# 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 text information associated with the posts. 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 that perpetuate 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 and densely connected social network provides opportunities for people to propagate misinformation through the grid which could lead to misinformation, massive unrest 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 post content, up votes and comments. Fake News consists of wide variety of classifications (Example - clickbait, conspiracy, bias ) which should be varying in their features, both in user communities, text information and activities surrounding the same.

## II. 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.

## III. DATA EXTRACTION

For our project, we wanted to work on the Reddit network using the official Reddit API. In order to simplify our work in python, we used the PRAW library which is a wrapper of the official API from Reddit. In order to get our post, we made around 20 searches per domain, for each post we fetched other information such as subReddit information and author information. Each request return a JSON object representing as many features as possible about a specific research. For instance, by fetching a specific user in the network, we will have his history on the network (up-votes, posts …) and his current status in the community (karma points, if the user is moderator …). Our data set was reduced from the beginning regarding the fact that Reddit administration has already made a pretty good job on moderating fake user. Most of the time, when we tried to fetch authors of specific posts, it appears that some of them were already banned, which gave us the inability to get information from those users. In order to avoid missing values in our data set, we chose to remove observation for which we were unable to fetch its user's information.

Finally, we chose a panel of relevant features of post, subreddit and author as follow :

* The Id of the post
* The link and the domain (this feature will not be kept for fitting models)
* The author name
* The title and text of the post
* The number and the top ten comments on a post
* The subreddit name
* The subreddit number of subscribers
* The number of active user in the subreddit.
* The number of upvotes
* Karma of the use (comment and post karma)
* If the user is gold or not (already made a donation to the community)
* if the user is the moderator of a subreddit.
* The adviser category and audience target of the specific user.

For all those research, we made a list of news web domain which was already labeled in previous projects (Truth value project, the open sources …), helping us to label our dataset thanks to the reliability of the domain (binomial labeling). At the end of our domain research, we ended with 800 different domain labeled as unreliable sources and 800 as relevant and reliable sources.

## IV. DATA OVERVIEW

The dataset that we managed to put together consist of 9855 observations of 17 features. Each observation has been classified between 2 labels: reliable (for posts having reliable news URL) or fake (for post having unreliable or doubtful news URL). We have 5512 (56 %) observations labeled as "fake" and 4343 (44 %)labeled as " reliable.

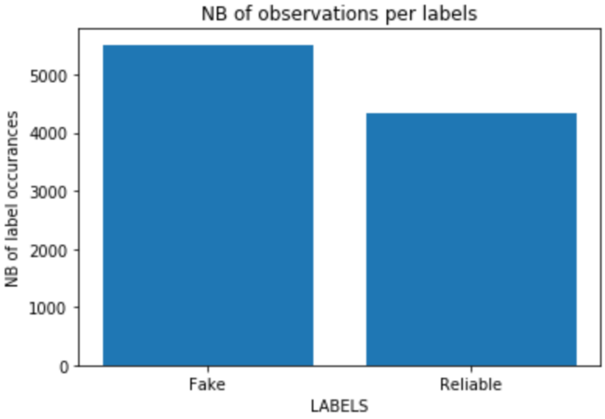
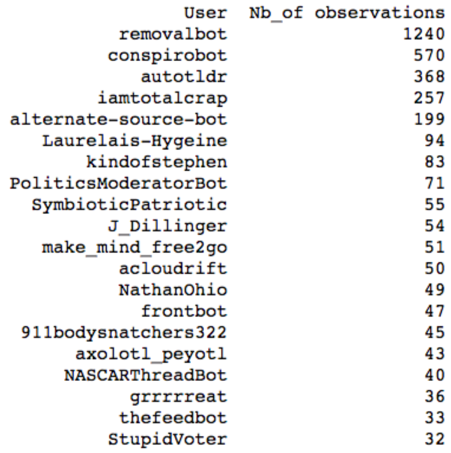


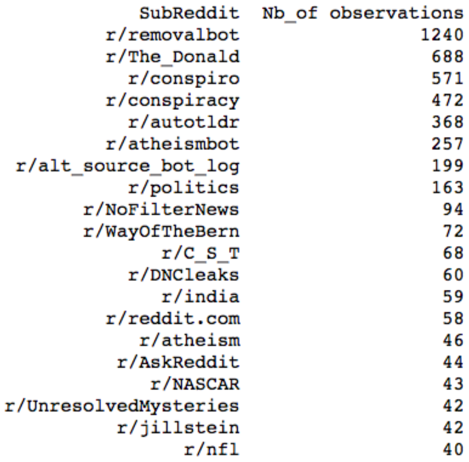
**Fig 1. Visualization of the data repartition per label**



```
                    User  Nb_of observations
             removalbot                1240
            conspirobot                 570
               autotldr                 368
           iamtotalcrap                 257
  alternate-source-bot                  199
       Laurelais-Hygeine                 94
          kindofstephen                  83
  PoliticsModeratorBot                   71
     SymbioticPatriotic                  55
             J_Dillinger                 54
        make_mind_free2go                51
              acloudrift                 50
              NathanOhio                 49
               frontbot                  47
    911bodysnatchers322                  45
            axolotl_peyotl               43
         NASCARThreadBot                 40
               grrrrreat                 36
             thefeedbot                  33
             StupidVoter                 32
```

**Fig 2. Top 20 users collected**

As we can see in the figure above, some of our main publishers are bots. This means that they may have a large influence on our dataset and therefore add some bias to it. We will have to take it into account for our further analysis



```
               SubReddit  Nb_of observations
            r/removalbot                1240
            r/The_Donald                 688
              r/conspiro                 571
            r/conspiracy                 472
              r/autotldr                 368
           r/atheismbot                  257
     r/alt_source_bot_log                199
              r/politics                 163
          r/NoFilterNews                  94
          r/WayOfTheBern                  72
                 r/C_S_T                  68
             r/DNCleaks                   60
                 r/india                  59
            r/reddit.com                  58
               r/atheism                  46
            r/AskReddit                   44
                r/NASCAR                  43
  r/UnresolvedMysteries                   42
             r/jillstein                  42
                   r/nfl                  40
```

**Fig 3. Top 20 subreddits**

As we can see on the figure above, the main subreddit is the removalbot one because it is a bot that post only in his sub-reddit. However, we can also notice that some of the other main subreddit are "The_Donald" , "conspiro" & "conspiracy" which are subreddits where a lot of fake news are posted.

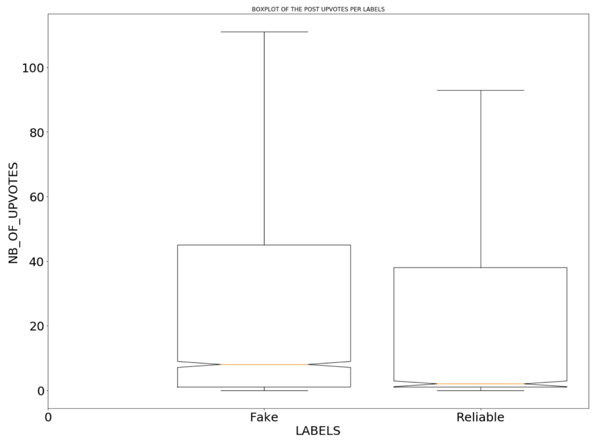


**Fig 4. Visualization of the up-votes score per label**

As we can see on the plot above, it would be difficult to attribute the number of up-votes to a certain label. However, we can notice that fake news up-votes have higher variance. This mainly means that some unreliable posts are highly vouched for.

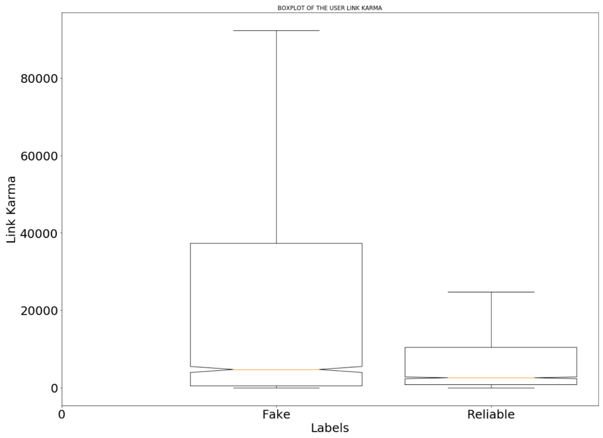


**Fig 5. Visualization of the link karma per label**

The plot enables us to see the boxplot of the link karma per label. We can notice that the mean is almost similar however, fake news observations tend to have again a higher variance.

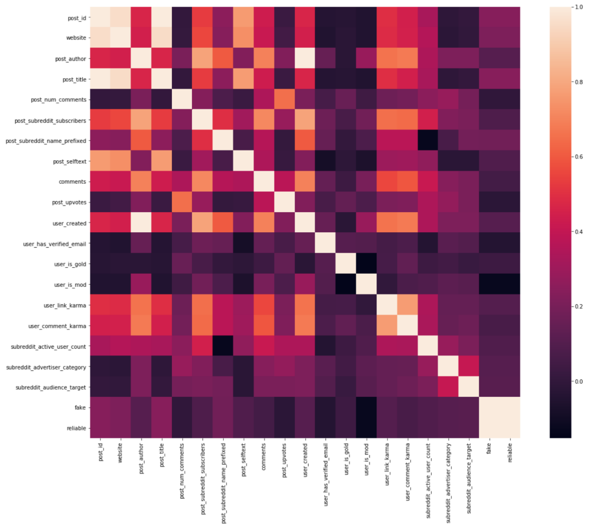


**Fig 6. Correlation matrix of the data features**

Looking at the last two columns (the ones that represent the labels) of the correlation plot, we can notice that there is very little correlation between labels and features. However, we will study the most relevant feature in order to fit a logistic regression in the next part.

From the data overview, it is not possible to infer that certain features enable to predict the reliability of a certain post. Indeed, in order to keep the clarity of the report, we only added the most relevant plots extracted from the study. However, we can notice the of observations generated by bots in our dataset. These observations may add some bias in our study. We will have to take it into account.

# V. MODELS CREATION USING NETWORK ANALYSIS AND NLP

We decided to perform our analysis in two different stages :

- Using the information from the Network to predict the reliability of a future post (based on user, post score and subreddit).
- Using natural Language processing of text features in order to predict the reliability of a post (comments, title, text of the post).

## A.Network Analysis

### 1. Approach

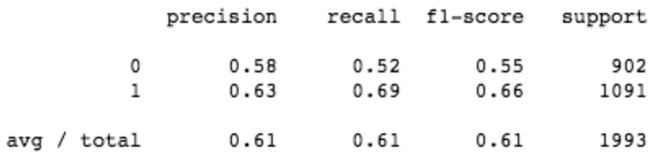
For this part of the project, we wanted to be focused on feature relative to the network such as karma of a user. For doing so, we removed text features from our original dataset (text and title of a post for instance). In order to fit some feature in our model such as subreddit audience target, we chose to create some dummy variables. After doing so, we processed our splitting as an 80% part of the data for the training and 20% part of the data for the testing dataset. Based on Sklearn library, we were able to work on features selection in order to avoid irrelevant features from the dataset. In order to select our features, we work with a recursive feature elimination based on the backward selection which consists of those steps :

* train the model on the training set using all predictors
* Compute the model performance
* Compute features importance or ranking
* Remove the worst predictor of the data and fit the model again.
* Perform those step until having no more predictors.
* Compute the performance profile for all of those models
* Use the optimal model.

In our case, the RFE algorithm selected 11 relevant features for our model.

### 2. Experiment

After working on the data and the model feature selection, we decided to fit a Logistic Regression using tuning parameter such as a lasso penalty with a regularization strength of 1. We chose to cross-validate our model based on KFold cross-validation of 20 splits. The cross-validation mean accuracy of our model is around 58.56 % for this model. We also decided to compute, accuracy, recall and F1 measure of our model over the testing dataset :



```
             precision    recall  f1-score   support

          0       0.58      0.52      0.55       902
          1       0.63      0.69      0.66      1091

avg / total       0.61      0.61      0.61      1993
```

Also creating this confusion matrix and ROC curve :

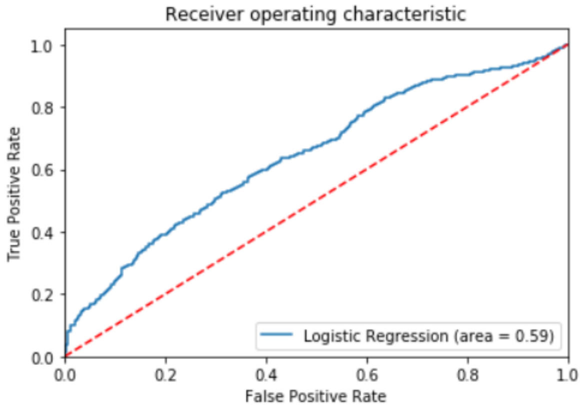


**Fig 7. ROC Curve of the Logistic Regression**

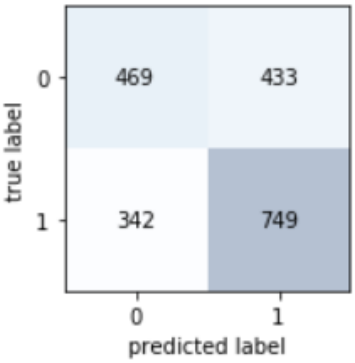


**Fig 8. Confusion matrix of test prediction**

Those final results prove the lack of performance for our logistic classifier. You can summarize from it that the score of a specific post joint with the score measure of the author and subreddit are not relevant to classify the reliability of a post.

Those observations can be explained by the way the Reddit community works. Most of this social network is composed of different communities represented in several subreddit. Most of the community try to put their posts and their subreddit in the foreground in order to be seen by users outside the circle of the subreddit community. This is why some users will up-vote posts, not on the content of the post itself but on how this post can have an influence on things (promote a specific cause for instance). A concrete representation of it will be the current hype of cryptocurrencies, there is one subreddit for each cryptocurrency. Most of the time people will spread a fake news and the community will support it in order to have an impact on the price of this cryptocurrency, which is known to have a high fluctuation on the market.

After studying this issue, we worked on a model based on text, title and comments of a post in order to see if the vocabulary used will help a model to predict the reliability of a post on Reddit.

## B. Natural Language processing on text features

### 1. Approach

We wanted to investigate whether there is a noticeable difference in the use of language that would serve as a good signal for identifying fake news Reddit Posts.

These are 5 major text features associated with the Reddit posts :

1. Author Name
2. Subreddit Name
3. Post Title
4. Post Text
5. Top 10 Post comments.

Upon initial analysis, we found that top Author names and Subreddits responsible for posting Fake News could be bots, which sometimes are evident in their titles. Hence these features are removed to avoid unnecessary bias. Later we found out that removing these features have not resulted in much changes to the accuracy measure.

Text analysis of Posts would be performed using Scikit Learn packages for implementing following steps in model building :

- Vectorizing and TF-IDF transformation and sparse matrix creation of features.
- Supervised text classification machine learning algorithms including Naive Bayes, Logistic Regression and SVM.
- Running combinations of tuning parameters and cross validation for finding model with highest accuracy, precision, recall and F1 values.
- Adding additional features and feature selection routines to improve accuracy.
- Identifying the top features associated with each label and analyzing the top miss classified documents for finding ways to improve precision and recall.

We will be plotting following graphs and plots for understanding performance of our model :

- Training size vs Accuracy
- Improvement in Accuracy vs Combinations of Tuning parameters
- Confusion Matrix
- Accuracy vs Feature selection

### 2. Experiment

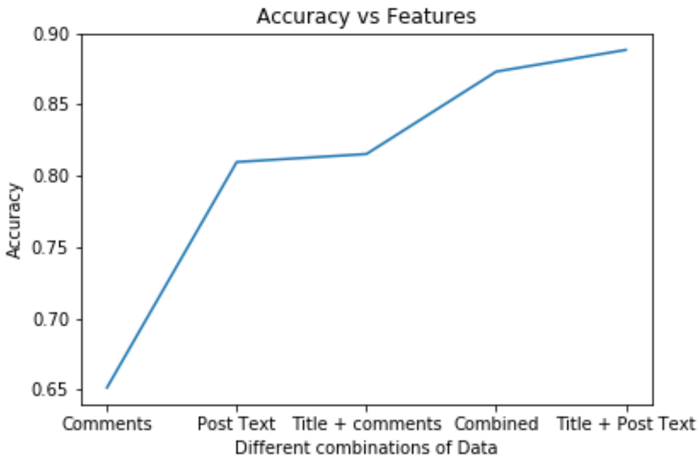


**Fig 9. Accuracy of the model based on features**

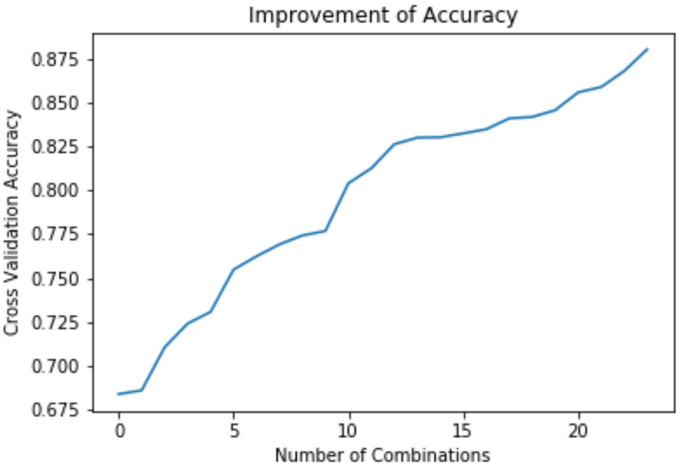


**Fig 10. Accuracy of the model based the number of combination of the tuning parameters**

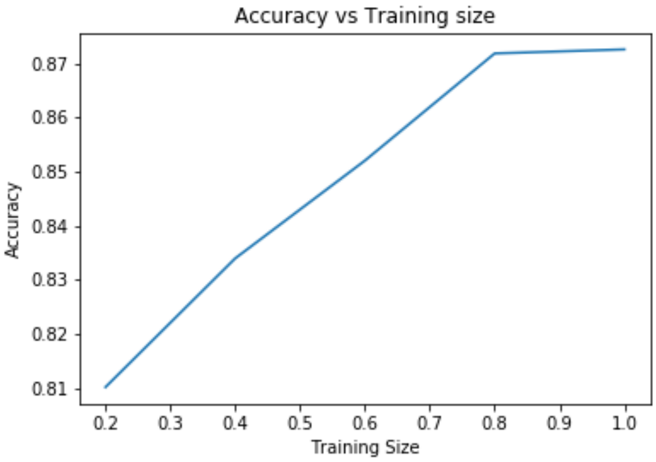


**Fig 11. Accuracy of the model based on proportion of training data**

We have built the model by starting from a simple Vectorization of words in the combined textual features with different combinations of tuning parameters and machine learning models. The best cross validation score was obtained for the below combination

*{'MinFreq': 2, 'Model': MultinomialNB(alpha=1.0, class_prior=None, fit_prior=True), 'NFold': 5, 'NGram': (1, 2), 'accuracy': 0.8561136478944699}*

When the top correlated terms for Fake and Reliable labels were examined, we could notice that there is lot of weightage given to the stop-words and other non frequent features that might affect model performance. This should be resolved by applying Term Frequency - Inverse Document Frequency for vectorizing the documents.

To improve the accuracy and identification to top features , we ran the model using TF-IDF vectorizer in Scikit learn along with the combinations of tuning parameters, and were to improve the cross validation accuracy for the below combination :

*{'MinFreq': 2, 'Model': LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,penalty='l2', random_state=None, solver='liblinear', tol=0.0001, verbose=0, warm_start=False), 'NFold': 5, 'NGram': (1, 2), 'SublinearTF': True, 'Use-IDF': True, 'accuracy': 0.8802638254693049}*

We compared the performance of the Machine learning models over all the iterations of run, and found the following distribution of accuracies.

| Model Name | Mean Accuracy per Model |
|---|---|
| LogisticRegression | 0.846385 |
| MultinomialNB | 0.773643 |
| SGDClassifier | 0.769419 |

**Fig 12. Mean accuracy of different models we used**

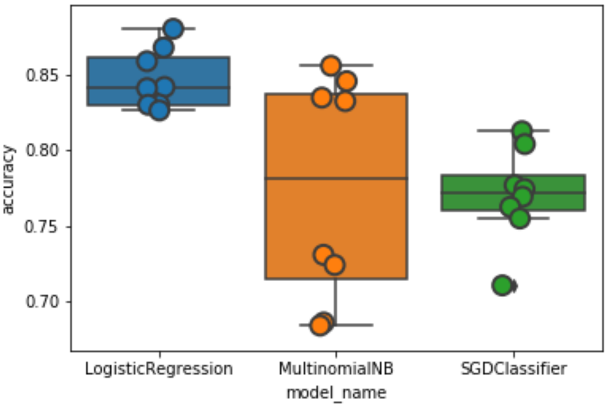


**Fig 13. Accuracies of different models**

Our best model gave us the following results :

```
              precision    recall  f1-score   support

   Reliable       0.90      0.81      0.85       879
       Fake       0.86      0.93      0.89      1092
avg / total       0.87      0.87      0.87      1971
```

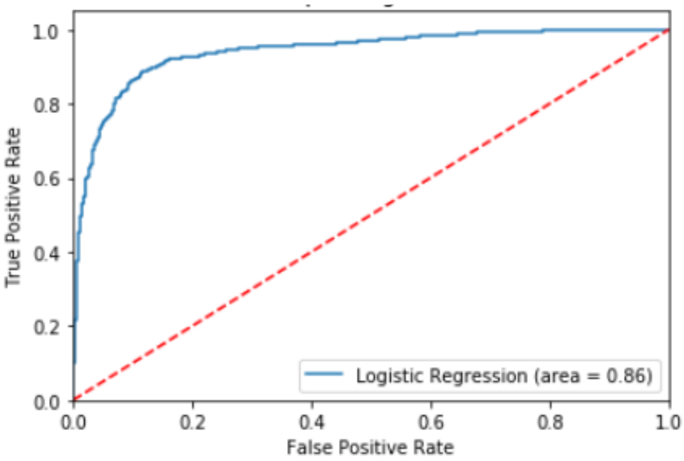


**Fig 14. ROC curve of our best model**

Interestingly, as you could see in the Accuracy vs Combinations of Data Graph, the model with combination of post title and post text scored above the combined data of post title , post text and comments. But we decided to keep the combined data since the features in comments might prove useful in avoiding overfitting.

The improved model has a great performance of 87 percent, with solid precision , recall and F1 scores. We also had plotted the training data set size vs accuracy , and got results with positive trend. From the confusion matrix, we could understand that 170 reliable posts were wrongly classified as Fake, while only 81 fake posts were wrongly classified as Reliable. Hence we decided to investigate the top correlated words and top misclassified posts based on the model.

| Top Reliable Posts Features | |
|---|---|
| Most correlated unigrams: | Most correlated bigrams: |
| news<br>. gizmodo<br>. com<br>. in<br>. uk<br>. contactmusic<br>. india<br>. bitcoin<br>. news24<br>. ibtimes<br>. for<br>. jpg<br>. undelete<br>. unreddit | com news<br>. news comments<br>. contactmusic com<br>. news24 com<br>. co uk<br>. com articles |

**Fig 15. Top reliable features representation**

| Fake Post Features | |
|---|---|
| Most correlated unigrams: | Most correlated bigrams: |
| org<br>. conspiracy<br>. trump<br>. russia<br>. obama<br>. presstv<br>. shills<br>. us<br>. hillary<br>. sputniknews<br>. clinton<br>. level<br>. muslim<br>. war | conspiracy comments<br>. original conspiracy<br>. sputniknews com<br>. discourse level<br>. link conspiracy<br>. conspiracy link |

**Fig 16. Top fake features representation**

We extracted most misclassified observation :

Predicted = Fake
Truth = Reliable
proba = 0.9473

SeaWorld Employee Infiltrated PETAsource link: **[SeaWorld Employee Infiltrated PETA](http://www.rttnews.com/2522638/seaworld-supposedly-infiltrates-peta.aspx)**

poster: **[RoachesWinInTheEnd](/u/RoachesWinInTheEnd)**, [original conspiracy link](/r/conspiracy/comments/3daqn7/seaworld_employee_infiltrated_peta/)

---

Predicted=Fake
Truth=Reliable
proba=0.9184

Obama Tells Netanyahu To Stop Pushing Congress Toward New Iran Sanctionssource link: **[Obama Tells Netanyahu To Stop Pushing Congress Toward New Iran Sanctions](http://www.chinatopix.com/articles/33819/20150123/obama-tells-netanyahu-stop-pushing-congress-toward-new-iran-sanctions.htm)**

poster: **[NOT_JTRIG](/u/NOT_JTRIG)**, [original conspiracy link](/r/conspiracy/comments/2tfs1n/obama_tells_netanyahu_to_stop_pushing_congress/)

---

Predicted=Fake
Truth=Reliable
proba=0.9132

ATTENTION UNDECIDED VOTERS: Last call for the Trump Train! NO BRAKES!'#NOT ONE LEAKED EMAIL WAS ABOUT HELPING US, THE AMERICAN CITIZENS!!!!', "It took me awhile, being a Bernie voter. But we can't do it. So many crimes she has committed, it's hard to keep track of. Here's the kicker...

#*Out of all the leaked emails, none of them were about how to help the American citizen.*

Not a single thought was put forward to improving our lives. Every single one was self serving in some way.", '[deleted]', '#THIS IS THE FINAL STRETCH BOYS, TOMORROW WE CELEBRATE OUR INDEPENDENCE DAY', 'NO BRAKES ON THE TRUMP TRAIN

---

Predicted = Reliable
Truth = Fake
proba = 0.8974

07-15 21:47 - 'Bitcoin = Nazi Money (!?)' (dailystormer.com) by /u/tinus42 removed from /r/Bitcoin within 1471-1476min[Bitcoin = Nazi Money (!?)](https://reddit.com//r/Bitcoin/comments/4svn5j)

[Go1dfish undelete link](http://r.go1dfish.me/r/Bitcoin/comments/4svn5j)
[unreddit undelete link](https://unreddit.com/r/Bitcoin/comments/4svn5j)

We could see that the top fake indicator features like ‘Trump’, Çonsipiracy’, ‘Russia’, and similar politically charged words , appear in the Reliable posts comments, at times as genuine news or sometimes mentioned in politically unrelated posts as part of spamming. Similarly , ‘bitcoin’ is in top Reliable post features, and hence the model was not able to classify one outlier post featuring bitcoin as fake news.

The following additional attempt were made to improve the accuracy.

- Addition of new features like number of capital letters, exclamation points, length of comments
- Feature selection using Chi squared test and SelectKBest in Scikit learn

But since the present accuracy is really high, further addition of features nor Feature Selection could improve the accuracy and F1 Measures of the present model.

## VI. CONCLUSION

The question of ‘How to curb Fake News propagation in Social Media through online network analytics ?’ is a simple , yet highly ambitious challenge which could be compared to the likes of ‘How can we solve the problem of spamming in Email ?’. The simple, yet highly intricate challenge is effectively tackled by Google, Microsoft and other major companies through spam filtering system. The challenge in Fake News problem, though it might seem simple at first look, is that Fake News evolves everyday, appearing in wide variety of contexts, has different kinds of sub classifications within them, and very difficult to track the original sources. Considering the complexity of problem, our efforts to understand the basic dynamics of Fake news posts in Reddit and whether there exists strong signals for classification of posts were indeed successful. Network Analysis of the Reddit posts detected average signals that shows that there is weak , yet detectable pattern in the nature of users and sub-reddit communities that are chains in fake news propagation. Natural language processing of the text information in Reddit posts shows highly strong signals of political and religious agendas that lurk behind the class of Fake News involved in our target dataset. Our studies definitely show that given a system of proper labelling of Social Media posts through a system of collective consensus, we would be able to use an predictive model to effectively classify and alert users about the quality of the posts. The future link of work to progress towards tackling challenge of Fake News includes extracting more contextual features to improve recall of Fake news, studies towards clustering of user communities based on their affiliations and bias, and ranking of users based on the quality of posts.

## 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/ )