# Applied Artificial Intelligence
# CS-514

## Project 3 – Spam Tweet Detector

Using Netica

Sherlock

# Abstract:

Spam tweet detector is a Bayesian Network that employs Bayes' theorem to determine whether a given tweet is a spam one using the observed features. 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 observed features are grouped into 3 parts to ensure the sparse structure and hence space variables including Account Features Suggestion, Tweet Feature Suggestion and Graph Feature Suggestion are used. Later these variables are used to decide on the likelihood of the tweet being a spam. The characteristics of a spam tweet are then manifested.

# Features and Variables:

The system observes 3 sets of variables and uses 5 intermediate variables to yield a decision on the likelihood of it being a spam.

1. 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,

2. 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.

3. 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.

4. Total Likes and Retweets – The idea here is that the total number of likes or retweets is lesser for a spam account compared to a typical user's account.

5. 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.

6. 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.

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

8. 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.

9. 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-ml-5be9c99293e5)

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

11. 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.

12. 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.

13. 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.

14. Graph Features Suggestion –

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,

15. 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.

16. 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.

17. Spam tweet likelihood – This is the ultimate intermediate node that guesses the likelihood of the tweet being a spam, considering it's origin account and the characteristics.

Adding to these, the characteristics of a spam tweet is then added in the causal direction.

18. 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.

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

20. Untrusted links: It is a classic feature of spam tweets to hold a questionable link.

21.Uncharacteristic content: Sometimes when a regular user's account is hacked to spread false information, a distinct difference in the pattern of the content is observed.

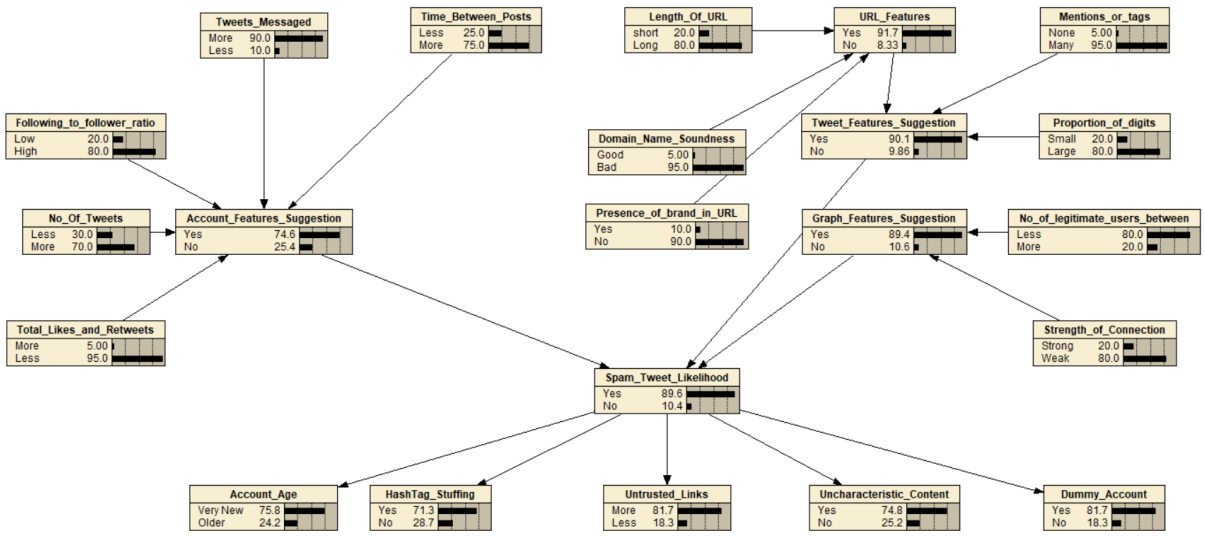
22.Dummy Account: Spam tweets are most likely to originate from fake accounts.

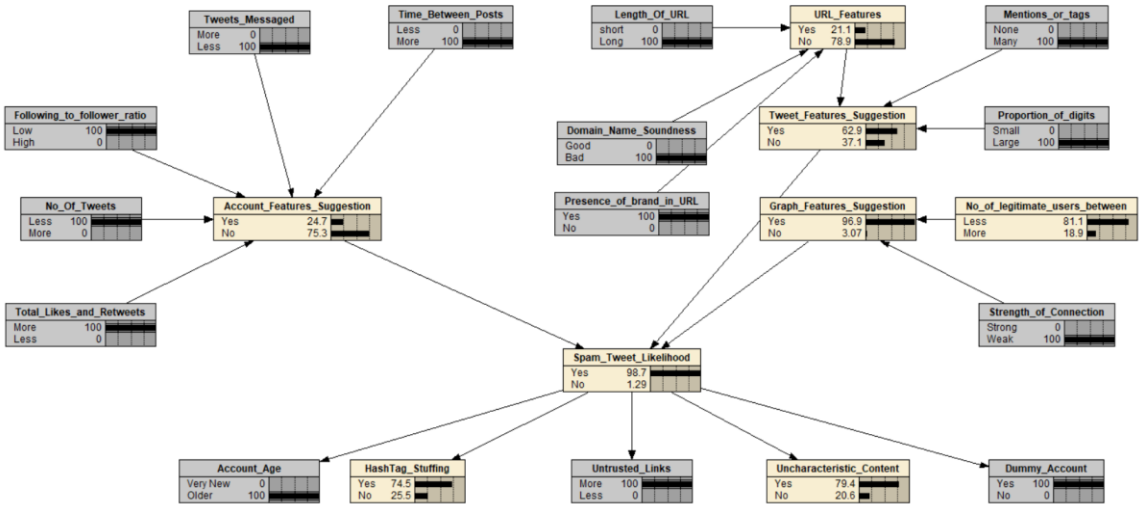
Changes in these variables propagates in the counter causal direction.

# User Manual:

- Copy the file Sherlock.neta in any location and open the file in Netica. Compile the network and input values by clicking on all the nodes without parents and see the changes flowing through the Bayesian Network

# Samples:






# 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