Examining Textual Features of Financial Social Media to Detect Cyber Attacks

Jinwei Liu*, Richard A. Aló*, Cong Liu†
*Department of Computer and Information Sciences, Florida A&M University, Tallahassee, FL 32307, USA
†School of Computer Science and Technology, Shandong University of Technology, Shandong, China
*{jinwei.liu, richard.alo}@famu.edu, †liucongchina@sdust.edu.cn

***Abstract*—With the proliferation of social media, cy- ber threats and attacks have significantly increased in complexity and quantity in financial market. Malicious hackers leverage the influence of social media to spread deceptive information with an intent to gain abnormal profits illegally or to cause losses. Measuring information content in financial social media helps identify these threats and attacks. In this paper, we propose CDetector, an ML- based approach to identifying social media features thatcorrelate with abnormal returns of the stocks of companies vulnerable to be targets of cyberattacks (e.g., cognitivehacking). To test our approach, we collected price data and the social media messages on multiple technology companies, and extracted features that contributed to abnormal stock movements. Preliminary results show that the top social media features associated with abnormal price movements are the terms that are simple, motivate actions, incite emotion, and uses exaggeration, and the selected features correlate with abnormal messages and abnormal returns of the stocks of companies.**

***Index Terms*—cyber threats, abnormal behavior, cyber-security, social media, feature extraction**

I. Introduction

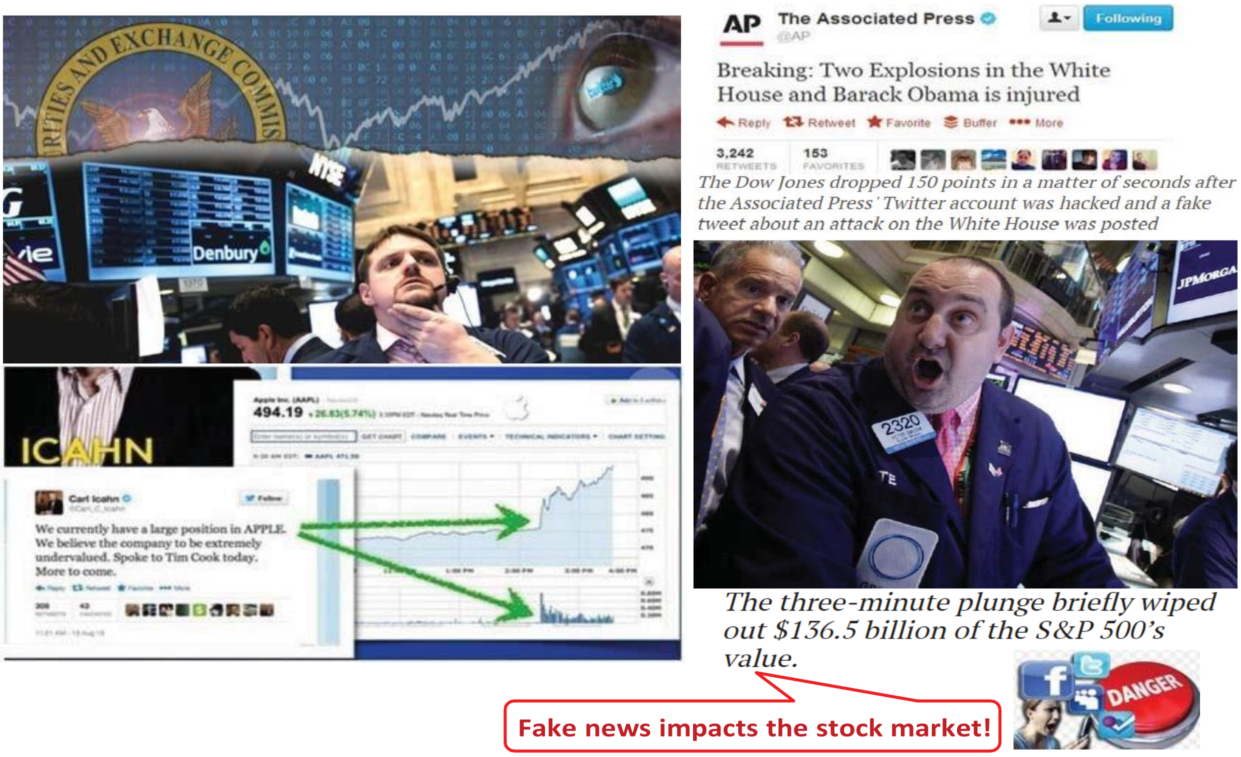
The prevalent use of social media facilitates the spreading of malicious messages that may inject misleading information to divert normal market operations [1]. Malicious hackers are increasingly exploiting the information environment and economic infrastructure to deceptively divert public sentiment to their own favors. The financial market thus becomes highly vulnerable to cybersecurity risks. Figure 1 shows examples of cyberattacks, and it reveals that misinformation (e.g., fake tweets) impacts the stock market. Cyber threats/attacks are still one of today's most challenging cybersecurity issues that are not well addressed by the commonly employed security solutions due to the rarity of confirmed positive threat cases and severe imbalance of ordinary datasets.

Machine learning attracts many interests and many studies utilize machine learning to identify cyber threats and attacks. Feature selection, as the process of selecting important features that contribute most to the learning model, hugely impacts the performance of the learning model. Efficient feature selection can significantly improve the accuracy and applicability of the learning and the process of classification [2]. In the data perspective, feature selection can be roughly categorized into four main categories: similarity-based methods, information-theoretical-based methods, statistical-based methods and sparse learning-based methods [3]. However, the fea- ture selection methods in these categories have some limitations [3]. To overcome the limitations of existing feature selection methods, we propose the hybrid feature selection (HFS). Financial social media data involves highly imbalance occurrences and the skewed distribution datasets significantly impact the performance of the learning model [4]. In this paper, we develop C Detector to identifying social media features that correlate with abnormal returns of the stocks of companies vulnerable to be targets of cyberattacks. C Detector first uses the resampling method to balance the dataset, then it uses HFS to identify the social media features that correlate with abnormal returns of the stocks. Finally, it uses the machine learning algorithm to perform binary classifica- tion for detecting cyberattacks by utilizing the identified features.

The remainder of this paper is organized as follows. Section II reviews the related work. Section III presents the problem formulation and the description of data collection in this paper. Section IV describes the proposed ML-based approach to identifying social media features that correlate with abnormal returns of the stocks of companies vulnerable to be targets of cyberattacks. Section V presents the experimental results, findings and analyses. Section VI concludes this paper with remarks on our future work.

II. Related Work

Cyberattack/threat detection has become an active research topic, and feature selection can possibly help to detect cyber threats. Many studies leverage feature selection to detect cyber threats. Below we review prior research on cyberattack/threat detection using feature selection in financial social media.



We currently have a large position in APPLE. We believe the company to be extremely undervalued. Spoke to Tim Cook today. More to come.

Breaking: Two Explosions in the White House and Barack Obama is injured

The Dow Jones dropped 150 points in a matter of seconds after the Associated Press' Twitter account was hacked and a fake tweet about an attack on the White House was posted

The three-minute plunge briefly wiped out $136.5 billion of the S&P 500's value.

Fake news impacts the stock market!

Fig. 1: Examples of fake tweets: fake news impacts stock market and can be dangerous.

Okutan *et al.* [5] used unconventional signals extracted from the Twitter and GDELT data sources and showed that they can be used to predict various cyberattacksfor the anonymized target entity KNOX. They applied feature technique to the data set with 10 x 10 folds cross validation to cross check their findings about the importance of each unconventional signal for each attack type. To balance the distribution of the majority and minority instances in an imbalanced data set, they developed SMOTE++ that uses a hybrid approach where under sampling, synthetic minority instance generation and reweighing techniques are used together. Bilge *et al.* [4] proposed a set of unique features (i.e., network, activity, user, and tweet content) based on tweets information, and they developed a supervised machine learning solution for detecting cyberbullying in the Twitter. They employed three feature selection algorithms (i.e., chi-square test, information gain and Pearson correlation) to determine the discriminative power of each feature, and they identified top 10 significant features. To handle the imbalanced class distribution in the data, they employeda combination of oversampling the minority (abnormal) class and undersampling the majority (normal) class and weight adjusting approaches *et al.* [6]. Specifically, they tested four classifiers (NB, LIBSVM, random forest and KNN) in four different settings: basic classifiers, classifiers with feature selection techniques, classifiers with SMOTE alone and with feature selection techniques, and classifiers with cost-sensitive alone and with feature selection techniques. Sabottke *et al.* [7] conducted a quantitative and qualitative exploration of the vulnerability-related information disseminated on Twitter. They used the mutual information for feature selection, and identified features that are useful for detecting exploits. Also, they designed and evaluated a detector for real-worldexploits utilizing features extracted from Twitter data.Yao *et al.* [8] formulated cyberbullying detection as a sequential hypothesis testing problem, and leveraged the feature selection to design an algorithm for reducing the time to raise a cyberbullying alert by minimizing thenumber of feature evaluations necessary for a decisionto be made.

However, most proportion of the prior feature selection methods is usually based on one chosen approach only without considering combining dissimilar feature selection approaches to enhance the performance. Also, many of the approaches neglect the nature of the financial social media data: highly imbalanced dataset. In this paper, we first use the re-sampling techniquesto balance the dataset and clean the data set. Then we leverage multiple dissimilar feature selection approaches to complement the errors made by singular mechanism, and propose a hybrid feature selection approach to detect cyber threats in financial social media.

III. Problem Statement and Data Description

In this section, we first formulate our problem, then we describe the data collection in this research.

A. Problem Statement

Given the training data from social media messages and stock prices of technology companies: Apple, Microsoft, Intel, Cisco IBM and VZ and social media messages, how to effectively identify social media features that correlate with abnormal returns of the stocks of companies vulnerable to be targets of cyberattacks?

B. Data Collection

We develop the data collection by downloading the social media messages and stock prices of six technology companies: Apple, Microsoft, Intel, Cisco, IBM and VZ from January 2019 to June 2019. The stock prices and social media messages were crawled once every five minutes on each US trading day.The messages were collected from the sites Twitter and StockTwits by using their public APIs. Also, wecollected numerical financial data of the aforementioned six companies from the following three sources: Google Finance, US Treasury and Yahoo Finance.

IV. Identifying Social Media Features

To detect cyberattacks in financial social media, we develop C Detector: an ML-based approach to identifying social media features that correlate with abnormal returns of the stocks of companies vulnerable to be targetsof cyberattacks. In this section, we first describe the calculation of abnormal returns, then we describe the approach we developed in this research.

A. Abnormal Returns Calculation

We computed the abnormal rate of return in the context of a market of S&P 500 stocks. The returns of these stocks were weighed by their market capitalizations. We generated a label for each five-minute scenario by computing the abnormal return for the stock portfolio i as $R_i - [R_f + \beta \cdot (R_m - R_f)]$ where R_i is the portfolio i's actual rate of return ($\frac{P_t - P_{t-1}}{P_{t-1}}$), and P_t is the weighted portfolio's price at time t; R_f is the risk-free rate of return normalized to the time span of the scenario; R_m is the market return based on S&P 500 index normalized to the time span of the scenario; β is the portfolio i's price volatility (over the past 36 months) relative to the overall market, and it is computed as the covariance of (Rate of Return of Stock, Rate of Return of Market)divided by market variance (where market is proxied by the S&P 500 index). Each scenario contains social media messages published during a five-minute window on a financial trading day (from 9:30 am to 4:00 pm). For each scenario, we assigned a label of 1 to the scenario if the absolute value of the abnormal rate of return is 0.5% or above; otherwise, we assigned a label of 0 to the scenario.

B. Extracting Social Media Features

We focus on the scenarios with abnormal return for identifying the associated social media features. We first remove URLs, punctuations and special characters, and then tokenize the messages, convert the messages to lowercase, and perform stemming.

Feature Identification using TFIDF: TFIDF is computed in Eqs. (1)-(3) below:

$$tf(t, d) = 0.5 + 0.5 \cdot \frac{f_{t,d}}{max\{f_{t',d} : t' \in d\}} \qquad (1)$$

$$idf(t, D) = log \frac{N}{|\{d \in D : t \in d\}|} \qquad (2)$$

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D) \qquad (3)$$

where $tf(t, d)$ is term frequency which measures the count of terms in a document, and $idf(t, D)$ is the inverse document frequency which measures how much information the word provides. In Formula (2), N is the total number of documents in the messages $N = |D|$, $|\{d \in D : t \in d\}|$ is the number of documents where the term t appears. If the term is not in the corpus, we adjust the denominator to $1 + |\{d \in D : t \in d\}|$ to avoid division-by-zero problem. In the experiment, we considered each message as a document, and all the messages in the entire collection of scenarios that had abnormal price movements.

Feature Identification using Information Gain: We also use Information Gain (IG) to identify the social media features that correlate with abnormal returns of the stocks. IG is defined as the expected reduction in entropy occurring when a feature is present versus when it is absent. Given a particular and arbitrary feature f_i, we compute the IG as follows [9]

$$IG(f_i) = E(I) - \sum_{v \in dom(f_i)} \frac{|I_v|}{|I|} \cdot E(I_v) \qquad (4)$$

where $E(I)$ is the value of the entropy of the data, I_v is the number of items in which the feature f_i, has a value equaling v, and $E(I_v)$ is entropy computed on data where the feature f_i assumes value v. In the experiment, we first computed the IG values of each term using Formula (4) and ranked the terms based on the IG value in descending order. Then we considered the terms with the top rank of IG values as the social media features that correlate with abnormal returns of the stocks.

Feature Identification using Gain Ratio Feature Selection: Gain ratio (GR) is an extension to information gain (IG) that reduces the bias of IG on highly branching features. GR considers the number and size of branches when it chooses an attribute. It corrects the IG by taking the intrinsic information of a split into account. Intrinsic information is entropy of distribution of instances into branches (i.e., how much information do we need to tell which branch an instance belongs to). As intrinsic information gets larger, the value of attribute decreases [10].

$$GR\ (Attribute) = \frac{Gain\ (Attribute)}{Intrinsic\_info\ (Attribute)} \qquad (5)$$

Feature Identification using Chi-squared Feature Selection: Chi-squared feature ranking evaluates the merit of each feature individually with the chi-squared statistical measure. Chi-squared attribute evaluation evaluates the worth of a feature by computing the value of the chi-squared statistic with respect to the class. This measure exams the absence of independence between a term t and a category c.

After calculating score of each feature using TFIDF, Information Gain, Gain Ratio and Chi-square, C Detector normalizes the scores to the same range ([0,1]). Denote R_1^* as the normalized TFIDF score of a feature, R_2^* as the normalized IG score of a feature, $R_3{}^*$ as the normalized GR score of a feature, and $R_4{}^*$ as the normalized Chi-square score of a feature. Thus the overall rank of the feature can be obtained as follows:

$$R = \Sigma_{i=1}^{m} w_i R_i^* \qquad (6)$$

where $m = 4$, and w_i $(i = 1, \cdots, m)$ are the weights.

To determine the weight of each factor (w_i) we use relative comparison method. Let $H = (h_{ij})_{m\times m}, i,j = 1, 2, \cdots, m$

$$h_{ij} = \begin{cases} 1, & if\ R_i^*\ is\ more\ important\ than\ R_j^* \\ 0.5, & if\ R_i^*\ is\ the\ same\ importance\ as\ R_j^* \\ 0, & if\ R_i^*\ is\ less\ important\ than\ R_j^* \end{cases} \qquad (7)$$

Obviously:

$$\begin{aligned} h_{ii} &= 0.5 \\ h_{ij} + h_{ji} &= 1 \end{aligned} \qquad (8)$$

Since the weight of each factor is related to the practical application, largely depending on the goal of the system, we can qualitatively assign weights to $R_1^*, R_2^*, R_3^*, R_4^*$ according to the goal of the system and determine the order of weight for each factor. The weight can be calculated by the following formula:

$$w_i = \frac{\sum_{j=1}^{m} h_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{m} h_{ij}} \qquad (9)$$

In this way, we quantify the qualitative weight of each factor in Formula (6).

C. The Design of C Detector

In this section, we present the design of C Detectorfor detecting cyberattacks in financial social media. C Detector first identifies social media features by using HFS, then it uses the machine algorithm to perform classification using the identified features. The HFS is an important component of C Detector. Figure 2 shows the architecture of C Detector. C Detector consists of four phases: data preprocessing (Phase 1), feature selection (Phase 2), training and testing (Phase 3), determining cyberattacks (Phase 4). Below we describe each phase.

In Phase 1, C Detector employs the SMOTEalgorithm [11] and the T-Link algorithm [12] to balance the dataset and clean the data. In Phase 2, C Detectorleverages HFS to extract social media features that correlate with abnormal returns of the stocks (seeSection IV-B). In Phase 3, C Detector performs training based on the identified features, and uses cross- validation for testing. In Phase 4, C Detector uses themachine learning algorithm to detect cybeattacks.

Algorithm 1: Pseudo code for CDetector

```
Input: Social media messages (tweets) and stock prices of
       four technology companies: Apple, Microsoft, Intel,
       Cisco, IBM and VZ
Output: The categories of social media tweets
1 Performs data cleaning (e.g., removing URLs, punctuations
    and special characters, and then tokenizing the messages,
    etc.)
2 Balance the dataset using the SMOTE algorithm and T-Link
    algorithm
3 Use HFS to identify top K features that correlate with
    abnormal returns of stocks // Section IV-B
4 Train the algorithm based on the identified top K features
5 Use cross-validation for testing
6 Predict if the tweets are cyber threats/attacks by predicting
    if the stock has abnormal return
```

Algorithm 1 shows the pseudocode for C Detector. The algorithm first uses the re-sampling techniques (i.e., SMOTE and T-Link) to balance the dataset and clean the data (line 1). Next, it generates the weights for different feature selection methods (line 2). After generating the weights for different feature selection methods, the algorithm uses HFS to identify top K features that correlate with abnormal returns of stocks (line 3). Next, the algorithm performs training based on the identifiedtop K features (line 4). Then, the algorithm uses cross-validation for testing (line 5). Finally, the algorithm use the machine learning algorithm to predict if the tweetsare cyber threats/attacks by predicting if the stock has abnormal return (line 6).

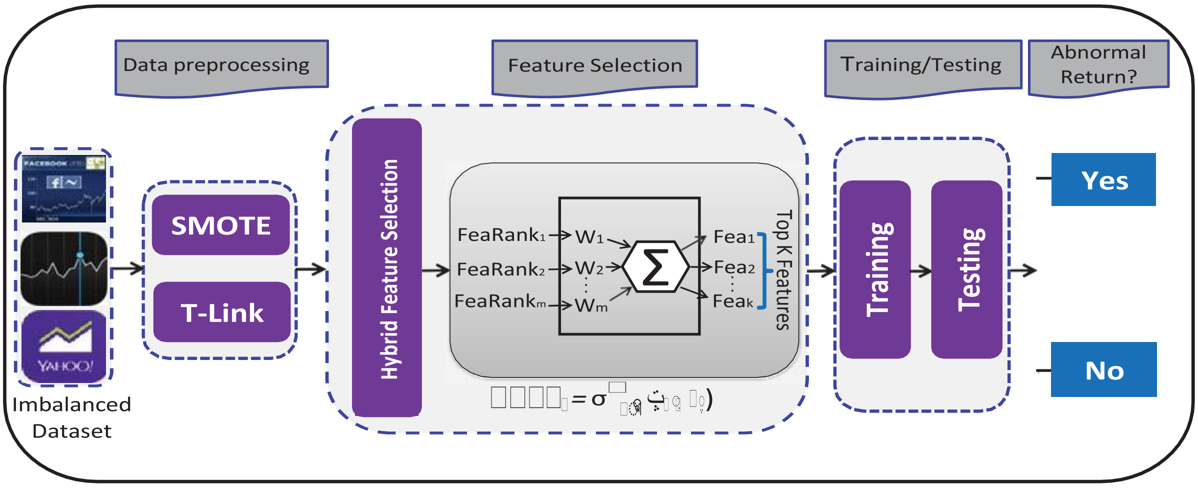


Fig. 2: The architecture of CDetector.

V. Experimental Findings

In this section, we present the experimental findings of an empirical study of the data collected from January 2019 to May 2019. The data consists of 2,302 scenarios (containing 105,402 messages). Then, we chose the scenarios with abnormal returns to analyze features that correlate with abnormal returns of the stocks, which consists of 69 scenarios (containing 1,053 messages). The proportion of the abnormal messages is 1%.

We conducted experiments to identify the social media features that correlate with abnormal returns of the stocks to detect cyberattacks based on different methods.For the implementation of TFIDF, we used the python libraries pandas, numpy, nltk for data processing, andwe used library scikit-learn to extract features. In the implementation of IG, we considered two categories: positive and negative. We first calculated the probabilities of different variables (e.g., prob. of a category C_i, prob. of a term, prob. of absence of a term, etc.) to compute the IG of each term. Finally, we calculated the IG values for each term and sorted the values indescending order and terms with top rank of IG values were selected as the target social media features that correlate with abnormal returns of stocks. Table 3 shows the representative messages from normal and abnormal scenarios.

Table II shows the top 20 features identified by each method. The last row shows features commonly identified by both methods. We see that the identified features are emotion-oriented and typically reflect the status/trend of stocks. These messages likely encourage or discourage people to take actions and therefore have the potential of incurring abnormal returns of stocks. Also, some features are identified by both methods, and hence are more likely to be associated with abnormal stock price movement.

TABLE I: Representative messages with abnormal stock returns and normal stock returns

Normal Messages
M1: $AAPL thanks god M2: $AAPL $AMZN like I said this morning.... M3: $AAPL has released the second iOS 11.3 beta for public testers. M4: $AAPL normal service resumed chaps M5: My shopping list today: $AAPL $MU $TXT $CY $JNJ $UNH $URI $CMI $LYB $DXJ $EEM $IWM $CSCO $ANET This is what happens when $AMZN use their own switches... M6: Time to load on some quality stocks $AMZN $GOOGL $FB $NFLX $INTC
Abnormal Messages
M1: $AAPL so strong.. sell the house and buy AAPL and can sleep well in the car! M2: $AAPL omg. Sell your kidney, first born and buy buy buy. M3: $BABA sell and buy $AAPL near/OTM calls and you'll have made all your money back in less than 10 days. M4: $AAPL cheap stock price is getting and getting more attractive to scoop up big! Then one of these days, it will gap up to 229+! Cheap Buy!!! M5: $INTC over sold on all metrics buy buy buy M6: $INTC Bogus press release from CEO on Friday starting to sink in longs? Extremely bearish! Sell your shares buy $AMD and thank me later!

TABLE II: Top 20 features correlate with abnormal returns of the stocks using different methods

Method	**Feature**
TFIDF	Like, Buy, Open, Free, Good, Weak, Earnings, Short, Nice, CEO, Bob, Great, Trade, Back, Active, Floor, Big, Trump, Coming, Low
IG	Bob, More, Free, Down, Close, Easte, But, Cut, Like, Open, Build, Floor, CEO, Good, Alert, Cleve, Low, Buy, Out, Great
TFIDF $\cup$ IG	Like, Buy, Open, Free, Good, CEO, Bob, Great, Floor, Low

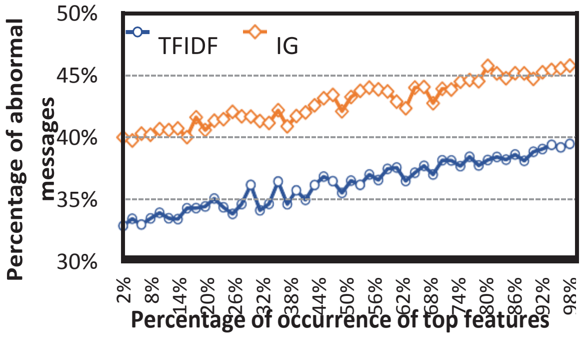


Fig. 3: Relationship between top features and abnormal messages.

TABLE III: Significance test of the correlation between the percentage of abnormal messages and the percentage of the occurrence of top features in different methods

Method	Pearson correlation coefficient	R Square	**p-value (5% level)**	**Statistical Significance (5% level)**
TFIDF	0.938	0.880	1.026E-23	Extremely Significant
IG	0.945	0.894	5.202E-25	Extremely Significant

They may be helpful to indicate risks of cyberattacks.

A. Relationship between Top Features and Abnormal Messages

To test if the identified features are correlated with abnormal returns of stocks, we also tested the correlation between the occurrence of top features identified by TFIDF and IG and the percentage of abnormal messages. Figure 3 shows the correlation between the occurrence of top features and the percentage of abnormal messages in TFIDF and IG. In Figure 3, we see that the percentage of abnormal messages increases as the percentage of the occurrence of top features increases. For the same percentage of occurrence of top features, the percentage of abnormal message in IG is relatively higher than that of TFIDF. This demonstrates that the identified features are correlated with abnormal returns of stocks.

To test the significance of the correlation between the percentage of abnormal messages and the percentage of the occurrence of top features, we conducted experiments and tested the P-value of the correlation for the three methods shown in Figure. Table III shows the significance test of the correlation between the percentage of abnormal messages and the percentage of the occurrence of top features in different methods. The experimental results show that there is a significant positive relationship between the percentage of abnormal messages and the percentage of the occurrence of top features,

$r(48)_{TFIDF} = 0.938$, $r(48)_{IG} = 0.945$, $p_{TFIDF} < 0.001$, $p_{IG} < 0.001$.

VI. Conclusions

In this paper, we develop C Detector, an ML-based approach to identifying social media features that correlate with abnormal returns of the stocks of companies vulnerable to be targets of cyber threats and attacks. Preliminary results show that terms that are simple, motivate actions, incite emotion, and use exaggerationare ranked high in the features of messages associated with abnormal price movements. We also provide selected messages to illustrate the use of these features in potential cyberattacks. This research should contributeto a new approach to understanding specific features that may associate with cyberattacks and to providing new empirical findings of examining social media and their abnormal effect on financial market. In the future, wewill expand the dataset and provide the experimental results of quantitative success metrics and give more solid evidence of the effectiveness of our approach. We will also compare our approach with state-of-the-art to fully verify the performance of C Detector.

References

[1] A. Sapienza, S. K. Ernala, A. Bessi, K. Lerman, and E. Ferrara. Discover: Mining online chatter for emerging cyber threats. In *Proc. of WWW*, 2018.

[2] K. Shaukat, S. Luo, V. Varadharajan, I. A. Hameed, and M. Xu. A survey on machine learning techniques for cyber security in the last decade. *IEEE Access*, 8:222310–222354, 2020.

[3] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu. Feature selection: A data perspective. *CSUR*, 50(6):94, 2017.

[4] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network. *Computers in Human Behavior*, 63:433-443, 2016.

[5] A. Okutan, S. J. Yang, and K. McConky. Forecasting cyber attacks with imbalanced data sets and different time granularities. *CoRR*, abs/1803.09560, 2018.

[6] X.-Y. Liu and Z.-H. Zhou. The influence of class imbalance on cost-sensitive learning: an empirical study. In *Proc. of international Conference on Data Mining (ICDM)*, 2006.

[7] C. Sabottke, O. Suciu, and T. Dumitras. Vulnerability disclosure in the age of social media: Exploiting twitter for predicting real-world exploits. In *Proc. of USENIX Security Symposium*, Washington, D.C., 2015.

[8] M. Yao, C. Chelmis, and D. Zois. Cyberbullying detection on instagram with optimal online feature selection. In *Proc. of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Barcelona, 2018.

[9] C. Musto, P. Lops, P. Basile, M. d. Gemmis, and G. Semeraro. Semantics-aware graph-based recommender systems exploiting linked open data. In *Proc. of ACM UMAP*, Halifax, Nova Scotia, Canada, 2016.

[10] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2001.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *JAIR*, 16:321–357, 2002.

[12] G. Batista, R. Prati, and M. Monard. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor*, 6(1):20–29, 2004.