TFC: Defending Against SMS Fraud via a Two-Stage Algorithm

Gaoxiang Li¹(✉), Yuzhong Ye¹, Yangfei Shi¹, Jinlin Chen^{2,3}, Dexing Chen^{2,3},
and Anyang Li⁴

¹ National Computer Network Emergency Response Technical Team/Coordination Center of
China Guangdong Branch, Guangzhou, China
jerishi@icloud.com

² Guangzhou Institute of Forensic Science, Guangzhou, China

³ Guangzhou Anti-Fraud Center, Guangzhou, China

⁴ Department of Mathematics, New York University, New York, USA
al5861@nyu.edu

Abstract. With the increasing of the telecom network fraud in China, SMS (Short Message Service) has become an important channel exploited by the criminals to contact victims. Due to the tiny amount compared with normal SMS, the high proportion of malicious adversarial characters, and the lack of knowledge to specific fraud types, it is still challenging to identify the fraud SMS efficiently. In this paper, we firstly conduct a measurement study to explore the characteristics of the fraud SMS. Based on the exploration, we propose a two-stage algorithm called TFC. TFC can quickly filter out normal SMS in the first stage with two indicator functions, and then easily identifies the category of fraud SMS in the second stage by combining the semantic deep features and the domain-knowledge based artificial features. We conduct two real-world SMS datasets for extensive experiments, and the results show that TFC successfully reduces calculation cost and achieves better performance in distinguishing various categories of fraud SMS.

Keywords: Telecom network fraud · SMS classification · Domain knowledge

1 Introduction

The evolution of the information society has brought great changes to the structure of crime, as traditional contact crimes keep declining, the new paradigm crime represented by telecom network fraud increases sharply in China. The average growing rate of telecom network fraud has reached 34% over the years [1], which has become the most prominent crime reported by the citizens. SMS is an important first-contact channel in many telecom network fraud cases, especially in the Loan Fraud, Part-time Fraud and Investing Fraud, etc. As shown in Fig. 1, if victims perform corresponding operations with fraud SMS, such as replying SMS or adding QQ, WeChat (the top 2 popular social media software in China) mentioned in the texts, the criminals would make further contact with victims in other communication channels where the real identities are

Fig. 1. Communication channel under different fraud procedures.

difficult to be traced. Therefore, the identification and prevention of fraud SMS has become a crucial issue.

Despite of some studies that have investigated detecting spam text already, these methods can not be applied in Chinese fraud SMS classification problem ideally due to the following reasons:

- Unlike the giant scale of normal SMS, the number of fraud SMS is relatively small. According to [2] and [3], the number of Smishing (SMS Phishing) for telecom network fraud only accounts for 0.05% of the whole SMS in China during 2020. It's important to design an effective method to filter out numerous normal SMS before the fraud SMS detection. Such scenarios are quite common and important for telecom operators, which are required to curb the fraud SMS by regulatory departments.
- Compared with English which has a relatively small alphabet, Chinese is logographic with a large set of characters. Besides, the battle between criminals and polices is extremely fierce, and the attack methods in fraud SMS, which has huge difference with English could be updated quickly in several hours. Typical attack methods contain Word Split (split a character into several characters), Homophony (different words or characters have the same PINYIN [4]) and Glyph Similar (obfuscated characters have similar structures with the original characters), etc.
- Differing from the traditional spam text (such as Gambling, Porn, Abuse, etc.) and Smishing (which mainly aims at collecting sensitive data from people), the domain-knowledge plays an important role in the identification of the specific type of fraud SMS. For example, there are lots of zeros in the Loan fraud SMS, but the SMS with many zeros could be the normal SMS of banking notice as well. The fraud domain-knowledge related studies are very limited in the previous work.

To tackle these challenging problems, we firstly conduct a measurement study to obtain a better understanding about the fraud SMS, then we design a novel two-stage algorithm called *TFC* to classify the Chinese fraud SMS. In the first stage, we utilize the key insights found in our measurement to quickly filter out the normal SMS. In the second stage, we exploit BERT-wwm [5] to obtain the semantic features, and thoughtfully designed 61-dim features base on the fraud domain-knowledge to capture the various attack methods and different kinds of abnormal SMS characteristics. By

conducting extensive experiments with real-world SMS dataset, we prove that *TFC* is able to achieve the best performance compared with five alternative algorithms, and the generated features are robust to many kinds of classification algorithms in the meanwhile.

The main contributions of this paper can be summarized in three aspects:

- We conduct a comprehensive analysis on the labeled telecom network fraud SMS and discover some interesting insights. To mitigate the problem of lack of Chinese fraud SMS datasets, we release part of our labeled dataset online [6] for research convenience.
- We propose a novel two-stage algorithm to address the fraud SMS classification problem. The algorithm can rapidly filter out normal SMS in Stage 1. After composing deep features and domain-knowledge related features in stage 2, five types of SMS (namely, normal SMS, Loan fraud SMS, Part-time fraud SMS, Investing fraud SMS and Gambling SMS) can be effectively classified with simple machine learning algorithm.
- We demonstrate our algorithm’s performance through real-world data driven experiments. Compared with various algorithms, our algorithm is able to obtain the highest F1-score and accuracy. Impacts of the Stage 1 and domain-knowledge keywords, and the performance of different ML classifiers with *TFC* are also included.

The rest of the paper is organized as follows. In Sect. 2, we review the previous work of SMS fraud and spam text related problems. In Sect. 3, we conduct a comprehensive measurement analysis with fraud SMS and obtain several interesting insights. We introduce our two-stage algorithm and detailed feature construction procedures in Sect. 4. Section 5 presents the results of experiments using real-world data. Section 6 concludes the paper.

2 Related Work

SMS fraud related problems have gained lots of attention for many years. *Joo et al.* [7] proposed an enhanced security model for detecting Smishing attack in smart devices. *Goel et al.* [8] designed a algorithm framework consist of three phases for Smishing attack detection. *Mishra et al.* [9] investigated Smishing detection problem through SMS Content, URL and Source Code, then a prototype of the proposed system with 96.29% overall accuracy is developed. *Pervaiz et al.* [10] investigate the scope and scale of the problem of SMS fraud in Pakistan, *Delany et al.* [11] and *Abdulhamid et al.* [12] reviewed and summarized the SMS spam filtering method. As for the fraud detection in the telecom area, previous works mainly use Call Detail Record (CDR) to identity the fraudulent user. For example, *Olszewski* [13] proposed an approach based on user profiling with Latent Dirichled Allocation (LDA), and detecting fraudulent behavior on the basis of threshold type classification with use of the KL-divergence. *Farvareh et al.* [14] employed a hybrid approach consisting of preprocessing, clustering and classification phases to identifying customers’ subscription fraud. The studies of SMS fraud from telecom operator’s perspective are very limited.

There are several attack methods in spam text, such as adversarial text, wrong-spell words etc. Another thread of research studied the related problem from different aspects.

Wang *et al.* [15] proposed an adversary-generation algorithm for increasing the variance of adversarial training data; Ebrahimi *et al.* [16] investigated the robustness of models with adversarial examples trained by the proposed two translation-specific types of attacks. Li *et al.* [17] presented an adversarial defense framework designed for Chinese based deep learning model, all the results of experiments show that these methods could improve model performance to defend against adversarial text attacks. Yeh *et al.* [18] presented a novel spelling error detection and correction method based on N-gram ranked inverted index; Xiong *et al.* [19] designed a unified framework HANSpeller for Chinese text spelling error detection and correction. Karan *et al.* [20] and Nobata *et al.* [21] investigated the abusive language detection problem with Cross-Domain method and NLP features based classification methodology separately. All these works focused on the characteristics of the spam text, and the differences between spam text and fraud SMS are not considered.

Differing from the previous work, our work focuses on two aspects. Firstly, we aim at addressing the fraud SMS classification problem from telecom operator's perspective, rather than identifying the fraud with CDR [13, 14] or detecting Smishing from the view of mobile devices [7–9]. Secondly, while previous works focus on English [15, 16] and Chinese [17–19] spam text related studies, our work studied the Chinese fraud SMS classification problem which can not be applied by many proposed methods in other languages due to the huge Chinese grammatical difference and the unique characteristics between fraud SMS and spam text.

3 Measurement Analysis

In this section, we mainly conduct the measurement analysis with fraud SMS dataset, which are reported by massive citizens and collected by the Guangzhou Anti-Fraud Center from 11/02/2020 to 11/11/2020, named **RD1**. Figure 2(a) plots the fraud SMS daily count over a week, and we can observe that the average fraud SMS count is larger in weekdays compared to the weekends. As shown in Fig. 2(b) (We use ‘#’ as the abbreviation of word ‘number’ in this paper), in terms of the fraud SMS types, Part-time fraud, Loan fraud and Investing fraud accounted for 85% of the whole **RD1**, and the Part-time fraud SMS has the highest proportion, nearly 50% .

Fig. 2. (a) Daily fraud SMS count of **RD1**. (b) # of fraud samples percentage.

Table 1. Examples of attack methods and corresponding # of *Uncommon Chinese Characters* and *Other Characters*

Attack Method	Example	# of the <i>Uncommon Chinese Characters</i>	# of the <i>Other Characters</i>
Word Split	加我微信→力口我微信	0	0
Homophony	微信→V信	0	1(V)
Glyph Similar	做事→做俸	1(俸)	0
Homophony+Glyph Similar	做事→作事→俸俸	2(俸,俸)	0
Homophony+Word Split	微信→V信→\./信	0	2(\,./)
Word Split+Homophony+Glyph Similar	加微做事→加V做事→+./俸俸	2(俸,俸)	3(+,./)

Before stepping into further analysis, we define five types of characters in the SMS as follows:

- *Common Chinese Characters*. Define the common words and sub-common words in the List of *Commonly Used Characters in Modern Chinese* [22] as *Common Chinese Characters*, and the dict size is 3500.
- *Uncommon Chinese Characters*. Define Chinese characters that not belong to the *Common Chinese Characters* as *Uncommon Chinese Characters*.
- *Numeric Characters*. Define the number characters from 0 to 9 as *Numeric Characters*.
- *English Characters*. Define the alphabet characters from a to z and their upper case A to Z as *English Characters*.
- *Other Characters*. Define the characters in SMS that do not falling into 4 types mentioned above are *Other Characters*, such as symbolic characters like \$! #, etc..

After we dive into the **RD1**, we obtained two interesting insights. One is that *Uncommon Chinese Characters* and *Other Characters* often appear in the fraud SMS. As shown in Fig. 3, to pass the keywords based spam detection methods and systems, the criminals tend to modify the SMS with several adversarial attacks, such as Word Split, Word Homophony, Glyph Similar, or combination of them. Therefore *Uncommon Chinese Characters* and *Other Characters* would have a considerable amount. Some examples of attack methods are presented in Table 1. Figure 4(a) plots the CDF (Cumulative Distribution Function) of the sum of *Uncommon Chinese Characters* and *Other Characters* in the fraud SMS. We can observe that only 22% of them have less than 10 *Uncommon Chinese Characters* and *Other Characters*. It even exceeds 30 in some fraud SMS samples, which proves that *Uncommon Chinese Characters* and *Other Characters* are frequently adopted in fraud SMS.

Fig. 3. Fraud SMS and Gambling SMS sample.

Fig. 4. (a) The CDF of sum of *Uncommon Chinese Characters* and *Other Characters*. (b) The CDF of SMS length in **RD1**.

Another insight is that there are often contained some contact information in the fraud SMS, which increases the count of *Numeric Characters* and *English Characters*. Table 2 shows that 99.41% of the fraud SMS have *Numeric Characters* and 69.54% of them have *English Characters*, it's reasonable that the proportion with *Numeric Characters* is much higher than the *English Characters*, because the fraud SMS often contains QQ or Wechat which represented by 5–11 digit numbers directly. From Fig. 4(b), we can observe that the minimal length of the fraud SMS in **RD1** is 17, and around 60% of the samples are longer than 50 characters. We argue that the length of the fraud SMS are commonly longer than normal one because of the expressions of contact information and the adversarial attacks (e.g., Word Split) in the fraud SMS.

Table 2. Statistics of *Numeric Characters* and *English Characters* in **RD1**.

Conditions	Percentage
# of <i>Numeric Characters</i> > 0	99.41%
# of <i>Numeric Characters</i> > 0 & # of <i>English Characters</i> > 0	69.02%
# of <i>English Characters</i> > 0	69.54%
# of <i>Numeric Characters</i> = 0 & # of <i>English Characters</i> = 0	0.03%

Given the above measurement results, the most intuitive idea is that we can design features based on the appearance of *Uncommon Chinese Characters* or *Other Characters*, *Numeric Characters* or *English Characters* to classify the fraud SMS. However, the authors found out that some other abnormal SMS have the same characteristics. Taking the Gambling SMS sample in Fig. 3 for example, the # of *Uncommon Chinese Characters* is 12, and the # of *Numeric Characters* and *English Characters* are also greater than 0. If only these two types of features are considered, too much Non-fraud SMS will be recalled in practice. What's more, we plan to classify fraud SMS in a fine-grained manner, i.e., classifying the Part-time fraud, Loan fraud and Investment fraud SMS separately. For this purpose, a much more accurate and robust algorithm needs to be developed.

4 Algorithm Design

4.1 Model Overview

The goal of fraud SMS classification problem is to design a classification algorithm $cls(\cdot)$ for any input text $x \in X$, $x = \{x_1, x_2, \dots, x_n\}$ and output its fraud category. In this paper, we propose an algorithm composed by two stages, which can be defined as:

$$cls(x) = cls_{ml}(V(cls_{s1}(x)); \theta) = \hat{y} \quad (1)$$

where Stage 1 $cls_{s1}(\cdot)$ is used to quickly filter out normal SMS, i.e. SMS not belongs to any fraud category. $cls_{ml}(V(\cdot); \theta)$ would be noted as stage 2, where $V(\cdot)$ is a feature extraction function that convert SMS into continuous or discrete features, $cls_{ml}(\cdot; \theta)$ is a machine learning classifier aims at obtaining the specific category of the SMS, and θ are the parameters of the classifier.

The procedures of our method are shown in Fig. 5. Concretely, Stage 1 is introduced Subsect. 4.2, and we presents the b) of Fig. 5 in the **Deep Model** of Subsect. 4.3. Figure 5's part c) and d) are explained in the **PINYIN and BISHUN** and **Hard Matching** of Subsect. 4.3 separately.

Fig. 5. Algorithm procedures.

4.2 Stage 1 - Normal SMS Filter

In Stage 1, we plan to use simple method to filter out massive normal SMS quickly while keeping the algorithm's recall at a high level. Based on the measurement analysis in Sect. 3, two indicators are considered as follows:

$$I_{nco}(x) = 1(c_{uc}(x) + c_{other}(x) > T_1) \quad (2)$$

$$I_{nen}(x) = 1(c_{num}(x) + c_{en}(x) > T_2) \quad (3)$$

where $c_{uc}(x)$ is the count of *Uncommon Chinese Characters* in SMS x , $c_{other}(x)$ denoted the count of *Other Characters* in x , $c_{num}(x)$ and $c_{en}(x)$ stand for the count of *Numeric Characters* and *English Characters* in x respectively, T_1 and T_2 are the threshold parameters. As mentioned in Sect. 3, differing from the normal SMS, the fraud SMS often uses obfuscated characters to avoid detection. Many of these obfuscated characters are homophony or glyph similar compared with original characters, and majority of them are *Uncommon Chinese Characters* and *Other Characters*. Given this context, we argue that when the sum of *Uncommon Chinese Characters* and *Other Characters* is bigger than 0, i.e. T_1 is 0, x could be a fraud SMS in a certain extent. On the other hand, criminals often need more contact channels to deliver information to victims other than SMS,

and the information of network communicating channels is frequently appeared in the SMS. So $I_{nen}(x)$ is introduced to catch these information indirectly, which has already been explained in Sect. 3. The final output of Stage 1 is given by Eq. 4

$$cls_{s1}(x) = \mathbf{x} \odot (I_{nco}(x) \wedge I_{nen}(x)) \quad (4)$$

i.e. only the SMS with *Uncommon Chinese Characters* or *Other Characters* appearance, and *Numeric Characters* or *English Characters* appearance simultaneously will be sent to the stage 2. For simplicity, we set T1 and T2 equal to 0, and the detailed analysis of threshold setting will be discussed in the ablation study.

4.3 Stage 2 - Fraud SMS Classification

In stage 2, we aim to classify the SMS provided by previous stage in a find-grained manner. Due to the impressive advantage of a huge corpus in text classification, we exploited a deep neural model trained from Chinese corpus to extract semantic information. Besides, for the reason that there are lots of attacked texts in the fraud SMS, which are able to fool the deep model trained by the normal SMS, we design an artificial extractor, which can match specific keywords defined by anti-fraud professionals. Finally, features from deep model and artificial extractor were combined and fed into a machine learning model, e.g. SVM (support vector machine) to obtain the final classification results. The entire progress of stage 2 can be expressed as:

$$V(x) = f_{deep}(x) \oplus f_{pb}(x, k_1) \oplus f_{hm}(x, k_2) \quad (5)$$

$$cls_{s2}(x) = cls_{ml}(V(x); \theta) \quad (6)$$

where cls_{ml} is a machine learning classifier, f_{deep} is deep model, f_{pb} is PINYIN and BISHUN similarity extractor, and f_{hm} is hard matching extractor. $\oplus$ represents feature combination. We define $f_{pb}(x, k_1) \oplus f_{hm}(x, k_2)$ as artificial feature because it depends on pre-defined keywords group k_1 and k_2 .

Deep Model. In the field of natural language processing, pre-trained deep language models have become a very important basic technology. We use the BERT-wwm, which released by Harbin Institute of Technology and IFlytek, for deep feature extraction [5]. BERT-wwm is one of the best open source deep models trained on Chinese wikipedia (including simplified Chinese and traditional Chinese) data and have better performance in formal text modeling. Specifically, we directly use the pre-trained model for deep feature extraction, i.e. taking the SMS as input, and use the corresponding output vector of the first symbol ('[CLS]') as the semantic representation.

PINYIN and BISHUN. Fraud SMS commonly use attacks against detection, and these transformed words could disturb the semantic meaning of sentences, affecting the accuracy of semantic features. Inspired by [23], we propose a strategy based on PINYIN [4] and BISHUN (stroke order) to recognize attack methods of some specific fraud keywords. Detailed procedures of our similarity judgement strategy are shown in Fig. 6.

Firstly, the algorithm will search the Word Split attack in the SMS, and make corresponding corrections if such obfuscated characters exists. Then the keywords generated by anti-fraud specialists will be searched in the SMS, and output a boolean vectors whether each keyword exists. We argue that two words can be regarded as same only when they are same or highly similar in some aspects. Concretely, two perspectives for judgement are considered in our strategy:

- *SAME PINYIN*. For homophony attack, one character would be transformed to others that has the same pronunciation. Since that Chinese characters can be pronounced base on their PINYIN, two words with the same PINYIN (without tones) could be regarded as the same word.
- *BISHUN similarity*. A Chinese character has five basic writing strokes, i.e. 横 (horizontal), 竖 (vertical), 撇 (left-falling), 捺 (right-falling) and 折 (turning). By citing each basic stroke as number 1–5 and arranging them in writing order, BISHUN (stroke order) can uniquely identify a number sequence of a Chinese character. For glyph similar attack, transformed character usually shares some similar structure with the original one. For this reason, two words with high BISHUN similarity would be regarded as the same word in our method. The BISHUN similarity is calculated by adding the length of left common substring and the length of right common substring, then dividing by the length of keyword character composition string.

To explain our strategy clearly, we construct an example to further illustrate the process. Given an attacked SMS ‘徽↑言’ and a keyword ‘微信’ (wechat). Our method will try to combine each two adjacent characters first. In this case ‘↑言’ would be corrected to ‘信’. Though the PINYIN of ‘徽’ (hui) is different to ‘微’ (wei), their BISHUN ‘徽’ (3322521554234134) and ‘微’ (3322521353134) maintain long left common substring (3322521) and right common substring (3134). We calculate the BISHUN similarity as Eq. 7:

$$(len(left) + len(right))/len(keyword) \approx 0.846 \quad (7)$$

Due to the reason of the high BISHUN similarity, the strategy would determine ‘徽↑言’ as an attack form of ‘微信’. Designed by anti-fraud specialists, we obtain 54 keywords in total from different fraud categories, therefore a 54-dim boolean features will be generated finally.

Hard Matching. Although PINYIN and BISHUN are able to catch the majority attack methods of words, there are still some attack methods that could bypass the detection, such as non-Chinese special characters. Due to the limited non-Chinese special characters which could be used for attack, we adopt a hard matching method to determine whether these keywords exist in the SMS. As shown in Table 3, we design 6 hard matching expressions to search pre-defined keywords. We also use the SMS length as an additional feature since the number of special characters is also related to the SMS length (which has been partially investigated in Fig. 4(b)), and thus obtain a 7-dim hard matching feature. Finally, these hard matching features are combined with PINYIN and BISHUN features, resulting in a 61-dim artificial feature.

Fig. 6. PINYIN and BISHUN similarity judgement procedures.

Table 3. Hard matching expressions.

Hard Matching expression	Keyword	Non-Chinese Attack Form	Design principle
Has Phone or QQ	5-11 digits	+Q, tel:	Criminals needs phone, QQ, wechat, website to further contact with victims
Has website	Website	.com, .net, .cc, ...	
Has Wechat	微信	+V, +\/, VX, ...	
# of 0	零	0,O,o,〇,°,o	Fraud SMS usually contains lots zeros to indicate the amount
# of 8	八	8,(8),°,s	Gambling SMS usually contains lots eights, which means ‘发’(lucky)
# of numeric numbers	1-9	①,②,③,④,⑤,⑥,⑦,⑧,⑨,x, X, ...	‘x’ or ‘X’ is used to replace numbers in some desensitization data

ML Classifier. After the deep features and artificial features extraction, each SMS is represented as continuous or discrete features. Any multi-class classifier can be deployed to learn feature distribution and give its prediction. In the experiments, we commonly use SVM in one-versus-all manner to train on SMS features with all hyper-parameter of the SVM set to default. The detailed analysis of ML classifier selection is also discussed in the ablation study.

5 Experiments

In this section, we conduct real-world data experiments to estimate the performance of our algorithm.

5.1 Dataset and Experiments Setting

Since we only have the fraud samples provided from citizens, the normal SMS and the gambling SMS samples need to be collected in another channel. We randomly sampled some normal SMS from [26] and [27] datasets and further gathered some gambling SMS from volunteers. Table 4 summarizes statistics of the two datasets explored in this paper, and we made **DATASET2** available [6] for research usage.

Table 4. Sample statistics of two datasets.

Dataset Name	Normal	Gambling	Fraud
DATASET1	31640	6793	10300
DATASET2	1317	646	517

As for the experiments with **DATASET1**, the fraud and gambling samples in training and testing set are separated based on the collected date, the normal samples are randomly distributed in the training and testing set. Due to the limited sample size of **DATASET2**, we choose the entire **DATASET1** as training set, and the whole **DATASET2** as testing set for **DATASET2** related experiments.

5.2 Comparison of Different Algorithms

For comparison with our proposed algorithm *TFC*, five strategies are considered as follows:

- *JWE* [24], in which exploiting stroke-level information for improving the learning of Chinese word embeddings. We get the sentence embedding vectors by averaging the word embedding vector obtained from the algorithm and Chinese word-split tools \cite{Jieba}.

- *cw2vec* [25], in which jointly embed Chinese words as well as their characters and fine-grained subcharacter components. The sentence embedding vectors are obtained with the same procedures in *JWE*.
- *BERT-www*, we feed the SMS into BERT-www model directly to get the deep features.
- *Stage 1 + BERT-www*, we utilize Stage 1 to filter some normal SMS, and feed the rest SMS into BERT-www model to get the deep features.
- *PyCor + BERT-www*, we use an open source Chinese text error correction tool Pycorrector [29] to correct the obfuscated characters in the SMS, and exploit BERT-www model to get deep features of the corrected SMS.

All the features calculated from *TFC* and the above five strategies will be sent to SVM with the same parameters to obtain the final classification result.

Table 5 presents the performance of different algorithms with **DATASET1** and **DATASET2**, among them our proposed algorithms (*TFC without Stage 1* and *TFC*) achieve best performance on both datasets. Particularly, we can observe that strategy with *BERT-www* has better performance compared with *JWE* and *cw2vec*. From the comparison of *TFC without Stage 1* and *PyCor + BERT-www*, we conclude that our artificial features outperform than *PyCor*, which is able to deal with the obfuscated characters more effectively. From the comparison of *TFC without Stage 1* and *TFC*, we can find that Stage 1 has negative effects on the overall performance. We argue that the result is reasonable, because the main purpose of Stage 1 is quickly filtering out the normal SMS, and it's inevitable that some fraud SMS samples may have 0 *Uncommon Chinese Characters* and *Other Characters*, so these samples will be misclassified as normal in Stage 1. Meanwhile, it's quite easy to identify the filtered normal SMS in stage 2, which further enhances the performance of *TFC without Stage 1*.

Table 5. Performance of different algorithm.

Algorithm	DATASET1		DATASET2	
	F1-score	Accuracy	F1-score	Accuracy
JWE	0.84	0.922	0.591	0.757
cw2vec	0.764	0.894	0.437	0.65
BERT-www	0.9942	0.982	0.833	0.906
Stage 1 + BERT-www	0.941	0.949	0.832	0.902
PyCor + BERT-www	0.94	0.982	0.835	0.906
TFC without Stage 1	0.952	0.986	0.851	0.923
TFC	0.948	0.956	0.843	0.915

Table 6 demonstrates the effectiveness of domain-knowledge related features. TFC uses three types of features, i.e. deep features from *BERT-www*, PB features from **PINYIN** and **BISHUN** extractor and HM features from **Hard Matching** principle. Deep features extracts semantic information of SMS and provides the basic classification performance. HM features and PB features further improve accuracy on **DATASET1** and

Table 6. Performance comparison of different domain-knowledge features.

Features	DATASET1		DATASET2	
	F1-score	Accuracy	F1-score	Accuracy
Deep	0.942	0.982	0.833	0.906
Deep + PB	0.942	0.982	0.840	0.910
Deep + HM	0.947	0.984	0.833	0.906
Deep + PB + HM	0.952	0.986	0.851	0.923

Table 7. Classification examples under different algorithms.

No	SMS Samples	Ground Truth	TFC	JWE	cw2vec	PyCor+ Bert
(a)	Fraud SMS Sample: ~由于您的购物记录良, 您来给我做事, 一日260-580请 +Q: 9213xxxx27 Non-Attack Sample: ~由于您的购物记录良, 您来给我做事, 一日260-580请 加Q: 9213xxxx27	Part-time Fraud	Part-time Fraud	Loan Fraud	Normal	Part-time Fraud
(b)	SMS Sample: 宝宝医疗费用预计xx万元	Normal	Normal	Loan Fraud	Loan Fraud	Normal
(c)	Fraud SMS Sample: 戒登宝, 做款 十扣 498xxx817秒下 Non-Attack Sample: 借货宝, 放款 加Q 498xxx817秒下	Loan Fraud	Loan Fraud	Gambling	Gambling	Gambling
(d)	Fraud SMS Sample: 嘿藉敢逃挂软件, 泊动零连盗刷, 技荒网催带每天赚50%, 溜 则砥滔汾苳. 乡稻康乡仔多仟赚多万溜寸核徐10-8888 +叩叩: 1060xxx102 Non-Attack Sample: 黑科技外挂软件, 自动下注盈利. 技术员私带每天赚50%, 前 期无需分红. 三百赚三千二千赚三万首充彩金10-8888 加QQ: 1060xxx102	Investing Fraud	Investing Fraud	Gambling	Gambling	Gambling

DATASET2, respectively. After combining all three types of features (i.e. *TFC without Stage 1*), our algorithm achieves better performance on both datasets.

Here we give a few examples to illustrate the effectiveness of *TFC*. As shown in Table 7, for example (a) and (b), only a few of characters in the sentence are changed, and the algorithm based on large corpus (like *TFC* and *pycorrector* + *BERT*) can effectively extract the semantic information of the sentence and obtain good performance. In example (c) and (d), almost each character in the sentence has been changed to its attacked form. It's challenging to correct them by the algorithm based on transition probability, which further effects the performance of deep features. The artificial features added by *TFC* can effectively identify the keywords in various attack methods, which could guide the classifier to make the correct choices.

5.3 Ablation Experiment

Recall that the main purpose of Stage 1 is filtering out normal SMS precisely and quickly, so we relabeled the Gambling and Fraud SMS as abnormal SMS. Therefore the output of Stage 1 can be viewed as a binary classification problem, i.e., classify normal and abnormal SMS. We studied the effects of T1 (the right hand side of Eq. 2) and T2 (the right hand side of Eq. 3) with **DATASET1** train and test set in Fig. 7 and Fig. 8. From Fig. 7 we can find out when T2 is 0, changing T1 from 0 to 6 has trivial effects on the F1-score and Accuracy, but the results dropped significantly if T1 is greater than 7. Figure 8 shows the similar results when T1 is 0, and we can conclude that the reasonable value

range for T2 is [0, 8]. Figure 9 plots the Stage 1's effect on total running time of *TFC* and # of samples. We can conclude that the total running time could be decreased 56.5% by simply setting T1 and T2 to be 0, which is quite important when the algorithm runs in the huge datasets. And the # of samples decreased 61.5% as well. The above results prove that the Stage 1 is able to effectively reduce the size of dataset while classifying the normal and abnormal SMS with satisfactory performance.

Given lots of domain-knowledge keywords are adopted in Stage 2, we also investigate their corresponding effects with **DATASET1**. We can change the problem into a binary classification problem for each fraud type, for example, Non Loan fraud SMS v.s Loan fraud SMS. Table 7 presents the results of binary classification with the impacts of the domain-knowledge keywords, we can observe that three types of fraud SMS have better performance when they exploit the related keywords. To explore the reason that both F1-score and accuracy are lower after utilizing the keywords for Part-time Fraud, the authors plot the percentage of samples that each fraud type's keywords appeared in the test set in Fig. 10. We can observe that the Part-time fraud keywords are appeared in 98.6% Part-time fraud SMS, but it also shown in 92.4% Investing fraud SMS and 75.5% Gambling SMS respectively, which means the uniqueness of the keywords is limited. As for the Investing fraud keywords, since the keywords appearance percentage in the Investing fraud samples is relatively higher, which helps the algorithm achieve bigger improving in both F1-score and accuracy (shown in Table 8).

Fig. 7. Accuracy and F1-score under different T1 (T2 = 0).

Fig. 8. Accuracy and F1-score under different T2 (T1 = 0).

Fig. 9. Accuracy and F1-score under different T2 (T1 = 0).

Table 8. Impacts of the domain-knowledge keywords.

Fraud types	With keywords		Without keywords	
	F1-score	Accuracy	F1-score	Accuracy
Loan fraud	0.970	0.988	0.949	0.980
Part-time fraud	0.955	0.963	0.969	0.974
Investing fraud	0.910	0.967	0.885	0.958
Gambling	0.974	0.975	0.971	0.972

We modify the final classification algorithm from SVM to RF (Random Forest), MLP (Multi-Layer Perceptron) and KNN (K-Nearest Neighbor) to study the different types algorithm’s performance, parameters of all the algorithms are used in default setting with scikit-learn [30], and Table 9 demonstrates that the features generated by *TFC* are robust to many algorithms, and the performances are quite close compared with each other. Researchers can apply advanced or parallel algorithms to further improve the performance or speed up the running time.

Fig. 10. The percentage of samples for each fraud type keywords.

Table 9. Performance of TFC with different ML algorithm.

ML algorithm	DATASET1		DATASET2	
	F1-score	Accuracy	F1-score	Accuracy
SVM	0.948	0.956	0.915	0.843
RF	0.931	0.937	0.877	0.779
MLP	0.933	0.944	0.879	0.807
KNN	0.939	0.948	0.773	0.693

6 Conclusion

This paper mainly investigates the fraud SMS classification problem. We conduct a comprehensive measurement study of the fraud SMS, and observe some interesting phenomena. In addition, we propose a two-stage algorithm called *TFC*, which is able to quickly filter out normal SMS in the first stage, and classifying fraud SMS precisely in the second stage with the semantic deep features and the domain-knowledge based artificial features. By comparing with other algorithms using real-world dataset, we demonstrate that our algorithm *TFC* achieves better performance on both F1-score and accuracy. In the future, we plan to study how to extract the domain-knowledge keywords precisely and automatically.

Funding. This work was supported by the Science and Technology Program of Guangzhou under Grant 2019030005.

References

- Supreme People’s Procuratorate sets up a steering group to crackdown the rapidly spread cybercrime. http://www.xinhuanet.com/legal/2020-04/08/c_1125829699.htm. Accessed 07 Apr 2021
- The survey report shows that three telecom operators have achieved nearly 100% scam SMS processing in September 2020. https://www.sohu.com/a/430647655_287480. Accessed 07 Apr 2021
- China public security department cracked 256,000 cases of telecommunications network fraud in 2020. http://www.gov.cn/xinwen/2021-01/02/content_5576223.htm. Accessed 07 Apr 2021
- Wikipedia, PINYIN. <https://en.wikipedia.org/wiki/Pinyin>. Accessed 07 Apr 2021
- Cui, Y., et al.: Revisiting pre-trained models for Chinese natural language processing. In: Proceedings of Conference on Empirical Methods in Natural Language Processing: Findings (2020)
- DATASET Repo. <https://github.com/leegdcert/TFCDATASET>. Accessed 07 Apr 2021
- Joo, J.W., Moon, S.Y., Singh, S., Park, J.H.: S-Detector: an enhanced security model for detecting smishing attack for mobile computing. *Telecommun. Syst.* **66**(1), 29–38 (2017). <https://doi.org/10.1007/s11235-016-0269-9>

8. Goel, D., Jain, A.: Smishing-classifier: a novel framework for detection of smishing attack in mobile environment. In: Bhattacharyya, P., Sastry, H.G., Marriboyina, V., Sharma, R. (eds.) *Smart and Innovative Trends in Next Generation Computing Technologies*, pp. 502–512. Springer, Singapore (2018). https://doi.org/10.1007/978-981-10-8660-1_38
9. Mishra, S., Soni, D.: Smishing detector: a security model to detect smishing through SMS content analysis and URL behavior analysis. *Futur. Gener. Comput. Syst.* **108**, 803–815 (2020)
10. Pervaiz, F., et al.: An assessment of SMS fraud in Pakistan. In: *Proceedings of 2nd ACM SIGCAS Conference on Computing and Sustainable Societies* (2019)
11. Delany, S., Buckley, M., Greene, D.: SMS spam filtering: methods and data. *Expert Syst. Appl.* **39**(10), 9899–9908 (2012)
12. Abdulhamid, S., et al.: A review on mobile SMS spam filtering techniques. *IEEE Access* **5**, 15650–15666 (2017)
13. Olszewski, D.: A probabilistic approach to fraud detection in telecommunications. *Knowl.-Based Syst.* **26**, 246–258 (2012)
14. Farvaresh, H., Sepehri, M.: A data mining framework for detecting subscription fraud in telecommunication. *Eng. Appl. Artif. Intell.* **24**(1), 182–194 (2011)
15. Wang, Y., Bansal, M.: Robust machine comprehension models via adversarial training. In: *Proceedings of NAACL-HLT* (2018)
16. Ebrahimi, J., Lowd, D., Dou, D.: On adversarial examples for character-level neural machine translation. In: *Proceedings of 27th International Conference on Computational Linguistics* (2018)
17. Li, J., Ji, S., Du, T., Li, B., Wang, T.: TEXTBUGGER: generating adversarial text against real-world applications. In: *Proceedings of Network and Distributed Systems Security (NDSS) Symposium* (2019)
18. Yeh, J., Li, S., Wu, M., Chen, W., Su, M.: Chinese word spelling correction based on N-gram ranked inverted index list. In: *Proceedings of Seventh SIGHAN Workshop on Chinese Language Processing* (2013)
19. Xiong, J., Zhang, Q., Zhang, S., Hou, J., Cheng, X.: HANSpeller: a unified framework for Chinese spelling correction. *Comput. Linguist. Chin. Lang. Process.* **20**(1), 1–22 (2015)
20. Karan, M., Snajder, J.: Cross-domain detection of abusive language online. In: *Proceedings of 2nd Workshop on Abusive Language Online* (2018)
21. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y.: Abusive language detection in online user content. In: *25th International Conference on World Wide Web* (2016)
22. Wikipedia, Commonly used words in Modern Chinese. <https://zh.wikipedia.org/wiki/现代汉语常用字表>. Accessed 07 Apr 2021
23. Li, J., et al.: TEXTSHIELD: robust text classification based on multimodal embedding and neural machine translation. In: *Proceedings of 29th USENIX Security Symposium* (2020)
24. Yu, J., Jian, X., Xin, H., Song, Y.: Joint embeddings of Chinese words, characters, and fine-grained subcharacter components. In: *Proceedings of Conference on Empirical Methods in Natural Language Processing* (2017)
25. Cao, S., Lu, W., Zhou, J., Li, X.: cw2vec: learning Chinese word embeddings with stroke n-gram information. In: *Proceedings of AAAI* (2018)
26. Chen, T., Kan, M.: Creating a live, public short message service corpus: the NUS SMS corpus. *Lang. Resour. Eval.* **47**(2), 299–335 (2013). <https://doi.org/10.1007/s10579-012-9197-9>
27. SpamMessage. <https://github.com/hrwhisper/SpamMessage>. Accessed 07 Apr 2021
28. Jieba. <https://github.com/fxsjy/jieba>. Accessed 07 Apr 2021
29. Pycorrector. <https://github.com/shibing624/pycorrector>. Accessed 07 Apr 2021
30. Scikit-learn. <https://sklearn.org>. Accessed 07 Apr 2021

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.