Classification of Illegal Drug Sales Posts using Clustering-Based Topic Modeling.

Anonymous ACL submission

Abstract

Drugs illegally traded online are causing social problems around the world wide. One of the ways to solve this problem is to automatically delete sales posts quickly even if they are uploaded. We propose new data on illegal drug sales posts in Korean collected directly from Twitter. There are about 100K collected data, and labels were added directly to each data. Supervised learning-based models generally show high performance, but label information is essential. It is difficult to add labels to all texts in situations where a large amount of text occurs. In this work, we propose a topic modeling-based classification model that can perform higher with even a small number of labels. As a result of the experiment, higher classification performance is shown when Topic modeling is used as a small number of data.

1 Introduction

Social problems caused by illegal drugs are becoming more serious. Recently, with the development of the Internet, the number of people using social media such as SNS is increasing(Demant et al., 2019; Kaplan and Haenlein, 2010). Social Network Service (a.k.a SNS) used by many people is used as one of the means of trading illegal drugs(Storrod and Densley, 2017). A recent survey estimates that the market using SNS is the largest among illegal online markets(Hall and Antonopoulos, 2016). People can easily contact illegal drug dealers on online and easily purchase illegal drugs. As such, detecting illegal drug sales posts to eradicate illegal drugs is very socially important. However, with huge amounts of text generated every day, major social media companies such as Twitter, Facebook, and Instagram have to spend huge amounts of time detecting illegal drug sales posts. Even if manpower is put in, it is not enough to solve the problem because there is a limit to the manpower put in.

Recently, some studies to solve this problem have attempted to solve the problem using a supervised learning-based model(Park and Fung, 2017; Kulsrud, 2019; Gunawan et al., 2018; Georgakopoulos et al., 2018; Anand and Eswari, 2019). However, since the supervised learning-based model requires a large number of labels, it is time-consuming and inefficient. In addition, there is a disadvantage of not being able to cope with new types of sales posts and slang.

In this paper, we perform a task of detecting illegal drug sales posts posted in Korean on Twitter using a topic modeling-based model. We directly collected Korean tweets from 2016 to 2018, and as a result, we were able to collect about 100K posts. Our model consists of a total of two steps. The first step is to classify a given document by Topic. Each of the classified topics has a specific topic, but it is not known whether the topic is related to illegal drugs. Therefore, we assume that there is some data with Label information, and then we train a classification model. The learned classification model determines whether each topic classified in the first step is related to illegal drug sales. We evaluated the performance of the model based on the data collected directly. As a result of the experiment, it was confirmed that the classification accuracy was higher than when only the existing classification was used. To summarize, our contributions are as follows:

- We propose 100K tweet data in Korean. Each tweet is tagged as to whether it is an illegal drug sales post. To the best of our knowledge, This is first Data to built on drug sales posts in Korean tweets.

- We propose a topic modeling-based classification model. Existing studies have mainly focused on supervised learning-based models, but our model shows higher performance than supervised learning with fewer labels by combining the advantages of topic modeling and classifiers.

2 Related Work

2.1 Text Classification

Although the topics to be discovered from the past are different, there have been studies that attempt to detect specific documents in several documents. For example, Grooming Attack(Michalopoulos and Mavridis, 2011; Gunawan et al., 2016), Toxicity Detection(Carta et al., 2019; Dixon et al., 2018), etc., have different topics, but they are all the same in that they all have the purpose of finding specific documents.

In order to classify a specific document in a document, the Supervised-learning method was mainly used. (Haralabopoulos et al., 2020) performed a task of classifying Toxicity in several documents. They constructed models with five different architectures using CNN(Kalchbrenner et al., 2014) and LSTM(Wang et al., 2015), and classified them by applying ensembles using these models. In addition, (Sulaiman and Siraj, 2019) performed text classification using machine learning models SVM and Naive Bayes based on the chat log. These studies applied Supervised-learning, and the advantage of this method is that it shows high accuracy. However, the downside of this approach is that labels are essential for each text, so labels are expensive. To compensate for these shortcomings, there are also attempts to classify text using the Semi-supervised method. (Rout et al., 2017) performed text classification based on Co-training, EM, and PU algorithm, which are semi-supervised learning models. Meanwhile, (Saraiva et al., 2021) created a graph by setting up a weighted-Edge that connects the sentiment and a specific token, and based on this, regularization was conducted to designate a label for each sentiment. Afterwards, they classified toxic comments using the Machine-learning model. This approach has the advantage of not requiring much labeled data, although it is relatively less accurate than a supervised learning-based method.

2.2 Topic Modeling

The purpose of topic modeling is to automatically discover a specific topic within a document. Topic modeling can be applied to multiple tasks, especially text classification. For example, (Pavlinek and Podgorelec, 2017) applied LDA-based topical modeling to self-training. In addition, (Chen et al., 2016) used a method of measuring the similarity between the two input sentences using the results of topic modeling. They classified documents by clustering with a KNN(Altman, 1992) model using similarity results. However, since these existing LDA-based methods give probability values to all term, words that do not conform to the subject (e.g. stop words) are also included in the calculation, potentially making it impossible to properly find the topic of the document. To solve this problem,(Lee et al., 2015) tried to exclude words that did not fit the subject by adding weight to the LDA. However, this method is not fundamentally a way to exclude words without information. (Angelov, 2020) proposed a clustering-based topic modeling method, away from the existing traditional topic modeling method. In particular, this method attempted to find a Topic Vector rather than giving statistical probabilities to each sentence. The Topic vector they suggested would be close to Term, which represents the topic of each sentence. In addition, this method performed better than the existing LDA and PLSA methods.

3 Drug Selling Post

3.1 Data Collection

We only collected about 100K drug sales posts on Twitter. In general, drug dealers use slang to trade with drug addicts, so drug sales posts cannot be found with common keywords. Therefore, after examining the slang used in the drug black market, we collected tweets with the following keywords. We investigated slang related to methamphetamine, which is popular in Korea, selected slang that is frequently used, and there are a total of five types. We collected tweets based on the selected words and were able to collect about 100K tweets. Hashtags on Twitter are used as a means of quick access to tweets related to one's interests. Therefore, we collected tweets using keywords to which hashtags were added and keywords without hashtags.

3.2 Add label to data

Label information is essential to learn the (semi) supervised learning model. We added label information directly to the collected data. We added labels while reading the text ourselves, and the types of labels were marked as 0,1. Label = 0 is a text, not a drug sales post, and label = 1 is a drug sales post. For the defect of label information, we performed cross validation and proceeded for about a month.

	hashtag + keywords		w/o hashtag + keywords	
	normal post	Drug sales post	normal post	Drug sales post
도리도리(MDMA)	137	-	27295	69
물뽕(GHB)	11	8	1389	3269
아이스(MDMA)	60	140	53163	216
크리스탈(MDMA)	405	12	18858	76
작대기(MDMA)	-	-	1340	-

Table 1: The table shows the number of tweets collected for each keyword. The meaning of each keyword is as follows. "도리도리(dori-dori)" : A type of methamphetamine called ecstasy, "물뽕(mulpong)" : GHB's slang, "아이스(ice)" : methamphetamine powder, "크리스탈(crystal)" : high-quality methamphetamine, '작대기(stick)': methamphetamine administered through syringes.

3.3 Characteristics

In Korea, not only slang but also words that people use on a daily basis are mainly used, and there are not many types of slang, but there are many cases where the words themselves are modified. For example, words such as '아이스' are intentionally used separately from consonants and vowels, such as '아ㅇ ㅣ스',' and 'ㅇ ㅏㅇ ㅣㅅ-'. In addition, most illegal drugs traded are methylenedioxymethylphetamine(MDMA) types.

Table 1 shows the number of texts collected for each keyword by label Among the collected data, the proportion of normal text is overwhelmingly high. There are a total of 106,648 normal texts, and 3,790 drug sales posts account for 3.56% of the total data. Class imbalance problems that appear in this data can interfere with model training.

4 Model Architecture

As can be seen in section 3.3, malicious tweets account for a very small proportion of all tweets. In addition, it is not easy to label all tweets in a situation where tens of thousands of tweets occur every day. We present a topic modeling-based classification model to solve this problem. Our model consists of a total of two steps.

Step 1 is a classification model consisting of a BERT+LSTM structure. We used pre-trained transformer-based BERT embedding model. The pre-trained model we used is the 'bert-base-multilingual-based' model released by Google.

In Step 2, topical modeling is performed using the data other than the data used in Step 1. We conducted topic modeling using the model used in Top2Vec(Angelov, 2020). Top2Vec is clustering-based topical modeling, where vectors are generated adjacent to each other by capturing the meaning of each document. Vectors thus generated are grouped with neighboring vectors using HDBSCAN algorithm(Campello et al., 2013), and as a result, a topic number is assigned to each cluster.

In the results of Step 2, the person should directly check and analyze what topic each post cluster has. However, the larger the number of clusters created, the more cost is consumed to check one by one. Therefore, we use the classification model learned in Step 1 to distinguish whether each cluster is a drug sales post or not. We classified each cluster in a simple way. We first performed a present for all posts belonging to each cluster. If more than half of the posts in each cluster were normal posts, it was determined as normal clusters, whereas if more than half were drug sales posts, it was determined as abnormal clusters. Using this result, each cluster was assigned a single label and the model was evaluated using the modified label.

5 Experiments

5.1 Datasets

We performed simple preprocessing on the data to be used. We removed all the special characters, emoticons, hyperlinks, and images included in each text. In addition, tokenization was performed using the Mecab python package, one of the morpheme analyzer.

We divided the data to be used for classification model and the data to be used for topical modeling into 3:7 ratios while maintaining the label ratio. For the classification model, we sampled 5000 of the normal posts to solve the class imbalance problem. In addition, we set the ratio of train:valid:test data to 8:1:1 to confirm the performance of the classification model. In the case of topic modeling, down sampling was not applied and used as it is.

	Precision	Recall	F1
BERT + LSTM	0.74	0.99	0.85
Our model	0.91	0.93	0.92

Table 2: Result of our model. Topic modeling.When topic modeling is additionally used for BERT-LSTM classification model, the F1 score is higher.

5.2 Experimental Results

We used AdamW as an optimizer when training the classification model, and the learning rate set to 1e-6. In addition, the Loss function uses Binary Cross Entry loss, and Epoch is set to a total of 10. The Top2Vec model used the code disclosed in github[1]. We compared the results of topic modeling with the results of not performing If topic modeling was not performed, the label of 70% of data previously divided was predicted. We used Precision, Recall, and F1 score as evaluation indicators of the model. Table 2 shows the performance of the model. If Topic modeling is not used, the Precision score is very low at 0.74. However, in the case of the model we proposed, the Recall score has fallen a little, but it can be seen that the preview score has improved significantly. The results of this experiment can be interpreted differently. In order to classify drug sales posts with high accuracy, it is better to use only the BERT+LSTM model with a higher Recall score. The model we presented does not filter more drug sales posts because the recall score is lower than the comparison target.

6 Discussion

This section describes the limitations of our study. We evaluated the model with the data collected directly. Due to the nature of illegal drug sales posts, the nature of the data is different from general text classification, such as the use of very many slang words and variations in words. Therefore, further research is needed for the model we present to be applied to other domains. In addition, if there is no text among some tweets and information related to drug sales is written in the attached picture, this model cannot be classified. To solve this problem, the information on the picture must be textured using OCR technology.

7 Conclusion

We collected about 100K drug sales posts posted on Twitter. The collected data was directly labeled and cross-validated. We presented a Topic modeling-based classification model to enhance the performance of the model using a small amount of labeled data, and confirmed that the classification performance was higher than that of the existing classification model.

A Ethics Statements

In this paper, we collected text generated from about 100K tweets to build drug sales post data. In the process of collection, we did not collect individual account IDs for anonymization of users. In addition, it was closed to protect personal information that may be included in the tweet itself. The collected data was used only for classification of drug sales posts. This study can contribute to blocking drug trade, a whole social issue.

References

Naomi S Altman. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.

Mukul Anand and R Eswari. 2019. Classification of abusive comments in social media using deep learning. In *2019 3rd international conference on computing methodologies and communication (ICCMC)*, pages 974–977. IEEE.

Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Salvatore Carta, Andrea Corriga, Riccardo Mulas, Diego Reforgiato Recupero, and Roberto Saia. 2019. A supervised multi-class multi-label word embeddings approach for toxic comment classification. In *KDIR*, pages 105–112.

Qiuxing Chen, Lixiu Yao, and Jie Yang. 2016. Short text classification based on lda topic model. In *2016 International Conference on Audio, Language and Image Processing (ICALIP)*, pages 749–753. IEEE.

Jakob Demant, Silje Anderdal Bakken, Atte Oksanen, and Helgi Gunnlaugsson. 2019. Drug dealing on facebook, snapchat and instagram: A qualitative analysis of novel drug markets in the nordic countries. *Drug and alcohol review*, 38(4):377–385.

[1] https://github.com/MaartenGr/BERTopic

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Spiros V Georgakopoulos, Sotiris K Tasoulis, Aristidis G Vrahatis, and Vassilis P Plagianakos. 2018. Convolutional neural networks for toxic comment classification. In *Proceedings of the 10th hellenic conference on artificial intelligence*, pages 1–6.

Fergyanto E Gunawan, Livia Ashianti, Sevenpri Candra, and Benfano Soewito. 2016. Detecting online child grooming conversation. In *2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS)*, pages 1–6. IEEE.

Fergyanto E Gunawan, Livia Ashianti, and Nobumasa Sekishita. 2018. A simple classifier for detecting online child grooming conversation. *Telkomnika*, 16(3):1239–1248.

Alexandra Hall and Georgios A Antonopoulos. 2016. *Fake meds online: The internet and the transnational market in illicit pharmaceuticals*. Springer.

Giannis Haralabopoulos, Ioannis Anagnostopoulos, and Derek McAuley. 2020. Ensemble deep learning for multilabel binary classification of user-generated content. *Algorithms*, 13(4):83.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.

Andreas M Kaplan and Michael Haenlein. 2010. Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68.

Halvor Bugge Kulsrud. 2019. Detection of cyber grooming during an online conversation. Master's thesis, NTNU.

Seonggyu Lee, Jinho Kim, and Sung-Hyon Myaeng. 2015. An extension of topic models for text classification: A term weighting approach. In *2015 International Conference on Big Data and Smart Computing (BIGCOMP)*, pages 217–224. IEEE.

Dimitrios Michalopoulos and Ioannis Mavridis. 2011. Utilizing document classification for grooming attack recognition. In *2011 IEEE Symposium on Computers and Communications (ISCC)*, pages 864–869. IEEE.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206*.

Miha Pavlinek and Vili Podgorelec. 2017. Text classification method based on self-training and lda topic models. *Expert Systems with Applications*, 80:83–93.

Jitendra Kumar Rout, Anmol Dalmia, Kim-Kwang Raymond Choo, Sambit Bakshi, and Sanjay Kumar Jena. 2017. Revisiting semi-supervised learning for online deceptive review detection. *IEEE access*, 5:1319–1327.

Ghivvago Damas Saraiva, Rafael Anchiêta, Francisco Assis Ricarte Neto, and Raimundo Moura. 2021. A semi-supervised approach to detect toxic comments. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1261–1267.

Michelle L Storrod and James A Densley. 2017. 'going viral'and 'going country': the expressive and instrumental activities of street gangs on social media. *Journal of youth studies*, 20(6):677–696.

Nur Rafeeqkha Sulaiman and Maheyzah Md Siraj. 2019. Classification of online grooming on chat logs using two term weighting schemes. *International Journal of Innovative Computing*, 9(2).

Xin Wang, Yuanchao Liu, Cheng-Jie Sun, Baoxun Wang, and Xiaolong Wang. 2015. Predicting polarities of tweets by composing word embeddings with long short-term memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1343–1353.