Twitter Hashtag Hijacking Detection

Shivani Matta
2018191
IIITD
shivani18191@iiitd.ac.in

Mukul K. Rajak
2018054
IIITD
mukul18054@iiitd.ac.in

Tejas Dubhir
2018110
IIITD
tejas18110@iiitd.ac.in

Anmol Kumar
2018382
IIITD
anmol18382@iiitd.ac.in

Rakshit Singh
2018079
IIITD
rakshit18079@iiitd.ac.in

1. MOTIVATION

Twitter is one of the most active and open platforms for sharing information and opinions on any subject matter. Since the onset of hashtags, online communication has become trend-driven. It has also helped in making the content on the platform somewhat structured. People can choose to follow any trend popular at any location in the world. Hashtags drive people's attention and affect their opinions, help them find people with similar interests, or even build an audience for marketing. Brands, political parties, and other organizations try to leverage hashtag promotions for campaigns and endorsements. But often relevant and meaningful twitter trends are targeted through spam and other content under the same hashtag pushed by a legitimate user. This is termed as Hashtag Hijacking. Spammers, tweet farms and bots try to bypass twitter restrictions and guidelines making people more exposed to privacy and security issues, fake news, etc. and users are always at a risk of being targeted for trolling and cyberbullying.

2. PROBLEM STATEMENT

In this project, we will try to classify legitimate and hijacked tweets/trends by analysing and evaluating text corpus and common spam account attributes under popular hashtags using NLP and machine learning techniques. We will be curating our own dataset for tweets using Twitter Search and Streaming API for both historical and real-time data. We further introduce a new attribute of scoring popular user sentiment and evaluating all the tweets under a particular hashtag/campaign to improve upon existing methods.

3. LITERATURE REVIEW

HashMiner does feature characterisation and analysis of hashtag hijacking using real-time neural network.[1] The paper characterizes the increases the cases of misuse of hashtags in the following ways:

1. Using trending hashtags to advertise some specific brand to gain popularity through the hashtag.
2. Users post unrelated content to distract the population from the main topic and distribute spam, fake and unwanted content.
3. Some of the hashtags are also hijacked to post political opinions which can sometimes be sensitive to the users.

For this paper, the authors have collected 10240 tweets and classified them for 10 hashtags in totality.

Jain, N. et al. presents a general algorithm to detect and analyze hijacked tweets from Twitter.[2] The paper proposes an algorithm that divides the collected datasets into various categories like entertainment, politics etc. This categorization is used to detect and analyze the hijacked tweets. The researchers collect tweets from Twitter and manual annotation of the tweets is done to prepare training and testing dataset. The dataset is used to find the top occurring words and assign tf-idf score to the top words in the dataset. These top scoring words are then saved into a dictionary corresponding to each category. A score is calculated for the new incoming tweets by matching the tokens in the tweet with words in these dictionaries. For each math a score of 1 is assigned. The algorithm outputs score for the tweets and high scoring tweets signify that the tweet is not hijacked whereas if a tweet has a low score then it is a misleading/hijacked tweet. The research paper however does not deal with URLs and is hence unable to detect hijacked tweets if they only contain a URL and no text.

VanDam, C. and Tan, P. 2016 have proposed a novel framework that combines information about the temporal distribution of hashtag frequencies and the content of their tweets and the users who posted the tweets to determine whether a hashtag has been hijacked.[3] They have selected around 2667 trending hashtags that have appeared in at least 100 tweets on the same day using Tweepy from 766,057 unique users and applied NLTK to 98,234 unique nouns from the data. They have generated 2 data matrix for each hashtag (i.e., X:a term frequency per day and U: a user frequency per day.) and normalized by frequency of tweets, which further decomposed into 3 latest factors (i.e., W: a terms by topic matrix, V: a user by topic matrix, H: a day by topic matrix) and detected hashtag hijack by employing a multimodal non-negative matrix factorization approach to learn

each hashtag's underlying topics by optimizing the following function:

$$\|XWH^T\|_F^2 + \alpha\|UVH^T\|_F^2$$

s.t. $W_{ij} \geq 0, V_{kj} \geq 0, H_{hj} \geq 0 \; \forall \; h, i, j, k$

They have also applied an alternating minimization approach to estimate the latest factors W, V, H. And then they have updated the matrix value until they converge using:

$$W_{ij} = W_{ij}(XH)_{ij}/(WH^TH)_{ij}, V_{kj} = V_{kj}(UH)_{kj}/(VH^TH)_{kj}$$

$$H_{hj} = H_{hj}(X^TW + \alpha U^TV)_{hj}/(HW^TW + \alpha HV^TV)_{hj}$$

This was followed by Hotelling's t2 (a popular change detection algorithm) test to look for the hijacked topic. The hashtag that fails this test has been considered hijacked. They have found that a window size of 3 days gave the best accuracy of 66.9% when user matrices are decomposed into 5 topics.

Spam detection and hashtag hijacking have many common research goals by principle. Chu, Z. et al. 2012. focuses on detecting spam-driven campaigns rather than focusing on individual tweets and accounts.[4] Their approach is more effective and robust as analyzing individual tweets is computationally more resourceful while hardly meeting stringent requirements. Also, spammers usually tend to make multiple accounts to stay under the radar to avoid detection, but their end goal is to get maximum reach and engagement in the form of content, which is hashtag or trend-driven in nature. In this study, the researchers collected 50 million tweets posted by 22 million users with the help of Twitter API and partitioned the tweets containing URLs based on campaigns. Spamming URLs with unrelated hashtags are one of the most common methods for the purpose. URLs are usually shortened by online tools to avoid spam blacklists with an added layer of multiple URL hops. For this, the researchers designed an extension that extracted the last hop URLs and compared them for affiliate/referral dissemination. Scripted automation along with human inspection for spam content / unrelated URL content/duplicate or similar content via multiple or single accounts was used for annotating the dataset. Further, many custom features based on Twitter parameters were designed, with the help of techniques like tf-idf scoring and 2nd order similarity scoring (to detect synonym interchanging). Ultimately, a Random forest with cross-validation scoring was chosen as the machine learning algorithm for its ensemble characteristics. Although the FPR and FNR are comparable, the method is highly efficient for real-time detection, considering the amount of tweets made per unit time on the platform. The main advantage of this approach is that it is trend or hashtag driven. But the major limitation is that only URL have been primarily been the focus of this study. Other parameters related to sentiments could have improved the performance metrics without increasing much of the complexity.

Hadgu, A. et al. 2013. present present the study on the polarization of hashtags on Twitter and then shows how certain jumps in polarity are caused by hijackers.[5] They have used retweets of labelled seed users (mainly political leaders) to obtain political orientation of users. By analyzing their hashtag usage, leaning to hashtags is assigned. "Change points" with a sudden jump in leaning are identified which is due to the activity of "hashtag hijackers". Their methodology first identifies seed users and then the tweets scanned for hashtags. Retweeting people are identified and filtered based on their location. After that, leaning with respect to a party is defined using lean formula and validated against wefollow.com, twellow.com and persecuting data. At last, outliers are marked as change points from which hijacker points are identified. The shortcomings of this paper are that URLs, images and sarcasm cannot be detected by their model and thus, we will have to find a way to include these too.

4. PROGRESS

We used the Twitter API to collect the tweets corresponding to the hashtags which are trending. For that, we had to apply for a Twitter's developer account which usually takes weeks to be approved. We were asked about the project and about how we were going to use the collected data, once we replied to them about the details of our work, we got access to the public tweets. These tweets were used to create the data-set. Another challenge we faced was that there could only be 100 tweets extracted through the API. The tweets were manually annotated into categories of whether the tweet is hijacked or not. Our goal is to collect 10,000 tweets and annotate them to create out data-set. Since tweets can be subjective and it depends on the person whether he/she considers it to be hijacked or not, we are following majority voting approach. Each of the tweet has been manually read and classified by each and every team member. Out of 5, if 3 members need to agree to classify a tweet hijacked. This way we are removing opinionated bias. Moreover, there are numerous hashtags and their respective popularity's are very dynamic. Currently, we have annotated over 6000 tweets and are aiming to finish the data annotation by the end of the week.

After the creation of sufficient size of the dataset, we proceed with dataset cleaning and using NLP techniques like tokenization, stop-words removal, and stemming/ lemmatization. Techniques like tf-idf, ranking algorithms, and vector space models are used for scoring words. In addition, we will be analyzing if the sentiment of the current tweet matches with the sentiment of the majority of tweets of that hashtag. Then the weighted cumulative score from both these parameters would help us in classifying the nature of the tweet.

We then try to apply various techniques on the obtained posts or tweets and obtain the sentiment, retweets/reposts, likes and dislikes, uploader, uploader's data (like followers, following, verified or not), related hashtags (their relative scores), the validity of URLs in the post, and the tf-idf score of the post. We then let a neural network run through the dataset we will manually prepare, which will contain a detailed examination of the posts and annotated classes (hijacked or not). After which we will use the obtained weights to validate the rest of the dataset which will ultimately return us an accuracy and precision measure.

4.1 Baseline results

As the tweets were being manually annotated, we could observe that the hijacked tweets usually are asking for retweets/ likes /comments in a given time frame, or trying to obtain more subscribers/followers on some given account, also known as spam. Some tweets are trying to advertise their products. They use the popularity of the trending hashtag as a way to get their motive fulfilled and waste the space if the hashtag where there could have been a relevant post instead. Firstly we preprocess the dataset for removing the stop-words, punctuations, extra spaces etc., converted each text to lowercase, and then we have tokenized each full-text. After that, we have lemmatized the tokenized words for bringing it to root words. And then convert it to vector space using countVectorizer method.

For baseline model, we have trained multinomial naive bayes on preprocessed full-text lemmatized words. Currently we trained our model on dataset which contains 1072 manually annotated dataset. We achieved the accuracy of 58.4% for now. Which will be improved when we'll finally run this model on the complete dataset after running sentimental analysis along with ranking algorithm on it.

4.2 Proposed method

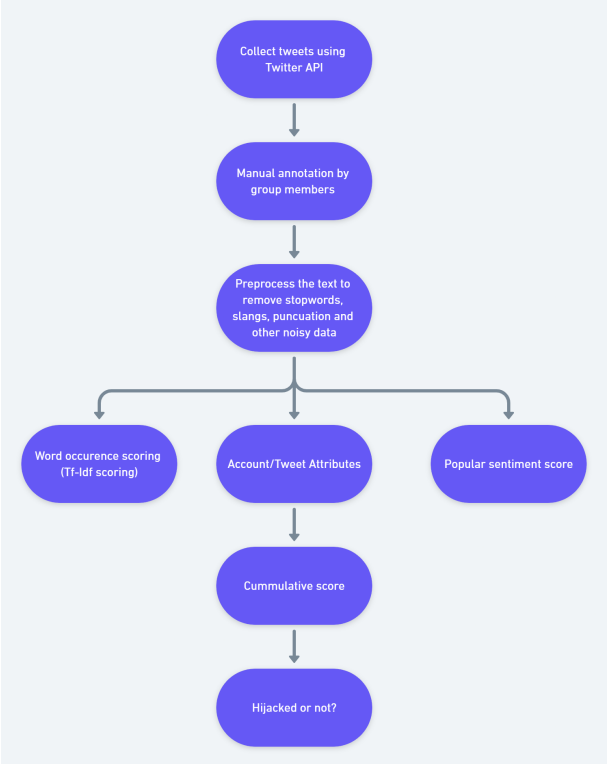

Figure 1: Procedure of our project

After the tweets have been preprocessed, we will divide the tweets into 4 categories, i.e., politics, entertainment, technology and others. A bag of words or vocabulary will be created for each of these categories using the training data. A different classifier is trained to classify tweets as hijacked or not-hijacked for each category using the training data(tweets) of that category. Whenever a new tweet will be given as input, it will be preprocessed and it will be classified into 1 of the 4 categories using the bag of words created for each category during training time. For an instance if the incoming tweet is classified as belonging to politics category, then the tweet will be given to the trained model for the politics category to classify whether it is hijacked or not hijacked. We'll combine this model with the result of sentiment analysis and ranking algorithm to finally obtain the required predictions.

4.3 Timeline

- Choosing the final team and finalizing the idea of the project: week 1-4.
- Application for Twitter's developer account to use Twitter API: week 4-6.
- Data collection using Twitter API on numerous hashtags : week 6-8.
- Data collection using Twitter API and manual annotation: week 8-9.
- Data pre-processing: week 9-10.
- Model preparation : week 10-11.
- Score Comparison : week 11-12.
- Final report submission : By week 13.

5. REFERENCES

[1] Virmani, D. et al. 2017. *HashMiner: Feature Characterisation and analysis of Hashtag Hijacking using real-time neural network*. Procedia Computer Science. 115, (2017), 786-793.

[2] Jain, N. et al. 2015. *HashJacker- Detection and Analysis of Hashtag Hijacking on Twitter*. International Journal of Computer Applications. 114, 19 (2015), 17-20.

[3] VanDam, C. and Tan, P. 2016. *Detecting hashtag hijacking from Twitter*. Proceedings of the 8th ACM Conference on Web Science. (2016).

[4] Chu, Z. et al. 2012. *Detecting Social Spam Campaigns on Twitter*. Applied Cryptography and Network Security. (2012), 455-472.

[5] Hadgu, A. et al. 2013. *Political hashtag hijacking in the U.S.* Proceedings of the 22nd International Conference on World Wide Web - WWW '13 Companion. (2013).

[6] Rodak, O. 2019. Hashtag hijacking and crowdsourcing transparency: social media affordances and the governance of farm animal protection. Agriculture and Human Values. 37, 2 (2019), 281-294.