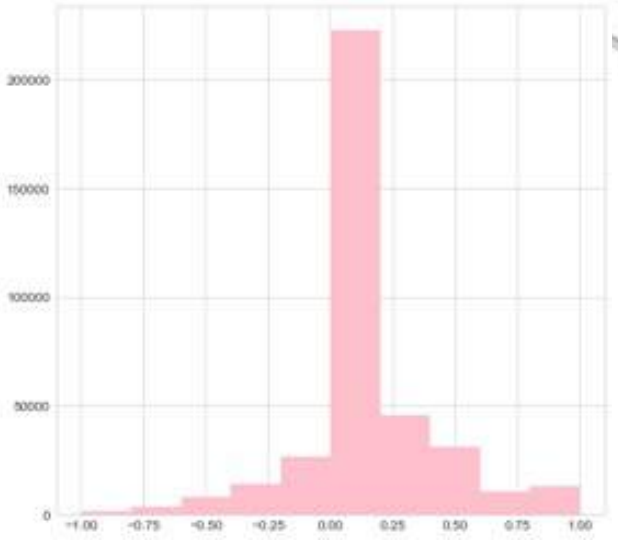
By Month



By Year



Average Sentiments of All Tweets Grouped by Time Broken by Topic



Sentiment Distribution of All collected Tweets

Modeling

Regression

Algorithm	Train/Test Score
Simple Linear Regression	AvgCV = 0.01
Decision Tree	0.99, -0.92
Random Forest	0.83, 0.07

Classification

Algorithm	Train/Test Score
Logistic Regression	0.66, 0.66
DecisionTree	0.71, 0.65
MLP	0.67, 0.66
MLP	0.77, 0.67

BEST MODEL

Layer (type)	Output Shape	Param #
Input_Layer (Dense)	(None, 128)	56704
dense_22 (Dense)	(None, 256)	33024
Output_Layer (Dense)	(None, 3)	771

Total params: 90,499
Trainable params: 90,499
Non-trainable params: 0

صندوق الاستثمارات العامة
Public Investment Fund

- ''' #PIF CONTRIBUTES TO THE DEVELOPMENT OF SAUDI ARABIAS ECONOMY BY INVESTING INDIVERSIFIED SECTORS, GEOGRAPHIES AND ASSET CLASSES, FORMING STRATEGIC PARTNERSHIPS AND LAUNCHING MAJOR INITIATIVES THAT MAXIMIZE SUSTAINABLE RETURNS IN LINE WITH THE GOALS OF #SAUDIVISION2030.

GET TOPICS MATRIX (LDA)

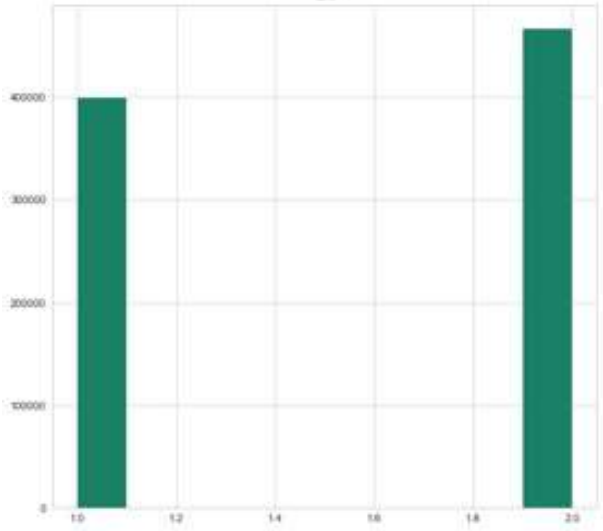
	0	1	2	3	4	5	6	7	8	9
0	0.046951	0.045885	0.045879	0.09988	0.358462	0.049083	0.045889	0.11382	0.125947	0.068204

CREATE DF OF ALL POSSIBILITIES (USERNAME & DATETIME)

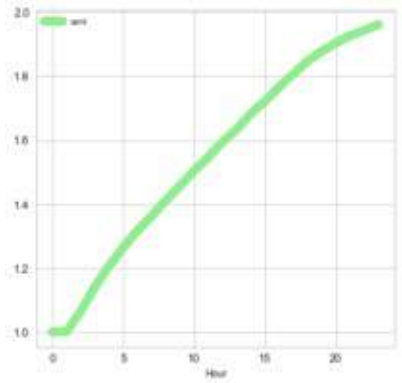
	month	day	hour	0	1	2	3	4	5	6	7	8	9
0	1	0	0	0.046951	0.045885	0.045879	0.09988	0.358462	0.049083	0.045889	0.11382	0.125947	0.068204
0	1	0	1	0.046951	0.045885	0.045879	0.09988	0.358462	0.049083	0.045889	0.11382	0.125947	0.068204
0	1	0	2	0.046951	0.045885	0.045879	0.09988	0.358462	0.049083	0.045889	0.11382	0.125947	0.068204
0	1	0	3	0.046951	0.045885	0.045879	0.09988	0.358462	0.049083	0.045889	0.11382	0.125947	0.068204
0	1	0	4	0.046951	0.045885	0.045879	0.09988	0.358462	0.049083	0.045889	0.11382	0.125947	0.068204

PREDICT PERCEIVED SENT

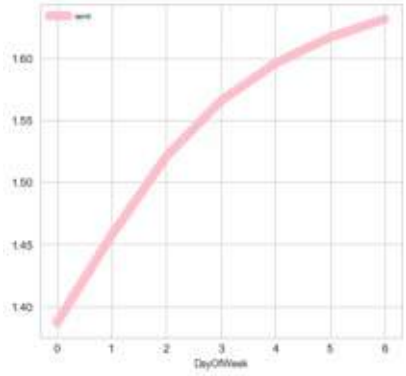
Sentiment Distribution of Users Towards Proposed Tweet



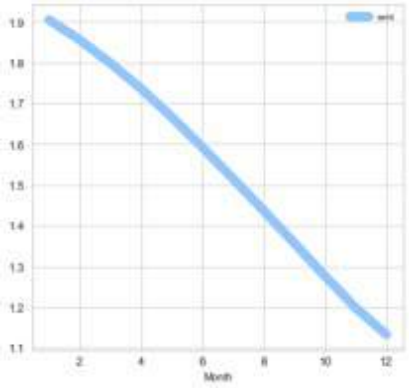
By Hour



By Day



By Month



Predicted Perceived Sentiment by ALL USERS

Challenges:

- Data Collection
- No Defined Metric
- Time
- Modelling based on non-apparent correlation Features
- Data Size

Future Work:

- Hashtags
- Pre-trained LDA topic model
- Time-Series based Sentiment Analysis
- Modelling
- Define a Metric (likes, RT)
- Tweet's sentiment (by head account)
- Resampling
- Wait for more followers to get bigger data

THANK YOU!

Lujain Felemban
General Assembly DSI Student
lifelemban@gmail.com