Spam Detection in Social Media Networks: A Data Mining Approach

Harshal S. Multani
BE Computer Engg.
KJEI's TAE, Pune

Amrita Sinh Marod
BE Computer Engg.
KJEI's TAE, Pune

Vinita Pillai
BE Computer Engg.
KJEI's TAE, Pune

Vishal Gaware
BE Computer Engg.
KJEI's TAE, Pune

ABSTRACT

The ubiquitous use of social media has generated unparalleled amounts of social data. Data may be – text, numbers or facts that are computable by a computer. A particular data is absolutely useless until and unless converted into some useful information. It is necessary to analyze this massive amount of data and extracting useful information from it. There are more active internet users on social networks than search engines. Social media networks provide an easily accessible platform for users who wish to share information with others. Information can be spread across social networks quickly and effectively, hence have now become susceptible to different types of undesired and malicious spammer/hacker actions. Therefore, there is a pivotal need for security in social media and industry. In this demo, a scalable and online social media spam detection system for social network security using TF-IDF algorithm is proposed.

General Terms

Porter Stemmer algorithm, TF-IDF algorithm.

Keywords

Spam, Social media networks, Security, TF-IDF.

1. INTRODUCTION

Generally, data mining is also known as data or knowledge discovery. It allows users to analyze data from various dimensions/angles, categorize it, and summarize the relationships that have been identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Social network has become very popular in the last decade. It is more affordable to access social network sites like Facebook, LinkedIn, Twitter, etc. through the internet. People rely on social network for news, information and opinions of other users on various subjects. The massive data generated is characterized by three computational issues viz.; size, noise and dynamism [7], which often make social network data extremely complex for manual analysis. This results in the relevant use of computational means of their analysis. Data mining provides an expansive set of techniques to detect useful knowledge from enormous datasets like rules, trends and patterns. Data mining techniques are used for machine learning, statistical modeling and information retrieval. These techniques include activities like data pre-processing and data interpretation during data analysis.

Spams can be in the form of images, text, videos etc. So, social media websites need to be untainted for long-term success. If there is a page representing a company or a brand on some social media, it has to be taken care that it is clean; else it might damage the reputation. Spams may contain virus links which could lead to personal or business loss [1]. There have been researches on detecting spam emails [8, 9], spam messages [4], spam images [10], spam video [11], web spam [12], spammers [13] [14] [15], etc.

The major advantages of the proposed system include:

1) A scalable database containing spam-related "stop words", which can be dynamically updated.

2) TF-IDF algorithm produces a score rank for the terms in the spam database.

3) The NLP Parser will semantically check for every occurrence of a spam word from the database.

4) A large amount of textual data can be handled by TF-IDF algorithm.

5) Java string tokenizer that will separate out the normal text words such as "is", "an", "the" from the spam words.

2. LITERATURE SURVEY

2.1 Data Mining Methodologies Used

NLP Parser

TF-IDF Algorithm

Porter Stemmer Algorithm

2.1.1 NLP Parser

NLP stands for Natural Language Processing. It is an arena in computer science, artificial intelligence (AI) and linguistics. It pertains to interactions between computers and human i.e. natural languages. It has its own vocabulary dataset against which it checks for the semantics of words present in a given document.

A natural language parser is a program that figures out the **grammatical structure of sentences**- for instance, which groups of words go together (as "phrases") and which words are the **subject** or **object** of a verb. The knowledge of language obtained from hand-parsed sentences is used by the probabilistic parsers which try to yield the *most likely* analysis of new sentences. The development of NLP parsers was one of the biggest advancements in natural language processing in the 1990s.

2.1.2 TF-IDF

TF-IDF stands for *term frequency-inverse document frequency*. TF-IDF weight is a weight used in text mining and information retrieval. It is a statistical measure which is used to ascertain how important a word is to a document in a collection or a corpus. The importance of a word is directly proportional to the number of times it appears in the document. There are certain variations of TF-IDF weighting scheme which are often used by search engines as a central tool to score and rank a document's relevance, when given a user query. Summing the TF-IDF for each query term in a query is one of the simplest ranking functions. TF-IDF can be successfully used for filtering of stop-words.

2.1.3 Porter Stemmer Algorithm [18][19]

Natural language texts typically contain many different variants of a basic word.

Morphological variants (e.g., COMPUTATIONAL, COMPUTER, COMPUTERS, COMPUTING etc.) are generally the most common, with other sources including valid alternative spellings, mis-spellings, abbreviation, etc.

In a typical IR environment, one has a collection of documents, each described by the words in thedocument title and possibly by words in the document abstract. Ignoring the issue of precisely where the words originate, it can be said that a document is represented by a vector of words, or \terms\. Terms with a common stem will usually have similar meanings.

For example:

CONNECT

CONNECTED

CONNECTING

CONNECTION

CONNECTIONS

Frequently, the performance of an IR system will be improved if term groups such as this are conflated into a single term. This may be done by removal of the various suffixes -ED, -ING, -ION, IONS to leave the single term CONNECT. The suffix stripping process will reduce the total number of terms in the IR system, and hence reduces the size and complexity of data in the system, which is always advantageous. [18].

3. PROPOSED WORK

3.1 GAD Clustering Algorithm

Clustering is a data mining technique widely used in numerous applications. It has also been studied in research areas such as biology, pattern recognition, statistics, machine learning, information retrieval market research and multimedia processing [2]. Many papers have been published for fast clustering on large data. Some develop fast core clustering algorithms; some develop pre-processing methods, such as sampling, sub- space and compression in order to reduce the data to smaller size to achieve speedup.

Within the GAD framework a set of algorithms have been designed for different scenarios:

(1) Exact GAD algorithm E-GAD, which is much faster than K-Means and gets the same clustering result.

(2) Approximate GAD algorithms with different assumptions, which are faster than E-GAD while achieving different degrees of approximation.

(3) GAD based algorithms to handle the “large clusters” problem which appears in many large scale clustering applications.

3.2 J48 Decision Tree Algorithm [3]

The data mining model using the decision tree J48 is created using WEKA. WEKA stands for Waikato Environment for Knowledge Analysis and has been developed by the University of Waikato, New Zealand. WEKA is a tool which comprises 60 machine learning algorithms. They developed a prototype system which detected spams on the Facebook. The application runs on the server to where a Facebook request is rerouted for carrying out the process of spam checking. The decision tree model J48 was used for classification. The selected attributes were the number of keywords, the number of links, the length of the post, and the average number of words in a post [3]. The relationships of the attributes were further to be scrutinized in the future. The goal of this work was to only demonstrate the use of data mining model in detecting spams in the Facebook application. The model trained 150 sample posts and tested 75 posts. The features used included the number of words, the length of the post, and the number of the links with the recall rate of 66%.

3.3 K-Means Algorithm [5]

K-Means algorithm is a simple iterative method for partitioning a given dataset into a user-specified number of clusters, say, k. It was discovered by many researchers belonging to different disciplines, especially, Lloyd (1957, 1982), Forgey (1965), Friedman and Rubin (1967). In database management, clustering of data is done. It is the process of dividing the data elements (input data) into groups so that the items in the same group are as similar as possible and items in different groups are as dissimilar as possible [5].

This method utilizes the K-means clustering algorithm to group the messages or emails based on the similarity of their attributes or features into K disjoint groups to improve the accuracy of spam detection. In classification, the objects are assigned to predefined classes; whereas in clustering, the classes are formed, in which, a data point can belong to only one cluster.

4. PROJECT FLOW

Messages, comments and reviews are extracted from the OSN database. Two parallel activities are carried out:

1) The extracted data is given to Keyword parser.

2) The extracted data is given to the NLP parser as well.

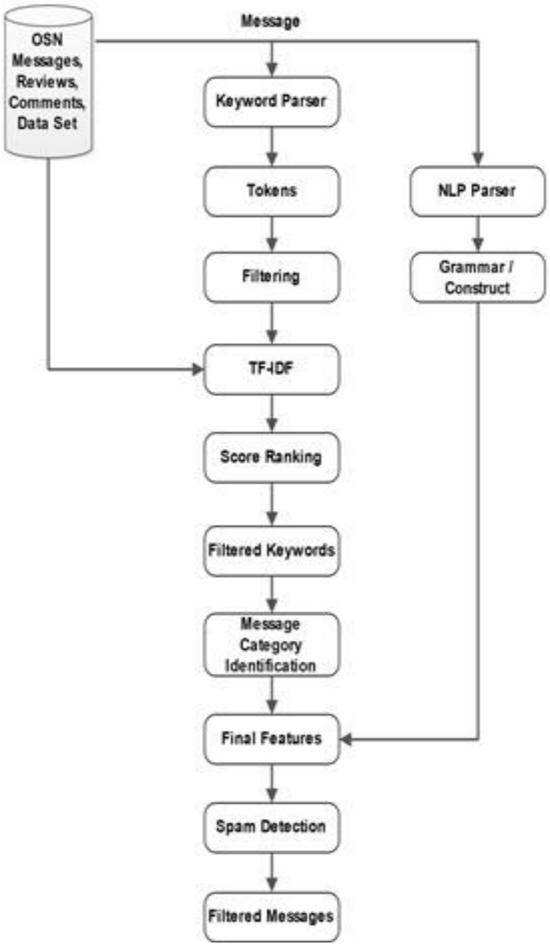

Fig 1: Project Flow Diagram for Spam detection in text feature.

4.1 Explanation

Messages, comments and reviews are extracted from the OSN database. Two parallel activities are carried out:

1) The extracted data is given to Keyword parser and

2) The extracted data is given to the NLP parser as well.

<u>**Activity 1**</u>: Java's in-built function **string tokenizer** will take data in the sentential form and break it into tokens (words, phrases, symbols and other meaningful elements).

-**Filtering** function will separate out the regular English words like “is”, “an”, “the” etc. from the given input tokens and produces output as “**filtered tokens**”.

-Filtered tokens act as “terms” for **TF-IDF algorithm**.

-TF-IDF algorithm produces a **score rank** for each and every query term.

-The score ranking will again apply filtering function which will eliminate all the words (terms) which never occurred or appear the least number of times.

<u>**Activity 2**</u>

-NLP Parser will check for the semantics of the queried words.

-It has its own vocabulary dataset against which it will check for the semantics of words present in a given document.

-From the results of activities 1 & 2, **spam detection** is done on the given input.

-Finally, a dataset of **filtered messages** is obtained.

5. CONCLUSION

The technical details of the system will be presented, including spam features, algorithms and efficient implementation. The hypothesis behind the design will be analyzed, especially on the scalability and accuracy issues, in order to show how this system can handle a huge number of posts and monitor real-time social activities in social media to identify spams. This method currently detects spam only in text but can further accommodate other features like images, video and social network features as well. The complexity of this approach is low and it can be used in reality easily.

6. REFERENCES

[1] Xin Jin, Cindy Xide Lin, Jiawei Han, JieboLuo - A Data Mining-based Spam Detection System for Social Media Networks

[2] GAD: General Activity Detection for Fast Clustering on Large Data ∗Xin Jin, Sangkyum Kim, JiaweiHan ,Liangliang Cao , Zhijun Yin ,University of Illinois at Urbana-Champaign

[3] Using a Data Mining Approach: Spam Detection on Facebook- M. Soiraya, S. Thanalerdmongkol, C. Chantrapornchai , Department of Computing, Faculty of Science , Silpakorn University, Thailand, 73000

[4] C. Shekar, S. Wakade, K. J. Liszka, and C.-C.Chan.Mining pharmaceutical spam from twitter. In ISDA, pages 813–817, 2010.

[5] K-Means Clustering Scheme for Enhanced Spam Detection-Nadir Omer FadlElssied and Othman Ibrahim Faculty of Computing, University Technology Malaysia, 81310, Skudai, Johor Bahru, Malaysia AlgerafSharq Technical College, Khartoum, Sudan

[6] Xiao-Li, C., L. Pei-Yu, Z. Zhen-Fang and Q. Ye, 2009. A method of spam filtering based on weighted support vector machines.Proceeding of the IEEE International Symposium on IT in Medicine and Education, pp: 947-950.

[7] A Survey of Data Mining Techniques for Social Media Analysis - Mariam Adedoyin-Olowe, Mohamed MedhatGaber, Frederic Stahl

[8] J. S. Kong, B. A. Rezaei, N. Sarshar, V. P. Roychowdhury, and P. O. Boykin. Collaborative spam filtering using e-mail networks. IEEE Computer, 39(8):67–73, 2006.

[9] K. Yoshida, F. Adachi, T. Washio, H. Motoda, T. Homma, A. Nakashima, H. Fujikawa, and K. Yamazaki.Density-based spam detector. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 486–493, 2004.

[10] B. Byun, C.-H.Lee, S. Webb, and C. Pu.A discriminative classifier learning approach to image modeling and spam image identification.In CEAS, 2007.

[11] F. Benevenuto, T. Rodrigues, V. Almeida, J. M. Almeida, C. Zhang, and K. W. Ross.Identifying video spammers in online social networks. In AIRWeb, pages 45–52, 2008.

[12] S. Webb, J. Caverlee, and C. Pu. Introducing the webb spam corpus: Using email spam to identify web spam automatically. In Proceeding of the Third Conference on Email and Anti-Spam (CEAS), 2006

[13] B. Markines, C. Cattuto, and F. Menczer.Social spam detection. In AIRWeb, pages 41–48, 2009.

[14] K. Lee, J. Caverlee, and S. Webb. Uncovering social spammers: social honeypots + machine learning. In Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval, SIGIR ’10, pages 435–442, 2010.

[15] S. Webb, J. Caverlee, and C. Pu. Social honeypots: Making friends with a spammer near you. In Proceeding of the Fifth Conference on Email and `Anti-Spam (CEAS), 2008.

[16] Gerard Salton and Christopher Buckley, Department of Computer Science, Cornell University, Ithaca, NY 14853, USA .Term-Weighting Approaches in Automatic Text Retrieval.*Information Processing & Management* Vol. 24, No. 5, pp.513-523,1988.

[17] Juan Ramos, Department of Computer Science, Rutgers University, 23515 BPO Way, Piscataway, NJ, 08855: Using TF-IDF to Determine Word Relevance in Document Queries

[18] Willett, P. (2006) The Porter stemming algorithm: then and now. Program:electronic library and information systems, 40 (3). pp. 219-223.