Identifying And Tracking Online Financial Services Through Web Mining and Latent Semantic Indexing

Kristen Bernard, Andrew Cassidy, Monica Clark, Kevin Liu, Katrina Lobaton, Drew McNeill, and Donald Brown, *Fellow, IEEE*

***Abstract*—As Internet usage has heavily increased within recent years, money launderers have started to take advantage of Online Financial Transaction (OFT) services to facilitate their money laundering activities. However, law enforcement has struggled to understand and detect OFT services that criminals use for money laundering. To assist law enforcement in its efforts to identify and monitor OFT services, we have designed the Online Financial Transaction Services Identification Tool (OFTSIT), which crawls the Internet and determines the probability that they are OFT services. OFTSIT analyzes a website's content and extracts textual features using latent semantic indexing (LSI). LSI is a text mining approach that can extract a small number (< 10) of features from more than 40,000 possible words on a website. OFTSIT inputs the LSI discovered features into a generalized linear model to produce the probability that a website is an OFT service. Testing showed that OFTSIT outperforms current method of manual searching. This paper describes the system architecture, algorithms employed to classify OFT services from other websites, and performance testing to demonstrate OFTSIT's operational relevance.**

I. Introduction

THE United States General Accounting Office defines money laundering as "the act of converting money gained from illegal activity, such as drug smuggling, into money that appears legitimate and in which the source cannot be traced to the illegal activity" [1]. While the exact level of money laundering activity remains unknown, estimates from 2001 put the global range between $600 billion and $2.8 trillion, or 2% to 9% of global GDP [2]. With the rapid growth of the Internet, money launderers have turned to Online Financial Transaction (OFT) services to facilitate their criminal activities.

OFT services, which began appearing in the late 1990's, enable users to transfer assets, such as money, credit card payments, or commodity backed currencies over the Internet. Uses for OFT services range from buying goods on eBay to playing online poker. While well-known services such as PayPal are regulated by standards similar to U.S. banks, other services have little to no regulation and often require a minimal amount of personal information before allowing transfers between users. These anonymous and unregulated transactions between two parties are extremely attractive to money launderers and other cyber-criminals. Evidence of money launderers using OFT services has been mounting in recent years, and law enforcement has turned its attention to this recent development. In 2007, an OFT service named E-Gold was indicted by the United States Department of Justice on charges of aiding in money laundering [3].

Despite increased scrutiny, law enforcement has struggled to understand and detect money laundering activities through OFT services [4]. The number of OFT services is a minuscule component of the Internet; therefore, locating new sites and changes in the OFT service landscape is difficult and costly. Our goal in building the Online Financial Transaction Service Identification Tool (OFTSIT) was to aid law enforcement agencies in understanding where money could be transferred online anonymously and how the OFT services change over time. To identify OFT services, OFTSIT crawls the web and then uses latent semantic indexing (LSI) combined with generalized linear models to obtain classifications of the crawled websites. The next section provides a brief literature review followed by Section III that shows the OFTSIT architecture. Section IV gives the details of the algorithms used by OFTSIT to classify websites and Section V describes the calibration of the algorithms. Section VI shows the testing results and Section VII provides conclusions and directions for future work.

II. Literature Review

There is an enormous growing body of literature on text mining, and overviews can be found in [5]**.** Text mining is a subset of data mining and seeks to discover patterns from unstructured textual data [5]. The text mining literature has focused on commercial and research applications rather than law enforcement [6]. One example from law enforcement particularly relevant to our work used text mining to detect money laundering by uncovering and building a network of criminal activity [6]. This research focused on automating the process of identifying illicit individuals and their financial patterns by drawing potential relationships between isolated online and offline events [6]. Much of the work

This work was partially supported by Booz Allen Hamilton.
Faculty Advisor: Donald Brown, *IEEE Fellow*, deb@virginia.edu

All authors are with the Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA 22904.

Authors: Kristen Bernard (kib97@virginia.edu), Andrew Cassidy (apc3n@virginia.edu), Monica Clark (mdc7y@virginia.edu), Kevin Liu (sl94sj@virginia.edu), Katrina Lobaton (kml6b@virginia.edu), and Drew McNeill (apm9t@virginia.edu)

using text mining for law enforcement focuses on identifying suspect persons through the characteristics and relationships of their Internet activities [7]. Text mining algorithms are used to extract information from available text. For example, concept space algorithms are currently used to connect various entities (persons, actions, places) present within the mined data [7]. Although some current applications of text mining in combination with knowledge extraction algorithms address law enforcement needs, there are no reported results of text mining to identify financial web services as facilitators of criminal activity [7].

III. OFTSIT System Architecture

The system architecture consists of four elements: the web-crawler, the text mining algorithm, the database, and the graphical user interface (GUI). The web-crawler, written in JAVA, crawls the web and saves text cleaned of HTML or XML. The text mining algorithm, written in R, analyzes the text using logistic regression and LSI and provides the probability of each website being an OFT service. The website's probability and other attributes are then saved into the MySQL database. The GUI is written in PHP and displays information from the database about the crawled sites. Included in the GUI is a page with information about high probability websites flagged by the text mining algorithm. The user can either accept or reject these potential OFT service entries into a list of verified OFT services. Confirmed websites are then saved into the database. After the system identifies the number of new sites found and user confirms its potential to be an OFT service, the text mining algorithm retrains by incorporating the verified list of OFT services into the training set for logistic regression. This refined model is used to improve future prediction. The focus of this paper is on the text mining algorithm, its design, and its performance.

IV. Text Mining Algorithm for Online Financial Service Discovery

The text mining system component produces a probability of being an OFT service for each website visited by the web crawler. "Bag of words" techniques are employed to identify new services thereby providing no attention to the order of the words or grammatical structure. Instead words are taken as independent data values associated with a website. Our process is made up of four linear steps:

- Preprocessing
- Creation of Term Document Matrix
- Latent Semantic Indexing (LSI)
- Logistic Regression

A. *Preprocessing and Creation of Term Document Matrix*

Preprocessing removes stop words and employs stemming techniques to clean up the text documents for analysis. Stop words are common words in the English language that do not carry significant meaning. Examples include "the", "is", and "that" [8]. Deleting stop words prior to analysis greatly increases computing efficiency without compromising the results. Stemming, the second step, recognizes and consolidates different forms of the same word into one. For example, the words "purchasing" and "purchased" are merely different forms of the word "purchase". They convey the same meaning and stemming allows them to be treated as the same word. Stemming improves both the computing efficiency as well as the quality of analysis. The stemming algorithm used in our system is Porter's stemmer [9].

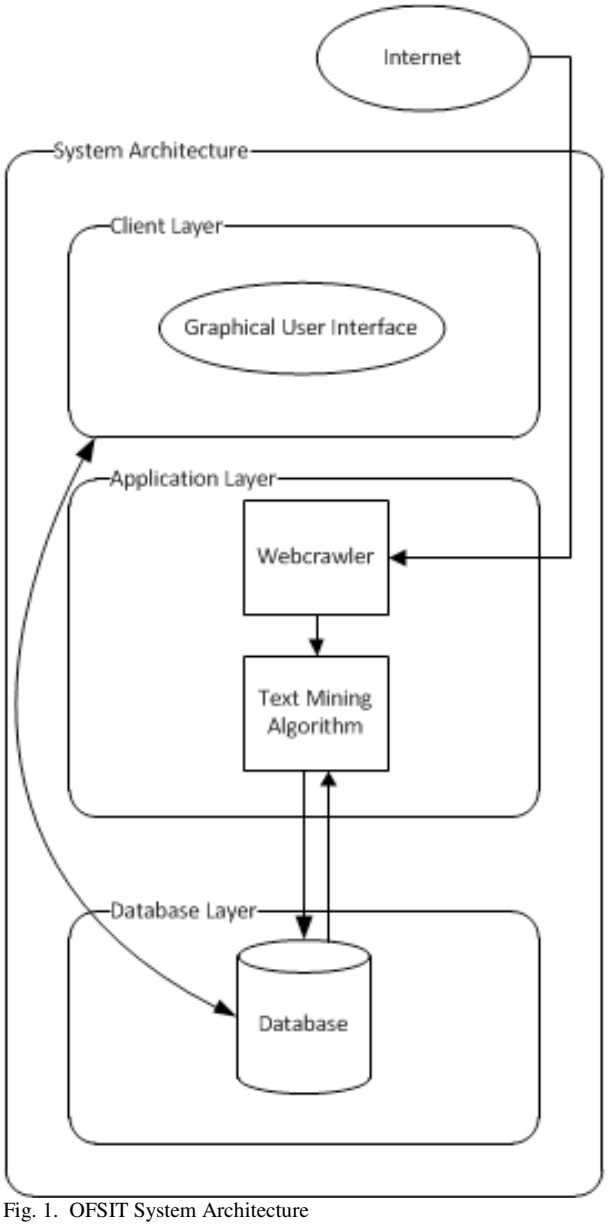

Fig. 1. OFSIT System Architecture

Next we construct an *n x m* term document matrix whose data entries correspond to the words that appear in the mined

websites. This matrix, which has *m* columns representing the collection of *m* documents and *n* rows representing all *n* words in the collection, is populated by the number of times each term appears in each document. The term document matrix is a sparsely populated matrix (the vast majority of entries are 0), meaning there are a large number of words on the web and only a select few appear on each website.

B. Latent Semantic Indexing

Latent Semantic Indexing (LSI) is a method that uses the linear algebra technique called Singular Value Decomposition (SVD) to identify concepts in a collection of text documents by looking at associations between terms as they occur in documents [10]. It is a popular technique used to identify the underlying topics or concepts associated with documents, which is critical for applications such as web searching and spam filtering [11]. The first step in LSI is to perform SVD where the term document matrix is decomposed into three different matrices.

$$\text{Original Matrix} = U \times \Sigma \times V^T$$

where U and V are orthogonal matrices and Σ is a diagonal matrix

The columns of U and V are the eigenvectors, also called singular vectors, associated with $D \cdot D^T$ and $D^T \cdot D$, respectfully. The diagonal matrix Σ contains the corresponding eigenvalues, or singular values. The matrices $D \cdot D^T$ and $D^T \cdot D$ contain the correlation between terms over documents and the correlation between documents over terms. Next, the *k* (a user selected value) highest values appearing in the Σ matrix are selected along with the corresponding *k* columns of *U* and the *k* rows of V^T. The new matrix appears as D*.

$$D^* = U^* \times \Sigma^* \times V^T$$

where Σ^* is a *k* by *k* matrix

In our algorithm, LSI is used as a dimension reduction technique, which reduces the sparsely populated term document matrix into small latent concept document matrix. As Srivastava and Sahmi stated, "the eigenvectors for a set of documents can be viewed as concepts describe by linear combinations of term chosen in such a way that the documents are described as accurately as possible using only *k* concepts" [12]. This lower dimensional space is given by $V x \Sigma^T$. Thus, we now have the *k* selected concept dimensions reflected over the *m* documents [13]. For our particular purposes, we refer to these *k* dimensions as concepts or indices.

C. Logistic Regression

The logistic regression utilizes the indices produced by LSI to produce a probability of the website being an OFT service. It does so by assigning weights to each of the indices in terms of how much each index reflects the topic of OFT service:

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}.$$

where
i = number of websites
k = the number of latent semantic indices
p_i *=probability that the site is an OFT service*
$x_{j,i}$ *=numeric projection of the document (i) into the different indices*
β_j *=weight on the* j^{th} *index*

The optimal set of weights is calculated using a training set of known financial and non-financial websites. As this sample increases in size, the accuracy of the logistic regression increases. Obtaining the training set and the steps for training the algorithm are discussed in the next section.

V. Model Design and Calibration

A sample set of known OFT services and other websites was collected to train and calibrate the algorithm. This list of 310 websites, 70 OFT services and 240 others, was obtained by manually searching the Internet and seeding the web crawler with known OFT services. The next step was to look at the links in the closely crawled space. Attempts were made to incorporate all types of websites, such as sports and news, as well as financial websites, in the sample set.

A. Model Design and Training

LSI was used on the training set to obtain the latent indices and to map each website onto the LSI space. The graph below shows each of the websites in the training set mapped onto the first two dimensions of a three dimensional LSI space.

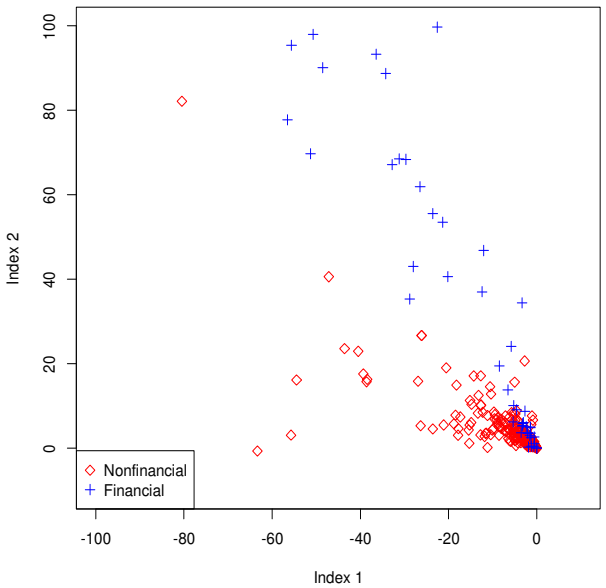

Fig. 2. Projection of websites onto LSI space

In an ideal situation, the OFT services and other websites would be completely separated, however, there is some overlapping in the two groups. Despite this overlap, there is enough separation between the two groups to build a model

that accurately distinguishes OFT services from the rest of the websites. In addition, the graph only shows two indices whereas many indices may be used to increase the accuracy of identification. The optimal number of indices for predictive power was found to be four. This process is discussed in detail later in this section.

The latent concept within each index can be deduced from looking at the words with the highest loadings within that index. Table 1 shows the thirty words with the highest weightings for Index 2. Many of the words in this list are related to precious metals. These words include "metal", "gold", "gold money", "silver", "palladium", "platinum", and "precious". This indicates that the latent concept in Index 2 could potentially be called "precious metals" or "currencies". Looking back at Fig. 2, one can see that OFT services were generally higher on Index 2, indicating that the distinguishing characteristics of OFT websites is their tendency to use precious metals as a medium of exchange or to back their e-currency with precious metals.

The indices are used to calibrate the logistic regression model to obtain the optimal set of weights for the indices. When the web crawler encounters a new website, the probability of being an OFT service is obtained by mapping the website onto the existing LSI space and inputting into the logistic regression equation.

TABLE I
WORDS WITH THE HIGHEST LOADINGS

Index 2		
metal	platinum	window
gold	payment	uk
goldmoney	server	invest
buy	service	bank
price	delivery	currency
rate	video	account
silver	hold	merchant
custom	report	secure
bar	certify	good
palladium	precious	article
metal	platinum	window
gold	payment	uk
goldmoney	server	invest
buy	service	bank
price	delivery	currency

B. *Measurement of Performance*

The performance of the algorithm is measured by its true positive rate (TPR), the fraction of true positives out of total positives, and false positive rate (FPR), fraction of false positives out of total negatives. To avoid biasing the test of performance, two-thirds of the data set was used to train the model while the remaining third was used to test the model. A Receiver Operating Characteristic (ROC) Curve applied to the testing set is reproduce in Fig. 3.

The values of the ROC curve for an ideal model are 0 for false positive rate and 1 for true positive rate. For any non-ideal model, there is an inherent trade-off between FPR and TPR that can be manipulated by changing the probability threshold used for flagging a site as an OFT service. The curves represent the FPR and TPR for different thresholds. From Fig. 3, the model can achieve a TPR of greater than 90% with a FPR of less than 10%.

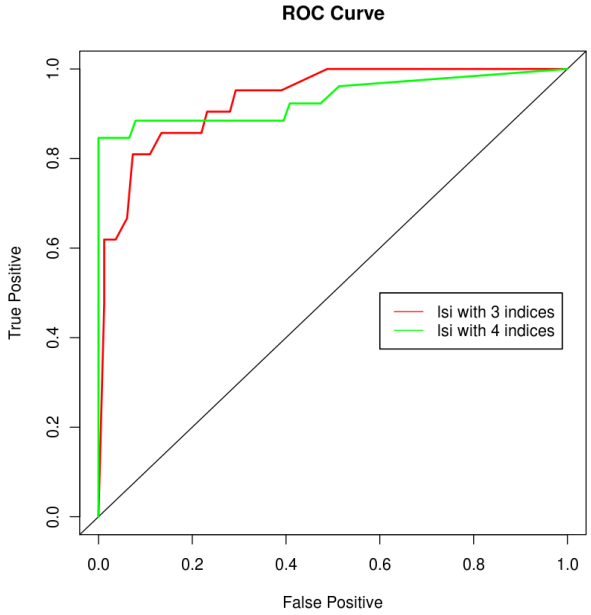

Fig. 3. ROC curve for two models, one with 3 indices and another with 4 indices

The TPR and FPR vary depending on which websites were randomly selected to be in the test set, therefore a technique called cross validation was used to obtain a more accurate measurement of the algorithm performance and determine its variance due to random sampling. Cross validation divides the sample set of 310 websites into ten random subsets. Nine subsets were used to train the model, while the last subset was used to test model performance. This procedure was repeated for nine more iterations until every subset was used for testing. The metric of performance in the cross validation technique was the total number of errors, or the sum of false negatives and false positives using a probability threshold of 0.5. The average error was obtained by averaging the total number of errors from all ten iterations. By conducting 100 simulations of cross validation, a distribution for the average error was obtained and reproduced in Fig. 4.

The cross validation results were also used to determine the optimal number of dimensions for the LSI space. Too few dimensions did not take advantage of the predictive power of the data; while too many dimensions resulted in over-fitting. Fig. 4 shows the distribution of average errors for models with different numbers of LSI dimensions. The models with four dimensional LSI spaces produced both the fewest average number of errors and the fewest median number of errors, so the final model was built using a four dimensional LSI space. Furthermore, the Kolmogorov-Smirnov test failed to reject the null hypothesis (under α level of .05) that the underlying distribution for average

errors in dimensions 3 & 4, the two highest performing dimensions, were the same. Therefore, the dimension with the smaller average error and median, 4, was chosen.

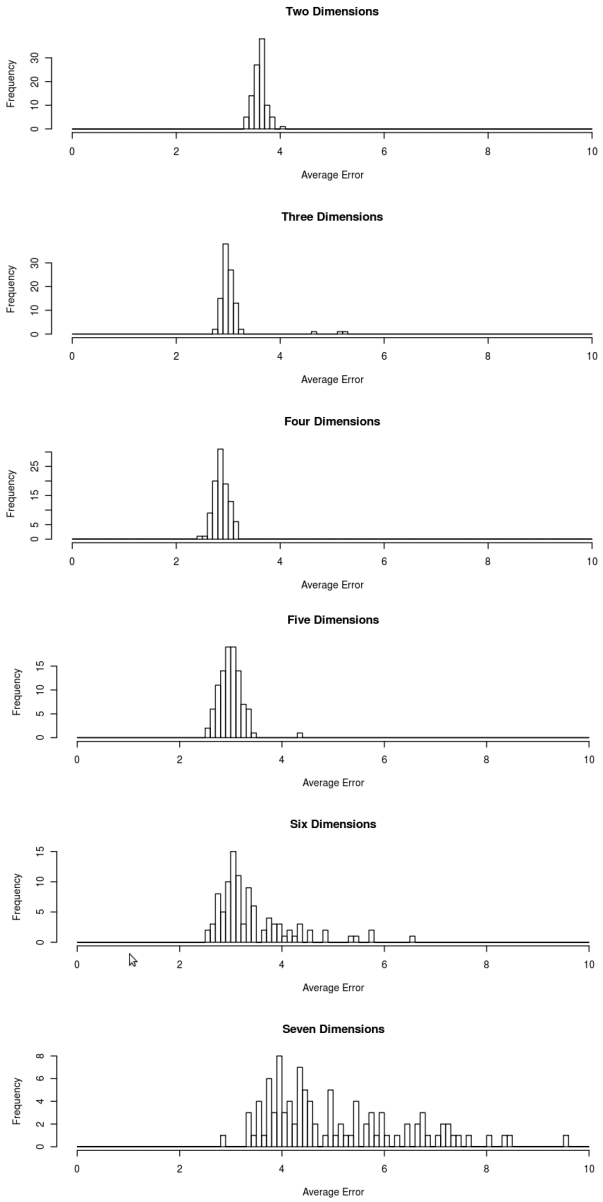

Fig.4: The distribution for the average error associated with cross validation simulations (310 divided by 10 gives a test set size of 31) shows that the mean, median, and variance of error all increase as the number of indices increases beyond 4. Four indices was the number selected for the final model.

VI. Testing and Metrics

We measured the performance of the Online Financial Transaction Service Identification Tool (OFTSIT) against the existing method of manually searching for and identifying OFT services by three metrics: cost, time, and accuracy. In terms of cost, each component of OFTSIT is open source and requires no cost, except the opportunity cost of constructing and developing the system. The only other capability requirement is a server to run OFTSIT, which can be considered negligible. In terms of measuring accuracy and time of each approach, we tested human subjects against the performance using two timed tests. The first test asked twelve users to determine whether a list of 40 websites were OFT services. The second test asked eleven users to manually find OFT services using an Internet browser within a five-minute time frame. The test results are shown in Table 2.

The team examined the OFTSIT's percent error in comparison to that of the testers to compare the user testing to the algorithm performance in terms of accuracy. Table 2 shows the testers had a lower average error but a wider range of variation in error compared to those of the system. While the testers had a smaller error, the sheer number of websites that the OFTSIT can visit makes it advantageous in terms of cost. On average, the webcrawler visited about 750 websites in a five-minute time frame, which is faster compared to the average time of 8:36 for users to go through 40 websites.

TABLE II
Test Results

Test 1 Results: Average Number of Identifications (12 subjects) Words with the Highest Loadings	
True Positive	9.5
False Positive	1.5
True Negative	1
False Negative	28

Test 1 Results: Time Summary in Minutes (12 Subjects)	
Average Time	13:37
Lowest Time	8:36
Highest Time	34:00

Test 2 Results: Average Number of Identifications (11 Subjects)	
Average Number	12.7
Lowest Number	4
Highest Number	25

Statistics: 95% Central Credible Interval of Percent Error		
Statistic	Testers	System (using P = 0.5)
Mean	0.05	0.1
Lower Bound	0.025	0.00309
Upper Bound	0.093	0.101
Range	0.068	0.00123

We measured the system's performance with respect to time in two ways. The first method measures the time to find required to visit websites. OFTSIT, seeded with a Google search of "money", was able to find two new OFT services within a 12-hour time period. The automatic OFTSIT had a shorter retrieval time over manual site browsing. Human subjects, with an average of 12.7 OFT services in five minutes, performed better at finding OFT services. While human subjects greatly outperformed OFTSIT at finding OFT services, they are also expensive. OFTSIT, on the other hand, is easily scalable and able to run continuously without incurring high costs over time.

VII. Implications and Future Steps

OFTSIT will allow law enforcement to efficiently monitor OFT services. It solves one of the biggest problems law enforcement encounters when attempting to understand and detect money laundering on the Internet: tracking the number of OFT services and their fast changing landscape. OFTSIT solves these problems by greatly reducing the time and cost necessary to search for and monitor OFT services.

OFTSIT makes it possible in the future for law enforcement to employ a risk-based approach in its fight against money laundering on the Internet. With an accurate and up to date knowledge of OFT services, law enforcement can focus its investigative efforts on the websites most at risk for facilitating illegal activity. Further research needs to be done to employ such an approach. The next steps include analyzing the characteristics of OFT services that increase risks for money laundering and developing automated tools to identify those OFT services with high-risk characteristics. The OFTSIT is an important first step in providing law enforcement with an understanding of the OFT service landscapes that had been prohibitively costly to acquire using traditional methods.

References

[1] The United States General Accounting Office. (1996). *Money Laundering: a framework for understanding US efforts overseas.* Washington, DC. http://www.gao.gov/archive/1996/gg96105.pdf.

[2] The World Bank. (2010). *World development indicators*. Retrieved November 8, 2010, from http://databank.worldbank.org/ddp/home.do?Step=12&id=4&CNO=2

[3] Digital Currency Business E-Gold Indicted For Money Laundering And Illegal Money Transmitting. Retrieved from http://www.justice.gov/criminal/cybercrime/egoldIndict.htm

[4] Wall, D. S. (2008). Cybercrime, media and insecurity: The shaping of public perceptions of cybercrime. *International Review of Law, Computers & Technology*, 22(1/2), 45-63.

[5] Delen, D. & Crossland, M. (2008). Seeding the survey and analysis of research literature with text mining. Expert Systems with Applications, 34(3), 1707-1720.

[6] Zhang, Z. (2003). Applying data mining in investigating money laundering crimes. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '03, 747.

[7] Chen, H., Chung, W., Xu, J., Yi Quin, G., Chau, M. (2004) Crime data mining: A general framework and some examples. Computer, 37(4), 50-56.

[8] Lewis, D., Yang, Y., Rose, T., and Li, F. (2004). RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research.* 5, 361-397.

[9] Porter, M. Snowball. (2001). A language for stemming algorithms. Retrieved from http://snowball.tartarus.org/texts/introduction.html

[10] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semantic analysis. Journal of the American Society of Information` Science, 41(6):391–407, 1990.

[11] Lee, S. Song, J. Kim, Y. An Empirical Comparison of Four Text Mining Methods. Journal of Computer Information Systems Vol. 51, 2010, pp. 1-10.

[12] Srivastava, A. N., & Sahami, M. (Eds.). (2009). Text mining: Classification, clustering, and applications (First ed.). Boca Raton, FL: Taylor and Francis Group, LLC.

[13] Feldman, R., & Sanger, J. (2007). The text mining handbook (First ed.). New York, NY: Cambridge University Press.