Archives of Suicide Research

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/usui20

Suicidality Detection on Social Media Using Metadata and Text Feature Extraction and Machine Learning

Woojin Jung, Donghun Kim, Seojin Nam & Yongjun Zhu

To cite this article: Woojin Jung, Donghun Kim, Seojin Nam & Yongjun Zhu (2021): Suicidality Detection on Social Media Using Metadata and Text Feature Extraction and Machine Learning, Archives of Suicide Research, DOI: 10.1080/13811118.2021.1955783

To link to this article: https://doi.org/10.1080/13811118.2021.1955783

ARCHIVES OF SUICIDE RESEARCH
https://doi.org/10.1080/13811118.2021.1955783

Suicidality Detection on Social Media Using Metadata and Text Feature Extraction and Machine Learning

Woojin Jung, Donghun Kim, Seojin Nam, and Yongjun Zhu

ABSTRACT

In this study, we implemented machine learning models that can detect suicidality posts on Twitter. We randomly selected and annotated 20,000 tweets and explored metadata and text features to build effective models. Metadata features were studied in great details to understand their possibility and importance in suicidality detection models. Results showed that posting type (i.e., reply or not) and time-related features such as the month, day of the week, and the time (AM vs. PM) were the most important metadata features in suicidality detection models. Specifically, the probability of a social media post being suicidal is higher if the post is a reply to other users rather than an original tweet. Moreover, tweets created in in the afternoon, on Fridays and weekends, and in fall have higher probabilities of being detected as suicidality tweets compared with those created in other times. By integrating metadata and text features, we obtained a model of good performance (i.e., F1 score of 0.846) that can assist humans in the real-world setting to detect suicidality social media posts.

KEYWORDS

Classification; feature extraction; machine learning; social media; suicidality detection

INTRODUCTION

In the world, about 800,000 people kill themselves every year which means that a single person commits suicide every 40 s (WHO, 2018). Many countries have made great efforts to reduce suicide rate and prevent suicide. For example, the U.S. Department of Health and Human Services is operating the *National Suicide Prevention Lifeline* (NSPL), a 24-h suicide prevention hotline in the U.S. The South Korean government has also been implementing many strategies to prevent suicide including the promotion of various suicide prevention campaigns and the operation of *Lifeline Korea*, which is a 24-h national suicide prevention lifeline. Despite these efforts, suicide is one of the highest causes of death in the world (WHO, 2018) and it is highly necessary to take further approaches into consideration to prevent suicide more proactively.

Suicide has several risk factors that are closely related with social and mental problems (Turecki & Brent, 2016). Detection of suicidality, which covers suicidal ideation, suicide plans, and suicide attempts, plays a significant role in suicide prevention (Goldsmith, Pellmar, Kleinman, & Bunney, 2002). Studies reported that people show warning signs before attempting suicide and those warning signs can be detected from

their behavioral patterns (Whitlock & Knox, 2007; Bailey et al., 2011) and words they use (Baddeley, Daniel, & Pennebaker, 2011; Fernández-Cabana, Caballero, Pérez, García-García, & Mateos et al., 2013; Fernández-Cabana et al., 2015). Therefore, suicidality can be detected from explicit and implicit warning signs. Detecting suicidality in the real-life is hard though, because people rarely talk about their suicidal ideation, plans, and attempts with others and rather try to hide them (Parekh & Phillips, 2014). However, on social media, people behave much differently. On social media, people share about their daily lives (Paul, 2014), exchange political opinions (Zúñiga, de Molyneux, & Zheng, 2014), and even talk about their suicidality (Jashinsky et al., 2014). Given the circumstances, using social media data to detect suicidal ideation for early intervention has been widely studied with advanced techniques such as machine learning and deep learning (Su, Xu, Pathak, & Wang, 2020; Wongkoblap, Vadillo, & Curcin, 2017). O'Dea et al (2015) tried to detect suicide-related posts from Twitter and showed a great possibility that social media can be used to detect suicidality. Later, multiple studies explored social media to study suicidality. For example, Vioulès, Moulahi, Azé, and Bringay (2017) classified Twitter texts into four degrees of suicidality. Extracting suicide risk factors or stressors that imply suicidality from social media texts was also studied by researchers (Cheng, Li, Kwok, Zhu, & Yip, 2017).

Most of the previous studies relied on exploiting textual information to detect suicidality (Su, Xu, Pathak, & Wang, 2020; Wongkoblap et al., 2017). Specifically, the studies extracted various features from social media texts (e.g., word count, POS tags) and built machine learning models using the extracted text features. Few studies explored and utilized metadata of social media posts such as tweet creation time (Lalrinmawii, Vanlalhruaia, & Debnath, 2020) and posting type (Burnap, Colombo, Amery, Hodorog, & Scourfield, 2017) for suicidality detection. Metadata of social media posts have been investigated on a very limited scale and there are still unexplored attributes with high potential to capture suicidality. Therefore, in this study, we aim to build a machine learning models that are based on a comprehensive feature set of metadata and text features to detect suicidality on Twitter. Specifically, we have the following research questions.

- RQ1: What metadata features can be used in detecting suicidality on twitter and what are their relative importance and contribution to the detecting task?
- RQ2: What text features can be extracted and used in detecting suicidality on twitter?
- RQ3: What are the best feature set of metadata and text features that yields the best performance in suicidality detection?

The rest of the paper is organized as follows. In the related work, we discuss recent studies and methods about detecting suicidality from social media posts. In the data section, we explain the data collection, preprocessing, and annotation processes followed by the methods section where we discuss the proposed methods. Following that, we present the experimental results and discuss findings. Finally, we conclude the paper with a summary and our contributions.

RELATED WORK

Detection of suicidality from social media texts has traditionally been studied as a machine learning-based classification problem. Learning from annotated social media texts, researchers train binary classification models to classify a text into one of the two possible labels. Two types of features: text features and metadata features have been investigated in previous studies.

Classification Using Text Features

Braithwaite, Giraud-Carrier, West, Barnes, and Hanson (2016) developed a machine learning model to classify Twitter posts into suicidal or non-suicidal with text features extracted using Linguistic Inquiry and Word Count (LIWC) (Pennebaker, Boyd, Jordan, & Blackburn, 2015). LIWC analyzes words of each tweet using 90 variables (e.g., part-of-speech (POS) tag, length, category of word meaning, etc.) and provides the percentage of total words in each text by the variables. For instance, if a tweet has 100 words and ten of these words are verbs, LIWC gives a score of 10 to the existence of verb category. The researchers used the scores of tweets as the features for their classification model. Cheng et al. (2017) investigated Sina Weibo, a social media platform in China, to detect users' suicide probability using Simplified Chinese Linguistic Inquiry and Word Count (SC-LIWC) dictionary, which is the Chinese version of LIWC (Gao, Hao, Li, Gao, & Zhu, 2013). They conducted a survey and used Suicide Probability Scale (SPS, Cull & Gill, 1982) to measure suicidality of respondents. A cutoff value was used to classify the posts into suicidal or not. If a respondent got 80 or a higher SPS score, the user's posts was labeled as having suicidal ideation. Each post was scored by SC-LIWC and the scores were used by their models in the same way Braithwaite et al. (2016) did. Du et al. (2018) adopted a word embedding technique to represent words in texts using vectors of numerical values. They mapped words in each text to one-hot vectors and converted the vectors to pre-trained Glove Twitter embedding (Pennington, Socher, & Manning, 2014) to use them as features of their suicidality detection models. GloVe Twitter embedding is an open-source word embedding which was trained with about 6 billion words from web text and Wikipedia. Each word in a tweet was mapped to a Glove embedding vector and a tweet was represented by concatenating vectors that represent each of the words in the tweet. The vectorized tweets were used to build their suicidality classification models. Vioulès et al. (2017) created customized sentiment dictionary of suicidality using Pointwise Mutual Information (PMI) (Krestel & Siersdorfer, 2013) to score words in tweets. They annotated tweets into four groups such as no distress, minimal distress, moderate distress, and severe distress based on the sentiment scores. Tweets labeled with minimal, moderate, and severe distress were consolidated into suicidality class and the remaining tweets were grouped into the non-suicidality class. They used the sum of each word's score in sentiment dictionary, frequencies of intensifying terms (e.g., very, extremely), first pronouns, and sum of sentiment dictionary scores in each tweet as features for detecting suicidality. In addition to using frequency of n-gram terms and sum of sentiment scores of words in each text, Patel, Shah, and Farooqui (2020) extracted emoticons from Tweets and classified them into positive or negative class and later used them as features. They replaced the emoticons with

words that express the meaning of them. For instance, ":)" was labeled as positive and replaced with the word "happy". The researchers considered frequencies of labeled emoticons in each text and used them as sentiment features in conjunction with Afinn Dictionary (Nielsen, 2011), an English sentiment dictionary. Including features such as word frequency, POS tags, variables of LIWC, word embedding vectors, Ji, Yu, Fung, Pan, and Long (2018) used allocated topics of each post computed using Latent Dirichlet Allocation (LDA), a topic modeling method that shows the probability of each word appears in a specific topic (Blei, Ng, & Jordan, 2003) to detect suicidality. Given a tweet, they calculated the sum of the included words' probabilities for each of the extracted topic and assigned the topic number in which the tweet has the highest score to the tweet. For example, if the aggregate probabilities of a tweet's words for the sixth topic is the highest, the number six was assigned to the topic and used as a feature for their classification models. Desmet and Hoste (2018) used named entity features in addition to the abovementioned text features to detect suicidality on Dutch-language forum posts. They extracted named entities using DBpedia Spotlight (Mendes, Jakob, García-Silva, & Bizer, 2011), a tool for annotating mentions of DBpedia resources in text. They created three features based on the annotated and extracted DBpedia named entities. One is a binary feature that denotes the existence of one or more named entities in the posts and the other two are integer features that indicate the number of named entities and unique named entities in the posts.

Classification Using Metadata and Text Features

Few studies used metadata of social media posts such as posting type and creation time along with text features to train classification models. Burnap et al. (2017) labeled each tweet based on whether it has "@" or not, a symbol used in twitter to mention users. A tweet is either a reply or a retweet if it has "@" in the body, which is different from a tweet that created by a user in response to nothing. They created two binary features to denote whether a tweet is a retweet from other users or not and whether it is a reply to other users or not. Huang et al. (2014) used two metadata features (i.e., posting type and posting time) to identify suicidal ideation from Sina Weibo posts. Posting type is a binary feature that indicates a post as either original or reposted. Posting time was grouped into four categories by comparing proportions of suicidality posts with that of non-suicidality posts within a given time period. Among the texts posted from 23:00 to 06:00, proportion of suicidal texts are much higher than non-suicidal texts. On the contrary, from 07:00 to 13:00, non-suicidal texts were posted much more than suicidal texts. Proportion of non-suicidality posts are slightly higher than suicidality posts from 14:00 to 18:00 while proportion of suicidality posts are slightly higher from 18:00 to 23:00. Based on the four manually identified time periods, a posting time-based categorical feature with four possible values was created. Huang et al. (2015) introduced a new social relationship feature in addition to the metadata features used in their previous work (Huang et al., 2014). The social relationship feature is a binary feature indicating whether other users were mentioned in the text or not.

From the previous studies, we can see that metadata of social media posts have been explored at a very limited scale and we have limited understanding about how

TABLE 1. The 19 elements of tweets.

Element	Description
has_media	Whether a tweet includes media (i.e., picture, video) or not
hashtags	List of hashtags in tweet
img_urls	List of images in tweet
is_replied	Whether a tweet has a reply or not
is_reply_to	Whether a tweet is a reply or not
likes	Number of likes
links	List of url links in tweet
parent_tweet_id	ID of a parent tweet
replies	Number of replies
reply_to_users	List of users who replied
retweets	Number of retweets
screen_name	Writer's login ID
text	Tweet text
timestamp	Posting time of tweet
tweet_id	Tweet ID
tweet_url	Url link of tweet
user_id	Writer's ID allocated by Twitter
username	Writer's nickname
video_url	List of videos in tweet

important they are in detecting suicidality. In addition, metadata features have not been well integrated into the widely studied text features to build a more robust suicidality detection model. Therefore, our study aims at investigating importance of metadata features in suicidality detection from Twitter and build a robust model by integrating metadata and text features.

DATA

Data Collection

Forty-five suicide keywords provided by Korean Suicide Prevention Center (KSPC) were used to search for suicide-related tweets. KSPC manually curated keywords such as *suicide*, *suicide methods*, *sleeping pills*, etc., that are related to suicide. We collected 457,947 tweets that include at least one of the 45 keywords and created between January 1, 2019 and December 31, 2019 using TwitterScraper 1.4.0 (Taspinar, 2019). The 19 elements of the collected tweets are shown in Table 1.

Data Preprocessing

While manually checking the collected tweets, we found that many of them are not related to suicide but about news articles, campaigns, entertainments, etc. Therefore, posts created by bot users (a tweet generating computer program), universities, campaigns, news publishers, and entertainments were removed because these users are not individuals and they mention suicide-related terms mostly for the purpose of suicide prevention. 30,210 tweets were deleted in the process and additional 39,015 tweets that include hyperlinks were removed because these hyperlinks mainly link to news articles. As a final step, 8,387 tweets that have hashtags related to bots (#**bot, #**auto), entertainment (#BTS, #V, #Taehyung), and marketing (#**sale, #saleabout**) were filtered out.

Data Annotation

20,000 tweets were randomly selected from the remaining 380,290 tweets and were annotated by three researchers in terms of the existence of suicidality. The researchers were instructed by a psychiatrist before the annotation to set up annotation guidelines. Among the 20,000 tweets, the annotators annotated 1,097 tweets as suicidal unanimously. As the first step, only tweets with unanimous agreement were annotated as suicidality tweets. With the arbitration of the psychiatrist, the researchers reannotated the tweets that two researchers had previously annotated as suicidality tweets. There were no additional suicidality tweets with unanimous agreement in the reannotation process and the annotators confirmed 1,097 tweets as suicidal.

METHODS

In this section, we describe selected metadata and text features as well as machine learning methods used in our suicidality detection models.

Feature Selection

From the initial elements shown in Table 1, we removed the following 11 features: *parent_tweet_id*, *user_id*, *username*, *screen_name*, *tweet_id*, *tweet_url*, *is_replied*, *likes*, *replies*, *retweets*, and *reply_to_users* when building the metadata feature set. The first six elements were removed because they are identifiers that should not be used as features. The last five features were not considered because they are not available at the time of tweet creation and affected by both tweet creation time and data collection time. For example, a tweet created in January has more time to get likes than a tweet created in December. Because both tweets were collected at the same time, it is not fair to use *likes* as a feature to compare them. Among the remaining eight elements, *text*, which is the body of a tweet, was used to derive text features and the other seven elements were used to derive metadata features.

Metadata Features

The seven elements (i.e., *has_media*, *hashtags*, *img_urls*, *is_reply_to*, *links*, *video_url*, and *timestamp*) were further divided into time and non-time features.

Time Features

Nine time-related features (i.e., *timestampMonth*, *timestampDayofweek*, *timestampIs_month_end*, *timestampIs_month_start*, *timestampIs_quarter_end*, *timestampIs_quarter_start*, *timestampIs_year_end*, *timestampIs_year_start*, and *is_weekday*) were derived from the *timestamp* element using the Python library—fastai (Howard & Gugger, 2020) because segmented features derived from *timestamp* enable us to consider the time aspect of tweets more sophisticatedly than just exploiting a single *timestamp*. We created two additional features from *timestamp*. We created a binary feature denoting whether a day is a holiday or not using the Python library—holidays

TABLE 2. Time features extracted from *timestamp*.

Feature	Description	Type
timestampMonth	Posting month of tweet	Categorical
timestampDayofweek	Posting day of week of tweet	Categorical
timestampls_month_start	Whether posting day is the first day of month or not	Binary
timestampls_month_end	Whether posting day is the last day of month or not	Binary
timestampls_quarter_start	Whether posting day is the first day of quarter or not	Binary
timestampls_quarter_end	Whether posting day is the last day of quarter or not	Binary
timestampls_year_start	Whether posting day is the first day of year or not	Binary
timestampls_year_end	Whether posting day is the last day of year or not	Binary
is_weekday	Whether posting day is weekday or not	Binary
is_holiday	Whether posting day is holiday or not	Binary
time_period	Time period of posting time	Categorical

0.10.4 (Montel, 2020). We further divided *timestamp* into four time periods (i.e., 00:00–05:59, 06:00–11:59, 12:00–17:59, and 18:00 ~ 23:59) and created a categorical feature. Table 2 shows the final list of time features.

Non-Time Features

The six elements (i.e., *has_media*, *hashtags*, *img_urls*, *is_reply_to*, *links*, and *video_url*) were converted into six non-time features. *has_media* and *is_reply_to* were converted into binary features while the remaining four features were converted into numeric features indicating the number of hashtags, images, url links, and videos in a tweet.

Text Features

Text features were derived from the *text* element. The following features were created.

Post Length

The number of characters in each tweet was calculated and used as a feature. Characters include whitespaces and punctuations. Post length of each tweet is represented as an integer.

Syntactic Features

POS tags of all the words in a tweet were extracted using KoNLPy 0.5.2 (Park & Cho, 2014), a Python library for Korean text analysis. The number of each of the 13 POS tags (e.g., noun, verb, adjective, etc.) in each tweet were used as syntactic features.

N-Gram Features

All tweets were tokenized into uni-, bi-, and trigrams. Frequency and tf-idf score of each n-gram in each tweet were computed using scikit-learn 0.23.2 (Pedregosa et al., 2011) and used as features. Tf-idf is a method that measures each word's importance in a text (Qaiser & Ali, 2018). Each n-gram is a feature with frequency or tf-idf score filled for given a tweet. 2,329 terms which appear in at least five posts were extracted as frequency features or tf-idf features. We conducted Principal Component Analysis (PCA) as a dimensionality reduction procedure to convert all possibly correlated terms into

linearly uncorrelated principal features (Howley, Madden, O'Connell, & Ryder, 2006). After PCA, each number of term frequency features and tf-idf features decreased to 652 and 755.

Topic Features

We extracted topic features from tweets using Gensim (Radim & Sojka, 2010), a Python library for text analysis and topic modeling. We used Mallet 2.0.8 (McCallum, 2002) model to identify the optimal number of topics by comparing different coherence values and extract topics. Six topics were extracted and the probabilities that a tweet belongs to each of the six topics were used as features values.

Polarity Lexicon Feature

We extracted polarity lexicon features using KnuSentiLex (Byung-Won, Sang, & Chulwon, 2018), a Python library for Korean sentiment analysis. KnuSentiLex categorizes a word into one of the five classes (very negative, negative, normal, positive, and very positive) and provides sentiment scores for each of the word in the lexicon from -2 to $+2$. The sentiment score of a tweet is the sum of sentiment scores of all words in the tweet.

Domain-Specific Lexicon Features

Domain-specific lexicon features were extracted from F code of KCDC7 (KOICD, 2016), a list of psychiatric disorders in Korean Standard Classification of Disease. Disorders include schizophrenia, depression and etc. The extracted disorders were used as binary features to denote whether the disorders are mentioned in tweets or not.

Machine Learning Methods

We used random forest (RF) and gradient boosting machine (GBM) as our suicidality detection methods. The two methods have been widely used in machine learning tasks that deal with tabular data. Random forest is an ensemble learning method based on bagging process (Rodriguez, Kuncheva, & Alonso, 2006). Each decision tree of random forest selects random features and/or samples and random forest provides the average performance of decision trees. Random forest has shown a good performance in a classification task of suicidality posts (Ji et al., 2018). GBM is a machine learning technique based on boosting process for classification and regression tasks (Natekin & Knoll, 2013). Boosting is an iterative process which boosts the previous decision tree's performance by reducing the value of loss function (Freund & Schapire, 1997). We used implementations provided in scikit-learn 0.23.2 (Pedregosa et al., 2011) and explored the performance of various combinations of features to build a robust model.

TABLE 3. Performance of suicidality detection models using base feature sets.

			Score			
Feature type	Feature set	Method	Recall	Precision	F1 (SD)	roc_auc
Metadata	Base	Random forest	0.916122	0.560618	0.553528 (0.016414)	0.60819
		GBM	0.917039	0.561169	0.554564 (0.012378)	0.605942
Metadata	Time	Random forest	0.559725	0.532833	0.532044 (0.016251)	0.552946
		GBM	0.559733	0.551619	0.55118 (0.015943)	0.567427
Metadata	Base + time	Random forest	0.652652	0.614136	0.619446 (0.015406)	0.661281
		GBM	0.724679	0.614916	**0.6641** (0.017198)	0.692521
Text	Word count	Random forest	0.726572	0.733139	0.730929 (0.013799)	0.772833
		GBM	0.824987	0.824619	**0.824434** (0.015295)	0.894292
Text	Tf-idf	Random forest	0.712018	0.709668	0.710441 (0.016917)	0.757259
		GBM	0.832193	0.803448	0.813892 (0.015516)	0.886331
Text	Others	Random forest	0.690133	0.693409	0.69151 (0.012)	0.732108
		GBM	0.778449	0.767317	0.770849 (0.013602)	0.840837
Text	Word count + others	Random forest	0.715563	0.71368	0.713923 (0.014361)	0.752738
		GBM	0.815838	0.830924	0.823867 (0.017176)	0.901086
Text	Tf-idf + others	Random forest	0.860592	0.789786	0.814912 (0.013211)	0.886109
		GBM	0.830442	0.818246	0.822531 (0.014502)	0.893685

RESULTS

In the experiments, we used 1,097 suicidality tweets and 1,097 non-suicidality tweets that had been randomly selected from the remaining 18,903 tweets. All the trained machine learning models were evaluated using a 10-fold cross validation.

Evaluation of Models Using Base Features

Eight base feature sets were created by grouping individual features. The base metadata feature set is a group of six non-time metadata features. The time metadata feature set is a group of 11 time-related features. Text features were divided into the word count feature set, tf-idf feature set, and others feature set that is composed of the rest of the text features. Machine learning models based on each of the feature set were trained using the default hyperparameter settings and the performance are shown in Table 3. Among the models based on metadata features, GBM with the whole metadata features achieved the highest F1 score of 0.66 is represented in bold. GBM with the word count feature set achieved the highest F1 score of 0.82 among text feature-based models is represented in bold.

Feature Importance of Metadata Features

We investigated feature importance of each metadata feature in detecting suicidality tweets. *is_reply_to* is the most important feature in the GBM models followed by *timestampMonth* and *num_hashtags*. *timestampMonth* has the highest feature importance in the RF models, followed by *timestampDayofweek* and *time_period*.

To further explore important metadata features, we used partial dependence plots to show how features affect the prediction of machine learning models (Brandon, 2017). We created partial dependence plots of features using PDPbox 0.2.0 (SauceCat, 2018), a Python library for creating partial dependence plots. Because GBM models had better performance as shown in Table 3, we selected the top five important features in GBM models and plotted them. Among the five top features, *num_hashtags* was excluded because although the feature is important, most of the tweets in the dataset do not have hashtags. Partial dependence plots of the other four features are shown in Figures 1–4. Figure 1 shows the partial dependence plot of *is_reply_to* from which we can see that if a tweet is a reply to other users, the probability that the tweet being suicidal is slightly higher. Among the 301 tweets that are not replies, less than 20% of them are suicidality tweets.

Figure 2 shows the partial dependence plot of t*imestampMonth*. More than 66% of tweets created in November were suicidality tweets while tweets created in January has the lowest probability of being suicidal.

According to the prediction distribution shown in Figure 3, tweets created on Fridays and weekends have higher probability of being suicidal than those created on the other days of the week. The lowest probability was shown in tweets created on Tuesday.

Figure 4 shows that more than 63% of tweets created between 12:00 and 18:00 are suicidality tweets while in the other time periods, we did not observe such a notable imbalance.

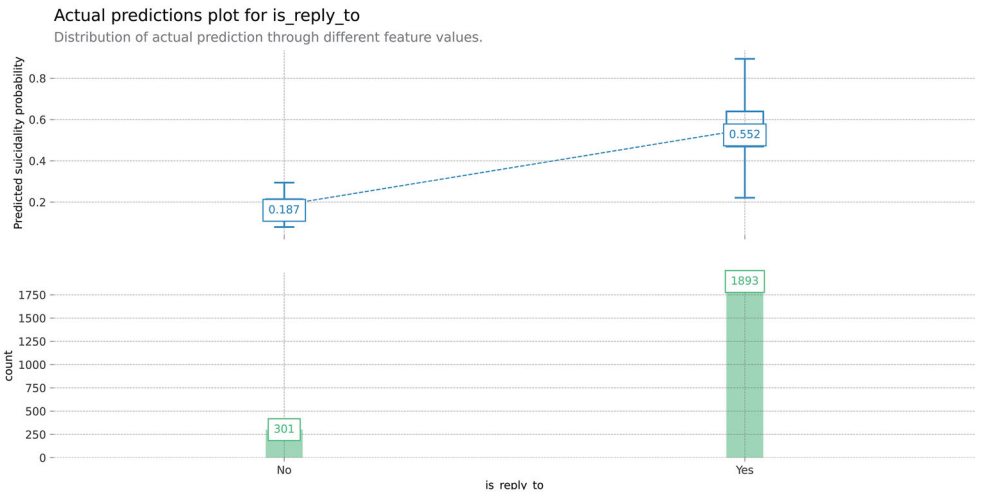

FIGURE 1. The partial dependence plot of is_reply_to.

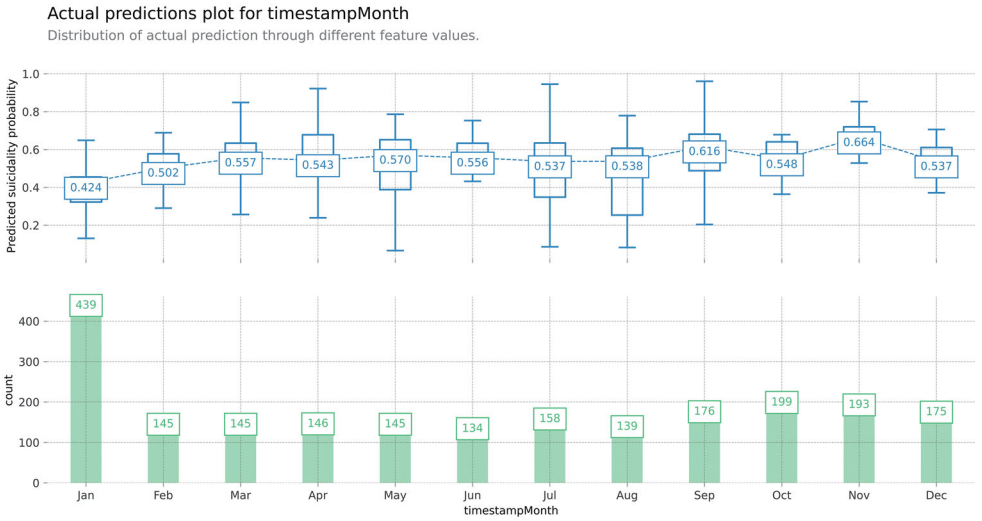

FIGURE 2. The partial dependence plot of timestampMonth.

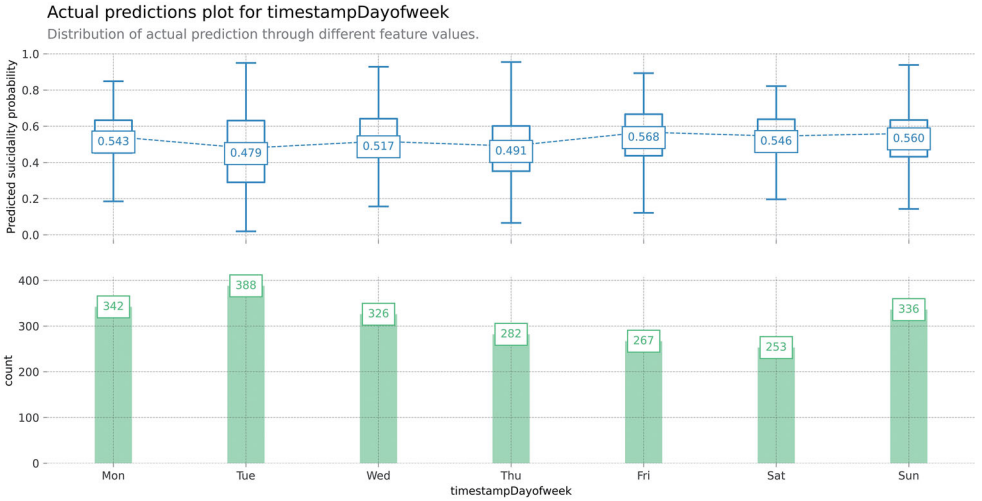

FIGURE 3. The partial dependence plot of timestampDayofweek.

Evaluation of Models Using Integrated Metadata and Text Features

We integrated metadata and text features to form combined feature sets and build better models. We used all the metadata features because the model trained using all the metadata features showed the highest performance among models trained using metadata features (Table 3). The metadata feature set was integrated with five different text feature sets. GBM with the default hyperparameter setting was used to compare the performance of the models based on integrated features sets (Table 4) with those based on base feature sets (Table 3) under the same condition. As shown in Table 4, the combined feature set of metadata and word count features showed the highest performance F1 score of 0.84 are represented in bold. Overall, the performance of the models based on integrated feature sets is higher than those based on base feature sets.

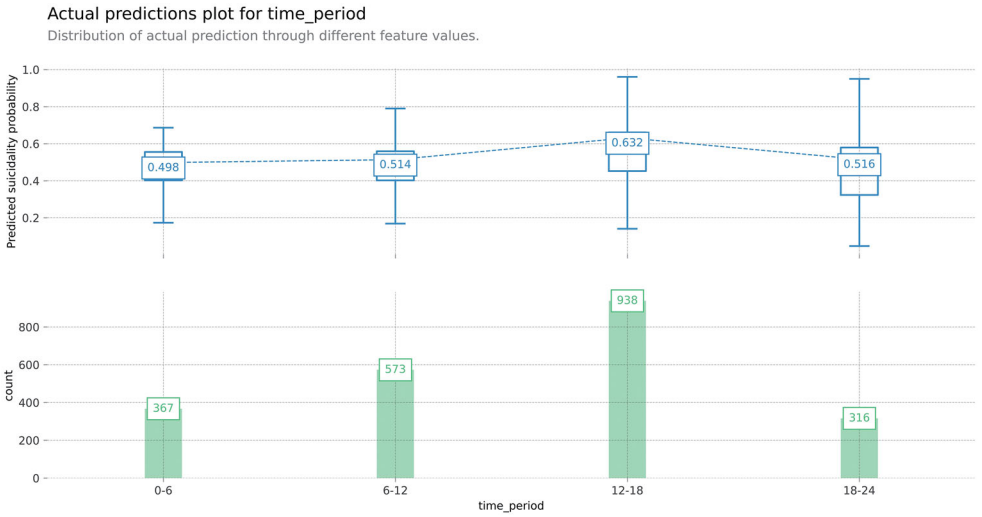

FIGURE 4. The partial dependence plot of time_period.

TABLE 4. Performance of suicidality detection models using integrated feature sets.

		Score			
Feature set	Method	Recall	Precision	F1 (SD)	roc_auc
metadata + word count	GBM	0.835054	0.838083	**0.835781** (0.015221)	0.906563
metadata + tf-idf	GBM	0.83859	0.81492	0.825904 (0.016443)	0.896372
metadata + others	GBM	0.791226	0.798007	0.792782 (0.012022)	0.868694
metadata + word count + others	GBM	0.830459	0.843948	0.835764 (0.015954)	0.910369
metadata + tf-idf + others	GBM	0.84221	0.82788	0.834288 (0.019021)	0.903498

To obtain models with better performance, we tunned hyperparameters of GBM. After several rounds of hyperparameter tunning, we obtained the best performing model with the F1 score of 0.846. The model was built using all the metadata features, word count based n-gram features and other text features.

DISCUSSION AND CONCLUSION

In this study, we have several meaningful findings regarding building effective suicidality detection models on social media. We explored the possibility and importance of metadata extracted from social media posts as features for suicidality detection models. While previous studies tended to only utilize social media texts as model features and overlook metadata, our findings suggest that various metadata features can be effectively utilized to detect suicidality of social media posts. Posting type was one of the most important features to detect suicidality posts. Our findings revealed that the probability of a social media post being suicidal is higher if the post is a reply to other users rather than an original tweet. Moreover, time-related metadata features showed their importance. Tweets created in in the afternoon, on Fridays and weekends, and in fall have

higher probabilities of being detected as suicidality tweets compared with those created in other times. Finally, by integrating metadata and text features and tunning the hyper-parameters, we obtained a model of good performance (F1 score of 0.846).

We expect that the implemented models can assist humans in the real-world setting to detect suicidality tweets and prevent suicide. By using the automated models as an initial step of identifying suicidality tweets with subsequent minimal human efforts on manual review, people who work for suicide prevention can proactively handle suicidality content that may link to suicide. This will save tremendous time of reviewing candidate tweets manually from the beginning. Implementing the models as on-the-fly systems, we can even monitor suicidality tweets in real time although privacy and related issues are challenging and need to be properly tackled. Nevertheless, it is worth mentioning that the proposed approach requires annotated data and the model's performance highly depends on the volume and the quality of annotated data. In the future work, we plan to exploit other external sources such as mental health-related knowledge graph to further reduce human efforts and at the same time improve the model's performance.

FUNDING

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ICT Creative Consilience program [IITP-2020-0-01821) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation); Ministry of Education of the Republic of Korea and the National Research Foundation of Korea [NRF-2019S1A5A2A03045289]; and SKKU-SMC Future Convergence Research Program Grant.

AUTHOR NOTE

Woojin Jung, Donghun Kim, Seojin Nam, and Yongjun Zhu, Department of Library and Information Science, Sungkyunkwan University, Seoul, Republic of Korea

Correspondence concerning this article should be addressed to Yongjun Zhu, Department of Library and Information Science, Sungkyunkwan University, Seoul, Republic of Korea. Email: yzhu@skku.edu

ORCID

Donghun Kim http://orcid.org/0000-0001-5441-1532

REFERENCES

Bailey, R. K., Patel, T. C., Avenido, J., Patel, M., Jaleel, M., Barker, N. C., … Jabeen, S. (2011). Suicide: Current trends. *Journal of the National Medical Association*, *103*(7), 614–617. doi:10.1016/S0027-9684(15)30388-6

Baddeley, J., Daniel, G., & Pennebaker, J. (2011). How Henry Hellyer's use of language foretold his suicide. *Crisis*, *32*(5), 288–292. doi:10.1027/0227-5910/a000092

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Braithwaite, S. R., Giraud-Carrier, C., West, J., Barnes, M. D., & Hanson, C. L. (2016). Validating machine learning algorithms for Twitter data against established measures of suicidality. *JMIR Mental Health*, *3*(2), e21. doi:10.2196/mental.4822

Brandon, M. (2017). pdp: An R package for constructing partial dependence plots. *The R Journal*, *9*, 421. doi:10.32614/rj-2017-016

Burnap, P., Colombo, G., Amery, R., Hodorog, A., & Scourfield, J. (2017). Multi-class machine classification of suicide-related communication on Twitter. *Online Social Networks and Media*, *2*, 32–44. doi:10.1016/j.osnem.2017.08.001

Byung-Won, O., Sang, P., & Chulwon, N. (2018). *KnuSentiLex*. https://github.com/park1200656/KnuSentiLex.

Cheng, Q., Li, T. M. H., Kwok, C. L., Zhu, T., & Yip, P. (2017). Assessing suicide risk and emotional distress in Chinese social media: A text mining and machine learning study. *Journal of Medical Internet Research*, *19*(7), e243. doi:10.2196/jmir.7276

Cull, J. G., & Gill, W. S. (1982). Suicide Probability Scale (SPS) Manual. *Western Psychological Services*. https://books.google.co.kr/books?id=BKe_PwAACAAJ.

Desmet, B., & Hoste, V. (2018). Online suicide prevention through optimised text classification. *Information Sciences*, *439-440*, 61–78. http://www.sciencedirect.com/science/article/pii/S002002551830094X. doi:10.1016/j.ins.2018.02.014

Du, J., Zhang, Y., Luo, J., Jia, Y., Wei, Q., Tao, C., & Xu, H. (2018). Extracting psychiatric stressors for suicide from social media using deep learning. *BMC Medical Informatics and Decision Making*, *18*(Suppl 2), 43–43. doi:10.1186/s12911-018-0632-8

Fernández-Cabana, M., Caballero, A. G., Pérez, M. T., García-García, M., & Mateos, R. (2013). Suicidal traits in Marilyn Monroe's fragments an LIWC analysis. *Crisis*, *34*(2), 124–127. doi:10.1027/0227-5910/a000183

Fernández-Cabana, M., Jiménez-Féliz, J., Pérez, M. T., Mateos, R., Gómez-Reino, I., & Caballero, A. G. (2015). Linguistic analysis of suicide notes in Spain. *The European Journal of Psychiatry*, *29*(2), 145–155. doi:10.4321/S0213-61632015000200006

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. doi:10.1006/jcss.1997.1504

Gao, R., Hao, B., Li, H., Gao, Y., & Zhu, T. (2013). Developing simplified Chinese psychological linguistic analysis dictionary for microblog. *Lecture Notes in Computer Science*, *8211*, 359–368. doi:10.1007/978-3-319-02753-1_36

Howard, J., & Gugger, S. (2020). Fastai: A layered API for deep learning. *Information*, *11*(2), 108. doi:10.3390/info11020108

Howley, T., Madden, M. G., O'Connell, M.-L., & Ryder, A. G. (2006). The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data. *Knowledge-Based Systems*, *19*(5), 363–370. doi:10.1016/j.knosys.2005.11.014

Huang, X., Li, X., Zhang, L., Liu, T., Chiu, D., & Zhu, T. (2015). *Topic model for identifying suicidal ideation in Chinese microblog*. Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, Shanghai, China.

Huang, X., Zhang, L., Liu, T., Chiu, D., Zhu, T., & Li, X. (2014). *Detecting suicidal ideation in Chinese microblogs with psychological lexicons*. 2014 IEEE 11th Intl Conf on Ubiquitous Intelligence and Computing and 2014 IEEE 11th Intl Conf on Autonomic and Trusted Computing and 2014 IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops, Bali, Indonesia. doi:10.1109/uic-atc-scalcom.2014.48

Jashinsky, J., Burton, S., Hanson, C., West, J., Giraud-Carrier, C., Barnes, M., & Argyle, T. (2014). Tracking suicide risk factors through Twitter in the US. *Crisis*, *35*(1), 51–59. doi:10.1027/0227-5910/a000234

Ji, S., Yu, C., Fung, S., Pan, S., & Long, G. (2018). Supervised learning for suicidal ideation detection in online user content. *Complexity*, *2018*, 1–10. doi:10.1155/2018/6157249

KOICD. (2016). *KCDC7*. http://www.koicd.kr/2016/kcd/v7.do#5&n.

Krestel, R., & Siersdorfer, S. (2013). *Generating contextualized sentiment lexica based on latent topics and user ratings*. doi:10.1145/2481492.2481506

Lalrinmawii, C., Vanlalhruaia, and Debnath, S. (2020). *Analysis of post centric suicidal expressions and classification on the Social Media Post: Twitter*. 2020 11th International Conference on

Computing, Communication and Networking Technologies (ICCCNT). pp. 1–5. doi:10.1109/ICCCNT49239.2020.9225638
McCallum, A. K. (2002). *MALLET: A machine learning for language toolkit*. Retrieved from http://mallet.cs.umass.edu/
Goldsmith, S. K., Pellmar, T. C., Kleinman, A. M., & Bunney, W. E. (2002). *Reducing suicide: A national imperative*. Washington, DC: The National Academies Press.
Mendes, P., Jakob, M., García-Silva, A., & Bizer, C. (2011). *DBpedia spotlight: Sh*edding light on the web of documents. doi:10.1145/2063518.2063519
Montel, M. (2020). *python-holidays*. https://pypi.org/project/holidays/.
Natekin, A., & Knoll, A. (2013). Gradient boosting machines: A tutorial. *Frontiers in Neurorobotics*, *7*, 21. doi:10.3389/fnbot.2013.00021
Nielsen, F. Å. (2011). *A new ANEW: Evaluation of a word list for sentiment analysis in micro-blogs*. https://arxiv.org/abs/1103.2903
O'Dea, B., Wan, S., Batterham, P. J., Calear, A. L., Paris, C., & Christensen, H. (2015). Detecting suicidality on Twitter. *Internet Interventions.*, *2*(2), 183–188. doi:10.1016/j.invent.2015.03.005
Parekh, A., & Phillips, M. (2014). *Preventing suicide: A global imperative*. Geneva: WHO.
Park, E. L., & Cho, S. (2014). *KoNLPy: Korean natural language processing in Python*. Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology, Chuncheon, Korea.
Patel, V., Shah, H., & Farooqui, Y. (2020). *Hybrid feature based prediction of suicide related activity on Twitter*. 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 00, 590–595. doi:10.1109/ICICCS48265.2020.9120876
Paul, M. (2014). The civic-social media disconnect: Exploring perceptions of social media for engagement in the daily life of college students. *Information, Communication & Society*, *17*(9), 1059–1071. doi:10.1080/1369118x.2013.877054
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.
Pennebaker, J., Boyd, R., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Retrieved from https://repositories.lib.utexas.edu/handle/2152/31333. doi:10.15781/t29g6z
Pennington, J., Socher, R., & Manning, C. (2014). *Glove: Global vectors for word representation* (Vol. 14). Retrieved from https://nlp.stanford.edu/projects/glove/
Qaiser, S., & Ali, R. (2018). Text mining: Use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications*, *181*(1), 25–29. doi:10.5120/ijca2018917395
Radim, Ř., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA.
Rodriguez, J. J., Kuncheva, L. I., & Alonso, C. J. (2006). Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(10), 1619–1630. doi:10.1109/TPAMI.2006.211
SauceCat. (2018). *PDPbox*. https://pdpbox.readthedocs.io.
Su, C., Xu, Z., Pathak, J., & Wang, F. (2020). Deep learning in mental health outcome research: A scoping review. *Translational Psychiatry*, *10*(1), 116. doi:10.1038/s41398-020-0780-3
Taspinar, A. (2019). *Twitterscraper 1.4.0*. https://github.com/taspinar/twitterscraper.
Turecki, G., & Brent, D. A. (2016). Suicide and suicidal behaviour. *The Lancet*, *387*(10024), 1227–1239. doi:10.1016/S0140-6736(15)00234-2
Vioulès, M., Moulahi, B., Azé, J., & Bringay, S. (2018). Detection of suicide-related posts in Twitter data streams. *IBM Journal of Research and Development*, *62*(1), 7:1–7:12. doi:10.1147/JRD.2017.2768678
Whitlock, J., & Knox, K. L. (2007). The relationship between self-injurious behavior and suicide in a young adult population. *Archives of Pediatrics and Adolescent Medicine*, *161*(7), 634–640. doi:10.1001/archpedi.161.7.634

WHO. (2018). *World health statistics 2018: Monitoring health for the SDGs, sustainable development goals*. Geneva: World Health Organization. https://apps.who.int/iris/handle/10665/272596.

Wongkoblap, A., Vadillo, M. A., & Curcin, V. (2017). Researching mental health disorders in the era of social media: Systematic review. *Journal of Medical Internet Research*, *19*(6), e228. doi: 10.2196/jmir.7215

Zúñiga, H., de Molyneux, L. G., & Zheng, P. (2014). Social media, political expression, and political participation: Panel analysis of lagged and concurrent relationships. *Journal of Communication*, *64*(4), 612–634. doi:10.1111/jcom.12103