Project Proposal: Identify Dominant and Suspicious Tweets in Trend Topics

Introduction:
As a social network site, Twitter is not only the showcase of people's daily life but also the melting pot of news and public opinion. It is an important platform for both private and public communication. We can see people actively participate in trend topics and express themselves. It is reasonable that we could gain insights from different aspects and be influenced by some dominant tweets or comments. And it would be better if we find a way to quickly get what is going on here and what other opinions are.

Meanwhile, there is the situation where tweets and comments might be produced by fake account intentionally to impose on the current discussion or to inflate the popularity of certain users and particular topics. And some spam tweets are to promote products or spread out fishing urls. We believe most of the users prefer the reliable information on social network. Therefore, it is necessary to identify those fake accounts and figure out what they are trying to sell.

The objective of our project is to first define and detect fake accounts on Twitter. Then, by analyzing tweets in trend topics, we want to discover the patterns in those tweets, such as the word co-occurrence statistics, which could summarize opinions related to the trend topic. Lastly, we will specifically find out the suspicious tweets produced by fake accounts. If our project works, the better experience would be provided to social network users. They will be able to learn the distribution of opinions and check the existence of fake/spam tweets so that they can form a deeper understanding of the trend topic.

Members:
Guanyu Chen, Tongke Wu, Jieyi Chen

Main Question:
What are the fake accounts and followers in Twitter right now?
What is the distribution of opinions in a trend topic?

Sub Questions:
What attributes distinguish a fake account from the real one?
What are the common features in tweets with more than 1000 likes?
Are there suspicious tweets/comments created by fake accounts?
Is there particular content in suspicious tweets?

Data:
1. Twitter accounts dataset provided by My Information Bubble. We are approved by the team to use it in our project. This dataset provides labeled Twitter accounts (genuine and fake accounts) and we will check whether all the account still exist and update account information. It includes two sub-datasets:

- genuine_accounts.csv: include all the human operated twitter accounts
- fake_followers.csv: include all the fake twitter accounts
    - Each csv files include:
        - user.csv: contains the user basic information, status, followers and friends counts
        - tweets.csv: contains the user id, their posts and comments, activity and timestamps.

2. Twitter API : we will interact with it to extract account information and stream of tweets.

Methodologies:

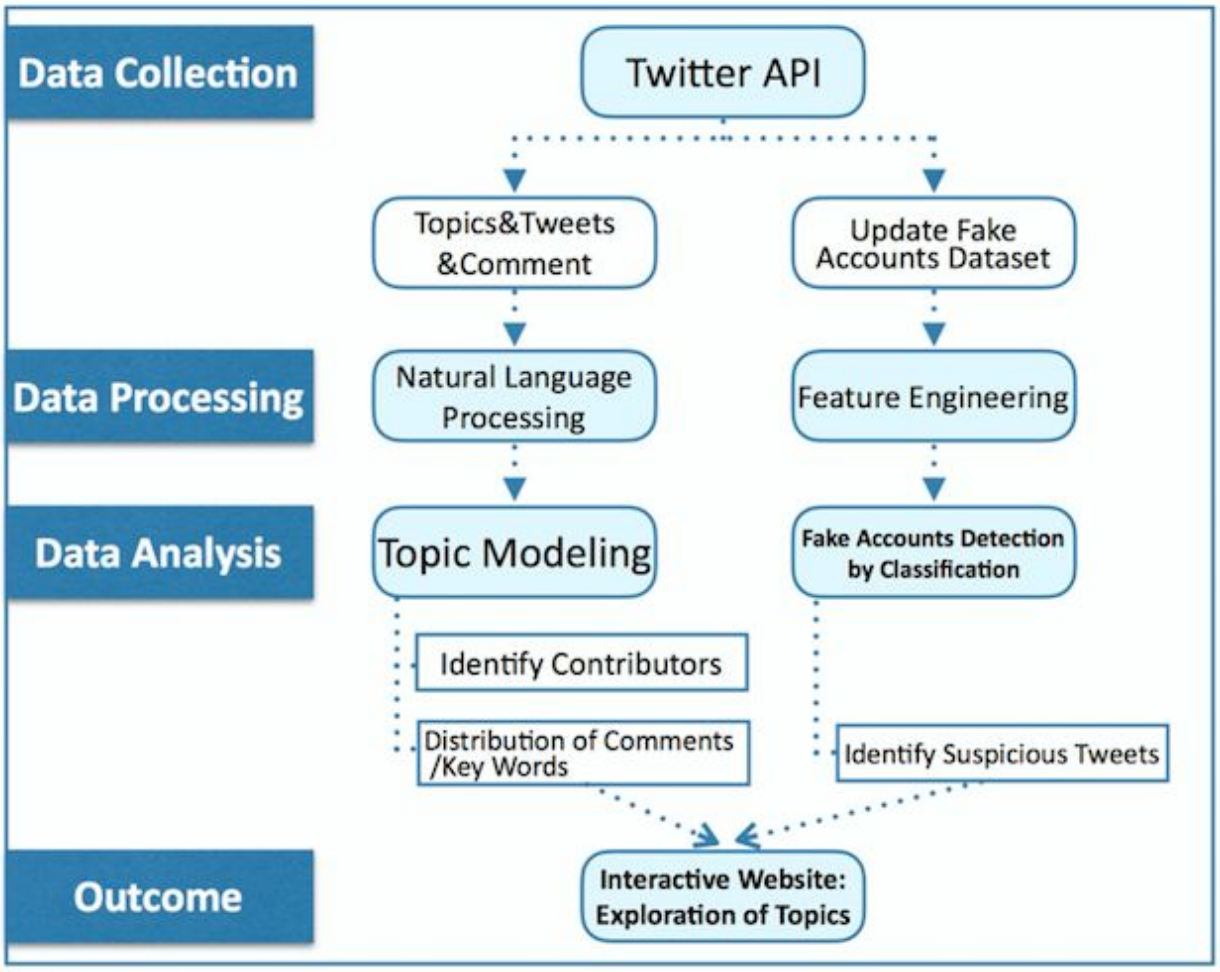
Latent Dirichlet Allocation is used for analysis of tweets in trend topics. We will apply this methods to get the word co-occurrence statistics. However, since tweets are short texts and unstructured with informal language, we need to do preprocessing work so that LDA could yield ideal results. And we will keep learning other models that can deal with short texts if we need.

We will build the detection model by several classifiers and improve this accuracy. Considering the evolving features of fake accounts, we will read relevant papers and find novel measurements to define fake accounts. In order to identify suspicious tweets, we will first obtain the list of users participated in a topic from Twitter API and then apply our detection model to see whether there are fake/spam accounts. At last, we will analyze the content produced by fake accounts.

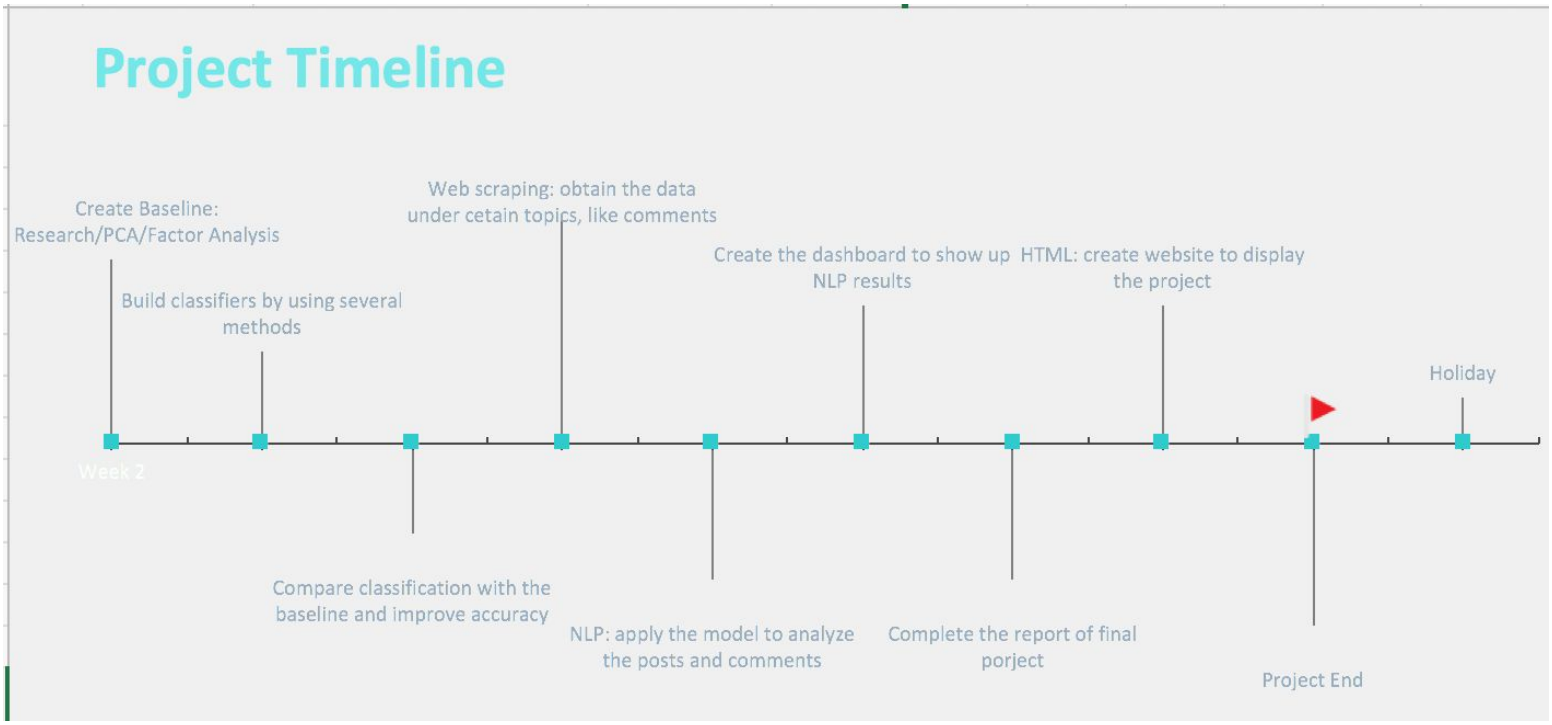
Potential Outcomes:

- Fake account detection model
- Distribution of word co-occurrence of tweets in trend topics

Workflow:


Data Collection
Twitter API
Topics&Tweets &Comment
Update Fake Accounts Dataset
Data Processing
Natural Language Processing
Feature Engineering
Data Analysis
Topic Modeling
Fake Accounts Detection by Classification
Identify Contributors
Distribution of Comments /Key Words
Identify Suspicious Tweets
Outcome
Interactive Website: Exploration of Topics

Timeline:



Skill Sets:

	Tongke	**Guanyu**	**Jieyi**
Skills	Machine Learning Web Scraping Network Analysis	Web scraping Time series	Natural Language Processing Web Scraping Logistic Regression

How will we combine different pieces of work/tasks?

We will use github for combining all the codes and work that we create.

Potential Problems:

- We have a limited understanding of the methods that we will use. Therefore, we might potentially apply a wrong method
- We might not be able to find the user information for every Twitter ID in the direct dataset
- We might not be able to find the popular topics from the post/comments
- Time management