Codewords Detection in Microblogs Focusing on Differences in Word Use Between Two Corpora

Takuro HADA
Graduate School of Informatics and Engineering
The University of Electro-Communications
Tokyo, Japan
hada.takuro@ohsuga.lab.uec.ac.jp

Yuichi SEI
Graduate School of Informatics and Engineering
The University of Electro-Communications
Tokyo, Japan
seiuny@uec.ac.jp

Yasuyuki TAHARA
Graduate School of Informatics and Engineering
The University of Electro-Communications
Tokyo, Japan
tahara@uec.ac.jp

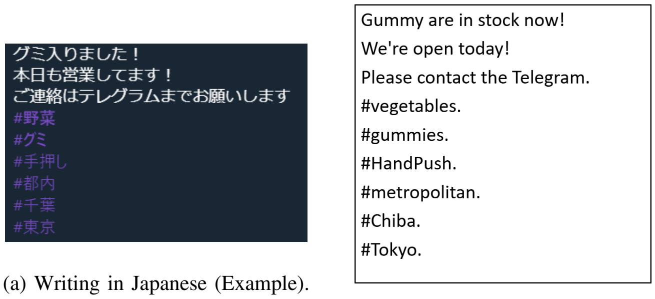
Akihiko OHSUGA
Graduate School of Informatics and Engineering
The University of Electro-Communications
Tokyo, Japan
ohsuga@uec.ac.jp

***Abstract*—In recent years, drug trafficking using microblogs has risen and become a social problem. A common method of cyber patrols for cracking down on crimes, such as drug trafficking, involves searching for crime-related keywords. However, criminals who post crime-inducing messages make maximum use of ”codewords” rather than keywords, such as enjo kosai, marijuana, and methamphetamine, to camouflage their criminal intentions. Research suggests that these codewords change once they become popular; therefore, searching for a specific word requires significant effort to keep track of the latest codewords. In this study, we focused on the appearance of codewords and those likely to be included in incriminating posts with aim to detect codewords with the high likelihood of inclusion in incriminating posts. We proposed new methods for detecting codewords based on differences in word usage and conducted experiments on concealed-word detection in order to evaluate method effectiveness. The results showed that the proposed method was capable of detecting concealed words other than those in the initial list and to better degree relative to baseline methods. These findings demonstrated the ability of the proposed method to rapidly and automatically detect codewords that change over time and blog posts that induce crimes, thereby potentially reducing the burden of continuous monitoring of codewords.**

***Index Terms*—codewords, detect, microblog, Twitter, word embedding**

I. INTRODUCTION

There have recently been numerous incidents related to enjo kosai (subsidized companionship) and illegal drugs promoted using microblogging. According to a previous study, enjo kosai ”A is the label used for young women who agree to meet strange men for dates, which might involve sex, in exchange for money or gifts” [1]. In particular, enjo kosai for young women under the age of 18 years of age has become a problem. Posters of enjo kosai and illegal-drug-related activity are wary of their posts and accounts being removed by cyber patrols from social networking service companies or their arrest by police. Therefore, only those aware of the meaning of codewords carry out illegal transactions (Fig.1).



グミ入りました！
本日も営業してます！
ご連絡はテレグラムまでお願いします
#野菜
#グミ
#手押し
#都内
#千葉
#東京

(a) Writing in Japanese (Example).

Gummy are in stock now!
We're open today!
Please contact the Telegram.
#vegetables.
#gummies.
#HandPush.
#metropolitan.
#Chiba.
#Tokyo.

(b) Translation of (a).

Fig. 1. Example sentences with codewords from Twitter.

For example, the word ”ganja” is popularly codeword used for marijuana, while the words ”es” and ”shabu” are mainly used for methamphetamine. Generation of keyword lists and use of countermeasures for their detection have limited success due to the frequency with which they are changed to avoid surveillance [2]. For example, in the case of marijuana, the words ”grass”, ”weed”, and ”joint” have previously been used. Similarly, for methamphetamine, the words ”ice” and ”crystal” have previously been used. Thus, cyber patrols need to continuously track new cryptograms and the possible addition of words to these cryptograms; thereby increasing the monitoring burden. Therefore, to support the prevention of crimes, such as drug trafficking and enjo kosai on microblogging sites (especially Twitter), we developed a method to detect crime-related tweets containing codewords. Previous studies of codeword use on the Internet, such as in bulletin boards, have been published; however, there are few studies targeting short sentences, such as microblogs, involving a limited number of characters in a single post, which makes it difficult to understand the meaning of sentences. This suggests that discovery of codewords in such sentences would be potentially

978-1-7281-6330-7/20/$31.00 ©2020 IEEE.

significant for crime prevention due to early detection. In this study, we focused on differences in usage of the same words between two corpora based on the idea that similar words appear interspersed with others in malicious communications. This method allows the detection of codewords likely included in crime-related tweets and among words likely to occur along with the codewords. This method builds a word distribution expression model using Word2vec [3] for each of the two corpora and detects codewords based on differences between words appearing higher in cosine similarity with respect to similar words.

II. Background

A. Increase in the number of crimes involving drug trafficking and enjo kosai

A news article based on a United Nations Office on Drugs and Crime report noted increases in online drug trafficking via Facebook, Twitter, Instagram [4]. Additionally, according to National Police Agency (Japan) data on enjo kosai, the number of smartphone victims of crimes and the number of children who fall victim to crimes originating from social networking sites (SNSs) increases annually, with the latter reaching an all-time high in 2019 (Fig.2). Moreover, among SNSs, Facebook and Twitter are often used [5],and Twitter reportedly has highest number of children victims (40%) (Fig.3). Therefore, in this study, we focused on Twitter.

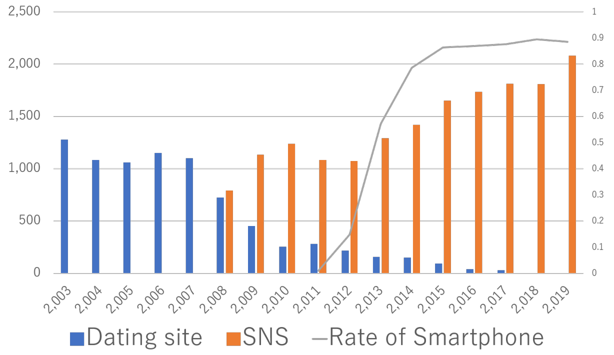


Fig. 2. Number of children victimized by SNSs according to data from the National Police Agency (JAPAN) [6] [7] and an undisclosed dating site (2018 and 2019)

B. Codewords

Codewords are used in various industries and expressed using various words. In this study, we defined codewords as those used in illegal transactions in order to avoid police surveillance. Different words with similar meanings can be used as codewords. For example, Yuan et al. [2] described a codeword as ”Dark Jargon”. In this study, we targeted the following codeword types.

1) Camouflaged words
These words are used to camouflage with an innocuous word that generally has a different meaning. For example, in the genre related to drug trafficking, codewords, such as ”vegetable” and ”grass”, are used for marijuana and ”ice” and ”crystal” for methamphetamine.

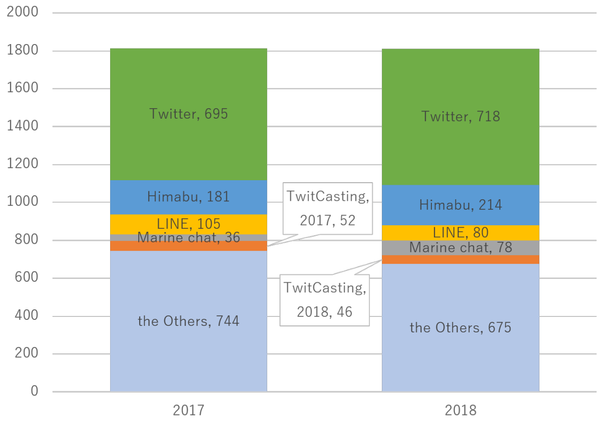


Fig. 3. Sites with multiple child victims according to data from the National Police Agency (Japan)) [8]

2) Words replaced by pronunciation
These include words replaced with other Chinese characters having the same sound when pronounced. Although the words are not generally used, they can be associated with the sound.
3) Words not commonly used
These words are unpopular, and even when used, is only understood by a limited number of people. For example, ”White Kush” and ”White Widow”, which are codewords for marijuana, fall into this category.

In this study, we focused on codewords related to drug trafficking and enjo kosai. Although such codewords from websites have been previously analyzed [9], they are not applicable to microblogging sites, such as Twitter. De la Rosa et al [10] described the following three features of microblogs.

- Short character lengths
  Microblogs comprise as little as a single word and at most less than a paragraph. In the case of Twitter, there is a limit of 140 characters per post.
- Informal and unstructured formats
  Microblogs contain slang, misspellings, and abbreviations.
- The text is semi-structured according to traditional NLP definitions

Therefore, it is difficult to maintain up-to-date dictionaries using prepared matching methods, given the frequency of codeword changes and accounting for slang use and misspellings. Furthermore, machine learning methods for detecting codewords are difficult to apply due to the limited length of the sentences and word presentation, which eliminates context. However, analysis of tweets in which codewords appeared indicated multiple cases involving the appearance of similar words, suggesting that codewords might be detected if a word-dispersion representation can be used to vectorize the words and identify similar words in their vicinity. In this study, we proposed a new method for detecting unknown words by focusing on the similarity of known words.

III. Related Works

Previous studies have addressed the increased number of crimes instigated by Twitter in an effort to reduce this number [11], [12]. These studies included detection of offensive and malicious words [13], [14], [15]. In criminal exchanges, codewords are sometimes used for transactions and cleverly hidden among common words in order to avoid their detection. Previous studies have been undertaken to detect such codewords. For example, on the dark web, cannabis is exchanged using the names ”popcorn” and ”blueberries”,’ and child pornography is referred to as ”cheese pizza”. Yuan et al. [2] proposed a method for automatically identifying jargon from the dark web. Because it is impossible to identify such jargon in a single corpus using Word2vec, they prepared multiple corpora and detected jargon according to semantic contradiction of terms appearing in two different corpora. However, this research does not cover Twitter with short sentences, slangs, and misspellings. Zhao et al. [16] focused on jargon used in cybercrime associated with the underground market in China and used unsupervised learning for its detection. They concluded that ”CBOW+NS” was the optimal setting for Word2vec, resulting in 20% higher rates of detection relative to an ”LDA” approach. However, these methods represent first-stage research [2]. Furthermore, Aoki et al. [17] detected nonstandard word usage involving definitions that differed from their original meaning. These words were not limited to use in crime-related contexts, and it is conceivable that crime-related codewords employ other methods for concealing the intent of a given communication.

IV. Approach

A. Core idea

Those who post crime-related codewords to camouflage their criminal intent tend to be clever; however, there are few differences in the contextual nature of back-and-forth exchanges. We hypothesized that words used in illegal negotiations are used in the same sense as their analogs. Therefore, we speculated that unknown codewords might appear as words similar to known codewords in a codeword corpus. Using data obtained from Twitter, we prepared a set of tweets focused on illegal trading purposes and divided them into two groups of tweets used for malicious purposes: the Bad Corpus is a collection of tweets containing one or more words in the word list, and the Good Corpus is a collection of all tweets not included in the Bad Corpus. We then performed word distribution analysis on each corpus [3] and calculated the cosine similarity using gensim [18]. We defined a word with a high cosine similarity as *W* and referred to a set of such words as ”Similar words” of *W*. For example, the word ”paper” is a codeword for ”LSD” as a type of methamphetamine. Similar words in each corpus are shown in Table I. Table 1 shows that for ”paper”, detection of similar words in the two corpora resulted in different results. Moreover, we found that six of the top 10 similar words in the Bad Corpus were codewords. To discover unknown codewords, we focused mainly on two points: 1) similar words, *W*, in the Good Corpus differ from those in the Bad Corpus; and 2) searches for similar words in the Bad Corpus result in similar metonymy and related maliciousness. We selected a list of codewords related to drug trafficking and enjo kosai.

TABLE I. Top 10 similar words to ”paper” in each corpus. (Bolded words are those defined as codewords.)

Rank	Good Corpus		Bad Corpus	
1	字詰め	Letter-writing	業販	Commercial Sales
2	試筆	Test Brush	市内	Within The City
3	便箋	Letterhead	営業中	Open (for business)
4	裏紙	Backing paper	メニュー	Menu
5	ハードカバー	Hardcover	**スカンク**	**Skunk**
6	アルシュ	Archetypal	**リキッド**	**Liquid**
7	用紙	Paper	**ノーザン**	**Northern**
8	断裁	Cutting	**グミ**	**Gummi**
9	模造紙	Imitation paper	**ハイレギュラー**	**High Regular**
10	方眼(紙)	Graph Paper	**ヘイズ**	**Hayes**

B. Procedure

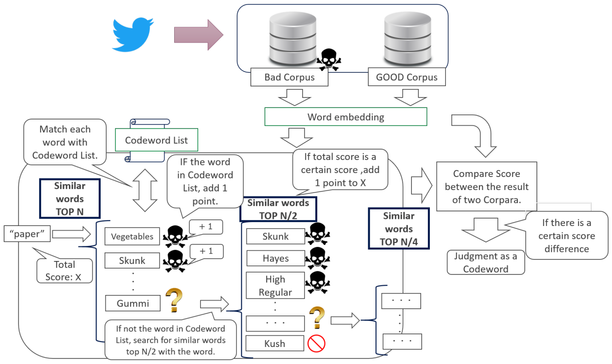

Bad Corpus
GOOD Corpus
Word embedding
Match each word with Codeword List.
Codeword List
IF the word in Codeword List, add 1 point.
If total score is a certain score ,add 1 point to X
Compare Score between the result of two Corpara.
Similar words TOP N
“paper”
Total Score: X
Vegetables
Skunk
Gummi
+ 1
+ 1
If not the word in Codeword List, search for similar words top N/2 with the word.
Similar words TOP N/2
Skunk
Hayes
High Regular
. . .
Kush
Similar words TOP N/4
. . .
. . .
If there is a certain score difference
Judgment as a Codeword

Fig. 4. Schematic of the system.

The outline of the system is shown in Fig. 4. The detailed flow of the method is illustrated in Algorithms 1 and 2.

1) For each word in the word list, calculate the score for each of the two corpora (Function *SIMILAR*).
   a) For each word, similar words up to the top *N* of the cosine similarity are searched using the pre-constructed word distribution expression model (Good_Corpus, Bad_Corpus) (*Get similar words*).In this experiment, we set *N* to 20.
   b) Match the retrieved *N* similar words individually against the matching list(Codeword_List).
   c) If a match is found in the codeword list, add points (X=X+1) to a maximum of 20 points.
   d) If the word, *W*, does not match any in the codeword list, it is possible that the word is not registered as a codeword; therefore, similar words up to the *N*/2 highest cosine similarity to the word *W* are

considered. If the score is greater than or equal to the threshold, points are added to *X* (*X*=*X*+1).

   e) For each of similar word to *W*, if the word does not match any in the hidden word list, *N*/4 similar words are searched based on the similar word, and word *W* is evaluated.

2) Calculate the difference (*Diff*) between the calculated Good (*Cnt_Good*) and Bad (*Cnt_Bad*) point totals.
3) If the Bad point total exceeds the threshold, it is identified as a codeword. If the *Diff* is above a certain level, the threshold value for the Bad point total decreases.

Algorithm 1 $Main$

```
Input: Word_List,N,Good_Corpus,Bad_Corpus
Output: Codewords
  for all $Word$ in $Word\_List$ do
    $Cnt\_Bad \Leftarrow SIMILAR(Word, N, Bad\_Copus, 1)$
    $Cnt\_Good \Leftarrow SIMILAR(Word, N, Good\_Copus, 1)$
    $Diff \leftarrow abs(Cnt\_Bad - Cnt\_Good)$
    if $(Cnt\_Bad/N >= 0.2)$ or $((Diff/N >= 0.15)$ and $(Cnt\_Bad/N >= 0.1))$ then
      $Codeword\_List.append(WORD)$
    end if
  end for
  $return(Codeword\_List)$
```

Algorithm 2 Function $SIMILAR$

```
Input: Word,N,Corpus,Loop_count
Output: Number of matches with codewords
  $X \Leftarrow 0$
  $Sim\_words \Leftarrow Corpus.Get\_similar\_words(Word, N)$
  for all $Sim\_word$ in $Sim\_words$ do
    if $Sim\_word$ in $Codeword\_List$ then
      $X \Leftarrow X + 1$
    else if $Loop\_count <= 2$ then
      $Y \Leftarrow SIMILAR(Sim\_word, N/2, Corpus, Loop\_count + 1)$
      if $Y/N >= 0.2$ then
        $X \Leftarrow X + 1$
      end if
    end if
  end for
  $return(X)$
```

V. Experiment

A. Summary of the experiment

We performed an experiment to detect codewords using 950 pre-annotated words and included 10 of 45 known codewords among the 950 words in a known-codeword list for matching. The system then evaluated the similarity of words to codewords in an attempt to identify the remaining 35 words. The experiment was performed using the following steps.

B. Experiment process

1) Data collection: Twitter data (47 days; 5.4 GB) were collected using the Twitter API, and only text was used for analysis. The following words not considered to be related to the pre-processed codewords were removed prior to analysis.

- Single-byte alphanumeric characters
- URLs
- Full-width symbols
- Line-feed characters
- Words frequently appearing on Twitter (e.g., ”RT”, ”Favorite”, etc.)

2) Creating corpora: The pre-processed Twitter data were analyzed to identify 10 words (words judged as having been posted for criminal purposes related to drug trafficking and/or enjo kosai) in each tweet, followed by classification into the following two corpora.

- Bad Corpus (8 MB)
  This represented a group of tweets containing one or more of the 10 words. We assumed that words from tweets related to illegal transactions were collected in this corpus.
- Good Corpus (4 GB)
  This represented a group of tweets not including words from the Bad Corpus. We assumed that most of these tweets were general interactions.

3) Morphological analysis: We focused on Twitter due to its use of short sentences, new words and slang, and limited sentence lengths. This suggested that some sentences might not be correctly separated. Moreover, the Japanese language has a unique sentence structure not separated by spaces; therefore, segmentation was necessary prior to word distribution. We segmented sentences from Twitter using Sudachi [19] for the following reasons:

- Sudachi is continuously improved and maintained, its dictionary is regularly updated, and it is expected to have the most up-to-date word list; and
- a word-division unit can be selected.

4) Word embedding: After split writing, word distribution was performed using Word2vec using parameters shown in Table II.

TABLE II. parameters of Word2vec

Parameter	***Value***
Size	200
min_count	20
window size	5
SG or CBow	Skip-Gram [20]

5) System execution: We created a word list from the corpus model generated by word distribution and extracted 940 nouns common between the two corpora models and 10 nouns from the Bad Corpus model. We then executed method using this word list.

C. Comparative Approach

To perform comparative analyses, we prepared a baseline method in which all nouns in tweets containing words used in a malicious exchange were defined as codewords. The results of comparison of the proposed method with the baseline method are shown in Fig. 5.

The baseline method for detecting codewords is described as follows.

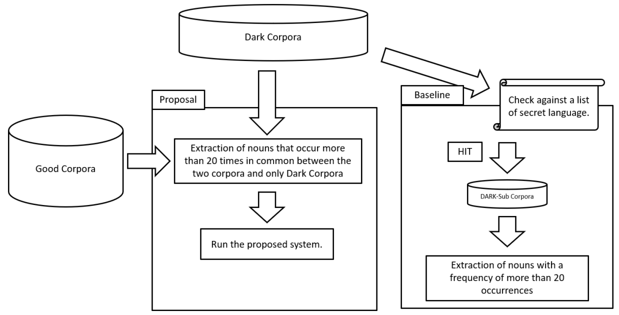

Dark Corpora
Proposal
Baseline
Good Corpora
Extraction of nouns that occur more than 20 times in common between the two corpora and only Dark Corpora
Run the proposed system.
Check against a list of secret language.
HIT
DARK-Sub Corpora
Extraction of nouns with a frequency of more than 20 occurrences

Fig. 5. Relationship between the proposed and baseline methods

- the Bad Corpus is used to analyse the codeword list used in the proposed method.
- All sentences containing words from the hidden word list were extracted, and a Bad subcorpus was created.
- from the Bad subcorpus, only nouns were extracted, with these considered hidden words.

D. Evaluation methods

The word list extracted from the corpora models comprised 950 nouns and was classified into the following three categories based on the tweets of two to three people with no prior knowledge of the words.

- “Codewords”
  These were words judged to have a meaning different from their original meaning.
- “Related words”
  Although these words could not be categorized as codewords, they tended to appear alongside codewords and were judged as rarely appearing in general tweets (e.g., ”stock” and ”price”).
- “Unrelated words”
  These were words not meeting the criteria of the previous two definitions.

E. Result

Annotation resulted in 45 codewords, of which 10 were prepared as a known-codeword list. The system was executed using 940 words that excluded these 10 words. Thirty-nine words were detected as predicted codewords, with 17 of these identified as true codewords. The results of the proposed method and the baseline method are shown in Table III. Table III, shows that the proposed method detected codewords at a higher rate than the baseline method; however, the number of codewords detected by the proposed method was lower than that by the baseline method. Furthermore, we determined four indicators of method performance (precision, recall, accuracy, and F-Score) (Table IV).

TABLE III. Evaluation results of evaluation

	All Words		Proposed Method		Baseline Method	
Clasified	Quantity	Rate	Quantity	Rate	Quantity	Rate
Codewords	35	3.7%	17	43.6%	23	5.7%
Others	905	96.3%	22	56.4%	379	94.3%
SUM	940		39		402	

TABLE IV. Details of the results

Evaluation Method	*Proposed*	*Baseline*	*Difference*
Precision	0.436	0.057	0.379
Recall	0.486	0.657	-0.171
Accuracy	0.957	0.584	0.373
F-score	0.459	0.105	0.354

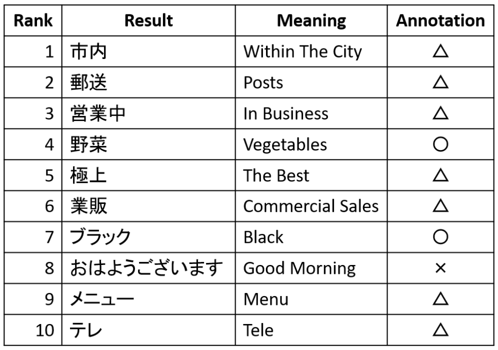
Table IV shows that the proposed method returned better results in terms of precision, accuracy, and F-Score relative to the baseline method. Moreover, the proposed method detected words, such as ”diesel”, ”skunk”, ”gummi”, ”lemon”, and ”joint”.

VI. Discussion

A. Detection Challenges

In this study, we developed a method to detect known codewords from short sentences, such as those used in tweets on Twitter. Although we anticipated lower recall in the proposed method relative to that in the baseline method, which uses a wide range of codewords, we observed minimal differences between the two results, and the proposed method showed higher precision, accuracy, and F-Score than the baseline method. These results indicated that the proposed method was able to detect codewords with higher accuracy than the baseline method. However, the proposed method was unable to detect ”typical” codewords, such as ”ice”’ and ”vegetable”. Therefore, we identified similar words to ”ice” in the word-distributed expression model created from the Bad Corpus (Table V).

TABLE V. Words similar to ”ice”.



Rank	Result	Meaning	Annotation
1	市内	Within The City	△
2	郵送	Posts	△
3	営業中	In Business	△
4	野菜	Vegetables	○
5	極上	The Best	△
6	業販	Commercial Sales	△
7	ブラック	Black	○
8	おはようございます	Good Morning	×
9	メニュー	Menu	△
10	テレ	Tele	△

Table V shows that multiple words used for malicious communications were identified as being similar to ”ice”. However, the number of matching words was small, because the number of words in the known-codeword list was too small. Therefore, using more words in the known-codeword list might improve the recall of the proposed method. Further, it was found that many related words defined in Section IV.D such as “Post” and “In Business” also appeared. Therefore, it can also be expected that the recall can be improved by introducing a mechanism for matching related words.

B. Detection of Related Words

Table VI shows results from including both codewords and related words. Table VII shows that the precision for the proposed method was high (0.718), and that multiple related words were identified, even if they were not codewords. Therefore, our future work will introduce a mechanism capable of detecting codewords and related words.

TABLE VI. Results evaluations involving related words.

	All Words		Proposed Method		
Clasified	Quantity	Rate	Quantity	Rate	DIFF
Codewords	35	3.7%	17	43.6%	+39.9%
Related word	119	12.7%	11	28.2%	+15.5%
Unrelated	786	83.6%	11	28.2%	-55.4%
SUM	940		39		

TABLE VII. Results generated by the inclusion of related words.

Evaluation Method	***Proposed***
Precision	0.718
Recall	0.182
Accuracy	0.854
F-score	0.290

C. *Applicability of the model to other languages*

Although we used the proposed method to analyse sentences written in the Japanese language, the method is versatile enough to apply to other languages.

VII. Conclusion

In summary, we developed a method to support cyber patrol detection of codewords in Twitter data. The method successfully detected codewords not included in the known-codeword list and outperformed a baseline method in terms of precision, accuracy, and F-Score. These findings suggest the efficacy of this method for automatic detection of frequently changing codewords.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Numbers JP17H04705, JP18H03229, JP18H03340, JP18K19835, JP19K12107, JP19H04113, and JST, PRESTO Grant Number JPMJPR1934.

References

[1] L. Miller, "Those naughty teenage girls: Japanese kogals, slang, and media assessments," 2004.

[2] K. Yuan, H. Lu, X. Liao, and X. Wang, "Reading thieves' cant: Automatically identifying and understanding dark jargons from cybercrime marketplaces," in *27th USENIX Security Symposium (USENIX Security 18)*. Baltimore, MD: USENIX Association, Aug. 2018, pp. 1027–1041.

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2013.

[4] "Asia-Pacific drug trade thrives amid the COVID-19 pandemic," https://www.reuters.com/article/us-asia-crime-drugs/asia-pacific-drug-trade-thrives-amid-the-covid-19-pandemic-idUSKBN22R0E0.

[5] J. Mangnejo, A. Khuhawar, M. Kartio, and S. Soomro, "Inherent flaws in login systems of facebook and twitter with mobile numbers," *Annals of Emerging Technologies in Computing*, vol. 2, pp. 53–61, 10 2018.

[6] "Juvenile delinquency, child abuse and sexual assault of children in 2019," https://www.npa.go.jp/safetylife/syonen/hikou_gyakutai_sakusyu/R1.pdf.

[7] "Current Situation and Measures for Children Victimized by SNS," https://www8.cao.go.jp/youth/kankyou/internet_torikumi/kentokai/40/pdf/s4.pdf.

[8] "Status of Children Victimized by SNS in 2018," https://www8.cao.go.jp/youth/kankyou/internet_torikumi/kentokai/41/pdf/s4-b.pdf.

[9] W. Lee, S. S. Lee, S. Chung, and D. An, "Harmful contents classification using the harmful word filtering and svm," in *Computational Science – ICCS 2007*, Y. Shi, G. D. van Albada, J. Dongarra, and P. M. A. Sloot, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 18–25.

[10] K. Dela Rosa and J. Ellen, "Text classification methodologies applied to micro-text in military chat," 12 2009, pp. 710–714.

[11] D. O'Day and R. Calix, "Text message corpus: Applying natural language processing to mobile device forensics," 07 2013, pp. 1–6.

[12] C. Kansara, R. Gupta, S. D. Joshi, and S. Patil, "Crime mitigation at twitter using big data analytics and risk modelling," in *2016 International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, Dec 2016, pp. 1–5.

[13] G. Xiang, B. Fan, L. Wang, J. Hong, and C. Rose, "Detecting offensive tweets via topical feature discovery over a large scale twitter corpus," 10 2012, pp. 1980–1984.

[14] G. Wiedemann, E. Ruppert, R. Jindal, and C. Biemann, "Transfer learning from LDA to bilstm-cnn for offensive language detection in twitter," *CoRR*, vol. abs/1811.02906, 2018.

[15] A. Hakimi Parizi, M. King, and P. Cook, "UNBNLP at SemEval-2019 task 5 and 6: Using language models to detect hate speech and offensive language," in *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 514–518.

[16] K. Zhao, Y. Zhang, C. Xing, W. Li, and H. Chen, "Chinese underground market jargon analysis based on unsupervised learning," in *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, 2016, pp. 97–102.

[17] T. Aoki, R. Sasano, H. Takamura, and M. Okumura, "Distinguishing Japanese non-standard usages from standard ones," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 2323–2328.

[18] R. Řehůřek and P. Sojka, "Software framework for topic modelling with large corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[19] K. Takaoka, S. Hisamoto, N. Kawahara, M. Sakamoto, Y. Uchida, and Y. Matsumoto, "Sudachi: a Japanese tokenizer for business," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018.

[20] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *CoRR*, vol. abs/1310.4546, 2013.