CMPS 287 - Final Project Report

Majd Alkawaas mma250@mail.aub.edu,
Melhem Rahmeh mfr08@mail.aub.edu,
Mohamed Louai Bouzaher mlb03@mail.aub.edu,
Nathalie Nassar nwn05@mail.aub.edu

May 2022

Bot accounts have been a significant problem for social media websites especially Twitter and Reddit, most bots are harmful and they seek to create chaos or generate useless content on popular channels of social media communication. Our solution is a set of different machine learning models that performs Reddit bot detection on different sets of data about the users. We have chosen this approach to boost the accuracy of the results.

We have chosen comment/post level detection because account-level approaches require increased amounts of user data and there is a scarcity of officially verified bot accounts, whereas comments and posts can be analyzed using Natural Language Processing techniques.

The best model we proposed is logistical regression, it is giving the highest accuracy for the 4 classification levels we are proposing:

- User Posts Titles : 90.4%
- Sub-reddits of User Posts: 92.3%
- User Comments : 91.3%
- Sub-reddits of User Comments : 90.5%

1 Introduction

Trolls and bots in social media are widespread, and it has been proven that they have unrecognized and significant effects on the users. The actions of a bot might affect our opinions, and thus the quality and accuracy of information we are getting. Moreover, some bots and trolls might be considered bad actors for having negative political, economic, and health effects.

Reddit, also known as "the front page of the internet," is a social media platform for news aggregation and discussions. On Reddit, people are able to form communities around shared interests or sometimes shared dislikes. The power these communities wield is constantly increasing, and it was shown, a year ago, by the r/wallstreetbets Sub-reddit; the participants of this Sub-reddit were able to manipulate the stock market and drive the prices of GameStop stock to a significant number. Furthermore, a huge number of bots has been detected by the users and moderators on this specific Sub-reddit. Another instance of bots interfering with the public opinion was annoucned in Reddit's Transparency report of 2017. In this report have released more than 950 Russian bots accounts. These accounts were controlled by Russian organizations and most of their activities were focused on political communities on Reddit mostly relating to the 2016 US Election. In addition, some bots exist to downvote or upvote specific content or that of a specific user for marketing, or economic purposes.

While bot detection is being explored on other high profile platforms such as Twitter with deep learning, there is extremely limited work on bot detection applications on Reddit, and in the wake of the ever-changing political and economic scene, there is an increasing need for accurate methods of detecting bots on such platform given that they become a source of information for millions of users. Moreover, The nature

of Reddit enables a very suitable environment for bots from all the existing social media platforms. The behavior of these bots, on Reddit, is turning some of the communities into unsafe and insecure online spaces that require termination.

2 Related Work

As we have mentioned there is not much work on the Reddit bot detection issue but some parties are making some great efforts in this matter. Even Reddit never announced any in deepth details on their efforts or progress with this issue but some work on the matter has been done by Reddit community members.

A group of graduate students at Stanford University worked on this problem for their thesis. They have applied deep learning to Reddit in an exploration of its viability in comment-level bot detection with a limited supervised dataset. They have performed a contextual analysis of the comments of these users using a variety of deep learning models and architectures (BERT and LSTM, RCNNs) they have achieved an AUC of 84.6% using an RCNN architecture on their very minimal dataset.

Medium project by Brandon Punturo, in this project the author dependent on the list of bots we are depending on. He built optimized random forest classifier and they final results and metrics were not promising.

Despite the fact that Reddit has significantly less labeled bot datasets than Twitter, the issue is gaining traction, and Reddit has escalated its attempts to combat it. The data of around 900 Russian troll accounts that submitted over 7000 total comments was revealed in Reddit's 2017 transparency report, producing an extremely limited but workable supervised dataset. This list or Russian bots was started by a professor at Boston University and then Reddit took over the process

3 Dataset

Our process of acquiring the data follows these steps:

1. Identify verified lists of bot accounts (Reddit 2017 transparency Report, autowikibot, botwatch)
2. Scrap the username of the account from their respective web pages.
3. Identify a list of normal users that were active during the same time the bot accounts were active.
4. Retrieve the accounts posts, and comments using the Reddit official API, and the PushShift API (Unofficial API).
5. Store the data in a MongoDB server given that a user has multiple comments and posts and storing this data as CSV means storing the parts of the data multiple times.
   That's the database diagram, and you can find the dataset explanation below.
6. Retrieve the required data depending on the model i.e training the comments body model we retrieve the comments and the label of all the users in our data.

The dataset consist of three main parts and each part has of the following features:

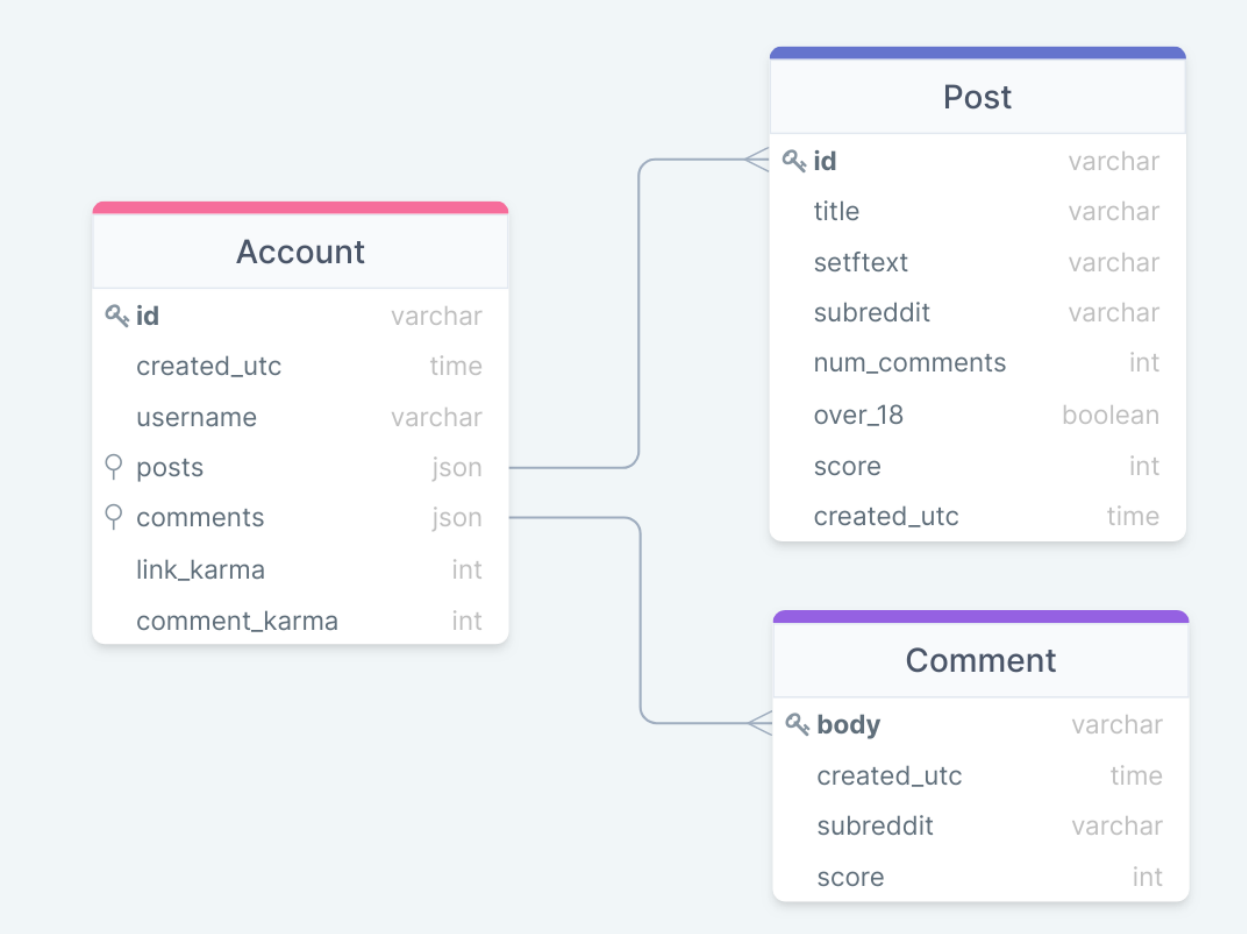


Figure 1: Database relational model

1. User Data
   (a) username
   (b) cakeday: (user account creation date)
   (c) comment karma: value of the sum of a given user's upvotes and downvotes.
   (d) post karma.
   (e) is bot: A boolean value added by us during the process of retrieval chosen depending on the source of the username (from list of bots or not).
2. Post Data
   (a) comment body
   (b) creation date: a timestamp of date and time.
   (c) Sub-reddit: the Sub-reddit at with comment was posted.
3. Comment Data
   (a) post title
   (b) post body
   (c) number of comments
   (d) creation date: a timestamp of date and time.

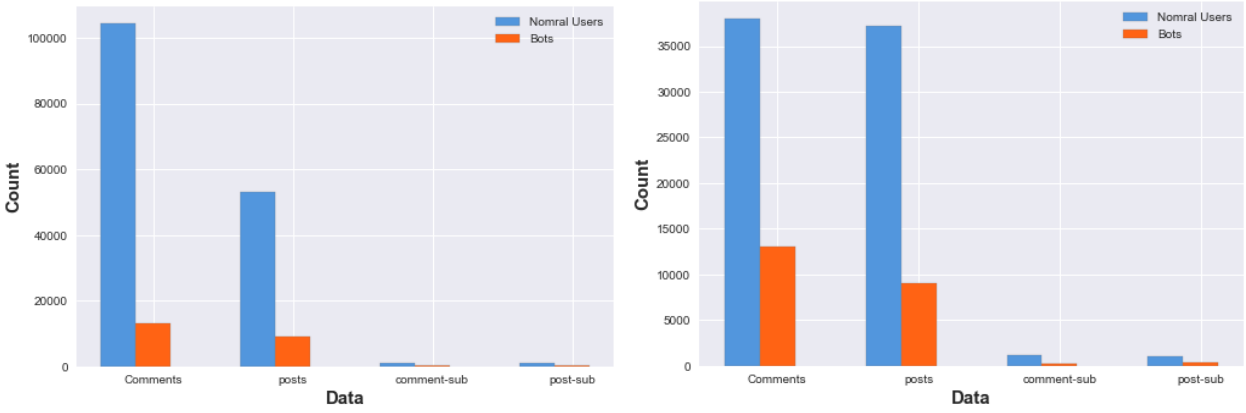
(e) Sub-reddit: the Sub-reddit at which the post was posted.

As for **data preprocessing**, we have used python and the Natural Language Toolkit (NLTK), we performed the following steps:

1. Removed all the special characters, stopwords, and all single characters, and handled spacing and uppercases latters.
2. Tokenization (for breaking phrases, and passages into tokens to help the computer understand the text). Experimented with both Stemming and lemmatization.
3. Document Representation with TF-IDF (Term-Frequency Inverse Document-Frequency), to be able to represent each document as a vector of terms.

Addressing imbalances in the dataset:

For the comments body data, we have an extreme imbalance between the normal user class and the bot user class the original ratio of normal:bot is around 11:1 as you can see in the figure so we performed under sampling on the normal users class, and the new ratio of normal:bot is around 3:1.



(a) Before Under Sampling

(b) After Under Sampling

4 Model

Our project follows the following pipeline.

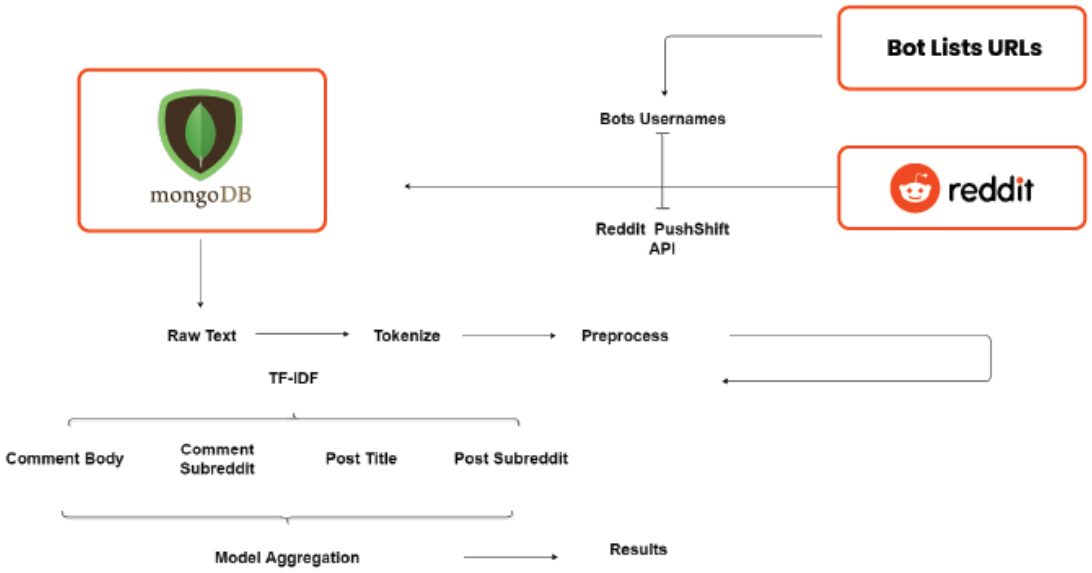


Figure 3: Database relational model

Given that we have different types of data (dates, comment/post title) we have four different models each handling a different form of data. This is also contributing to boosting the accuracy given that the models have varying F1 Scores.
For each subset of the data We have a the following models

1. comment Sub-reddit model
2. comment body model
3. Post title model
4. Post Sub-reddit model

For each one of the mentioned models, we experimented with different combinations of the following models: SVM , Logistical Regression, Neural Networks , KNN and Naive Bayes.

- **SVM**: The SVM model was proposed by Cortes and Vapnik. It is a supervised learning model mostly used for classification and regression problems. The SVM aims at calculating the weights w and bias b such that the hyperplane found perfectly separates the data while maximizing the margin.

  For an efficient non-linear transformation, we used kernels.

  The first kernel used was linear and the second was polynomial.

- **Logistical Regression** : Logistic regression is a statistical model used for binary classification, and it can be generalized to multiclass classification.
  Logistic regression uses the Sigmoid Function, that is $\frac{1}{1+e^{-x}}$

  It outputs a value between 0 and 1, that's the probability if the positive label class. Here, it's the $isBot$ label, and if it's greater than 0.5, the $isBot$ will be equal $True$, else the $isBot$ will be equal $False$.

Logistic Regression try to minimize the "cross-entropy error", that is:
$\frac{1}{N}\sum_{n=1}^{N} ln(1 + e^{-y_n w^T x_n}) = 1$

It is an extensively employed algorithm for classification in industry because of it's simplicity and good results.

- **Neural Networks** Neural network is based on contextual long short-term memory architecture that exploits both content and metadata to detect bots at the post level: the pre-processed data was fed as input to the network in order to see whether it consists of a bot or genuine user.

  After the training using the ReLU activation function for the deep layers and Sigmoid activation function for the output layer knowing it consists of a binary classification.

  After training for 100 epochs, the accuracy on the test dataset was 0.89.

- **KNN**: Briefly, K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

- **Naive Bayes**: Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem.

4.1 Comment Subreddit

The comment subreddit is the subreddit at which the comment was posted.
During the classification of the subreddit model, we used three models in order to evaluate their efficiency in correctly predicting bots from normal users. For that, we used: Support Vectors Machine(SVM), Logistic Regression, and Neural Networks. For logistical Regression and SVM, we performed hyperparameter tuning using grid search, and random search for neural networks.

Model	Accuracy	F1 Score
SVM Linear	0.86	0.83
SVM Polynomial	0.86	0.84
Logistical Regression	0.87	0.84
Neural Networks	0.84	0.82

4.2 Comment Body

During the classification of the subreddit model, we used three models in order to evaluate their efficiency in correctly predicting bots from normal users. For that, we used: Support Vectors Machine(SVM), Logistic Regression, and Neural Networks. For logistical Regression and SVM, we performed hyperparameter tuning using grid search, and random search for neural networks.

Model	Accuracy	F1 Score
SVM Linear	0.90	0.88
SVM Polynomial	0.91	0.90
Logistical Regression	0.91	0.91
Neural Networks	0.90	0.90

4.3 Post Title

The post title is the text of the post title.
We did not consider the posts bodies as it has no limit, and can include videos, images, tables, index tables... During the classification of the Post Title model, we used four models in order to evaluate their efficiency in correctly predicting bots from normal users. For that, we used: Support Vector Machines, Logistical Regression , Naive Bayes and KNN. For logistical Regression, SVM, KNN, we performed hyperparameter tuning using grid search, random search for neural networks, and no hyperparameter tuning for Naive Bayes, as there are parameters to tune.

Model	Accuracy	F1 Score
SVM Linear	0.89	0.89
KNN	0.87	0.84
Naive Bayes	0.61	0.66
Logistical Regression	0.91	0.89

4.4 Post subreddit

The comment subreddit is the subreddit at which the post was posted.
During the classification of the subreddit model, we used three models in order to evaluate their efficiency in correctly predicting bots from normal users. For that, we used: Support Vectors Machine(SVM), Logistic Regression, and Neural Networks. For logistical Regression and SVM, we performed hyperparameter tuning using grid search, random search for neural networks.

Model	Accuracy	F1 Score
SVM Linear	0.91	0.90
SVM Polynomial	0.89	0.89
Logistical Regression	0.90	0.90
Neural Networks	0.90	0.89

4.5 Disregarded experiments:

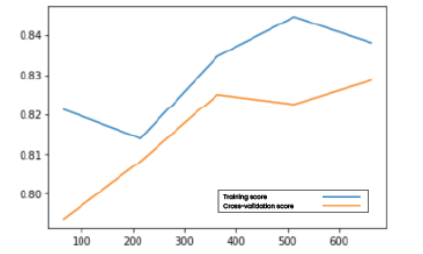
We have experimented with model aggregation, we performed weighted average vote using the following formula:

$$AggregatedProbability = \frac{(f1_A * model_A) + (f1_B * outProb_B) + (f1_C * outProb_c) + (f1_D * outProb_D)}{4}$$

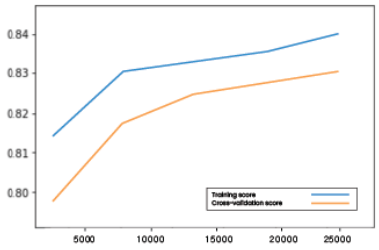
This experiment did not yield any good results given the scarcity of the data because not all normal users and bots have enough posts and comments, most posts and comments belong to a smaller subset of the collected accounts.
We have also experimented with compost features regarding the users' behavior (frequency of posting and comments, times of posting and commenting). This experiment also did not yield any promising results, and with a further investigation of our data and its source we discovered that an analysis conducted on the same dataset at Boston University discovered that all the Russian bot accounts were mirroring the behavior of the normal users.

5 Results

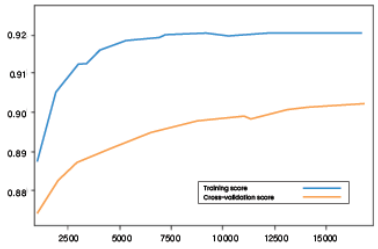
5.1 Learning curves



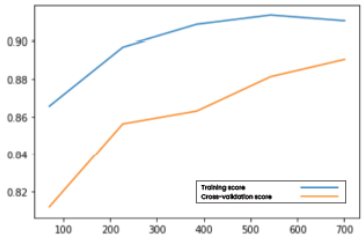
Logistic Regression - Subreddits of User Comments



Logistic Regression - User Comments



Logistic Regression - User Posts



Logistic Regression - Subreddits of User Posts

5.2 Classification Reports / Confusion matrices

The best model for all 4 detection level is logistical regression.

Model / Metric	F1 Score	Accuracy	Precision	Support
Comment Subreddit	0.90	0.91	0.90	553
Comment Body	0.91	0.91	0.92	20238
Post Title	0.89	0.90	0.90	24917
Post Subreddit	0.92	0.92	0.93	585

6 Conclusion

Detecting bots by analyzing their behavior was a difficult task given the scarcity of the data and the mirroring behavior of most bot accounts in our dataset, since they can behaved approximately like any other user. However, inspecting the language they use to post and comment and the textual patterns they follow to do so, proved to be quite effective and yielded high F1 scores and high Accuracy.

The next step is to further experiment with word embedding models and contextual analysis to boost our metrics (F1, Accuracy). Our aim is to squeeze out more features from the existing dataset to compensate for the small amount of data about bots. In addition to that deploy the models and provide a user interface for Reddit users to try our tool.