Georgetown University

Master of Science in Data Science and Analytics
Fall 2021

ANLY580

NATURAL LANGUAGE PROCESSING

PROJECT PROPOSAL
CL1486,

GIT: https://github.com/Cholianlll/2021-Fall-ANLY-580-Final-project
SUBMISSION DATE: 1 NOVEMBER 2021
PAGES = 13 (excl. Title page, Table of contents, Reference & Appendix)

Contents

1. Project Summary

Question: High level description of what you plan to do, and why you think it is interesting.

"Crypto" related things have grown into a public-hitting topic. With a few years' explosion in this area, the traditional crypto topic, concentrated in discussing market quotes and pure fanaticism, has derived into several different directions, including "NFT", "Bitcoins", "Ethereum", "Defi", "Metaverse" and so on. This project is aimed to explore some insights hidden in the massive topic squares platform, such as the Twitter platform. The first and most important task is to explore whether there are any gaps within this huge area, implying a trend from a single topic to diversified areas. Then, for every single sub-topic, what special theme is will be the second task. Finally, after building a successful model for detecting and exploring different topics, the final stage is aimed to build the auto-enlabeling system on each new tweet, classifying each tweet into the right sub-topic.

There will be several practical contributions to this tweets auto-enlabeling system. Firstly, by detecting and clarifying the number of sub-topics related to the "crypto", it can provide a clear guide on assimilating into the "crypto" stream. For example, if a person wants to implement a project related to the "NFT", with the assistance of this project, he can learn it from searching with the keyword "airdrop" rather than the keywords "bitcoin", saving a lot of time. Then, the model built during this project is a good text classification algorithm attempt for some crypto-related news platforms, where it can help push the proper news to the right audience.

2. Project Description

2.1. Modeling approach

Question: What modeling approach do you intend to use?

There will be three stage along with three different models.

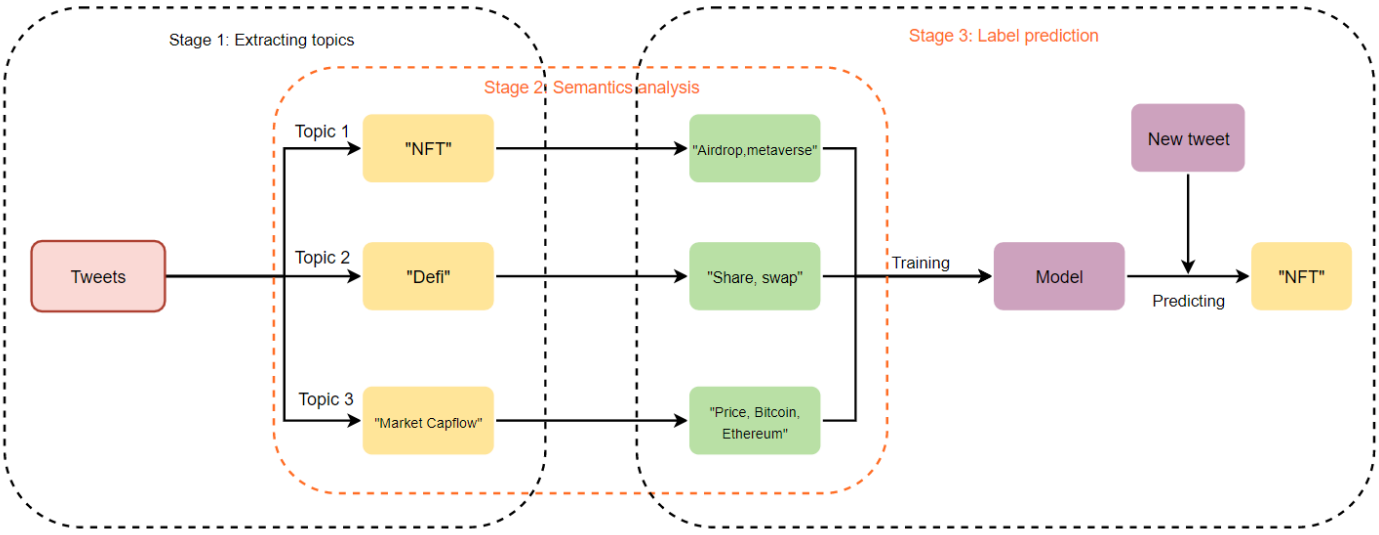

Figure 1 Modeling Approach

The first stage will focus on extracting N topics. Since all the tweets are collected together, the first thing is to decide

how many topics or sub-areas we should be clarified. The appropriate models will cover pLSA, LDA, and NMF models. Ideally, the model will help us extract the main keywords from a large corpus database.

The second stage will focus on distributional semantics analyzing, exploring certain topics extracted in the first stage. The appropriate model will in this stage cover some distributional semantics analyzing technical, such as a continuous bag of words, Skip-gram skills. Beyond that, as shown in Lab-05 "Visualizing tweets with Word2Vec", there will be also a similar word cloud for the audience to show the relationship within each topic.

The final stage is to predict the label of new tweets. After building the label from the previous two stages, it is worth building a model to help predict the label of each new tweet. In this stage, some advanced models will be introduced because of the high accuracy, including "CNN", "RNN", and even some attention-based frameworks.

2.2. Data preparation

Question: What data do you intend to you use?

There will be two main resources of the dataset. Firstly, to support the dataset of this project, there are two available Twitter developer accounts for this project to listen to the timely tweets and extract raw tweets dataset. Secondly, downloading from the Kaggle. Since there is an upper limit on each Twitter developer account, one of the good methods to obtain the dataset is in the Kaggle, where there are so many raw update-to-day datasets.

There are 40119 rows and 35 columns in our current collected dataset. Most of the data are collected by monitoring Twitter via Twitter developer account API. As this project is focusing on text analysis, the main feature the project will use is major relying on the "text" column.

Index	user_id	status_id	created_at	screen_name	text	source	display_text_width	reply_to_status_id	reply_to_user_id
1	1.348E+18	1.4368E+18	1631404785	Alexa170880617	@crypt0e	Twitter fo	187	1.43683E+18	1.39893E+18
2	1.427E+18	1.4368E+18	1631404783	asamifire	#BAKECO	Twitter fo	284	NA	NA
3	1.39E+18	1.4368E+18	1631404783	Ethurfer	2021-09-	bitcoAuto	137	NA	NA
4	1.379E+18	1.4368E+18	1631404783	daeevii	#BAKECO	Twitter fo	284	NA	NA
5	43170080	1.4368E+18	1631404777	INUmto3	#financial	Twitter fo	260	NA	NA
6	1.413E+18	1.4368E+18	1631404768	SpawnofSatoshi	Buying	Twitter W	150	NA	NA
7	1.427E+18	1.4368E+18	1631404753	cyberneus	#BAKECO	Twitter fo	284	NA	NA
8	9.979E+17	1.4368E+18	1631404746	RixxTech	Daily top	Rixx	162	NA	NA
9	9.979E+17	1.4368E+18	1631404745	RixxTech	4 hour top	Rixx	163	NA	NA
10	1.005E+18	1.4368E+18	1631404735	Doan_minh15	@FishyTa	Twitter W	268	1.43511E+18	1.39025E+18

Figure 2 Sample data

2.3. System evaluation and evaluation criteria

Question: How will your system be evaluated and what are the evaluation criteria?

To evaluating the system, there will be different stage measurement.

Table 1 System evaluation

Stage	Evaluation	Method	Criteria
Extracting topics	Divergence between topics	K-L divergence, Cross entropy	Highly distinction between each topic
Semantics analysis	-	-	
Label prediction	Prediction accuracy	F1 score, AUC,ACC, etc.	high prediction accuracy

In stage 1, the main method of evaluating is to the divergence between different topics. Therefore, some measurements, where it can be compared between two corpora, will be set as the criteria, including K-L divergence and Cross entropy. Both indicators try to achieve a higher score.

In stage 3, the major evaluation will depend on the label prediction accuracy. Some indicators will be considered, such as the F1 score, AUC, ACC, and even confusion matrix. Within this system evaluation method, the model will try to achieve the highest accuracy with less error.

2.4. Computational and hardware considerations

Question: Are there any special computational/hardware considerations?

- **GPU acceleration**

Since the project will implement some advanced models in stage 3, including the RNN and CNN, it may need to use GPU accelerating during the model training stage.

- **Then operational system**

As some python packages have been not adopted the M1 yet, it is potential that we may need a compatible server like Ubuntu or Windows.

2.5. Potential Obstacles

Question: What are the biggest unknowns that might dictate the success or failure of this project?

- **Limited Twitter API account**

There are only two Twitter API accounts available now. To increase the generalization of a NN model, it is necessary to train it under a large dataset. However, there is a limitation on our API account, where Twitter only allows us to listen to 50,000 tweets for each account per month. Beyond that, the downloaded data from Kaggle just cover a certain period. Without the continuous monitor data sources, it could reduce the generalization on our model, because of the time gap, where there is a weak representation capability between our training data and test data.

3. Project presentation

Question: How will the results of your work be presented? Will this be a live demo, a written report, a slide deck + oral presentation? Any of these are acceptable! Demos can be given along with reports/presentations.

3.1. Presentation type:

Slide deck + oral presentation

3.2. Presentation content:

- Stage 1

Top keywords of each topic; code demo; model evaluation indicators.

- Stage 2

Word cloud; visualization of words.

- Stage 3

Model evaluation indicators; code demo; model structure.