International Conference on Engineering & Technology ( ICET-2020)

Available at: International Journal of Scientific Research and Engineering Development

RESEARCH ARTICLE OPEN ACCESS

# Detection and Removal of Publishing Fake Posts & Users Irrelavent Comments in Social Media

Mrs. R. Bhuvaneswari[1], J. Prasanth[2], P. Santhanakumar[3] and R. Sukumar[4]
[1]Head of the Department, [234]UG Students
[1234]Department of CSE, Selvam College of Technology, Namakkal.
hod.cse@selvamtech.edu.inprasanthjayakumar36@gmail.comsanthanamcse73@gmail.comsukucse717@gmail.com

----------------------------------------------------************************************----------------------------------------------------

**Abstract**

In this research, we show online social networks can be used to study crime detection problems. Crime is defined as an act harmful not only to the individual involved, but also to the community as a whole. It is also a forbidden act that is punishable by law. Crimes are social nuisances that place heavy financial burdens on society.

Here we look at use of data mining followed by sentiment analysis on online social networks, to help detect the crime patterns. Twitter is an online social networking and micro blogging service that enables users to post brief text updates, also referred to as "tweets". These updates can convey important information about the author.

A filter was designed to extract tweets from cities deemed to be either the most dangerous or the safest in the United States (US). A geographic analysis revealed a correlation between these tweets and the crimes that occurred in the corresponding cities. Over 100,000 crime-related tweets were collected over a period of 20 days. Sentiment analysis techniques were conducted on these tweets to analyse the crime intensity of a particular location. This type of study will help reveal the crime rate of a location in real-time. Although the results of this test helped in detecting crime patterns, the sentiment analysis techniques did not always guarantee the proper results.

We conclude with applications of this type of study and how it can be improved by applying media to text processing techniques. we also added to the projects are detected the current location of user.

----------------------------------------------------************************************----------------------------------------------------

## 1. INTRODUCTION

National security concern is the primary goal of any nation. Criminology studies focus on identifying criminal characteristics. The application of data mining techniques can help with this identification. Crime analysis, a part of criminology, is a law enforcement function that involves the systematic analysis of identifying and analyzing both patterns and trends in crime and disorder.

In the current world, the criminals are becoming technologically sophisticated, often expressing their emotions on the web. The World Wide Web's phenomenal growth has resulted in more users expressing their opinions

online. Customers use these opinions to buy a product, conduct market analysis, and so forth.

Twitter is one of the most popular online social networks to date, where users post their opinions in short text called "tweets". These tweets are typically limited to 140 characters. Twitter has approximately 500 million users; approximately 340 million tweets are sent every day. Twitter is used, primarily, for the following four reasons.

Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis, and computational linguistics to identify and extract subjective information in source materials. It is used to determine an author's attitude, with respect to a particular topic or the overall contextual polarity in the text. The rapid growth of Social media has spurred interest in sentiment. Various forms of online expressions (e.g., opinions-like reviews, ratings, and recommendations) have become major sources of information for businesses looking to market their products and manage their reputations.

The challenge of detecting crime patterns lies in geographically analyzing crime-related tweets and then performing sentiment analysis to identify crime prone zones in nearly real-time. Most of the studies that focused on crime pattern detection [8, 9] used data mining techniques to better understand historic data. This study used online social media to detect crime prone areas in almost real-time.

## 2. EXISTING SYSTEM

Much work has been done to understand the information dynamics in social networks, including theoretical models and empirical studies. However, most of the existing studies assume that each piece of information spreads independently regardless of the interactions between contagions. In the real world, multiple contagions may compete or cooperate with each other when they spread at the same time.

For example, the news about the banning of Samsung Galaxy S7 in airports may promote the spreading of the battery explosion events of S7, but suppress the news that Samsung is releasing other exciting products. Thus, in this example, the contagion-contagion interaction can be seen as a "competition" between the popularity of two pieces of information. Taking the interactions into account is crucial to address the question of how much a user would like to adopt a contagion. Recent diffusion models have started to consider interactions between contagions however, in most cases, the interactions they learned are latent factors and thus are difficult to understand.

For example, in the interactions it considered are between latent topics, making the promotion or suppression effects difficult to interpret it considered are between latent topics, making the promotion or suppression effects difficult to interpret. Specifically, given two contagions that are of unrelated content or subject matter, it is difficult to infer whether they will interact with each other when they spread simultaneously.

Actually, what interests us is the interactions among explicit categories, namely whether contagions belonging to one category (say food) would have some positive/negative effects on the spreading of contagions belonging to another category (say health). These interactions can be used to design viral marketing strategies to promote or suppress some products or news. For example, if it can be inferred that contagions belonging to sports usually have positive effects on the adoption of energy drinks, advertisements on energy drinks can be exhibited alongside with sports news in a user's input stream of posts to promote the sales.

## 3. ARCHITECTURE

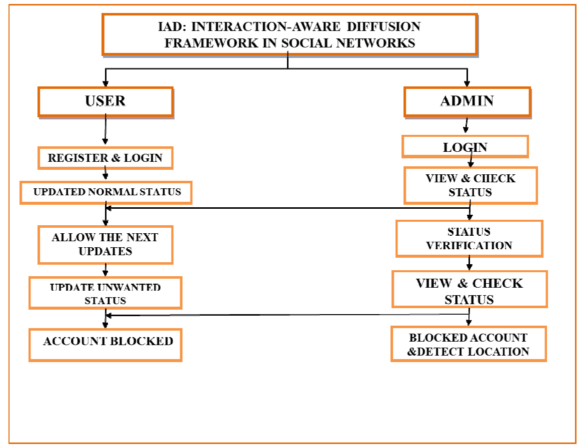


Fig. Architecture Diagram

### 3.1 USER

A user is a person who utilizes a computer or network service. Users of computer systems and software products generally lack the technical expertise required to fully understand how they work.

### 3.2 ADMIN

A system administrator is a person who is responsible for the upkeep, configuration, and reliable operation of computer systems; especially multi-user computers, such as servers.

### 3.3 STATUS VERIFICATION

Admin checks the status of the post which is posted by the user. This process is used for checking any irrelevant post is posted.

### 3.4 ACCOUNT BLOCKED

A blocked account refers to an account that does not allow for the indiscriminate withdrawal but instead has certain restrictions or limitations on when, how much, and by who, capital can be withdrawn.

### 3.5 DETECT LOCATION

The location detection is used for detecting the location of the user. It can be used by the higher officials if any suspicious activities are found.

## 4. PROPOSED SYSTEM

### 4.1 CRIME DISTRIBUTION

The main challenge behind crime data mining is to understand patterns in criminal behavior in order to predict crime and prevention. Any research that can assist in solving crimes is preferred to protect individuals. A number of studies examined data obtained from either a sheriff's office or a Crime Analysis Unit.

Clustering and Series Finder algorithms, respectively, were applied to the data in an effort to predict crime. Twitter, a powerful online social network, was used in this study to detect crime in almost real-time. The top ten most dangerous cities in the United States, as listed by Forbes magazine, were chosen for examination; the top ten safest cities were also examined for comparison.

**4.2 SENTIMENT ANALYSIS**

Sentiment analysis was used to determine a writer's/speaker's attitude with respect to either a topic or the overall contextual polarity of a text. Researchers use this analysis to measure emotions in online texts. The rise of social media fueled interest in using sentiment analysis to identify public opinions and interests.

Several open source software tools utilize machine learning, statistics, and natural language processing techniques to automate sentiment analysis on a large collection of texts that have been gathered from various sources Sentiment analysis is a two-step process that includes both subjectivity classification and sentiment classification. The term subjectivity classification is defined as distinguishing factual sentences from those used to present opinions, before analyzing sentiments. Paragraphs that present facts are typically removed so the researcher can focus on those paragraphs in which the author expresses opinions. Both naive Bayes classification and Cut Based classification are used for subjectivity classification.

The term sentiment classification is defined as detecting sentiment polarity of the subjective sentences. This sentiment classification is also divided into two categories: binary sentiment classification and multi-class sentiment classification. Binary sentiment classification involves classifying sentiments either positive or negative. Multi-class sentiment classification involves classifying sentiments into one of five categories: strong positive, positive, neutral, negative and strong negative.

The most common machine learning techniques used for sentiment classification include naive Bayes, maximum entropy, and support vector machine. Most sentiment analysis algorithms use simple terms to express sentiment. However, the cultural factors, linguistic nuances, and differing contexts prevent researchers from drawing the sentiment accurately.

**4.3 ANEW BASED APPROACH**

The ANEW is being developed to provide a set of normative emotional ratings for a large number of words in the English language. This was developed to aid researchers when studying emotions; it is often used to determine a tweet's sentiment.

Siddharth and Dr. Healey adopted a dictionary-based approach for determining the sentiment of tweets. They used the ANEW dictionary to provide pre-existing, normative emotional ratings for 1034 words along the three dimensions of valence, arousal and dominance. They used an independent matching technique to map all of the words in a tweet that were found in ANEW. They used two approaches (the arithmetic mean and normal distribution) to calculate both the mean valence and arousal.

**4.4 DEEP LEARNING FOR SENTIMENT ANALYSIS**

Richard, Alex, Jean, and Jason introduced the Recursive model, a state-of-the-art in sentiment analysis. They also introduced both the Recursive Neural Tensor Networks (RNTN) and the Stanford Sentiment Treebank. The Treebank includes fine-grained sentiment labels for over 200,000 phrases in the parse trees of over 11,000 sentences. When the RNTN model is trained on the new Treebank, it outperformed all previous methods on several metrics.

This approach follows the multi-class sentiment classification; it predicts five sentiment classes: very positive, positive, neutral, negative and very negative. The sentiment prediction's accuracy can reach 80.7%. An example of the RNTN accurately predicting.

**5. RELATED WORKS**

**5.1 INFORMATION DIFFUSION**

In recent years, researchers have extensively studied the information diffusion in social networks. A collection of models is proposed to explain the diffusion process from various perspectives, while some other models are proposed to predict whether a piece of information will diffuse.

However, most of the prior models assume the spreading of each piece of information is independent of others, e.g., the Linear Threshold Model the Independent Cascade Model SIR and SIS Model. The diffusion of multiple contagions has been covered in several recent works. The scenario discussed by these works is that one contagion is mutually exclusive to others, i.e., only involving the competition of contagions. In an agent-based model is employed to study whether the competition of information for user's finite attention may affect the popularity of different contagions, but this model does not quantify the interactions between them.

Our research is not limited to the mutual exclusivity condition, but instead, the proposed model assumes that a user can adopt multiple contagions. We offer a comprehensive consideration for the inter-relations of the contagions in online social networks. The most related work to ours is the IMM model proposed in [8], which statistically learns how different contagions interact with each other through the Twitter dataset. It models the probability of a user's adoption of information as a function of the exposure
sequence, together with the membership of each contagion to a cluster. However, this model doesn't consider the user roles, which has been proved to play an important role in information diffusion in our work. In addition, the clusters in this model are latent variables without real-world meanings. In contrast, our proposal can infer interactions among explicit categories, which are easy to interpret. IMM model is implemented as a state-of-the-art baseline to compare with. Other studies also neglect the influence of user roles' interactions, and do not discover the interactions of actual categories of contagions.

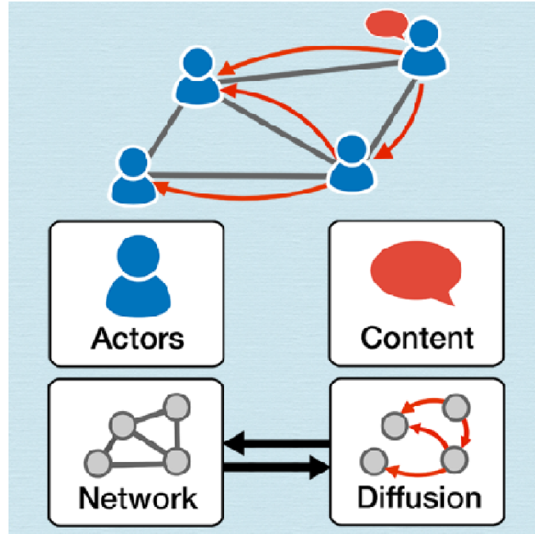

Fig. Information Diffusion

**5.2 SENTIMENT ANALYSIS**

Although recent works suggest the sentiments in the contents can play important roles in various applications such as product and restaurant reviews stock market prediction few existing studies quantify the effects of contents sentiment on the dynamics of information diffusion. Empirical analysis on German political blogosphere indicates that people tend to participate more in emotionally charged (either positive or negative) discussions.

A recent study on Twitter exhibits the effect of sentiment on information diffusion and reveals different diffusion patterns for positive and negative messages respectively. However, different from our IAD model, these works still treat each contagion in isolation and thus do not take the interactions into account. Besides, most of the previous studies try to extract only the sentiments. However, sentiments polarities are often dependent on topics or aspects. Therefore, detecting on which topics of the users are expressing their opinions is very important. Several models have been proposed to infer the topic and sentiment simultaneously propose the TSM model which can reveal the latent topical facets in a Weblog collection, the subtopics in the results of an ad hoc query, and their associated sentiments propose a novel probabilistic modelling framework based on LDA, called joint sentiment/topic model (JST), which detects sentiment and topic simultaneously from a text.

This model assumes that each word is generated from a joint topic and sentiment distribution, and hence doesn't distinguish the topic word and opinion word distributions propose a topic-adaptive sentiment classification model which extracts text and non-text features from twitters as two views for co training propose a LDA based model, Foreground and Background LDA (FB-LDA), to distil foreground topics and filter out longstanding background topics, which can give potential interpretations of the sentiment variations.

There are some other topic models considering aspect specific opinion words A recently proposed TSLDA model can estimate different opinion word distributions for individual sentiments for each topic and has been successfully applied to stock prediction. One weakness of TSLDA is that it divides a document into several sentences and sample the topic and sentiment of each sentence. Therefore, its performance is limited when it is applied to Weibo where most of the messages have only one or two sentences. The other weakness is that it lacks prior information, making it difficult to achieve good results for short texts. To address the aforementioned problems, we propose a variation of TSLDA model, namely LDA-S, to make it work for short texts such as Weibo and Twitter.


Fig. Sentiment Analysis

## 6. CONCLUSION

A crime pattern can be detected, nearly in real-time, when online social media is monitored. Crime can occur anywhere at any time. Previous statistics do not accurately identify the crime intensity of a specific location. More accurate results can be drawn from social media. Results from geographic data analysis conducted on various tweets provided a clear picture of the criminal trends in several different cities. The crime intensity day-wise positively correlated with crime statistics from cops, which ultimately prove the hypothesis. The Ferguson shooting case study clearly differentiates the city's safe and dangerous pattern. To be more precise, we analyzed the specific twitter accounts which tweet only about the crime scenarios happened in the city based on sheriff data and visualized.

The results gathered from this study were positive. An advanced sentiment analysis algorithm will aid in differentiating a sinister murderer from tweets within a specific location. Video-to-text processing, image-to-text processing, and data from various online sources would also help improve accuracy. This type of study would help with informing others of the crime pattern both within and around their location, ultimately assisting them with staying in a safe zone. Monitoring various social media outlets (e.g., Facebook, Google+, Tumblr, and Myspace) would improve accuracy.

## 7. REFERENCE

[1] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2003.

[2] J. Goldenberg, B. Libai, and E. Muller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," Marketing letters, vol. 12, no. 3, pp. 211–223, 2001

[3] H. W. Hethcote, "The mathematics of infectious diseases," SIAM review, vol. 42, no. 4, pp. 599–653, 2000.

[4] M. E. J. Newman, "The structure and function of complex networks," SIAM review, vol. 45, no. 2, pp. 167–256, 2003.

[5] S. Wang, X. Hu, P. S. Yu, and Z. Li, "MMRate: inferring multi-aspect diffusion networks with multi-pattern cascades," in Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2014, pp. 1246–1255.

[6] Y. Tang, Y. Shi, and X. Xiao, "Influence maximization in near-linear time: a martingale approach," in Proc. ACM Int. Conf. Special Interest Group Manage. Data, 2015.

[7] L. Weng, A. Flammini, A. Vespignani, and F. Menczer, "Competition among memes in a world with limited attention," Scientific Reports, vol. 2, 2012.

[8] S. A. Myers and J. Leskovec, "Clash of the contagions: Cooperation and competition in information diffusion," in Proc. 12th IEEE Int. Conf. Data Mining, 2012.

[9] X. Rong and Q. Mei, "Diffusion of innovations revisited: from social network to innovation network," in Proc. 22th ACM Int. Conf. Inf. Knowl. Manage.,2013.

  

[10] Y. Bi, W. Wu, and Y. Zhu, "Csi: Charged system influence model for human behavior prediction," in Proc. 13th IEEE Int. Conf. Data Mining, 2013.

[11] M. Coscia, "Competition and success in the meme pool: a case study on quickmeme. com," in Proc. Int. AAAI Conf. Weblogs Soc. Media, 2013.

[12] I. Valera and M. Gomez-Rodriguez, "Modeling adoption and usage of competing products," in Proc. 15th IEEE Int. Conf. Data Mining, 2015.

[13] N. Pathak, A. Banerjee, and J. Srivastava, "A generalized linear threshold model for multiple cascades," in Proc. 10th IEEE Int. Conf. Data Mining, 2010.

[14] B. A. Prakash, A. Beutel, R. Rosenfeld, and C. Faloutsos, "Winner takes all: competing viruses or ideas on fair-play networks," in Proc. 21rd Int. Conf. World Wide Web, 2012.

[15] B. Karrer and M. E. J. Newman, "Competing epidemics on complex networks," Physical Review E, vol. 84(3): 036106, 2011.

  