*Korean Journal of English Language and Linguistics*, Vol 23, June 2023, pp. 431-447
DOI: 10.15738/kjell.23..202306.431

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

ISSN: 1598-1398 / e-ISSN 2586-7474

http://journal.kasell.or.kr

# Detecting Suicide Notes with the Probability of Positive Sentiment and Interquartile Range

**Yong-hun Lee**(Chungnam National University)



Received: Arpil 3, 2023
Revised: May 17, 2023
Accepted: June 7, 2023

Lee, Yong-hun
Instructor, Dept. of Linguistics, Chungnam Naional University
Tel: 042) 821-6391
Email: yleeuiuc@hanmail.net

## ABSTRACT

**Lee, Yong-hun. 2023. Detecting suicide notes with the probalility of positive sentiment and interquartile range. *Korean Journal of English Language and Linguistics* 23, 431-447.**

This paper proposes a new algorithm for detecting suicide notes using sentiment analysis. As suicides increase nowadays, it is important to detect the suicide signs before the actual suicides are committed. Detecting suicide signs is not so easy, because suicide notes are usually short. This study proposes a modified algorithm of sentiment analysis which is based on the probability of positive sentiments (PPS), not on the categorical classifications. The original BERT model is revised so that the model calculates the PPS values for each sentence. A total of 8 corpora are constructed, among which 4 are the corpora of suicide notes, and the others are novels. For each sentence in the corpora, the PPS values are calculated using the revised BERT model. Then, the distributions of PPSs are statistically analyzed with Interquartile Range (IQR). The suicide notes are distinguished with more than 30 IQR values. In the experiments with the corpora of suicide notes and ordinary texts, the developed method achieves about 86% of accuracy. The proposed algorithm can make use of the sentiment properties of suicide notes, and it is effective not only for large-size corpora but also for small-size suicide notes.

## KEYWORDS

suicide notes, sentiment analysis, BERT, probability of positive sentiment, IQR

## 1. Introduction

As suicides increase nowadays, it is important to detect the suicides signs before the actual suicides are committed. Automatic detection of suicide signs is not so easy, however, because suicide notes are usually short in length. Traditional methods in forensic linguistics may have difficulty detecting suicide notes with such a short length. In addition, those who decide to commit suicide usually leave a suicide note which says the motivation for suicide and their psychological or emotional states.

It is also important to judge the trustworthiness of suicide notes because of the following two reasons. First, fake documents may be used for various forms of criminal activities. Accordingly, it is necessary to analyze the trustworthiness of the suicide notes, because they can be a piece of important evidence in the identification of authenticity, and they can be used as a piece of evidence in various judicial proceedings or individual cases. Second, the suicide notes give us clues about the motivation for suicide and the psychological or emotional states of suicide committers. Many previous studies on suicide notes have been conducted mainly from a psychological point of view (Leenaars 1988, Sheidman 1963), but the studies from a linguistic point of view are still insufficient (Giles 2007, Roubidoux 2012, Shapero 2011).

This paper tries to provide a new algorithm by which we can effectively detect suicide notes using sentiment analysis with deep learning architecture. As for the data, a total of 8 corpora are employed, among which 4 are the corpora of suicide notes and the others are ordinary texts. The 4 corpora for suicide notes include the actual suicide notes of 3 people who actually committed suicide (Virginia Woolf wrote two suicide notes). We also include, in the dataset, four literary works by Virginia Woolf. The reason why Virginia Woolf's literary works are included is that they can be used to investigate whether the proposed algorithm can clearly identify the suicide notes from the other types of ordinary writings, even in the case where the same person writes both types of writings (suicide notes vs. ordinary texts).

The deep learning model which is adopted in this paper is the Bidirectional Encoder Representations from Transformers (BERT, Devlin et al. 2019) model. Among the BERT models, a $\text{BERT}_{\text{LARGE}}$ model was downloaded from the Hugging Face site and it was trained with the IMDB dataset.[1] Although most of the previous studies on sentiment analysis utilize a binary (*positive* or *negative*) or ternary classification (*positive*, *neutral*, or *negative*), this paper employs the probability of the positive sentiment (PPS) for each sentence which can be obtained during the process of determining the sentiment label (*positive* or *negative*). That is, the PPS value is calculated for each sentence in the corpora using the modified $\text{BERT}_{\text{LARGE}}$ model. Then, the distributions of PPS values are numerically analyzed with the IQR values.

To demonstrate the combination of the PPS values plus the IQR values can provide a better result, an experiment is conducted. In the experiment, 200 files are constructed from two kinds of corpora (suicide notes and ordinary texts), and the proposed model is evaluated with precision, recall, *F*-score, and accuracy.

The following are the contributions of this paper:

(1) Contributions of this paper
   a. It proposes a more efficient technique for detecting suicide notes, which utilizes the advanced Transformer model (Vaswani et al., 2017).
   b. The method identifies suicide notes more accurately (about 86%) than previous studies.

[1] https://huggingface.co/

 

c. The method is also applicable to small-size of texts as well as large-size of corpora.
d. The method reveals the basic emotional properties of suicide notes.

Considering that the text size of the suicide notes is relatively small, the method presented in this paper has the advantage of practical application to various types of texts including Twitter and SNS.

## 2. Previous Studies

### 2.1 Forensic Linguistics

Most of the early studies on suicide notes are based on forensic linguistics. Forensic linguistics is a subfield of applied linguistics, which applies linguistic information, analysis techniques, and insights to the setting of areas including law, criminal investigation, court proceedings, judicial procedure, etc. It is particularly a subfield of corpus linguistics, and its purview extends to authorship identification as well as numerous aspects of criminal investigations and judicial processes (such as authorship verification, authorship profiling, and authorship attribution).

The term 'forensic' has historically been used to describe the use of scientific techniques in criminal investigations, which includes the legal requirements for admissible evidence and criminal procedure. The duties of forensic scientists are to gather, preserve, and analyze the evidence. To gather evidence, some forensic scientists go to the scene of the crime; however, others analyze artefacts in a lab. The evidence consists of a variety of tangible and intangible items, including fingerprints, hair, DNA testing, and transfusion analysis. On the other hand, when digital data become the focus of an inquiry, digital forensics gathers and examines the digital data that is closely used in daily life. Usually, digital data is communicated through a network after being saved on digital media. These data are gathered, preserved, and examined by experts in digital forensics, who then present them as evidence in court.

Professor Jan Svartvik introduces the term 'forensic linguistics' in 1968 when he examines Timothy John Evans' writings (a prominent murder suspect). He examines four texts with a variety of linguistic features and discovers significant differences across the texts. It is suggested that the authors of the writings may not be the same. The International Association of Forensic Linguists (IAFL) is established in 1993, and an international journal The Law and the International Journal of Law, Language, and Discourse is started to be published in 1994.

The use of linguistic expertise in forensic contexts can be divided into roughly three categories: (i) providing linguistic evidence for court rulings; (ii) comprehending language use in the judicial processes; and (iii) understanding lexis and languages in the law. More specifically, forensic linguistics deals with authorship attribution and plagiarism, forensic phonetics, speaker identification, courtroom interaction, the language of legal documents, the language of the police and law enforcement, interviews with minors and other vulnerable witnesses in the legal system, linguistic evidence in courtrooms, and expert witness testimony (Coulthard et al. 2016). The field of forensic linguistics is not uniform in nature, and a variety of specialists and researchers from other fields are involved.

By applying linguistic knowledge to a specific (social) setting (i.e., a legal scenario), forensic linguistics is defined by Olsson (2004) as being situated at the intersection of language, crime, and law. Olsson (2008) asserts that any spoken or written words that are referenced in legal or criminal proceedings may be considered forensic texts. The study also points out that 'an analysis of the suicide notes' must be included in the investigation of

 

forensic linguistics since a suicide note typically contains lines that suggest a means to kill oneself. Investigating the veracity and intent of suicide notes requires the application of linguistic knowledge to the examination of suicide notes.

### 2.2 Studies on Suicide Notes Using Forensic Linguistics

Numerous investigations on suicide notes have been conducted as forensic linguistics advances. Suicide is divided into four types of categories by Durkheim (1951), based on integration and regulation. The first type is referred to as 'egoistic suicide,' which happens when those who commit suicide are not accepted by their social group and do not receive any support from it. People who commit this kind of suicide frequently feel alone or despondent, when they are facing challenging circumstances. The second type, known as 'altruistic suicide,' takes place when there is a very low level of social cohesion. It is a type of suicide that is often committed by people who have close ties to a community and are valued for their unique identities. People who commit this kind of suicide typically disregard their own needs in favor of the interests of the group. The third type of suicide is known as 'anomic suicide,' which happens when there is insufficient social control, when people experience chaos or an unanticipated economic collapse, when they lose the moral standards or socially accepted values that guide their behavior, or when they lose a lover due to a breakup or death. People who commit this kind of suicide do so because a significant life shift has overtaken them. The fourth option is 'fatalistic suicide,' which happens as a result of extreme tyranny of despair. People believe they have no future and that authorities are stifling their freedom.

Suicide may come from unmet psychological demands, according to Sheidman and Faberow (1963). The psychological research divides the causes of suicide into five groups: (i) dissatisfied love or possessions; (ii) disorganized or disordered control power (related to their own accomplishment); (iii) insulted image of oneself; (iv) failure in relationships with others that make one feel depressed; and (v) excessive rage and aggression that is brought on by unfulfilled desire. He contends that 'extreme psychache,' or intolerable psychological distress, is the root cause of all suicides.

Chaski (2012) lists a few words that frequently appear in suicide notes. They express apology (*I'm sorry* or *Please forgive me*), love (*I love you* or *I cannot live without you*), anger (*I cannot please you* or *I hope you are happy now*), complaint (*The situation is not acceptable* or *I can no longer tolerate*), or psychological shock (since the divorce). According to Chaski (2012), suicide notes only contain one to four different writing styles rather than the whole form of writing.

According to Sboev et al. (2015), psycholinguistic cues can be used to determine the psychological state of the author(s) at the time of writing. These markers include pronouns, nouns, adjectives, verbs, and adverbs, as well as word counts (also known as word tokens in corpus linguistics), the average number of words used in a sentence (also known as the mean sentence length), the ratio of nouns to verbs, the number of exclamation points, the number of emoticons, and others.

The suicide notes have typical (linguistic) characteristics from a forensic linguistics perspective. It is necessary to consult all of the theoretical linguistics in order to analyze suicide notes in forensic linguistics (including phonetics, phonology, morphology, syntax, semantics, pragmatics, and discourse analysis). The majority of the early linguistic research on suicide notes is based on a corpus assembled (Sheidman and Faberow 1963). There are 66 writings in the corpus, which is a mix of real and fictional suicide notes. The messages are linguistically studied using discourse analysis techniques, the usage of different auxiliary verbs (including modals), or verbs that distinguish between the genuine and the fake.

 

Some studies have used discourse analysis (Edelman and Renshaw 1982) or semantic space (Leenaars, 1988) to examine suicide notes. Using the so-called Syntactic Language Computer Analysis, Edelman and Renshaw (1982) examines not just syntactic elements like parts of speech (POS), but also the semantic qualities of nouns, verbs, adjectives, and other words. (Pestian et al. 2008, Pestian et al. 2010).

There are currently multiple studies using corpus linguistics and machine learning methods to analyze suicide notes. These studies usually involve statistical analysis of different linguistic units such as personal pronouns, past tense verbs, nouns, and various semantic categories (Olsson 2008). For the detection of suicide notes, there has been a strong trend in recent years to rely on automated corpus analysis tools. Scholars use corpus-analysis methods (Shapero 2011) or automated machine-learning techniques to recognize and categorize typical suicide notes (Pestian et al. 2010).

Such experiments produced the Linguistic Inquiry and Word Count (LIWC) program (Pennebaker et al. 2001, Tausczik and Pennebaker 2010). Numerous academic fields, including computer science and psycho-linguistics, have made extensive use of the application. The software uses 72 linguistic features to assess the writings of regular people and gather statistics about certain semantic word patterns (Pennebaker and King 1999). The Standard Linguistic Dimension, Psychological Process, Relativity, and Personal Concerns categories can be used to categorize the variables (i.e., linguistic components) in the LIWC. The program divides the 72 variables into the aforementioned four categories based on their frequencies. The ratios of words in each category reveal the psychological processes and states of the writing's author. LIWC includes more information. The percentages of words in each category show the writer's psychological state and writing processes. LIWC includes more than 3,000 content words that are used often in daily life, as well as various different kinds of function words (article, preposition, first/second/third personal pronouns), word length, and word types (Pennebaker et al. 2001). It is feasible to compile the analysis results in numerous individual variations in the psychological sectors, which cannot be obtained from the prior studies, by counting function words and pronouns.

### 2.3 Machine Learning/Deep Learning Approaches to Suicide Notes

Machine learning means "fields of study that give computers the ability to learn without being explicitly programmed" (Samuel 1959). Mitchell (1997: 2) also mentions as follows: "A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*." By combining these two well-known definitions, machine learning can be defined as an area of computer science, particularly within artificial intelligence, that enables a machine (a computer) to automatically learn from training data (which can be referred to as *E*) and perform a class of tasks *T* (either classification or regression) with the performance metric *P*.

There are some previous studies which utilized machine learning/deep learning methods to investigate suicide notes. First of all, Pestian et al. (2012a) reports on the construction of a suicide note corpus that can be used for machine learning. The corpus includes 1,278 notes, which are written by people who commit suicide.

Pestian et al. (2012b) reports on a shared task involving the assignment of emotions to suicide notes. This study is unique compared to earlier shared projects in the biomedical field in two ways. One is that it leads to the creation of a corpus of entirely anonymous clinical writing and annotated suicide notes. The fact that the work needs categorization in relation to a broad range of labels is another important aspect of it. Much more people participate in this endeavor than in any previous biological challenge. They provide a preliminary analysis of the outcomes and outline the data creation process and evaluation metrics. Many systems perform at levels that are close to the inter-coder agreement, indicating that existing technologies are capable of achieving human-like performance on

 

this test.

Yang et al. (2012) discusses the development of automatic algorithms that can recognize the affective text of 15 particular emotions at the sentence level. This study highlights their efforts in developing automatic systems that can recognize the emotive text of 15 various emotions from suicide notes at the sentence level. They make a hybrid model that combines various NLP methods, such as lexicon-based keyword detection, CRF-based emotion cue identification, and machine learning-based emotion categorization. Different vote-based merging algorithms are used to combine the results produced by various methodologies. With a micro-averaged *F*-measure score of 61.39% in textual emotion recognition, the automated system does well compared to the manually annotated gold standard and produced positive results, placing first out of 24 participating teams in this challenge. The outcomes show that efficient emotion recognition by an automated system is feasible in the presence of a sizable annotated corpus.

McCart et al. (2012) focuses on sentiment analysis, determining if 15 emotions (labels) were present at the same time in a collection of suicide notes spanning more than 70 years. They investigate several strategies combining conventional expression-based rules, statistical text mining (STM), and a strategy that applies weights to text while considering multiple labels. The top entry utilized a combination of rules and STM models to produce a micro-averaged *F1* score of 0.5023, which was just over the average of the 26 competing teams.

Wang et al. (2012) presents their solution for the i2b2 sentiment classification challenge. Their hybrid approach combines rule-based classifiers and machine learning. They study several lexical, syntactic, and knowledge-based features for the machine learning classifier and conduct experiments to demonstrate how much each feature affects the classifier's performance. They suggest an approach to automatically extract useful syntactic and lexical patterns from training data for the rule-based classifier. According to the experimental findings, the rule-based classifier outperforms the machine learning classifier that uses unigram characteristics as a starting point. The hybrid system achieves the greatest micro-averaged *F*-measure (0.5038), which is superior to the mean (0.4875) and median (0.5027) micro-average *F*-measures among all participating teams, by merging the machine learning classifier and the rule-based classifier.

Desmet and Hoste (2013) explores whether recent developments in sentiment mining and natural language processing can be utilized to precisely identify 15 different emotions that might be a sign of suicidal conduct. Binary support vector machine (SVM) classifiers are used to create a system for automatically detecting emotions. Using bootstrap resampling, the ideal feature combination for each of the various emotions is identified. The accuracy of the classification varies amongst moods, with values reaching 68.86% of the *F*-score. They demonstrate that classifier tuning and a combined lexico-semantic feature representation are beneficial for fine-grained autonomous emotion recognition. They conclude that suicide prevention may one day benefit from the use of natural language processing tools.

In order to perform emotion identification on the well-curated dataset, Ghosh et al. (2020) establishes a fine-grained emotion-annotated corpus (CEASE) of English suicide notes. The corpus (version 1) includes 2,393 sentences from approximately 205 suicide notes that were gathered from a variety of sources. A set of 15 different fine-grained emotion labels (forgiveness, happiness, peacefulness, love, pride, hopefulness, thanksgiving, blame, anger, fear, abuse, sadness, hopelessness, guilt, information, and instructions), are used to mark each line with a certain emotion class. They create an ensemble architecture for the evaluation, including the three supervised deep learning models as the basic models: Convolutional Neural Network (CNN), Gated Recurrent Unit (GRU), and Long Short-Term Memory (LSTM). They achieve the maximum cross-validation accuracy of 60.32% and test accuracy of 60.17%.

 

## 3. Research Method

### 3.1 Corpus Compilation

In order to detect suicide notes using sentiment analysis, two different kinds of corpora are necessary. One is the corpus of suicide notes, and the other is a corpus of ordinary texts against which suicide notes are compared.

For this purpose, in this paper, the following 8 texts are employed. Table 1 enumerates the number of sentences in each text/corpus. Because sentiment analysis is conducted on each sentence, it is necessary to examine the number of sentences in each text/corpus.

**Table 1. Number of Sentences per Corpus**

| Corpus | # of Sentences | Corpus | # of Sentences |
|---|---|---|---|
| 01.Noh | 14 | 05.TheVoyageOut | 8,088 |
| 02.Cobain | 34 | 06.NightAndDay | 7,531 |
| 03.Woolf-S | 13 | 07.Jacob'sRoom | 3,413 |
| 04.Woolf-H | 22 | 08.MondayOrTuesday | 1,142 |

Among these 8 corpora, the first 4 in the left part of the table are corpora for suicide notes, and the other 4 in the right part of the table are corpora for ordinary texts. All of the corpus files came from the study by Lee and Joh (2019). Among these corpora, corpora 03~08 are texts which were written by the same author, Virginia Woolf. The reason why they are included in the study is that it is necessary to examine whether the suicide notes demonstrate different properties from the ordinary texts which are written by the same author.

The details of each corpus are as follows. 01 is a suicide note written by a Korean politician.[2] It is originally written in Korean, but it is translated into English for investigation. 02 is a suicide note of Kurt Cobain, who is a musician and commits suicide in 1994. 03 and 04 are suicide notes by Virginia Woolf. The former is a letter written to her sister and the latter to her husband. 06~09 are literary works (novels) written by Virginia Woolf.

From this table, we can make the following two hypotheses.

(2) Hypotheses
   a. Hypothesis 1: The corpora 01~04 will demonstrate different properties from the corpora 05~08.
   b. Hypothesis 2: The corpora 03~04 will show different properties from the corpora 05~08, though they are written by the same author.

We will return to these hypotheses in Section 6.

### 3.2 Ternary Classifications of Sentiment Analysis

The early approach to sentiment analysis produces the outputs of either binary or ternary classifications. The binary classifications group the sentiment in either *positive* or *negative*. In the ternary classifications, sentiment analysis classifies emotions into *positive*, *neutral*, and *negative*. More recent studies classify the sentiment into multi-classes.

[2] '01.Noh' was a hyman translation. Since the length of the suicide note was short, it was manually translated.

 

In this paper, a ternary classification of sentiment analysis is also employed for the comparisons with the analysis method in Section 3.3. A ternary classification has been one of the traditional analysis methods in sentiment analysis, and it was also presented for the comparisons of this method and the analysis method in Section 3.3 For this purpose, a pre-trained BERT model is downloaded from the Hugging Face site, and all the sentences in the corpora are classified into three groups (*positive*, *neutral*, and *negative*). Then, to be represented in a box plot, each label is converted into numbers: positive into 1, neutral into 0, and negative into -1. Then, the results are represented in a box plot.

### 3.3 Sentiment Analysis with the PPS Values

The method that this paper proposes is not the sentiment analysis using the ternary classifications but using the probability of positive sentiment (PPS). This section presents how this analysis method is made.

First, a pre-trained $\text{BERT}_{\text{LARGE}}$ model has to be downloaded from the Hugging Face site.[3]

Second, this deep learning model was trained with the IMDB dataset, which is called the $\text{BERT}_{\text{LARGE}}$-IMDB model.

Third, the PPS value is calculated for each sentence in the test set of the IMDB dataset by the algorithm in Section 3.4.

Fourth, the $\text{BERT}_{\text{LARGE}}$-IMDB model was evaluated as follows. If the PPS value is less than 50, the given sentence is labelled with *negative*. If not, the given sentence is labelled with *positive*. Then, the obtained labels are compared with the original labels in the IMDB dataset. We obtain over 98% of accuracy.

Fifth, the PPS values are calculated for all the sentences in the corpora of Table 1.

Finally, the distributions of PPS values are represented with a box plot, and the IQR is calculated for each sentence.

### 3.4 Algorithm for Obtaining the Probabilities of Positive Sentiment

The fundamental architecture of the BERT model serves as the algorithm's starting point. Following the processing of the input sentence by the BERT model, the model generates a class label that is either TRUE or FALSE.

[3] https://huggingface.co/bert-large-uncased

 

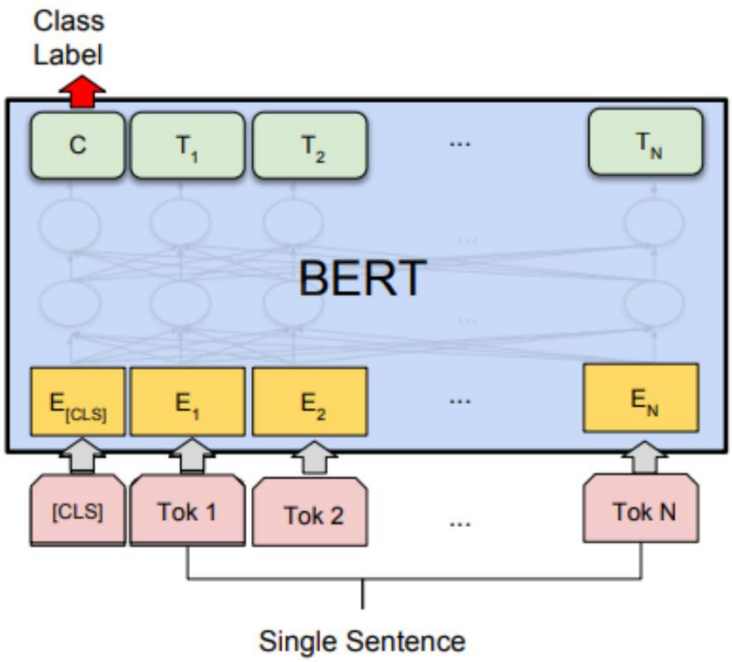


**Figure 1. BERT Model with Single Sentence**

In this study, the final output section is changed such that the model can now return not a class label but the probability that the given sentence will be TRUE (*positive*). During the determination of the class label, the $\text{BERT}_{\text{LARGE}}$-IMDB model returns a tensor which contains the probability of FALSE (0) and TRUE (1). Usually, they are *logit values*. Then, they are converted into probability using the *inverse logit function*. Then, the probability of TRUE is taken, and this value is converted into the percentage value, which ranges from 0 to 100.[4]

## 4. Analysis Results

### 4.1 Ternary Classifications plus Box Plot

The following is the box plot which shows the distributions of sentiments of ternary classifications. Remember that the first 4 are corpora for suicide notes and the others are corpora for ordinary texts. Here, the black dots are used for the mean values, and the thick black lines indicate the medians. Note that the sentiment labels *positive*, *neutral*, and *negative* are inverted into numbers 1, 0, and -1 respectively (Section 3.2).

[4] A similar algorithm was also employed in Lee (2021) and Lee (2022), but the syntactic acceptabilities, not the sentiment, were measured in these papers.

 

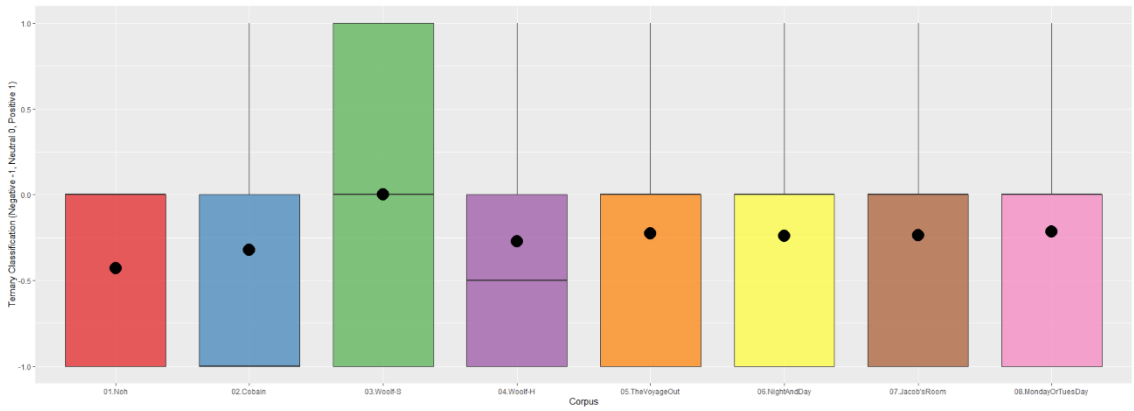


**Figure 2. Box Plot for Ternary Classifications**

As shown in the box plot, though the distributions of sentiment labels in suicide notes (01~04) are slightly different from those in ordinary texts (05~08), it is difficult to identify the former from the latter. There are no typical properties which can be used for the identification of suicide notes in this box plot. It is also difficult to use the mean values or medians since there is no consistent tendency in the box plot for suicide notes. Especially, note 03. As shown in the box plot, sentiments are evenly distributed around 0, and the mean and median are also located around 0.

### 4.2 PPS Distributions plus Box Plot

The following box plot displays the sentimental distributions of the PPS values.

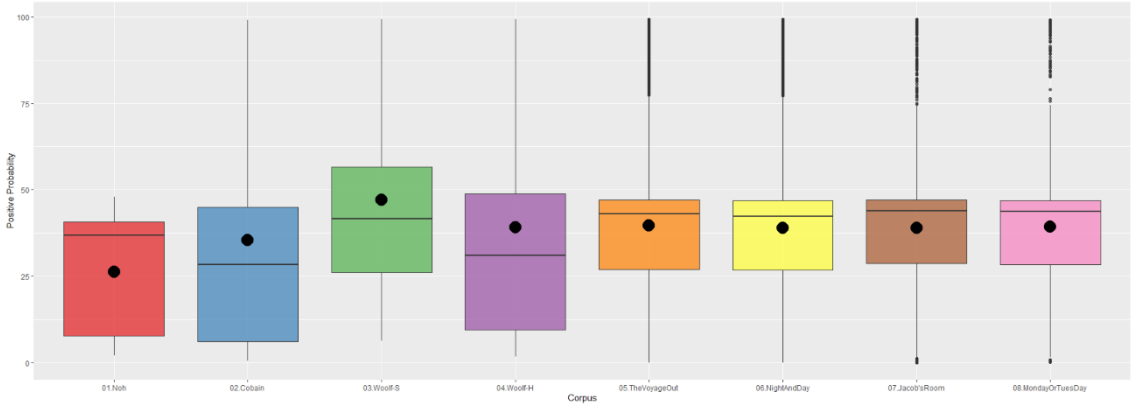


**Figure 3. Box Plot for PPS Values**

Here, two interesting properties are observed in the suicide notes (01~04). First, the medians of the suicide notes are generally lower than those of the ordinary texts (05~08). It indicates that pessimistic sentences are used much more in suicide notes. Second, the box sizes (from the 1st quantile to the 3rd quantile) are bigger in the suicide notes than in the ordinary texts. It implies that there are more emotional fluctuations in suicide notes than in ordinary texts. Also, note that the mean and median values of 03 in Figure 3 are slightly lower than those in Figure 2. It indicates that the overall emotional distribution is positively skewed a little in the box plot of Figure 3.

 

However, it is also difficult to utilize the mean values or medians to identify suicide notes. As for the mean values, even though the mean values of the ordinary texts (05~08) are similar to each other, the mean values of suicide notes are around the mean values of the ordinary texts. Some values (such as 03) look higher than the values of the ordinary texts, but others (such as 01 or 05) are located below the means of ordinary texts.

### 4.3 Suicide Notes and Statistical Analysis

There are two fundamental problems to use statistics in the detection of suicide notes: normality and data size.

In the ternary classification, since the output is categorical, the data do not follow the normal distribution. Even when the PPS values are adopted in the analysis, the data do not follow the normal distribution, as the distributions in Figure 3 indicate. It prevents us to use parametric tests in the statistical analysis.

If non-parametric tests are employed, another problem arises: data size. As shown in Table 1, since the size of suicide notes is very small, most statistical values which are obtained from the suicide notes are hard to satisfy the conditions of many non-parametric tests. For example, it is doubtful to compare the mean OR median values of such small-size corpora (01~04) with those of large-size corpora (05~08).

Then, isn't there any method which can be used for detecting suicide notes? Probably, the interquartile range (IQR) can be used for comparison, because (i) it is a statistic which can be used for non-parametric tests and (ii) it can be applied regardless of data size. The following table enumerates the IQR values of each corpus.

**Table 2. IQR per Corpus**

| Corpus | IQR | Corpus | IQR |
|---|---|---|---|
| 01.Noh | 32.971 | 05.TheVoyageOut | 20.188 |
| 02.Cobain | 38.760 | 06.NightAndDay | 20.120 |
| 03.Woolf-S | 30.605 | 07.Jacob′sRoom | 18.320 |
| 04.Woolf-H | 39.367 | 08.MondayOrTuesDay | 18.434 |

As shown in this table, the IQR values of suicide notes are greater than 30, whereas those of ordinary texts are around 20. Actually, the comparisons with the IQR values utilize the property that the distances from the 1st quantile to the 3rd quantile are much bigger in suicide notes than those in ordinary texts (See Figure 3). These differences between the two groups are statistically significant, and this fact indicates that the IQR values may be adopted to numerically distinguish suicide notes from ordinary texts. It also implies that the sentiment analysis with the PPS values has to be conducted to get the distributions, rather than the analysis with ternary classifications.

## 5. Experiment

### 5.1 Dataset

The detecting algorithm of suicide notes using the IQR of PPS values obviously identifies the suicide notes from ordinary texts. One significant problem is the question of whether it is possible to compare the analysis results of the small-size texts (01~04) with those of the large-size corpora (05~08). As noted in Table 1, the number of sentences in the former types of corpora is less than 50, whereas that of the others are thousands or more than ten thousand. It may be unreasonable to compare the analysis results of 01~04 with those of 05~08. To solve this

 

problem, we design an experiment.

From each type of corpus, 20 sentences are randomly selected, and the extracted sentences form a single file. The process continues until each type of corpus contains 100 files. By this procedure, two groups of files are collected (100 files for suicide notes and the other 100 files for ordinary texts), and each text includes exactly 20 sentences. Then, the PPS values are calculated for all the sentences in each file using the algorithm in Section 3.4. Finally, the IQR is calculated for each file, and the analysis results of suicide notes are compared with those of ordinary texts. During the process, the proposed algorithm is evaluated with precision, recall, *F*-score, and accuracy.

### 5.2 Evaluation with IQR values

The distributions of IQR values of two types of corpora are as follows.

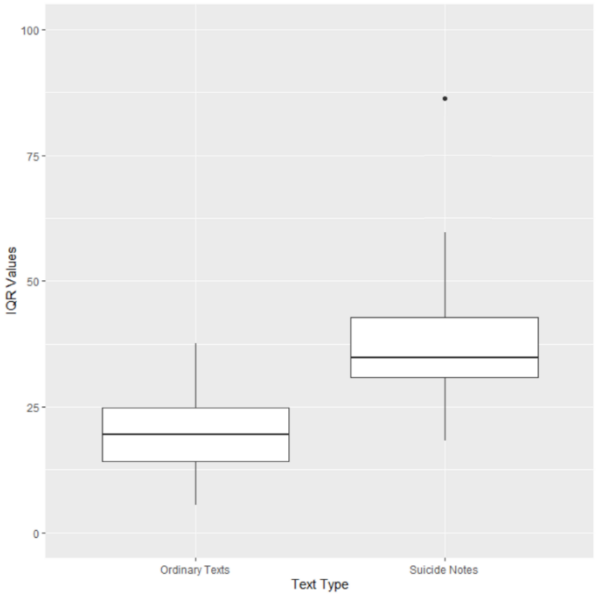


**Figure 4. Box plot for IQR Values**

As shown in this box plot, the IQR values of suicide notes are much higher than those of suicide notes (*mean*(ordinary texts)=20.190, *mean*(suicide notes)=39.147, *median*(ordinary texts)=19.535, *mean*(suicide notes)=34.831). In order to examine the differences between the two distributions, a Mann-Whitney test is conducted, and the result is that these two distributions are significantly different ($U = 9277$, $p < .001$).

Then, based on the following table, precision, recall, *F*-measure, and accuracy are measured.

**Table 3. Evaluation with IQRs**

| | | Predicted | |
|---|---|---|---|
| | | Suicide Notes | Ordinary Texts |
| Actual | Suicide Notes | 79 | 21 |
| | Ordinary Texts | 7 | 93 |

 

The evaluation metrics are calculated from this table as follows: *precision* 0.790, *recall* 0.919, *F-measure* 0.849, and *accuracy* 0.860. That is, although each file contains only 20 sentences, the proposed algorithm using PPS values plus IQR values correctly divides the texts into two groups (suicide notes and regular texts) with an accuracy rate of 86.0%.

### 5.3 Comparison with Previous Studies

The following table compares the performance of the proposed method with those of previous studies.

**Table 4. Evaluation Summary**

| Model | *F*-score | Accuracy |
|---|---|---|
| Yang et al. (2012) | 61.39 | |
| McCart et al. (2012) | 50.23 | |
| Wang et al. (2012) | 50.38 | |
| Desmet, M., & Hoste (2013) | 68.86 | |
| Ghosh et al. (2020) | | 60.71 |
| **PPS Values + IQR** | **84.90** | **86.00** |

As this table illustrates, the detecting algorithm of suicide notes with PPS values shows a much higher *F*-score and accuracy. It implies that the proposed algorithm performs very well in detecting suicide notes from ordinary texts.

## 6. A Discussion

### 6.1 Implications

In this paper, a new method is proposed which is based on the PPS values. After the $\text{BERT}_{\text{LARGE}}$ model is trained with the IMDB dataset, the PPS values are obtained by converting the logit values into the PPS values. Then, the PPS values are graphically represented with a box plot. Then, this study adopted the IQR for the detection of suicide notes. As Table 2 indicates, the IQRs of suicide notes are greater than 30, while those in ordinary texts are less than 30. This criterion utilizes the property of suicide notes that suicide committers demonstrate a high fluctuation of sentiment (or a wider range of sentiments), which is clearly illustrated in the box plot of Figure 3.

This criterion for detecting suicide notes (with the IQR value) is not only available in large size of corpora but also applicable to small-size of suicide notes. In Section 5, the datasets are constructed from two different corpora. One is from the suicide notes, and the other is from the ordinary texts. From these two types of corpora, 20 sentences are randomly selected respectively and they compose a single file. A total of 100 files are constructed in the experiment for each corpus. Then, the composed files are analyzed with IQR values. If the IQR value is less than 30, the given file is sentenced to be an ordinary text. If not, it is classified as a suicide note. In the experiment, we obtained 86.00% of accuracy by this criterion. It implies that this criterion for detecting suicide notes (the IQR values of PPS) is also applicable to small-sized suicide notes in addition to the big size of corpora.

The proposed method in this paper utilizes or actually reveals some important properties of suicide notes. The first is that the overall sentiments of suicide notes are more negative than those of ordinary texts. This tendency is predictable. The suicide committers may have very pessimistic toward their environments when they decide to

 

commit suicide, and this pessimistic tendency is reflected in the emotional distributions of suicide notes. The negative sentiment can be detected in the box plot in Figure 3. The boxes of suicide notes (01~04) are located a little lower than those of ordinary texts (05~08).

The second property of suicide notes is that they may show a wide range of sentiment spectrum. Although most sentences are emotionally *negative* in suicide notes, some sentences may be highly positive. Such sentences may appear when suicide committers remember happy days in their life. Accordingly, the distributions of sentiment may have a wide range of the spectrum. Although this property is visually shown in the box plot in Figure 3 (where the boxes for suicide notes are bigger), it is more effectively represented in the IQR values. As mentioned earlier, the IQR values of suicide notes are greater than 30, whereas those of ordinary texts are smaller than 30. These numerical values indicate that suicide notes can express a wide range of emotions.

Another interesting property is that, despite being written by the same author, the suicide notes exhibit different characteristics from regular writings. As mentioned in Section 3.1, Virginia Woolf is the same author of the works 03~04 and 05~08. As seen in Figure 3, the emotional patterns of 05~08 are notably different from 03~04. In addition, the former's IQR values are greater than 30, whilst the latter's are fewer than 30 (Table 2). Even though these two very different sorts of texts are produced by the same person, it shows that the emotional patterns of suicide notes are very distinct from those of ordinary texts.

As noted in Section 2, earlier studies on suicide notes are investigated with forensic linguistics (Coulthard and Johnson 2016, Olsson 2004, Olsson 2008, Svartvik 1968). Some studies such as Chaski (2012) make use of expressions such as apology, love, anger, complaint, or psychological shock. Pennebaker and King (1999) adopts some kinds of linguistic units, and some scholars employ discourse analysis (Edelman and Renshaw 1982) or semantic space (Matykiewicz et al. 2009). Pestian et al. (2008) and Pestian et al. (2010) utilize the LIWC which includes various kinds of linguistic features. These kinds of methods, however, may not be available in the investigations of suicide notes, since the length of the text is usually short in suicide notes. Some linguistic features of forensic linguistics may appear in such a short text, but many linguistic features may not be included in the analysis owing to the short length of the texts. However, sentiment can be measured for every sentence regardless of the length of text, and the detecting algorithm with sentiment analysis may extract more reliable data from such a short text. In addition, still further studies are necessary to investigate how the distributions of sentiments (the distributions of PPS values) can be related to the classifications in Durkheim (1951) or Shneidman (1996).

Now, return to the hypotheses in Section 1. The first hypothesis is that the suicide notes will demonstrate different properties from the ordinary texts. This hypothesis is supported by the IQR values in Table 2. The suicide notes have IQR value greater than 30. Ordinary texts, on the other hand, have less than 30 IQR values. The second hypothesis is that, although they are written by the same author, the suicide notes will show different properties from ordinary texts. This hypothesis is also supported. In Figure 3, the emotional patterns and the IQR values of 03~04 are significantly distinguished from 05~08. It indicates that the emotional patterns of suicide notes are highly different from those of ordinary texts, even if these two different types of texts are written by the same author.

### 6.2 Applications and Limitations

As mentioned in the previous sections, the developed algorithm in this paper is applicable not only to the large size of corpora but also to the small size of suicide notes. Then, the proposed method can be used not only to detect suicide signs in offline texts but also to identify the suicide signs in online texts such as Twitter or SNS. As mentioned in Section 1, suicide notes are usually short in length. Because of such a short length, it is not easy to

 

apply some linguistic properties of forensic linguistics (Section 2.1 and Section 2.2). Since the developed algorithm in this paper is available also for the small size of texts, it is very useful to apply the algorithm to detect suicide (warning) signs in online texts.

Next, note that the first corpus in our study 01 is a Korean politician's suicide letter. Although it was initially written in Korean, a translation into English was done for research. As noted in Figure 3 and Table 2, this corpus demonstrates the typical characteristics of suicide notes. Then, it implies that the proposed algorithm is also available to the English translations of other languages. It opens a way that the detection algorithms can be applied to English translations.

The developed algorithm in this paper, however, has two important shortcomings. First, there are some gaps between precisions and recalls in the evaluation: precision 0.790 and recall 0.919. Low precision here indicates that there are many false positives (the texts which are actually suicide notes but wrongly classified as ordinary texts). The reason seems to be originated from the short length of the texts. Accordingly, it is necessary to develop new methods which can overcome this problem. Second, in this paper, we examine the data with the IQR values. Though this method is workable, it would be better to have a method by which we can directly compare the distributions of PPS values among corpora. Since the PPS values do not follow the normal distribution, we cannot use the *F*-tests to compare the distributions of data, which are parametric tests. it would be better to have a non-parametric counterpart of *F*-test, which can be applied to the multi-modal distributions and which is not available in current statistics.

## 7. Conclusion

This paper proposes a new type of algorithm for detecting suicide notes using sentiment analysis. For this purpose, 8 corpora are constructed, among which 4 were the corpora of suicide notes. For each sentence in the corpora, the probability of positive sentiment (PPS) was calculated using the $\text{BERT}_{\text{LARGE}}$ model. Then, the PPS values are visually represented with a box plot and numerically analyzed with the IQR.

The results show that, while ordinary texts have less than 30 IQR values, suicide notes have more than 30 IQR values. It was also discovered that the proposed algorithm in this study can be used to analyze texts of small size, not just available to huge corpora. Another intriguing discovery is that, despite being written by the same person, suicide notes differ from other ordinary texts in terms of style. We think that the proposed method can be utilized to detect suicide signs of various kinds of online and offline suicide notes. We also hope that more advanced deep learning technologies will contribute to saving more lives.

## References

Chaski, C. 2012. Author identification in the forensic setting. In L. Solan and P. Tiermsa, eds., *The Oxford Handbook of Forensic Linguistics*, 333-372. Oxford, UK: Oxford University Press.

Coulthard, M. and A. Johnson. 2016. *An Introduction to Forensic Linguistics*. Cambridge, MA: Cambridge University Press.

Desmet, M. and V. Hoste. 2013. Emotion detection in suicide notes. *Expert Systems with Applications* 40(16), 6351-6358.

Devlin, J., M. Chang, K. Lee. and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for

 

language understanding. arXiv Preprint arXiv:1810.04805.

Durkheim, E. 1951. *Suicide*. New York: The Free Press.

Edelman, A. and L. Renshaw. 1982. Genuine versus simulated suicide notes: An issue revisited through discourse analysis. *Suicide and Life-Threatening Behavior* 12(2), 103-113.

Ghosh, S., A. Ekbal, and P. Bhattacharyya. 2020. CEASE: A corpus of emotion-annotated suicide notes in English. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 1618-1626. Marseille, France, 11-16 May 2020.

Giles, S. 2007. *The Final Farewell*: *Using a Narrative Approach to Explore Suicide Notes as Ultra-social Phenomenon*. Unpublished doctoral dissertation, University of Liverpool, Liverpool, England.

Lee, Y. 2021. English Island Constraints Revisited: Experimental vs. Deep Learning Approach. *English Language and Linguistics* 27(3), 21-45.

Lee, Y. 2022. Negative Polarity Items in English: A Deep Learning Model and Statistical Analysis. *Korean Journal of Linguistics* 47(1), 29-56.

Lee, Y. and G. Joh. 2019. Identifying suicide notes using forensic linguistics and machine learning. *The Linguistic Association of Korean Journal* 27(2), 171-191.

Leenaars, A. 1988. *Suicide Notes*. New York: Human Sciences Press.

Matykiewicz, P., D. Wlodzislaw. and J. Pestian. 2009. Clustering semantic spaces of suicide notes and newsgroups articles. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, 179-184. Boulder, Colorado, USA, 5 June 2009.

McCart, J., D. Finch, J. Jarman, E. Hickling, J. Lind, M. Richardson, D. Berndt. and S. Luther. 2012. Using ensemble models to classify the sentiment expressed in suicide notes. *Biomedical Informatics Insights* 5(1), 77-85.

Mitchell, T. 1997. *Machine Learning*. New York: McGraw Hill.

Olsson, J. 2004. *Forensic Linguistics*: *An Introduction to Language, Crime, and the Law*. London, UK: Continuum.

Olsson, J. 2008. *Forensic Linguistics*: *An Introduction to Language, Crime, and the Law*, 2nd edition. London, UK: Continuum.

Pennebaker, J. and L. King. 1999. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology 1999* 77(6), 1296-1312.

Pennebaker, W., E. Francis. and J. Booth. 2001. *Linguistic Inquiry and Word Count (LIWC)*: *LIWC2001*. Mahwah, NJ: Lawrence Erlbaum Associates.

Pestian, J., P. Matykiewicz, J. Grupp-Phelan, A. Lavanier. J. Combs. and R. Kowatch. 2008. Using natural language processing to classify suicide notes. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing* (*BioNLP'08*), 96-99. Columbus, Ohio, USA, 19 June 2008.

Pestian, J., P. Matykiewicz. and M. Linn-Gust. 2012a. What's in a note: Construction of a suicide note corpus. *Biomedical Informatics Insights* 5(5), 1-6.

Pestian, J., P. Matykiewicz, M. Linn-Gust, B. South, O. Uzuner, J. Wiebe, B. Cohen, J. Hurdle. and C. Brew. 2012b. Sentiment analysis of suicide notes: A shared task. *Biomedical Informatics Insights* 5(1), 3-16.

Pestian, J., H. Nasrallah, P. Matykiewicz, A. Bennett. and A. Leenaars. 2010. Suicide note classification using natural language processing: A content analysis. *Biomedical Informatics Insights* 3(3), 19-28.

Roubidoux, S. 2012. *Linguistic Manifestations of Power in Suicide Notes*: *An Investigation of Personal Pronouns*. Unpublished doctoral dissertation, University of Wisconsin at Oshkosh, Oshkosh, Wisconsin, USA.

Samuel, A. 1959. Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal* 3, 210-229.

Sboev, A., D. Gudovskikh, R. Rybka, and I. Moloshnikov. 2015. A quantitative method of text emotiveness

 

evaluation on base of the psycholinguistic markers founded on morphological features. *Procedia Computer Science* 66, 307-316.

Shapero, J. 2011. *The Language of Suicide Notes*. Unpublished doctoral dissertation, University of Birmingham, Birmingham, UK.

Sheidman, E. and N. Faberow. 1963. *Clues to Suicide*. New York: McGraw-Hill.

Shneidman, S. 1996. *The Suicidal Mind*. Oxford, UK: Oxford University Press.

Svartvik, J. 1968. *The Evans Statements*: *A Case for Forensic Linguistics*. Gothenburg, Sweden: University of Gothenburg Press.

Tausczik, Y. and J. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29(1), 24-54.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser. and I. Polosukhin. 2017. Attention is all you need. arXiv Preprint arXiv:1706.03762.

Wang, W., L. Chen, M. Tan, S. Wang. and A. Sheth. 2012. Discovering fine-grained sentiment in suicide notes. *Biomedical Informatics Insights* 5(1), 137-145.

Yang, H., A. Willis, A. De Roeck. and B. Nuseibeh. 2012. A hybrid model for automatic emotion recognition in suicide notes. *Biomedical Informatics Insights* 5(1), 17-30.

Examples in: English
Applicable Languages: English
Applicable Level: Tertiary

 