**Project Title: Twitter**

-Project statement. Based on the project description, state a well-defined question that

you’ll address in the project.

- Literature review: Please provide basic background and relevant references on your

project topic. A paragraph description should suffice.

- Preliminary EDA. It is important that at this stage you are already working with the

data for your project, and should demonstrate some basic exploratory results working

with your data. Explain your preliminary data exploration. You may include simple

visualizations or just a verbal description.

# Problem statement

With spread of social media and their increasing impact on the news the topic of bots and spread of fake news is one of the hot topics now. The goal of this project is to use machine learning techniques which we have learned during this semester to detect the tweets which were generated by Twitter bots from human users.

# Literature Review

Bots have a heavy presence in the social media. “Of all tweeted links to popular websites, 66% are shared by accounts with characteristics common among automated “bots,” rather than human users.”[[1]](#footnote-1).

these environments. A recent review [20] proposed a taxonomy

describing three types of approaches: (a) methods based on social

network; (b) systems based on crowd-sourcing and human computation;

(c) algorithms based on predictive features that separate

bots from humans. Our framework falls in the last category.

# Data Resources

Data is mined and collected from the Twitter developer API (using tweepy)

Automated accounts play a prominent role in tweeting out links to content across the Twitter ecosystem. The Center’s analysis finds that an estimated 66% of all tweeted links to the most popular websites are likely posted by automated accounts, rather than human users.

Certain types of sites – most notably those focused on adult content and sports – receive an especially large share of their Twitter links from automated accounts. Automated accounts were responsible for an estimated 90% of all tweeted links to popular websites focused on adult content during the study period. For popular websites focused on sports content, that share was estimated to be 76%.[[2]](#footnote-2)

Broadly speaking, Twitter bots are accounts that can post content or interact with other users in an automated way and without direct human input.

Bots are used for many purposes. This study focuses on a particular kind of bot behavior: bots that tweet or retweet links to content around the web. In other words, these are bots that post or promote specific websites or other online content.

Many bots do not identify themselves as bots, so this study uses a tool called [Botometer](https://botometer.iuni.iu.edu/" \l "!/) to estimate the proportion of Twitter links to popular sites around the web that are posted by automated or partially automated accounts. One [study](https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15587/14817)suggests [Botometer](https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15587/14817) is about 86% accurate, and Pew Resesarch Center conducted its own independent validation tests of the Botometer system. To acknowledge the possibility of misclassification, we use the term “suspected bots” throughout this report. For details on how Botometer functions, see the [methodology](https://www.pewinternet.org/2018/04/09/bots-in-the-twittersphere-methodology/).[[3]](#footnote-3)

* Netflix Bot (@netflix\_bot) automatically tweets when new content has been added to the online streaming service.
* Grammar Police (@\_grammar\_) is a bot that identifies grammatically incorrect tweets and offers suggestions for correct usage
* Museum Bot (@museumbot) posts random images from the Metropolitan Museum of Art
* The CNN Breaking News Bot (@attention\_cnn) is an unofficial account that sends an alert whenever CNN claims to have breaking news
* The New York Times 4th Down Bot (@NYT4thDownBot) is a bot that provides live NFL analysis.
* PowerPost by the Washington Post (@PowerPost) is a bot that provides news about decision-makers in Washington.

EDA

Statuses Count

• Followers Count

• Friends Count

• Favorites Count

• Listed Count

• Default Profile

• Geo Enables

• Profile Uses Background image • Verified

• Protected

Similarly, for tweet level classification we use only 6 features, apart from the tweet content itself:

• Retweet Count

• Reply Count

• Favorite Count

• Number of Hashtags • Number of URLs

favorite\_count', 'followers\_count', 'friends\_count',

'in\_reply\_to\_status\_id', 'retweet\_count'

**References**

1. We will use some references from CS109A/Twitter\_project references
   1. Tweepy Python Library
   2. Twitter’s developer resources: developer.twitter.com
2. CS109B class contents
3. Other papers about twitter and bots

1. “Twitter Bots.” [↑](#footnote-ref-1)
2. <https://www.pewinternet.org/2018/04/09/bots-in-the-twittersphere/> [↑](#footnote-ref-2)
3. <https://www.pewinternet.org/2018/04/09/bots-in-the-twittersphere/> [↑](#footnote-ref-3)