Applied Artificial Intelligence
CS -514

Project 4 – Decision network to report spam tweets.

Using Netica

Sherlock

Abstract:

This system is a decision network to detect and report spam tweets. The probability of the occurrence of the observable variables flow through the system in the direction of causality and the counter causal direction to estimate the likelihood of a given tweet being a spam one.

The characteristics of a spam tweet take form of the nature nodes and the decision node of whether based on the observed characteristic, what action should be performed. The utility node measures the quality of the future tweets assuming the action being done.

Features and Variables:

There is a link between all the characteristics and the action node as the decision maker can observe all the variables from a given tweet that he or she is presented with. It is also crucial to note that various variables such as graphical features suggestion, which could be used to detect the likelihood of the tweet being a spam one cannot be connected to the action node as they are non-observable features and cannot contribute to the decision made.

The utility node here measures the quality of the tweets concerning to the user satisfaction which will be directly impacted by the action of the user and analyzing the characteristics of the tweet.

The action node can suggest on one of the three decisions based on the observed variables

1. Report the User – On analyzing the characteristic of the tweet, a decision of reporting the user to twitter could be done which is an extreme decision to make but in turn if proved to be a spam account and hence eradicated will substantially increase the quality of the future tweets.
   However before choosing this as an action one must understand that just reporting the user alone is not going to eradicate the user's contents from appearing on your feed. Twitter will examine the case and will ask for more evidence about the same and only if the severity of the problem is higher, the spam account will be removed from twitter. Also, it is essential to understand that false positive decisions will involve in the wrong reporting of a legit account and hence the user has to carefully consider all the factors before bluntly reporting the issue to twitter. Thus, it is safe to say, that only when an account continuously posts spam tweets and causes a higher degree of threat, any user is likely to choose this option to tackle spam tweets.

2. Unfollow the account – This decision is less severe and a subjective decision. When an account that the user follows post irrelevant content or appears to be a fake account, the

user is of the liberty to simply unfollow that account instead of having to escalate the issue and report it. It is essential to understand that false positive decisions in this category will potentially involve in the user unfollowing legit account and thus is prone to missing out on information or a valid connection in the user's network which decrease the quality of good tweets. Thus, the user satisfaction becomes lower as the quality of tweet decrease due to missing out of useful content.

3. Continue following the account – This decision of continuing to follow the user considering the characteristics of the tweet and the account could be a wise decision in certain cases. However, the false positive decision is likely to increase spam content in the feed which would decrease the user satisfaction as the quality of tweet is low and the risk of the user falling trap to the spams is high.

The other nodes in the network are,

1. Account Age: Spam tweet generating accounts generally tend to be recently opened accounts and the probability of it being a mature account is very less. Generally, they are old in the order of days or hours.

2. Hashtag Stuffing: To attract more attention, a spam tweet is filled with trending hashtags even though they might not be relevant to the content.

3. Untrusted links: It is a classic feature of spam tweets to hold a questionable link. Reporting accounts that tweet untrusted links instead of just unfollowing would be a preferred idea.

4. Unrelated content: Sometimes, the content provided by the account might not match with the interest of the user or the user might not even be aware of content.
   Say for example – A non-technical person encounters a very specific tweet that concerns a programming language from an account that he has been following for a long time and the tweet is filled with links and tags.

5. Dummy Account: Spam tweets are most likely to originate from fake accounts. Some dummy accounts are harmless while other could be spam generators. New accounts could appear to fake at first and hence cannot be dismissed immediately without investigation.

The other nodes are

6. Account Feature Suggestion – This takes into consideration variables from the observable environment. Most of these features are user controlled i.e. the user will be able to tweak these variables and is under their control.

The intermediate variable Account Feature Suggestion – analyses various characteristics of the account that posts a tweet and bears the likelihood of the account being a source of a spam tweet.

The factors observed are,

7. Following to follower ratio – It is assumed that for a spam generating account (cases of bots and fake accounts) the number of followers is lesser compared to the number of following people. This translates to the fact that the following to follower ratio is higher for spam account with this assumption.

8. No of Tweets – Generally, a normal user controlled account which is non-spamming in nature has only certain number of tweets (around 22 tweets / day / account) https://blog.hubspot.com/blog/tabid/6307/bid/4594/is-22-tweets-per-day-the-optimum.aspx /) and the basic assumption is that, a spam account would post significantly more number of tweets as compared to a normal account.

9. Tweets Messaged – Not just posting the tweet but messaging the tweet is also an important way the spam propagates. The number of tweets messaged from a spam account is likely to be more than the regular exchange of tweets that a normal user would send since the spam targeted as personal messages have a higher chance of being noticed. Adding to this, the number of replies a spam tweet gets is greatly lesser than the response achieved by a normal tweet.

10. Time between posts – Spam accounts focus on spreading more in shorter time. Hence the idle time is lesser when compared to an authentic account.

There are so many other features like the age of the twitter account, the profile picture, the name which could help us analyze the trustworthiness of an account.
Using the above explained factors we derive at a value which carries the likelihood of the account being a spam account.

11. Tweet Feature Suggestion: This intermediate variable computes a factor which suggests the likelihood of a tweet being a spam.

12. URL Features:

URLS in a tweet play a major factor in propagating spam and act as click bait in tricking the user. Thus, we introduce another intermediate variable to understand the intricacies with respect to the URLs in the tweet.

13. Length of URL - Research suggests that URLS longer with lots of random gibberish content tend to be click baits. (https://towardsdatascience.com/phishing-domain-detection-with-ml5be9c99293e5)

14. Soundness of domain name – Clickbait tend to have funny or uncommon domain names compared to authentic tweets.

15. Presence of a brand name in the URL – Tweets are widely recognized to be associated with a brand name such as a news provider or with an influential person. So, links that might have association with a popular entity has lesser chances of being a spam.

16. Mentions or tags – In order to gain more attention, the spam tweets are loaded with tags and hashtags compared to a normal tweet which focuses more on the content and only partially on tags.

17. Proportion of numbers – Interestingly, the number of digits involved in a spam tweet tends to higher than a regular tweet. Cash prize amount, phone numbers are commonly found on spam tweets.

18. <u>Graph Features Suggestion</u> –

This intermediate node is a manifestation of the two unobservable variables in the environment. All other variables discussed so far could be manipulated by the spam account to prevent from being identified but these are not under the user's control hence, unobserved. Under this category we consider two variables,

19. Number of legitimate users between – This considers the number of users that the user has in common between the account that produces the tweet which we are analyzing. A larger number implies that the account is authentic, and a smaller number could be because the user is not having a good follower base which is a prime symptom of a spam account.

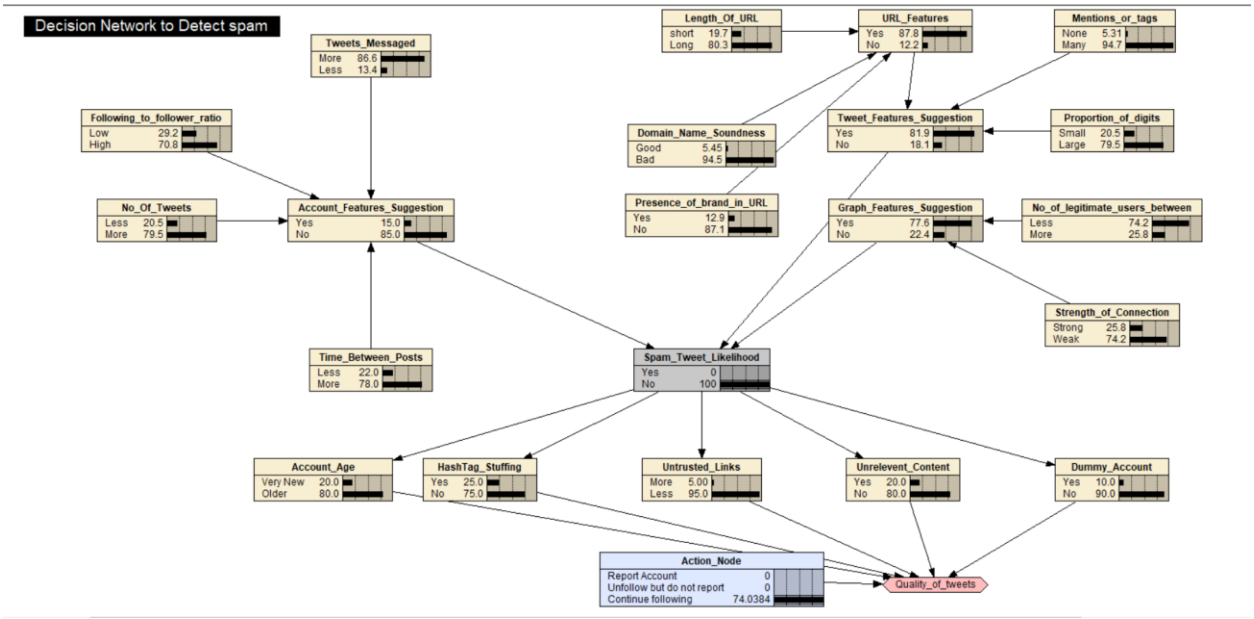
20. Strength of connection – the degree of connections – the extent in which first, second and third-degree connections are connected in the network is another sign of denoting a genuine account.
21. Spam tweet likelihood – This is the ultimate intermediate node that guesses the likelihood of the tweet being a spam, considering its origin account and the characteristics.

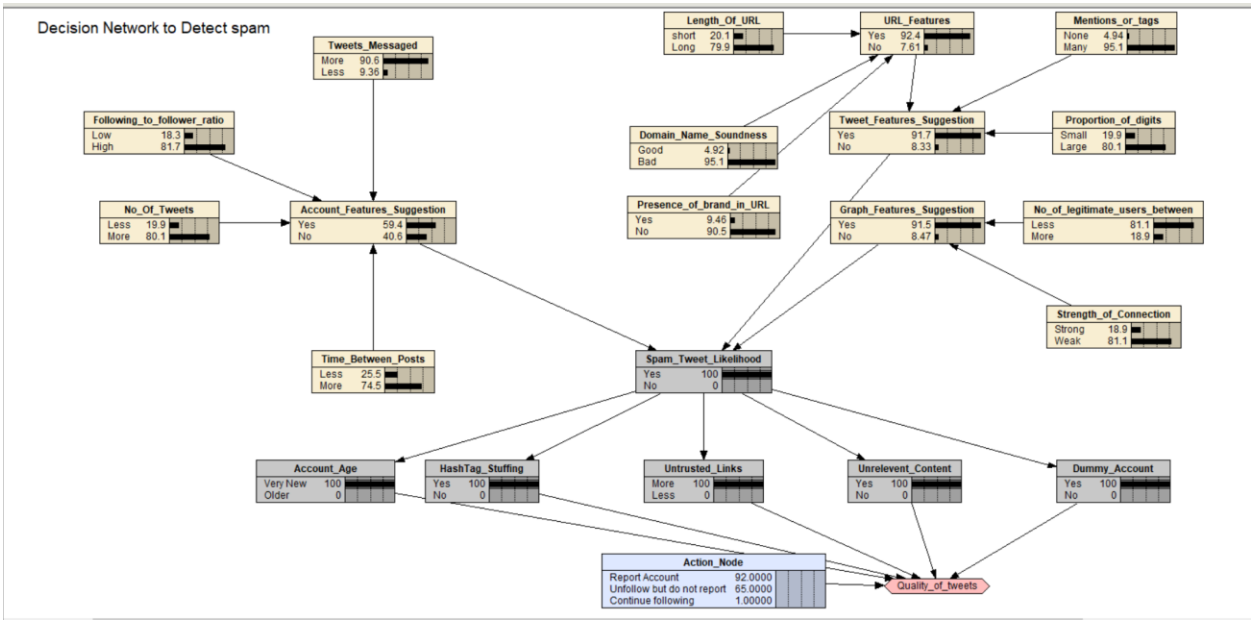
Changes in these variables propagates in the counter causal direction.

User Manual:

- Copy the file Sherlock.neta in any location.
- Open the file in Netica.
- Compile the network
- Edit input values by clicking on all the nodes that are connected to the action node and observe the decisions.
- If for some reason, the values in the action node don't change relevantly and appear to be stuck or display weird results, please compile again and try out with the same probabilities for right decisions.

Samples:


Decision Network to Detect spam
Tweets_Messaged
More 86.6
Less 13.4
Following_to_follower_ratio
Low 29.2
High 70.8
No_Of_Tweets
Less 20.5
More 79.5
Account_Features_Suggestion
Yes 15.0
No 85.0
Time_Between_Posts
Less 22.0
More 78.0
Length_Of_URL
short 19.7
Long 80.3
URL_Features
Yes 87.8
No 12.2
Mentions_or_tags
None 5.31
Many 94.7
Domain_Name_Soundness
Good 5.45
Bad 94.5
Tweet_Features_Suggestion
Yes 81.9
No 18.1
Proportion_of_digits
Small 20.5
Large 79.5
Presence_of_brand_in_URL
Yes 12.9
No 87.1
Graph_Features_Suggestion
Yes 77.6
No 22.4
No_of_legitimate_users_between
Less 74.2
More 25.8
Strength_of_Connection
Strong 25.8
Weak 74.2
Spam_Tweet_Likelihood
Yes 0
No 100
Account_Age
Very New 20.0
Older 80.0
HashTag_Stuffing
Yes 25.0
No 75.0
Untrusted_Links
More 5.00
Less 95.0
Unrelevent_Content
Yes 20.0
No 80.0
Dummy_Account
Yes 10.0
No 90.0
Action_Node
Report Account 0
Unfollow but do not report 0
Continue following 74.0384
Quality_of_tweets


Decision Network to Detect spam
Tweets_Messaged
More 90.6
Less 9.36
Following_to_follower_ratio
Low 18.3
High 81.7
No_Of_Tweets
Less 19.9
More 80.1
Account_Features_Suggestion
Yes 59.4
No 40.6
Time_Between_Posts
Less 25.5
More 74.5
Length_Of_URL
short 20.1
Long 79.9
URL_Features
Yes 92.4
No 7.61
Mentions_or_tags
None 4.94
Many 95.1
Domain_Name_Soundness
Good 4.92
Bad 95.1
Tweet_Features_Suggestion
Yes 91.7
No 8.33
Proportion_of_digits
Small 19.9
Large 80.1
Presence_of_brand_in_URL
Yes 9.46
No 90.5
Graph_Features_Suggestion
Yes 91.5
No 8.47
No_of_legitimate_users_between
Less 81.1
More 18.9
Strength_of_Connection
Strong 18.9
Weak 81.1
Spam_Tweet_Likelihood
Yes 100
No 0
Account_Age
Very New 100
Older 0
HashTag_Stuffing
Yes 100
No 0
Untrusted_Links
More 100
Less 0
Unrelevent_Content
Yes 100
No 0
Dummy_Account
Yes 100
No 0
Action_Node
Report Account 92.0000
Unfollow but do not report 65.0000
Continue following 1.00000
Quality_of_tweets

References:

http://thesai.org/Downloads/Volume8No3/Paper_5A_Survey_of_Spam_Detection_Methods_on_Twitter.pdf

https://ieeexplore.ieee.org/document/7582694

https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7937783