Reddit and Fake News Detection

Manu Sreekuma A20405120, Student, IIT, Martin Berger A20410866, Student, IIT, Antoine Gargot A20410860, Student, IIT

[1]IIT, Illinois Institute of Technology, 3300 S Federal St, Chicago, IL 60616

Abstract – Reddit is considered as the front page of internet , where millions of news articles in various topics gets posted and discussed everyday. The platform attracts lots of conspiracy and fake news which does not have good reliability score to show its validity. We are trying to solve the problem to providing a reliability score for the news URLs posted by redditors based on several key metrics like up-votes, user profile and past activity. The problem is very interesting because a heuristic solution to the problem of fake news propagation would curb the spread of misinformation and negative mob effects through social media.

I. PROBLEM OVERVIEW

In today's age of information, it is not always easy to figure out if a given post on social media is reliable or not. The vast social network provides opportunities for people to propagate misinformation through the grid which could lead to misinformation and grave implications. We are planning to conduct an investigation on the Reddit platform to understand whether there is a pattern among users who share the hoax content and kind of activity that surround the post, in terms of content, votes and comments.

II. DATA

For our project, we will use the Reddit's API in order to find most of our needed information regarding users and posts (https://www.reddit.com/dev/api/). We will be using the Python Reddit API Wrapper which will simplify our requests to the Reddit API. Each API request return JSON object which gives maximum information. For instance, when you made an overview request of a specific user, you will have many pieces of information about his comment, post and up-votes history which will be very useful in our study. For each post, there are many informations such as the score, the number of report of this post from other users, the link to the news, comments, and likewise. Each of those useful information will be stored in instances and be used for prediction.

Data points that we plan to extract through Reddit API are :

- Id of the post (will work as the identifier)
- The truth value project score of the link
- the link by itself
- score of the post (number of up-votes)
- number of comment
- up-vote percentage for the post
- post sub-reddit name (it is known that some sub-reddit are more reliable than others)
- number of readers
- user name
- user post karma
- user comment karma
- number of years on Reddit
- is user moderator of some sub-reddit (moderators tend to be more reliable than other users)
- posts of this users
- score of each posts of the user

We are also for to use the API which of an open source project called the truth value project. This API will give us more information about a specific link based on the rating of the news site and also by the sentiment analysis of the post.

We will also use the open sources group database composed of different websites (close to 1000 observations) and which gives a label to them (satire, fake, bias, unreliable) in order to filter our data, the . Thanks to all those data, the main purpose of our project will be to use a Reddit post and study the history of the user who shared that link and make some prediction thanks to all those different features and observations.

III. METHOD

We are planning to perform analysis of reddit data in two stages :

1. Reddit Network Analysis for classifying the publishers of news in Reddit.
2. Natural Language analysis of the posts to extract possible features.

The data extraction would be carried out using Reddit API . We are planning to pipeline the two stages to perform the classification of Reddit Posts.

A - Network Analysis

Network analysis is performed to roughly classifying the network of users who have higher proclivity towards propagating fake vs real news, by the following steps :

- Creating a training set of users labelled as trolls or genuine users.

The users will be labelled solely by their activity of sharing real vs fake news. The following subsets would be incorporated as part of the same:

1. A set of users will be extracted using Reddit API based on the content they have shared (real vs fake news domains and score of what they shared)

2. All the posts shared by these users will then be extracted to understand the ratio of real vs fake news domains that had been shared till date.

3. We are planning to devise an algorithm similar to Boolean Label crowdsourcing algorithm in "Some Like it Hoax" research paper to label users based on their activity.

- Creating a classification model based on Social profile and interaction of labelled users in training set (sub-reddit, karma points , activity , personal information). This could be implemented using multiple classification algorithms - Logistic Regression, K-Means and Random Forests.

B - Natural Language analysis of posts

The training set used would be the set of fake and authentic news sources available from FakeNews Git repo, snopes and Polifact ,if the news article is part of the post. If the article is an external link, then the data points like comments and up-votes / down-votes + post heading would be used as features. The analysis would be performed using scikit - learn python packages. We are planning to use bag of words model and apply ML algorithms like Naive Bayes , Support Vector Machines and GridSearch CV to understand the patterns.

K-fold cross validation would be used on the acquired data set to perform training and prediction of posts. Confusion matrix, precision, recall and AUC curve would be used as baselines for analyzing the prediction accuracy of the models.

IV. PRELIMINARY EXPERIMENTS AND RESULTS

We wanted to study the impact of fake news domain on Reddit before working on a real model in order to see if website known as fake or unreliable sources have a big impact on this social network. For that, we extract the 800 domains from the open source project and assign those with their respective labels: Bias, Conspiracy, Unreliable, Hate, Clickbait or rumor for instance.

After doing that, we used this database in order to extract fraction of Reddit post related to those web domain (by doing a search of that domain in all sub-reddit in the network). When this work is done, we made a data frame of all those observations with specific features such as source, authors, number of up-votes, number of down-votes, number of reports and studied the score of posts based on the domain label. An interesting information is that website with a biased label tends to have a bigger number of comments. On other observation is that, regarding the fact that most of that website are labeled as unreliable sources, there is a huge amount of up-votes among posts related to them. In fact, our first assumption was that controversial post will tend to be more irrelevant, however, most of those links have a huge number of up-votes regarding the fact that most of them are shared inside communities that share the same political point of view. For instance, sources with known political bias are more likely to be shared in sub-reddit which has many supporters of one specific cause in order to emphasize their point of view (many irrelevant news and links are shared in the sub-reddit r/The_Donald which have a huge number of Donald Trump supporters). In another side, some sub-reddit are specialized in non serious posting such as the sub-reddit r/SyrianCirclejerkWar with attempt at humor. Those kind of observations reinforced the fact that we need to consider the sub-reddit name as a feature of our classifier.

V. RELATED WORK

A - With Facebook, Blogs, and Fake News, Teens Reject Journalistic "Objectivity" [1]

In this paper Regina Marchi clearly explain the shift that is happening in the way that the new generation consume news. Indeed, while a lot of people think that the new generation is uninterested by news, because they don't follow the traditional channels, Regina explains us that it is nothing like. The younger generation is in high demand of news, and they think that the role of journalist is very important for a society. However, since the advent of social media and mobile apps, younger generation completely changed the way that they access information. They want to be able to select the subject and they often follow the information that propagate through their network.

Even Though this paper is not technical, it is interesting because it underlines the importance of having a way to access the reliability of a news . If we don't manage to find a way of doing content checking on social media, it opens some space to crowd manipulation. This could have consequences that have never been witnessed in human history.

B - Fake News Detection on Social Media: A Data Mining Perspective [2]

In this paper the author defines clearly what is a fake news and the entire logic behind it's propagation. It lays down a clear definition of the problem and the challenge that it represents. The authors follow up by giving us an overview of the different ways to do fake news detection. They explain us that this task can be split two: Feature Extraction & Model Construction. The multiple ways to extract features is very interesting. This paper clearly explains that we can do feature extraction based on different characteristics: Linguistics, Visual, User, Post & Network. By focusing on certain aspect of a fake news, we can create different models. Models can also be seen differently according on whether we focus on the news content or its social context.

This paper enables us to have a clear overview of the options that we have when it comes to analyse fake news. It enables us to clarify what we are willing to do and helps us to adopt the right jargon when it comes to fake new detection. Our project will mainly be based on a Feature Extraction and a Model Construction based on the social context.

C - Automatic Deception Detection: Methods for Finding Fake News [3]

This research paper is a bit dated but it emphasize the fact that their a two main different approach to tackle fake news detection. The first one is a linguistic approach that has been studied for a long time and the second one is a network based approach. In this paper, we can notice that the network based approach wasn't giving really good results so long ago.

This paper is interesting because it confirms the fact that a network based approach is still a new way to detect efficiently fake news.

D - Some Like it Hoax:Automated Fake News Detection in Social Networks [4]

This paper explains how they managed to categorize Facebook posts in two different categories (hoax and non-hoax) by analyzing who liked it. They used to different methods to analyze the data that they collected. They used logistic regression and boolean label crowdsourcing (BLC). They end up with an accuracy higher than 99% with the two methods. However, the BLC methods seems to be more efficient because it can transfer information between pages more efficiently.

This paper is very interesting for our project because, just like us, it focuses more on the users rather than on the content of the posts to detect the truthfulness of it.

E - Fake News Research Project [5]

This project is about the study of fake news during the 2016 election. The authors focused their studies on twitter users who tweeted fake news domains. They remind us that in social media there is a homophile effect where people tend to follow like minded friends. They also explain that they could witness a Pareto principle in the URL that were shared. Indeed, only a small amount of URLs were widely shared. When analyzing the fake news URL, they could notice that each link would have a political bias in the ideas as well as in the lexical semantic Furthermore, the emphasize on the fact that it is not an easy task to see the difference between a bot and a non-bot user.

This analysis is interesting because even though they do not try to make any prediction, it provides us with very useful insight about the fake news phenomenon. They also mainly focus on the link between the URL and the users aspects. While we were collecting the first sample we could witness the homophile effect in the data that we collected.

F - Reddit was a misinformation hotspot in 2016 election, study says [6]

Reddit, being less popular than other social media, didn't get the same attention than the other website when it come to fake news. This article explains clearly that even if it is a smaller website, it isn't protected from fake news.

VI. PROJECT ATTRIBUTION

In order to organize the work in the team, we are splitting the workload using the ASANA website. We have invited you (the professor) in our team if you want to follow to work progress. However, we can say that the project can be divided in 6 main parts:

Data Collection, Network Analysis Model Construction, Linguistic Analysis Model Construction, Models Evaluation, Merging the two models and making an overall evaluation, report & presentation.

The breakdown of these steps and the person in charge can be found on the asana website.

https://app.asana.com/

VII. PROJECT TIMELINE

- Data Collection - Due: 24 March 2018
- Network Analysis Model Construction - Due: 31 March 2018
- Linguistic Analysis Model Construction - Due: 31 March 2018
- Models Evaluation - Due: 7 April 2018
- Merging the two models and making an overall evaluation - Due: 10 April 2018
- Report & Presentation - Due: 13 April 2018

For a more detailed view, the timeline can be found with the work split on the ASANA website.

References :

[1] With Facebook, Blogs, and Fake News, Teens Reject Journalistic "Objectivity" (http://journals.sagepub.com/doi/abs/10.1177/0196859912458700)

[2] Fake News Detection on Social Media: A Data Mining Perspective (http://www.kdd.org/exploration_files/19-1-Article2.pdf)

[3] Automatic Deception Detection: Methods for Finding Fake News (https://pdfs.semanticscholar.org/939f/eec48ae1abb222cf9881932680b7ec3c68a7.pdf)

[4] Some Like it Hoax:Automated Fake News Detection in Social Networks (https://arxiv.org/abs/1704.07506)

[5] Fake News Research Project (https://pdfs.semanticscholar.org/efef/ebcccc2e7c1ed2252ce09d37d367c930b14c.pdf)

[6] Reddit was a misinformation hotspot in 2016 election, study says (https://www.cnet.com/news/reddit-election-misinformation-2016-research/)