**SPAM DETECTION**



MINOR PROJECT REPORT

MADE BY- PALAK AGARWAL(14503022)

MAHAK SHARMA(14503014)

BATCH-B13

**ABSTRACT**

As the popularity of content sharing websites such as YouTube and Flickr has increased, they have become targets for spam, phishing and the distribution of malware. On YouTube, the facility for users to post comments can be used by spam campaigns to direct unsuspecting users to bogus e-commerce websites. In this project, we demonstrate how spam campaigns uses the vast data of set of comments on youtube videos to predict whether the particular comment is a spam or ham comment. The usage and popularity of content sharing websites continues to rise each year. YouTube now receives more than three billion views per day, with forty-eight hours of video being uploaded every minute; increases of 50% and 100% respectively over the previous year. Unfortunately, such increases have also resulted in these sites becoming more lucrative targets for spammers hoping to attract unsuspecting users to malicious websites, where a variety of threats such as scams (phishing,e-commerce) and malware can be found. Spam detection method will prevent user to visit malicious website and pages.Youtube spam, which is referred to as unsolicited comments containing malicious link that directs victims to external sites containing malware downloads, phishing, drug sales, or scams, set of irrelevant comments which is completely different from the video content etc, not only interferes user experiences, but also damages the whole Internet.

We have used vast data set of comments in which after filteration and cleaning procedures , algorithms such as TFIDF ,LDA,Cosine similarity etc are applied to test data for spam content.The spam flag on each set of comment has helped to implement machine learning algorithms such as K neighbour classifier, multi naïve baye’s classifier, decision tree classifier ,rain forest bound classifier etc on the test data and predicted their accuracy score and f-score of being a spam or ham comment.

We have designed a algorithm in which term related comments are clustered in one set and the count of each comment is increased for every upcoming set of terms. Furthur cluster formation is done on the basis of related comments and ouliers are found and rejected for similar kind of data. The comparisons between different algorithms used for spam comment detection is done and their accuracy values are checked so that we can use the best one to predict the spam.

**DIVISION OF PROJECT**

1. PALAK AGARWAL 50%

2. MAHAK SHARMA 50%

**PROBLEM STATEMENT**

Youtube is a social media platform with massive number of users. Three hundred videos are uploaded to YouTube every single minute, available to more than 1 billion YouTube users in 75 countries and in 61 languages. Talk about a global reach. It allows sharing of videos where user can upload, share or view content of different videos available on the site. Today, YouTube is the largest user-driven video content provider in the world; it has become a major platform for disseminating multimedia information and thus it is more attacked by spam or fake content from all over the world. Social media websites allow users freely distribute and share information to friends. Information can spread very fast and easily within the social media networks. Because of this, such websites expose to various types of unwanted and malicious spammer or hacker actions. There is a crucial need in the society and industry for a security solution in social media. Social media websites need to be clean for long term success. A company/brand page on social media also needs to be clean to reduce the risk of damaging its reputation. Virus links from the spams could lead to personal or business loss and damage.

**RESEARCH PAPER SUMMARY**

* TITLE - Network Analysis of Recurring YouTube Spam Campaigns.

PUBLISHED BY- O'Callaghan, Derek.

* TITLE- TopicSpam: a Topic-Model based approach for spam detection.

PUBLISHED BY-Li, Jiwei, Claire Cardie, and Sujian Li.

* TITLE- Detecting "Smart" Spammers On Social Network: A Topic Model Approach.

PUBLISHED BY- Liu, Linqing.

* TITLE- Finding Valuable Yelp Comments by Personality, Content, Geo, and Anomaly Analysis.

PUBLISHED BY- Koven, Jay, Hossein Siadati, and Ching-Yung Lin.

* TITLE- Spam Detection using Clustering, Random Forests, and Active Learning.

PUBLISHED BY- DeBarr, Dave, and Harry Wechsler.

* TITLE- Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm.

PUBLISHED BY - Harisinghaney, Anirudh.

* TITLE- Identifying video spammers in online social networks.

PUBLISHED BY- Benevenuto, Fabricio.

**RESEARCH PAPER INTEGRATION**

1. **Network Analysis of Recurring YouTube Spam Campaigns:**

Comments filtering techniques such as removing stop words, punctuations, stemming, removing non ascii(latin words) characters ,lowercase letter conversion and removing all the null content comment from the dataset is done .All of the data filteration and converting the data into simpler form for spam detection.

1. **TopicSpam: a Topic-Model-Based Approach for Spam Detection:**

LDA-based topic modeling approach for fake review detection. Our model can aptly detect the subtle differences between deceptive reviews and truthful ones and achieves accuracy on review spam datasets, outperforming existing baselines by a large margin .We have set a threshold values for terms in comments to predict and classify spam and ham comments.

1. **Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm.:**

We have used the techniques of machine learning of KNN, multinominal naïve bayes and further more techniques to check the accuracy score between training and test data set of our comments containing spam flag values.

**ALGORITHS USED AND IMPLEMENTED:**

**1.TF-IDF:** Tf-idf stands forterm frequency-inverse document frequency, and the tf-idf weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Variations of the tf-idf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query.

One of the simplest ranking functions is computed by summing the tf-idf for each query term; many more sophisticated ranking functions are variants of this simple model.

Tf-idf can be successfully used for stop-words filtering in various subject fields including text summarization and classification.

**2.LDA:** Latent Dirichlet allocation (LDA) is a [topic model](http://en.wikipedia.org/wiki/Topic_model) that generates topics based on word frequency from a set of documents. LDA is particularly useful for finding reasonably accurate mixtures of topics within a given document set. This walkthrough goes through the process of generating an LDA model with a highly simplified document set. This is not an exhaustive explanation of LDA. The goal of this walkthrough is to guide users through key steps in preparing their data and providing example output.

It involves importing the documents, cleaning of document is done by tokenization ,stemming,removal of stop words .In the next step a document term matrix is generated and than the LDA model is constructed.Set of hot trending topics are taken out according to the content of our crime patrol data which are useful for detecting spam comments i.e which are completely different and have no relation from the given topic name.

**3.cosine similarity:** The cosine similarity between two vectors (or two documents on the Vector Space) is a measure that calculates the cosine of the angle between them. This metric is a measurement of orientation and not magnitude, it can be seen as a comparison between documents on a normalized space because we’re not taking into the consideration only the magnitude of each word count (tf-idf) of each document, but the angle between the documents. It calculates the cosine similarity values between set of comments in our Dataset.

**4.KNN**- KNN is an non parametric lazy learning algorithm. That is a pretty concise statement. When you say a technique is non parametric , it means that it does not make any assumptions on the underlying data distribution. This is pretty useful , as in the real world , most of the practical data does not obey the typical theoretical assumptions made (eg gaussian mixtures, linearly separable etc) . Non parametric algorithms like KNN come to the rescue here.

It is also a lazy algorithm. What this means is that it does not use the training data points to do any generalization. In other words, there is no explicit training phase or it is very minimal. This means the training phase is pretty fast . Lack of generalization means that KNN keeps all the training data. More exactly, all the training data is needed during the testing phase. (Well this is an exaggeration, but not far from truth). This is in contrast to other techniques like SVM where you can discard all non support vectors without any problem.  Most of the lazy algorithms – especially KNN – makes decision based on the entire training data set (in the best case a subset of them).

**5.Multinomial Naive Bayes**:It is a specialized version of **Naive Bayes** that is designed more for text documents. Whereas simple **naive Bayes** would model a document as the presence and absence of particular words, **multinomial naive bayes** explicitly models the word counts and adjusts the underlying calculations to deal with in.We have used this classifier technique to calculate how accurate is our test data according to given set of training data.

**6.Random Forest Classifier:** [Random](https://gerardnico.com/wiki/number/random) [forest](https://gerardnico.com/wiki/tree/forest) (or random forests) is a trademark term for an [ensemble](https://gerardnico.com/wiki/data_mining/ensemble) [classifier](https://gerardnico.com/wiki/data_mining/classification) that consists of many [decision trees](https://gerardnico.com/wiki/data_mining/decision_tree) and outputs the [class](https://gerardnico.com/wiki/data_mining/class) that is the [mode](https://gerardnico.com/wiki/number/function/mode) of the classes output by individual trees.

Random forests are collections of trees, all slightly different.It randomize the algorithm, not the training data. It generally improves decision trees decisions.

Unlike single [decision trees](https://gerardnico.com/wiki/data_mining/decision_tree) which are likely to suffer from high [variance](https://gerardnico.com/wiki/data_mining/variance) or high [Bias](https://gerardnico.com/wiki/data_mining/bias) Random Forests use averaging to find a natural balance between the two extremes.

A random forest is a meta estimator that fits a number of classifical [decision trees](https://gerardnico.com/wiki/data_mining/decision_tree) on various [sub-samples](https://gerardnico.com/wiki/data_mining/bootstrap) of the dataset and use averaging to improve the predictive accuracy and control [over-fitting](https://gerardnico.com/wiki/data_mining/overfitting).

Each [decision tree](https://gerardnico.com/wiki/data_mining/decision_tree) is constructed by using a [Random](https://gerardnico.com/wiki/number/random) subset of the [training](https://gerardnico.com/wiki/data_mining/training) data.

**RESULT**

**CONCLUSION AND FUTURE WORK:**

YouTube spam campaigns typically involve a number of spam bot user accounts controlled by a single spammer targeting popular videos with similar comments over time. We have shown that dynamic network analysis methods and algorithms are effective for identifying the recurring nature of different spam campaign strategies, along with the associated user accounts. We can use a characterization of YouTube users in terms of motifs in the comment network to highlight the users in question. While the YouTube comment scenario could be characterized as a network in a number of ways, we use a network representation comprising user and video nodes, user-video edges representing comments and user user edges representing comment similarity.

For future experiments, it will be necessary to annotate the data set with spam/non-spam labels, or perhaps a more extensive annotation that considers the associated campaign strategies. Feature selection of a subset of motifs could then be performed along with subsequent user classification. The use of a subset of motifs is attractive, as it would remove the current requirement to count all motif instances found in the user egocentric networks, which can be a lengthy process.

The LDA data generated for useful topics can be further used for clustering hot topics for a given type of data set and removing outliers. Sentiment and geological analysis for the dataset can also be done and some other machine learning classification techniques can be used to work on the SPAM FLAG prediction which provides hint for being a spam data.

**REFERENCES:**

* O'Callaghan, Derek, et al. "Network analysis of recurring youtube spam campaigns." arXiv preprint arXiv:1201.3783 (2012).
* Koven, Jay, Hossein Siadati, and Ching-Yung Lin. "Finding Valuable Yelp Comments by Personality, Content, Geo, and Anomaly Analysis." Data Mining Workshop (ICDMW), 2014 IEEE International Conference on. IEEE, 2014.
* Liu, Linqing, et al. "Detecting" Smart" Spammers On Social Network: A Topic Model Approach." *Ar Xiv preprint arXiv:1604.08504* (2016).
* Koven, Jay, Hossein Siadati, and Ching-Yung Lin. "Finding Valuable Yelp Comments by Personality, Content, Geo, and Anomaly Analysis." Data Mining Workshop (ICDMW), 2014 IEEE International Conference on. IEEE, 2014.
* - DeBarr, Dave, and Harry Wechsler. "Spam detection using clustering, random forests, and active learning." Sixth Conference on Email and Anti-Spam. Mountain View, California. 2009
* - Harisinghaney, Anirudh, et al. "Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm." Optimization, Reliabilty, and Information Technology (ICROIT), 2014 International Conference on. IEEE, 2014.
* Benevenuto, Fabricio, et al. "Identifying video spammers in online social networks." Proceedings of the 4th international workshop on Adversarial information retrieval on the web. ACM, 2008.