



Twitter Sentiment Analysis

Abitbol Charlotte
Choukroun Gabriel
de Pape Alexandre
Szalavec Clémentine
Siles Daniel

Goals

Generate real time stocks prediction based on Twitter posts for one company chosen (Apple)

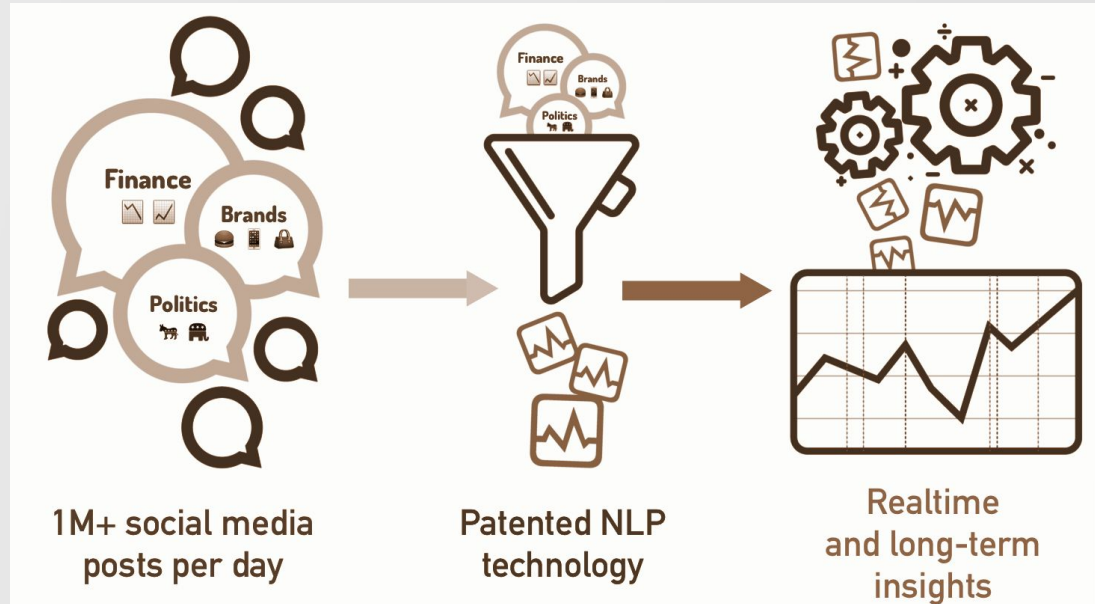
01 *Twitter sentiment analysis*

02 *Time series model*



Business Use Cases

- For portfolio management, leverage social network data to increase risk assessment performance
- To develop new investment strategies
- For venture capital, get new insights about startups



01

Dataset

Description of our dataset

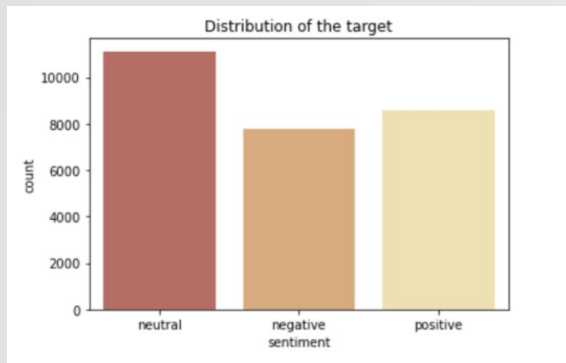


Description



27, 481 tweets

Dataset from Kaggle



The three classes to predict are :
positive, negative, neutral

3 classes



The dataset is balanced



280 characters

A tweet is composed of a text of, at most, 280 characters and it can contain a lot of words that are not in the English dictionary

Exploration of the dataset

Most common words in the dataset of tweets

Tree of Most Common Words



Common words for positive tweets

Common_words count

0	love	781
1	^i	730
2	day	704
3	good	644
4	i'm	434
5	happy	404
6	like	391
7	^happy	386
8	great	361
9	get	361
10	hope	346

Common words for negative tweets

Common_words count

0	^i	1032
1	i'm	636
2	like	462
3	get	428
4	miss	409
5	go	366
6	insult	321
7	don't	314
8	it's	301
9	can't	300
10	going	299



02

Preprocessing

1. Basic Tweet Cleanup

INPUT: @Diablo @KristinaSOSKi #ACTIVISION #BLIZZARD I got a ban and nobody comments on what happened, and I just want my account back. <https://t.co/aRPHu10JFN>

OUTPUT: I got a ban and nobody comments on what happened, and I just want my account back.



2. Insults Replacement

INPUT: Sons of ****, why could not they put them on the release we already buy??

OUTPUT: Sons of -INSULT-, why could not they put them on the release we already buy??



3. Contractions Extraction

INPUT: I've been sick for the past few days and thus, my hair looks weird. if I didn't have a hat on it would look...

OUTPUT: I have been sick for the past few days and thus, my hair looks weird. if I did not have a hat on it would look...



4. Remove Emojis, punctuation, words with numbers

INPUT: I have been sick for the past few days and thus, my hair looks weird. if I did not have a hat on it would look...

OUTPUT: I have been sick for the past few days and thus my hair looks weird if I did not have a hat on it would look...



5. Lemmatization

INPUT: I have been sick for the past few days and thus, my hair looks weird if I did not have a hat on it would look

OUTPUT: I have be sick for the past few day and thus , my hair look weird if I do not have a hat on it would look



6. TFIDF Vectorizer

- Consists in creating a new dataset where each row is a **tweet** and each column is a unique **word** present in all of the documents of the dataset.
- Values are assigned to **each** word present in the dataset.
- Takes into consideration the number of **occurrences** of the word in the tweet and in all of the documents of the dataset.
- The new dataset has **20000+ columns**, because every unique word in all of the documents is represented by a column.





03

Model performance

Baseline Model for predicting sentiment analysis

0.536

Of accuracy

GaussianNB

0.671

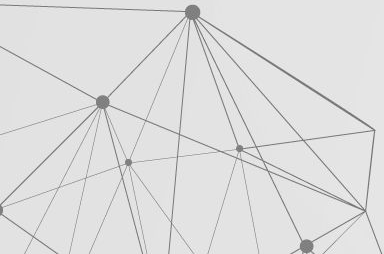
Of accuracy

Logistic Regression

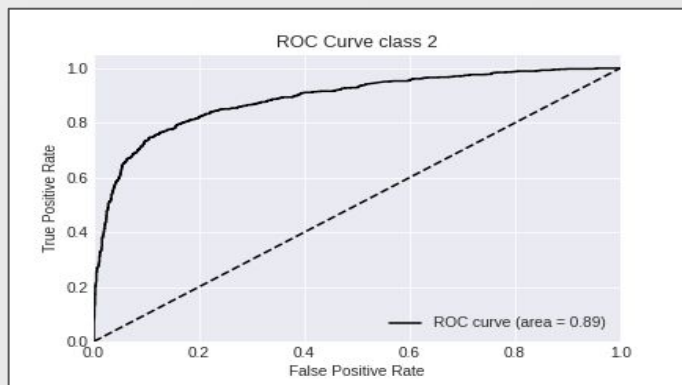
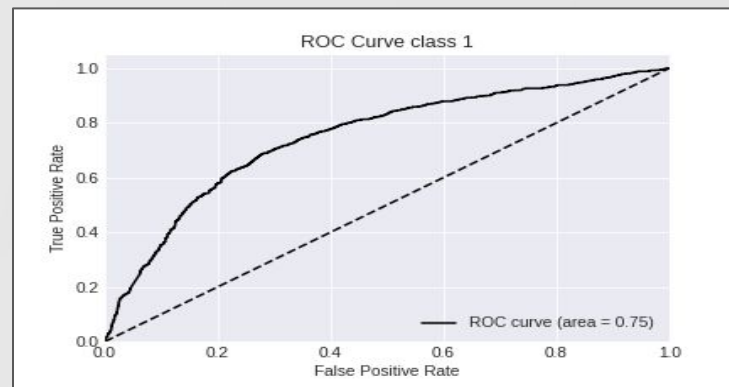
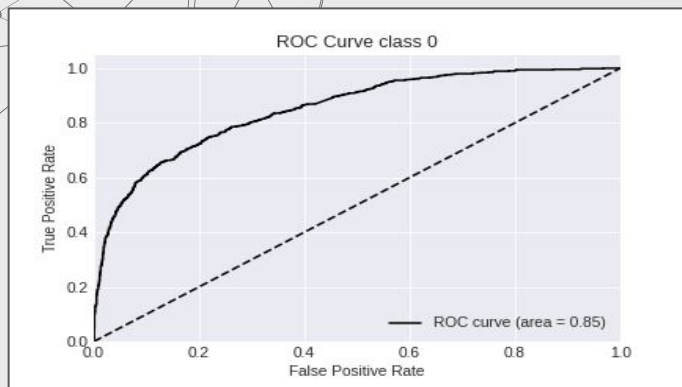
0.687

Of accuracy

SVM



ROC Curves for SVM



----- Classification Report SVM -----

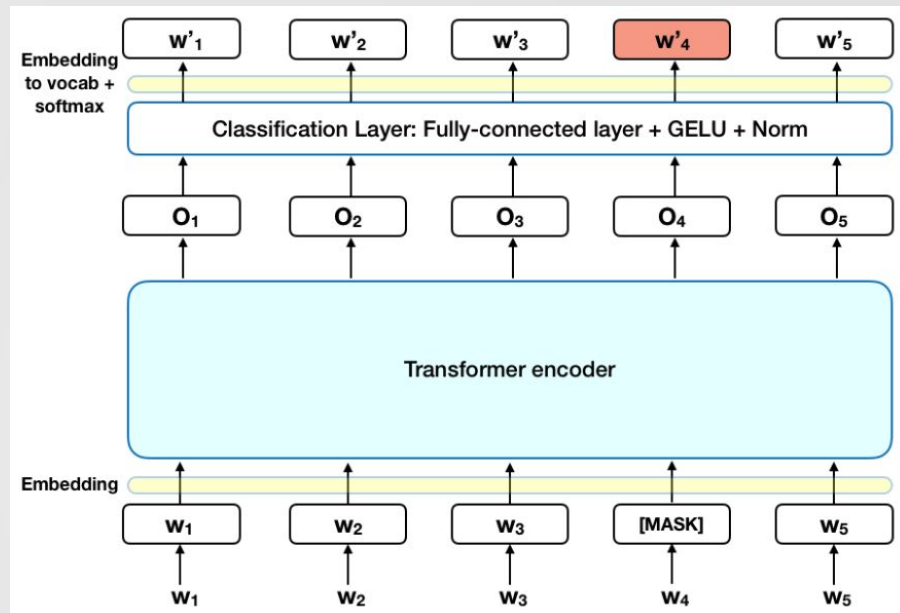
	precision	recall	f1-score	support
0	0.66	0.60	0.63	748
1	0.65	0.70	0.67	1141
2	0.75	0.73	0.74	859
accuracy			0.68	2748
macro avg	0.69	0.68	0.68	2748
weighted avg	0.68	0.68	0.68	2748

Bert Model

Transfert Learning

- USE transfer learning from a pre-trained model of BERT transformer.
- Achieve with BERT after tuning hyperparameters :

0.8
of accuracy



Examples of tweets

Test set : 118,350 Apple tweets from the year of 2016

Twitter Sentiment Analysis

Twitter :

I really like the new Apple watch

Positive!

Submit

Powered by Le Directoire Group

Twitter Sentiment Analysis

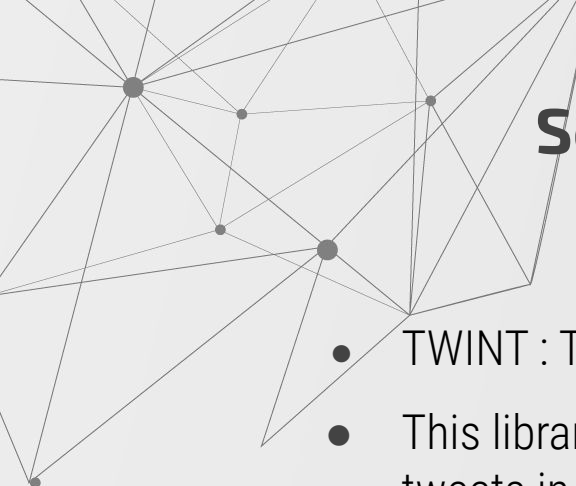
Twitter :

what be this new macbook I suck

Negative

Submit

Powered by Le Directoire Group



Scraping real time tweets to prepare

- TWINT : Twitter Intelligence Tool
- This library allowed us to scrape more than 100,000 tweets in the past 3 days choosing a specific company :
 - Company chosen : APPLE
- This dataset will be used as a test set for the stocks prediction part.





Next Steps

- Extract stock price for Apple and join the dataset with our Apple sentiment analysis dataset
- Train a time series model using VAR + LSTM
- Test our model on real time tweets to forecast Apple stocks



04

Technical Challenges



Technical Challenges

- Accuracy - Improve the accuracy of our sentiment analysis model
- Twitter API technical limitations - 7 day window, limited number of tweets etc.
- Scraping - Get relevant tweets among the noise

Conclusion

Next Steps

- Add a stock market trend predictor
- Work on our accuracy and evaluation of our model

Hindsight

- Can we really predict stocks or is it something random?
- Is public mood correlated or even predictive of economic indicators?
- How long would it even take for a tweet to impact the market?
- Twitter is only one example of social media, we might need to aggregate more of them and gather news feed to perform a more relevant prediction