RESEARCH ARTICLE OPEN ACCESS

Predictive Scrutiny and Acquaintance based Network Analysis by Sentiment of Opinions with Supervised Learning Algorithms in Social Media

[1]K.C. Anju,[2]N. Rojaramani
[1]Research Scholar, Department of Computer Science, AnnaiVailankanni Arts and Science College, Thanjavur-613007
[2]Assistant Professor, Department of Computer Science, AnnaiVailankanni Arts and Science College, Thanjavur-613007

Abstract:

In this paper meet some problem in current Social Networks is to assign users the influence to manage the messages posted on their private space to avert that unnecessary content. The unwanted data may contain political, In social media our account is hacked with unknown person and they post fake news, for avoiding this can use a method of text analysis with based on our interest and about us, careers and before posted posts, categorized the words in black list and white list using this method 90% white list included message is post, otherwise a Onetime Password (OTP) send to the email and check the person is right person., as opposed to database applications, which use very structured data. In the implementation section use a college website and make department wise posts are posted by comparing the words with which department is posted check with dictionary of that subject, e.g. It can check the post from physics department, tokenized the post and based on that tokens that all tokens compared with physics dictionary, 90% tokens from physics that will be posted, like this method all department posts are checked and make black list and white list and checked. Through the sentiment analysis the comments posted by the users are analyzed by two categories like black list and white list. Then it checks with IP section for the first time login verification to check whether authorized person login or hacker hacks any information.

Keywords — **Social Network, Private Space, Text Analysis, One time Password (OTP), tokenized Post, Sentiment Analysis, Signup Verification.**

I. INTRODUCTION

There is increasing recognition that the Web represents a valuable source of national security-relevant intelligence and that computational analysis offers a promising way of dealing with the problem of collecting and analyzing data at Web scale. As a consequence, tools and algorithms have been developed which support various security informatics objectives. To cite a specific example, we have shown that blog network dynamics can be exploited to provide reliable early warning for a class of extremist-related, real world protest events.

Making sense of online content at Web scale is both important and technically challenging. For instance, discussions on social media sites such as blogs and forums often reflect the sentiments and opinions of individuals and groups about security-relevant topics, and thus can represent valuable intelligence data. However, these

views are typically expressed as informal communications and are buried in vast volumes of irrelevant discourse, so that efficiently and accurately extracting them is usually quite difficult. While powerful analysis techniques have been derived for traditional forms of content, less has been done to develop strategies that are well-suited to the particular characteristics of content generated in social media. Consider the important task of deciding whether a given post expresses positive or negative opinion toward a topic of interest. The informal, multilingual nature of social media content poses [1, 2, 3] a challenge for language-based sentiment analysis. While statistical learning-based methods often provide good performance in unstructured settings like this, obtaining the required labeled instances of data, such as a lexicon of sentiment-laden words for a given domain or a collection of "exemplar" blog posts of known polarity, is labor-intensive and time-consuming for Web applications.

1.2 PROBLEM DESCRIPTION

Social media analysis has not yet been considered as a means to understand everyday experiences of security technologies. When considering why might need new methods to specifically capture naturalistic everyday experiences, it is important to note the two types of knowledge that researchers may aim to elicit from users through research: explicit knowledge, and tacit knowledge.

Explicit knowledge comprises that which is easy to transmit to another person e.g. the number of siblings a person may have, or the number of passwords that a person uses. Much of what we know about the world however is tacit knowledge; knowledge that is difficult to transfer to another person e.g. how to use complex equipment (where experience has guided learning), or reasons for why a system feels secure.

Accessing tacit knowledge typically requires researchers to apply a mix of methods in their interactions with people (e.g. observations, surveys, diaries etc.) to identify and capture instances of interesting behaviors, and may even include working together with participants to highlight that interesting behaviors exist in the first place, but also to understand why those behaviors come about. Relatively few methods [6] in usable security go beyond traditional interviews and surveys to elicit this tacit knowledge; these classic methods best serve as tools to capture explicit knowledge (except where dialog is explicitly supported), as these do allow respondents to reflect on their own behaviors, however, within an experimenter defined framework. The use of social media posts as a lens on everyday security experiences would immediately present a number of methodological benefits:

- People are free to use their own vocabulary to describe their feelings and practices.
- Posts are created in naturalistic contexts and in the course of everyday activities.
- Security would be positioned as social and collective, rather than a personal and secretive practice.
- There is a large volume of social media posts to study, and they are typically short in length.

II. REVIEW OF LITERATURE

Xiao and Tao say "Personalized privacy preservation" prove that *l*-diversity always guarantees stronger privacy preservation than *k*-anonymity. Though several important models and many efficient algorithms have been proposed to preserve privacy in relational data, most of the

existing studies can deal with relational data only. Those methods cannot be applied to social network data straightforwardly. Anonymizing social network data is much more challenging than anonymizing relational data.

Hay etal.,"Resisting structural identification in anonymized social networks" considered a simple graph model, in which vertices and edges are unlabeled. They addressed vertex identifier attacks, and proposed a vertex clustering approach. Three models of external information were considered as the possible background knowledge of an adversary. These models represent a range of structural information that may be available to an adversary, including complete and partial descriptions of vertex neighborhoods, and connections to hubs in the network. The authors [7, 8, 9] formalized the structural indistinguishability of a vertex with respect to an adversary with external information about the local neighborhood of the vertex. Specifically, background knowledge of adversaries is modeled using the types of queries.

David L. Chaum used the public key cryptography. There are two keys, one public key known to others and one private key known to the intended receiver. The encryption of data is done by adding some random bits and then encrypting with public key to provide more security to the data. The scheme has two assumptions:

a)No one can determine anything about the correspondence between a set of sealed items and the corresponding set of unsealed items, or create forgeries without the appropriate random string or private key.

b)Anyone may learn the origin, destination(s), and representation of all messages in the underlying telecommunication system and anyone may inject, remove or modify messages.

In the paper, Yanfei Fan, Yixin Jiang, Haojin Zhu andXuemi (Sherman) Shen [5] proposed an efficient privacy preserving scheme against traffic analysis for network coding. It uses a light weight homomorphic encryption method on Global Encoding Vectors (GEVs). The scheme proposed significant privacy features i.e. packet flow intractability and message content confidentiality which obstructs traffic analysis attacks such as flow tracing.

Rack off and Simon's approach introduced two effectual protocols to provide security against adversary, which are secure multiparty computation and non-interactive zero-knowledge proof. The paper mentioned three adversaries: external adversary, passive internal adversary and active internal adversary. It also deals with the solution of the above adversaries against traffic analysis. The approach is to defend against traffic analysis. It passes the message through various mixes in such a way that it should completely differentiate from the resulting permutation. The quickness of the process depends upon the rapidity with which certain Markov processes converge on their stable distribution. Unfortunately it is practically limited, synchronous in nature, at most two messages are sent through mix-node process at a time and routes are inhibited. It is not easily implement to real world.

Presented by William Conner, TarekAbdelzaher and KlaraNahrstedt, the paper brings up an approach to prevent against traffic analysis in target tracking sensor network [11] by using the combination of indirection and data aggregation. It makes use of a decoy sink protocol. In this, a decoy sink node will get messages from all the sensors, aggregates them and a summary is to be made, and then send it to the real sink node. This approach leads to more traffic near decoy sink node and less near the real sink node due to

aggregation. This process protects the location of real sink node from adversaries to perform traffic analysis.

III. PROPOSED SYSTEM

The rising tide of attacks on social networks, according to, tells us that "social networks and their millions of users have to do a lot more to protect themselves from organized cybercrime, or risk failing to identity theft schemes, scams, and malware attacks. Understanding these risks and challenges should be addressed to avoid potential loss of private and personal information." Also, as says, "The area of internet information security is well developed and evolves continuously in response to new threats" and so it must evolve with social media too." The amount of personal information posted should be limited, and not post home addresses or private contact information. This, and information about your likes and daily routine can all be pieced together by a cybercriminal. Also, think of the Internet as public. Even if privacy settings are in place, information posted can still get out there, through friends reposting, and it is stored on servers that can be hacked. Be comfortable with the public seeing whatever you are posting on social network sites. Also be skeptical and beware of strangers. Not everyone is who he or she claims to be, and they could have stolen someone's identity to commit cybercrime.

3.1 Sentiment Analysis

Sentiment analysis is the multidisciplinary field of study that deals with analyzing people's sentiments, attitudes, emotions and opinions about different entities such as products, services, individuals, companies, organizations, events and topics and includes multiple fields such as natural language processing (NLP), computational linguistics, information retrieval, machine learning and artificial intelligence. It is set of computational and NLP based techniques which could be leveraged in order to extract subjective information in a given text unlike factual information, opinions and sentiments are subjective. Despite the recent surge of interest in sentiment analysis since the term was coined by the demand for information on sentiment and opinion during decision-making situations dates back to long before the widespread use of the World Wide Web. Opinions are central to almost all human activities as they could influence our behaviors specially when making a decision.

3.2 Language Processing Based Methods

Meanwhile, some other works have addressed sentiment analysis from two different aspects, namely, lexicon-based, and linguistic analysis. The most obvious yet important indicators of sentiments are sentiment or opinion words such as good, amazing, poor, and bad as well as some phrases and idioms which are used to express positive or negative opinions. A sentiment lexicon is the list of such words and phrases and is necessary but not sufficient for sentiment analysis. In addition to exploiting lexicons, linguistic based approaches also use the grammatical structure of the text for sentiment classification. There are two kinds of lexicon generation methods, namely, dictionary based and corpus-based approaches. The first category starts with a small set of opinion words and expands the lexicon through bootstrapping a certain dictionary while the second category generates the opinion lexicon through learning the dataset. For example, it proposes to employ predefined syntactic templates to capture links between opinion words and their targets. The links can then be used to infer more opinion words.

In social media our account is hacked with unknown person and they post fake news, for avoiding this we can use a method of text analysis with based on our interest and about us, careers and before posted posts, categorized the words in black list and white list using this method 90% white list included message is post, otherwise a OTP send to the email and check the person is right person. Message filtering systems are designed for unstructured or semi-structured data, as opposed to database applications, which use very structured data. In the implementation section use a college website and make department wise posts are posted by comparing the words with which department is posted check with dictionary of that subject, e.g. We can check the post from phisics department, tokenized the post and based on that tokens that all tokens compared with phisics dictionary, 90% tokens from phisics that will be posted, like this method all department posts are checked and make black list and white list and checked. Through the sentiment analysis the comments posted by the users are analyzed by two categories like black list and white list. Then it checks with IP section for the first time login verification to check whether authorized person login or hacker hacks any information.

3.3 PROPOSED ALGORITHM

3.3.1 Supervised Learning Algorithms for Sentiment Analysis in Text

The pre-processed input data is a collection of features (words). The positive, negative and neutral frequencies of the features in test data is obtained using the sentiment dictionary and then using various methods the probability of the input belonging to all the classes namely positive, negative and neutral class is calculated. The input data is then considered to be classified in the class which has the maximum probability.

Sentiment Analysis will require the following pre-processing:

1. Noise Removal - Cleaning the data from irrelevant news as well as advertisements/bio (if you have collected data by web crawling)

2. Classification - Categorizing the news data to different domains - "Markets", "Economy", "Industry", "and Technology" and so on. It is as necessary as the algorithm because you will have different set of features for different domains and thus, each domain should have different classifier.

3. Named Entity Recognition - This is the most important part of sentiment analysis as the objective of sentiment analysis is (In words of Bing Liu):

"Given an opinion document, discover all the opinion quintuples - entity, aspect, sentiment on aspect of the entity, opinion holder and the time/context of opinion."

For example, Sentiment analysis on political news to predict elections will obviously have to extract political entities - NarendraModi/Rahul Gandhi and the aspects of their campaign - secularism/minority upliftment from the news and then, tag them as positive or negative.

4. Subjectivity Classification - Classifying sentences as subjective or objective since subjective sentences hold sentiments while objective sentences are facts and figures.

5. Feature Selection - The features can be unigrams and/or bigrams or higher ngrams with/without punctuation and with/without stop words with presence (boolean)/count (int)/tfidf(float) as accompanying feature scorer for each sentence/paragraph/file. Filtering Stop words reduces accuracy. Adverbs and

determiners that start with "whey" can be valuable features, and removing them as English Stop words causes dip in accuracy. Similarly, punctuation helps in detecting sarcasm and exclamation.

IV. IMPLENTATION OF THE PROPOSED SYSTEM

1. New Candidate Registration

In this module performs if anyone wants to friendship for communication, that candidate registered in relevant details. Candidate gives the name, age, date of birth, gender, address, Mobile Number, place and needed information in new candidate registration. The given details finally administrators receive the information's and precede the further activity for applicants.

2. Login

Login system is the module which checks for a valid candidate when the user enters his loginid, user-id, password and link to the main page. As user id rules the system so a person is known by his uniqueness of his user id. As it is provide the candidate system, the user id is being validated with password in different cases to validate the genuinely of the candidate.

3. Data Acquisition

To obtain data for analysis "screen-scraper" program that automatically associated project pages from the Scratch website. Add online safeguarding issues to your current strategy, policies and procedures for safeguarding and user protection, retention and management of personal information, use of photographs, and codes of conduct/behavior. Organizational reporting procedures should also include the reporting of potentially illegal/abusive content or activity, including user's sexual abusive messages and online grooming.

4. Onion Model

Adapted onion model also revealed interesting trends in another two statistics that computed but which were not used to categorize Scratch users. First, the average number of friends showed a steep upward trend across categories, suggesting a trend of increasing social involvement. Second, members of higher categories have a higher average duration of activity than members of lower categories (defined as the time between a Scratch user's first and most recent postings of a project), suggesting that Scratch users started at the lower levels and sequentially progressed over time to higher levels.

5. Privacy Levels

Privacy and safety settings available across all aspects of the services - for photos, blog entries and image galleries - and set the appropriate level of privacy. Think about your target audience and who you wish to see the content. Failing to set appropriate privacy levels could result in messages which are defamatory, libelous or obscene appearing on your profile before you have a chance to remove them. This may result in significant personal distress, risk to the reputation of the individual or the organization, and require the intervention of the organization, the service providers and possibly the police.

6. Tagging

In this module performs the privacy settings on information's sharing websites. If you or a friend are tagged in an online messages (face book, flicker) the whole information's may be visible to their friends, your friends and anyone else tagged in the messages. Does not have to be friends with anyone to be tagged in their information album, if adults are tagged in a message can remove the tag but not the information's. When over the age of 18 the website will only look into issues that contravene their terms and conditions.

V. CONCLUSION & FUTURE WORK

One final approach for reducing boredom might leverage the community in order to increase the intensity and frequency of Scratch users' online activities. At present, the Scratch environment provides few incentives to drive users toward creating increasingly sophisticated information's. This stands in contrast to the afterschool club setting, where the Scratch designers report, "Every two or three months, we organized a Scratch-a-thon during which all members of process.".

, there is no agreed upon definition of the term "fake news". To better guide the future directions of fake news detection research, appropriate clarifications are necessary. . Though Centrality measures have been studied to find the spreading trace of misinformation within a social media site, a global analysis based on different web sites can further facilitate recovering the trace and seeking the real provenance of misinformation. Since budget of competing misinformation is often limited, efforts should be paid on the most destructive rumors. An effective way to estimate potential impact of misinformation will be very useful to control the negative influence.

REFERENCES

[1] Chen, H., C. Yang, M. Chau, and S. Li (Editors), *Intelligence and Security Informatics*, Lecture Notes in Computer Science, Springer, Berlin, 2009.

[2] US Committee on Homeland Security and Government Affairs, *Violent Extremism, the Internet, and the Homegrown Terrorism Threat*, 2008.

[3] Bergin, A., S. Osman, C. Ungerer, and N. Yasin, "Countering Internet Radicalization in Southeast Asia", ASPI Special Report, March 2009.

[4] Colbaugh, R. and K. Glass, "Predictive Analysis for Social Processes I: Multi-Scale Hybrid System Modeling, and II: Predictability and Warning Analysis", *Proceedings 2009 IEEE Multi- Conference on Systems and Control*, Saint Petersburg, Russia, July 2009.

[5] Pang, B. and L. Lee, "Opinion Mining and Sentiment Analysis", *Foundations and Trends in Information Retrieval*, Vol. 2 , pp. 1- 135, 2008.

[6] Das, S. and M. Chen, "Yahoo! for Amazon: Extracting Market Sentiment from Stock Message Boards", *Proceedings APFA*, 2001.

[7] Dhillon, I., "Co-Clustering Documents and Words Using Bipartite Spectral Graph Partitioning", *Proceedings ACM International Conference on Knowledge Discovery and Data Mining*, San Francisco, August 2001.

[8] Kim, S. and E. Hovy, "Determining the Sentiment of Opinions", *Proceedings International Conference on Computational Linguistics*, 2004.

[9] Sindhwani, V. and P. Melville, "Document-Word Co-Regularization for Semi-Supervised Sentiment Analysis", *Proceedings 2008 IEEE International Conference on Data Mining*, Pisa, Italy, December 2008.

[10]http://www.cs.cornell.edu/People/pabo/movie-review-data/.

[11] Ramakrishnan, G., A. Jadhav, A. Joshi, S. Chakrabarti, and P. Bhattacharyya, "Question Answering via Bayesian

Inference on Lexical Relations", *Proceedings ACL*, 2003.

[12] http://www.borgelt.net/bayes.html.

[13] www.mediaislam-bushro.blogspot.com.