2016 12th International Conference on Signal-Image Technology & Internet-Based Systems

The Forensic Algorithm On Facebook Using Natural Language Processing

Mahasak Ketcham[1], Thittaporn Ganokratanaa[2] and Sasiprapa Bansin[1]
[1]Department of Information Technology Management, Faculty of Information Technology
King Mongkut's University of Technology North Bangkok, Thailand
[2]Department of Electrical Engineering, Chulalongkorn University, Bangkok, Thailand
mahasak.k@it.kmutnb.ac.th, charisma_sbunny@hotmail.com and s5707021857203@email.kmutnb.ac.th

***Abstract*—These days, social media has played a significant role in daily life of all people and ages in order to communicate as well as express their thoughts and feelings. In this paper, the authors have studied user data from social media (Facebook) whose shared posts are positive, and also the negative side posts that may lead to negative affect personally or can be further extended to the community and nation level. The purposes are to identify users who have commented on the negative side that may be a lawbreaker on Computer related crime. On this which beneficial about investigation for legal proceeding and it facilitate for the police or people who take a part in the operation on law. It also contributes in the community at large to peacefulness. The effective Naïve-Bayes classifier is used in order to classify these two user groups. It significantly shows that analyzing social media data by using Naïve Bayes model presented sharing positive and negative views accurately as well as reflects satisfied results.**

Keywords-Forensic Analysis · Social Media · Facebook · Naïve Bayes

I. INTRODUCTION

Digital development in modern society and the use of Information Technology which are very useful and creative become a big part in our daily life. On the other hand, it may lead to negative affect when it is used improperly in person or can be further extended to the community and nation level where there is a very high possibility of the transaction.

In recent years, the social media has significant role to everyday life of all people and ages to communicate as well as sharing their thoughts and feelings. Forensic Analysis is the application of scientific principles and techniques to the legal process in order to investigate, testify evidence, and lead offenders to criminal justice.

Without forensic evidence, many offenders in complex criminal cases are still free and may repeat offending against the laws and harm the others.

Thus, in many developed countries such as Japan, European countries and the US have applied existing scientific knowledge and technology to identify evidences in criminal trials to gain true scientific results which is very useful in criminal investigations. Particularly in Japan, shows it having over 90% of the murderers are arrested by applying scientific equipment and invented and innovated technology to be used in forensic science. In fact, if we bring information technology and social network process into

transaction management in database, and then apply to the forensic science, it will lead greatly to the benefits even in forensic process, fairness human right, as well as order in society.

Recently, many researches present the social media data that are analyzed in or-der to apply for correlation analysis and also the further result of existing data. Particularly from the study of researchers in forensic field, data analysis in forensic [1-8] shows that the social media data can be divided into major groups to be used in the study such as the researches that analyze data in text form, and the researches that analyze data in image form. In Thailand todays, the use of social media is un-closed that the information of social media users in text form is popular and it also reflects the use of social network such as Facebook is highly accepted and its growth rate is multiple.

Therefore, the authors propose to use data from social network (Facebook) applied with the field of forensic, analyzing with Naïve-Bayes classification algorithms in order to clarify data into two specific groups. The first group relates to sharing their thoughts and feelings on the positive side, and the second one relates to negative

posting. The purposes are to identify users who have commented on the negative side which may cause unwanted effect personally or can be further extended to the community and nation level

II. LITERATURE REVIEW

Uraz Yavanoglu, Busra Caglar, Ozlem Milletsever, Medine Colak, Semra Cakir and Seref Sag_roglu proposed the Intelligent Approach for Identifying Political Views over Social Networks. It is a research-based analysis of political views by analyzing Social Network data through Artificial Neural Networks: ANN model and Data mining. The data used in this research is taken from Twitter which is a public data. Therefore, this work helps to analysis thoughts and ideas from Twitter users both supporting or opposing the government [1].

Chris Howden, Lu Liu, Zhijun Ding, Yongzhao Zhan and KP Lam proposed the Moments in Time: A Forensic View of Twitter which the Twitter is carried through the Python IDLE client with MySQL database and display the data from Twitter via the SQL Statement [2].

Shankar Setty, Rajendra Jadit, Sabya Shaikh, Chandan Mattikalli and Vma Mudenagudi proposed the Classification of Facebook News Feeds and Sentiment Analysis which presented a system for classification of Facebook news feeds and

A learning based classifier is built using various classification algorithms such as Bi-nary Logistic Regression, Naive Bayes, Support Vector Machine (SVM), Bayes Net and J48 and This experiments on the

978-1-5090-5698-9/16 $31.00 © 2016 IEEE
DOI 10.1109/SITIS.2016.103

live news feeds showed that the proposed approach could achieve significantly improved performance for structuring the data on Facebook using SVM classifier learning model [3].

Zhaochun Ren, David van Dijk, David Graus, Nina van der Knaap, Hans Henseler and Maarten de Rijke proposed the Semantic Linking and Contextualization for Social Forensic Text Analysis which is a research-based analysis of the connective data between two social networks and after that analysis data connection setting
within the context [4].

Noora Al Mutawa, Ibtesam Al Awadhi, Ibrahim Baggili and Andrew Marrington proposed the Forensic artifacts of Facebook's instant messaging service which Facebook Chat conversations in Latin and Arabic character set were conducted using three major web browsers, and then forensically retrieved [5].

Reza Soltani and Abdolreza Abhari proposed the Identity Matching In Social Media Platforms which is a research-based analysis of the relationship of data users on Facebook, Twitter and Linkedin by dividing the data into three classes; Personal Identities (uses String Matching Algorithm and the Google Maps APIs), Social Identities, such as posted information and video files (NLP and Youtube APIs are used in this class), and Relational Identities such friend's information and group membership [6].

Aniello Castiglione, Giuseppe Cattaneo and Alfredo De Santis proposed A Forensic Analysis of Images on Online Social Networks which is a research-based analysis of visual information that may violate the copyright law or engage in illegal activity on social networks. In this research, the analysis mainly focuses on processing the uploaded images and what changes are made to some of the characteristics. The pixel resolution and related metadata are studied together [7].

Norulzahrah Mohd Zainudin and Madjid Merabti proposed the Online Social Networks As Supporting Evidence A Digital Forensic Investigation Model and Its Application Design. The application is designed to automatically search on social network, and the searched data is stored in order to analyze the evidence [8].

III. FORENSIC SCIENCE

Forensic science (often shortened to forensics) is the applications of scientific principles to matters of legal problems and criminal investigation [9]. The Law (the rule) is a legal principle enacted by authoritarians to force people to follow unconditionally, and any person who fails to comply with the law will be provided adverse actions. The Laws may be enacted to define the rules of the relationships between people or people and the State or in governing the country. Furthermore, the law may be established from accepted customary within a particu-lar social setting.

IV. PROPOSED METHODOLOGY

Social Media is a tool of communication used to share thoughts and ideas as well as build and maintain relationship. However, people in various moral background both good and bad generally use social media to communicate their behaviors and it may lead to problem if that tool is not used properly. These days the tendency of Facebook users has grown quickly and it's expected to become more popular. In fact, APIs is very important because they dictate how developers can analysis Facebook data in order to use it effectively, therefore Facebook creates its own APIs. It is very useful if we can use data from Facebook users to analysis and use it for the forensic process in order to build peaceful coexistence in today's society:

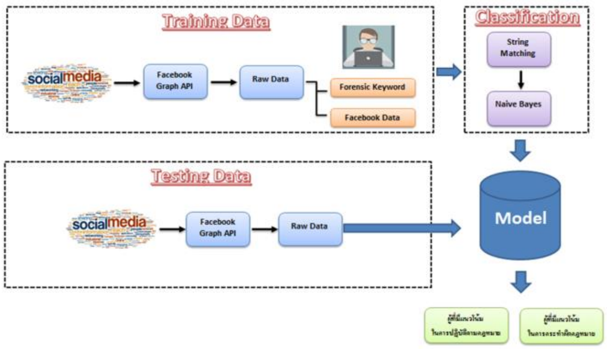


Figure1. Diagram of Forensic Analysis using Social Media Process

Step 1: Prepare data from Social Media. In the first step, the researchers provide data from social media in order to be analyzed. As mentioned, the social media data is taken from Facebook and divided into two categories: 1) Facebook user data (User detail) and 2) Details of sharing comment through the wall of social media (Opinion Detail). The data has shown on Table 1.

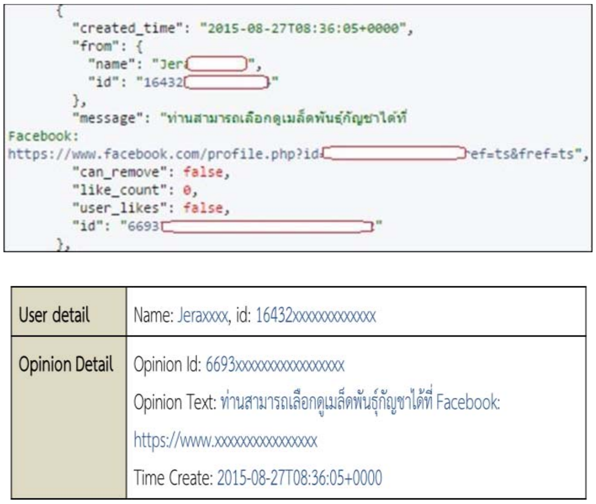
TABLE I. TWO CATEGORIES OF SOCIAL MEDIA DATA USED FOR ANALYSIS

Data Type	Attribute
User Detail	User Id, User Name
Opinion Detail	Opinion Id, Opinion text, Time Create

Table 1 shows attribute data of each Data Type. It's categorized for suitable rea-son in the analysis and maximize to the accurate results. The first Data Type is to manage the user data such as User ID and User Name. The second Data Type is a detail of sharing comment through Facebook wall which includes Opinion ID, Opinion Text and Time Create. Proceedings are:

1. The data of 100 accounts from Facebook are chosen to be collected; 50 ac-counts are users who have shared their views positively, and another 50 accounts are users who have shared their views negatively.

2. Manage the provided social media data into the first Data Type, which relates to user data and the second Data Type which is about the views of each user.



```
    {
      "created_time": "2015-08-27T08:36:05+0000",
      "from": {
        "name": "Jer[redacted]",
        "id": "16432[redacted]"
      },
      "message": "ท่านสามารถเลือกดูเมล็ดพันธุ์กัญชาได้ที่
Facebook:
https://www.facebook.com/profile.php?id[redacted]ef=ts&fref=ts",
      "can_remove": false,
      "like_count": 0,
      "user_likes": false,
      "id": "6693[redacted]"
    },
```

User detail	Name: Jeraxxxx, id: 16432xxxxxxxxxxxx
Opinion Detail	Opinion Id: 6693xxxxxxxxxxxxxxxxxx Opinion Text: ท่านสามารถเลือกดูเมล็ดพันธุ์กัญชาได้ที่ Facebook: https://www.xxxxxxxxxxxxxxxx Time Create: 2015-08-27T08:36:05+0000

Figure2. Facebook Data

Step 2: Prepare forensic keyword. Prepare 100 Forensic keywords from 100 Facebook accounts. This preparation is included 50 accounts are users who have shared positive views, and another 50 accounts are users who have shared negative views

Step 3: Define word weight for forensic keyword. To define Forensic keyword and its weight, the data is separated into two sets. The first data set presents 50 positive Keywords which are from those sharing positively on Facebook and it's defined on number 1-50. The second data set presents 50 negative Keywords which are from those sharing negatively on Facebook and it's defined on number 51-100.

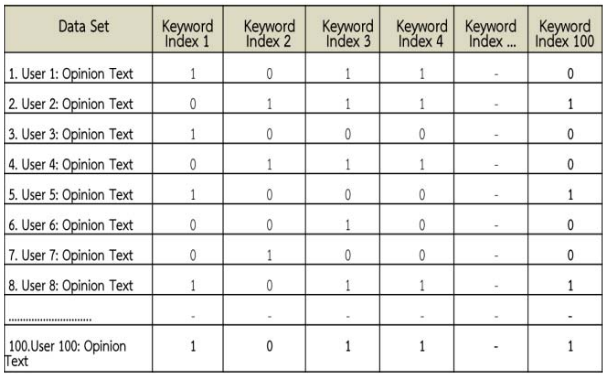
TABLE II. EXAMPLE FORENSIC KEYWORD

Forensic Keyword	Weight	Forensic Keyword	Weight
ความรัก (love)	1	คิดถึง (miss)	8
น้ำใจ (kindness)	2	ชอบ (like)	9
ขอบคุณ (thanks)	3	หัวใจ (heart)	10
ปราณี (mercy)	4	ยิ้ม (smile)	11
ยินดี (joy)	5	สวย (pretty)	12
ความสุข (happy)	6	คุณธรรม (moral)	13
สวัสดี (hello)	7	ความหวัง (hope)	14
โง่ (fool)	51	ซ่อง (brothel)	58
โรคจิต (pervert)	52	ขายบริการ (prostitute)	59
ฆ่า (kill)	53	แอบถ่าย (paparazzi)	60
แม่งเอ้ย (holy shit)	54	ระเบิด (bomb)	61
กัญชา (marijuana)	55	ปืน (gun)	62
ยาบ้า (amphetamine)	56	มีด (knife)	63
เฮโรอีน (heroine)	57	นรก (hell)	64

Manage the prepared User Detail, Opinion Detail and Forensic Keyword into a database.

Step 4: Search and Data Extraction. At this stage the divided Forensic keyword data is processed to search, compare, and find out the Frequency of Forensic keyword that are shown by the user reviews. To process on this, the String Matching technique is used in mapping Forensic Keyword together with Opinion Text data in order to prepare for data analysis in next step

TABLE III. EXAMPLE DATA SET FOR NAÏVE BAYES TRAINING



Data Set	Keyword Index 1	Keyword Index 2	Keyword Index 3	Keyword Index 4	Keyword Index ...	Keyword Index 100
1. User 1: Opinion Text	1	0	1	1	-	0
2. User 2: Opinion Text	0	1	1	1	-	1
3. User 3: Opinion Text	1	0	0	0	-	0
4. User 4: Opinion Text	0	1	1	1	-	0
5. User 5: Opinion Text	1	0	0	0	-	1
6. User 6: Opinion Text	0	0	1	0	-	0
7. User 7: Opinion Text	0	1	0	0	-	0
8. User 8: Opinion Text	1	0	1	1	-	1
....	-	-	-	-	-	-
100.User 100: Opinion Text	1	0	1	1	-	1

Step 5: Data analysis: Naïve Bayes. In this procedure, it conducts to data analysis process by classifying with Naïve Bayes technique [10] that a principle of probability measure is applied to explain the research method as below.

$$P(C \mid A)=(P(A|C) \times P(C))/(P(A)) \qquad (1)$$

- Posterior probability or P(C|A) is the probability of class C (target) given predictor A (attribute).

- Likelihood or P(A|C) is the probability of training data which is predictor A (attribute) given class C (target) where $A = a\_1 \cap a\_2 \cdots \cap a\_M$ and M is the number of attribute in the training data.

- Prior probability or P(C) is the probability of class C. But the probability that attribute $A = a\_1 \cap a\_2 \ldots \cap a\_M$ occurs in the training data is very low or there is no any attribute pattern at all. Therefore, the principle that each attribute must be independent, and its P(A|C) equation is changed to

$$P(A \mid C)=P(a\_1 \mid C)\times P(a\_2 \mid C)\times \cdots P(a\_M \mid C) \quad (2)$$

From this research, two classes of result are classified as follows.1. Class Positive: Social media users whose views are positive. 2.Class Negative: Social media users whose views are negative.

Thus, the processes are operated as follows.

1. Calculate the probability that attributes are operated on Opinion = Positive.

P(Opinion = positive|A) = P(Wordweight1 = 1| Opinion = positive) × P(Wordweight2 = 2| Opinion = positive) × …× P(Wordweight50 = 50| Opinion = positive)

2. Calculate the probability that attributes are operated on Opinion = Negative.

P(Opinion = negative|A) = P(Wordweight51 = 51| Opinion = negative) × P(Wordweight52 = 52| Opinion = negative) × …× P(Wordweight100 = 100| Opinion = negative)

In using Naive Bayes model finds that the probability of some attributes is 0, so it indicates that there is no any pattern of this attribute occurred in the training data. Since in using a model that the probability is 0 normally produces the predictive value of 0, therefore it must add 1 to the count for every attribute which is known as Laplace Smoothing.

Step 6: Getting result of analysis. In this procedure, it is an analysis of the results from using Naïve Bayes technique in analyzing social media users are those share positive or creative views, or the negative shared posts which may cause unwanted effect personally or can be further extended to the community and nation level.

V. Experimental Result

The study carried out the Forensic in analyzing social media data, the researchers have conducted the Naive Bayes model and other methods in order to analysis through RapidMiner tool for sentiment analysis and text mining. In this paper, two groups of users who share post on positive side and negative side are presented in Table 4.

TABLE IV. Comparison the training model

ID	Model	Accuracy	Classification Error
1.	Naive Bayes	90.89	9.11
2.	AutoMLP	85.00	15.00
3.	Support Vector Machine	82.00	18.00
4.	KNN	40.00	60.00

From Table 4, the same set of data is analyzed by using Naive Bayes, AutoMLP, Support Vector Machine, KNN and Linear Discriminant Analysis. It is found that Naive Bayes is the most effective model in analyzing that achieve the maximum accuracy 90.89% and Classification Error 9.11%.

VI. Conclusion

In this research, the purposes are to identify users who have commented on the negative side that may be a lawbreaker on Computer related crime, on this which beneficial about investigation for legal proceeding from Facebook's user information and it facilitate for the police or people who take a part in the operation on law. It also contributes in the community at large to peacefulness. The researchers have studied in order to maximize social media data in forensic. Facebook API is implemented to help in preparing the unclosed or public information. Half of accounts are the posts of users who share their positive and creative views, on the other hands, the rest are those whose sharing posts reflect their negative views. Moreover, the forensic keyword of both clarified users are used in data analysis equally. To carry out this research, the forensic social media data analysis by using Naive Bayes, AutoMLP, Support Vector Machine, KNN and Linear Discriminant Analysis with the same data set, presents that Naïve Bayes is the most effective model with its accuracy of 90.89%. As it benefits, we can maximize the study and the development of this research by extending the size of analyzing data and forensic keyword for the effective accuracy of data analysis. Hopefully, the researchers are confident that this research should be further developed for the benefits of forensic process as well as building peaceful coexistence in today's society.

Acknowledgment

This research was funded by King Mongkut's University of Technology North Bangkok. Contract no. KMUTNB-60-GEN-018.

References

[1] Uraz Yavanoglu, Busra Caglar, Ozlem Milletsever, Medine Colak, Semra Cakir, Seref Sag_roglu. Intelligent Approach for Identifying Political Views over Social Networks. IEEE 2013.

[2] Chris Howden, Lu Liu, Zhijun Ding, Yongzhao Zhan, KP Lam. Moments in Time: A Forensic View of Twitter. IEEE 2013.

[3] Shankar Setty, Rajendra Jadit, Sabya Shaikh, Chandan Mattikalli and Vma Mudenagudi. Classification of Facebook News Feeds and Sentiment Analysis. IEEE 2014.

[4] Zhaochun Ren, David van Dijk, David Graus, Nina van der Knaap, Hans Henseler, Maarten de Rijke. Semantic Linking and Contextualization for Social Forensic Text Analysis. IEEE 2013.

[5] Noora Al Mutawa, Ibtesam Al Awadhi, Ibrahim Baggili and Andrew Marrington. Forensic artifacts of Facebook's instant messaging service. IEEE 2011.

[6] Reza Soltani, Abdolreza Abhari. Identity Matching In Social Media Platforms. IEEE 2013.

[7] Aniello Castiglione, Giuseppe Cattaneo, Alfredo De Santis. A Forensic Analysis of Images on Online Social Networks. IEEE 2011.

[8] Norulzahrah Mohd Zainudin, Madjid Merabti. Online Social Networks As Supporting Evidence A Digital Forensic Investigation Model and Its Application Design. IEEE 2011.

[9] Office of the Royal Society. [Online]. Dictionary. Source: http://www.royin.go.th/

[10] Eakasit Pacharawongsakda. [Online]. Data Mining Trend. Source: http://dataminingtrend.com/2014/naive-bayes/