Finding People with Emotional Distress in Online Social Media: A Design Combining Machine Learning and Rule-based Classification (Research Note)

Abstract: Many people face the problems of emotional distress and suicidal ideation. About 9.2% of people worldwide have had suicidal ideation at least once in their lifetime and 2% have had that in the past 12 months (Borges et al. 2010). There is increasing evidence that the Internet and social media can influence suicide-related behavior (Luxton et al., 2012). In particular, a trend appears to be emerging in which people leave messages showing emotional distress or even suicide notes on the Internet (Ruder et al., 2011). Identifying distressed people and examining their postings on the Internet are important steps for health and social work professionals to provide assistance, but the process is very time-consuming and ineffective if conducted manually using standard search engines. Following the design science approach, we present the design of a system called KAREN, which identifies individuals who blog about their emotional distress in the Chinese language, using a combination of machine learning classification and rule-based classification with rules obtained from experts. A controlled experiment and a user study were conducted to evaluate the performance of the system in searching and analyzing blogs written by people who might be emotional distressed. The results show that the proposed system achieved better classification performance than the benchmark methods, and that professionals perceived the system to be more useful and effective for identifying bloggers with emotional distress than benchmark approach.

Keywords: social media, emotional distress, suicide research, design science, classification

Finding People with Emotional Distress in Online Social Media: A Design Combining Machine Learning and Rule-based Classification (Research Note)

1. Introduction

There has been an increasing research interest in opinion mining and sentiment analysis (Liu, 2012; Pang & Lee, 2008) in the business or political domains and the applications range from finding customer opinions and evaluations regarding products or services to gathering the public's opinions on political events or candidates. However, it has not been widely used in addressing public health-related issues, which may potentially have significant personal and social consequences.

Emotional distress and suicide are prevailing complex social and public health problems in modern societies. About 9.2% of people worldwide have had suicidal ideation at least once in their lifetime and 2% have had that in the past 12 months (Borges et al., 2010), and around 804,000 individuals took their own lives around the world every year (World Health Organization, 2014). Many suicides and attempted suicides are preventable if their suicide intentions and emotional distress are discovered and timely, appropriate assistance provided (Smith et al., 2008).

Traditionally, a person's suicide intention and emotional distress are often noticed by his or her family and friends because they may share their feelings with them rather than professionals. With the advances in information and communication technologies, many people, especially adolescents, like to express their emotions, positive or negative, in social networking sites, blogs, and forums. There is increasing evidence that the Internet and social media can influence suicide-related behavior (Luxton et al. 2012). In particular, a trend appears to be emerging in which people leave messages showing emotional distress or even suicide notes on the Internet (Ruder et al., 2011). In Hong Kong, about 30% of the students who committed suicides had expressed their intentions on social media (Hong Kong Education Bureau, 2016).

Expressions of emotional distress indicate that the person may be in need of help due to problems such as depression or suicidal ideation. In view of this, some non-governmental organizations (NGOs) have started to actively search for these distressed and negative self-expressions in online social media to identify potential severely depressed people in order to provide help and follow-up services. Such online searching is regarded as a proactive and engaging way to identify the high-risk groups. Nevertheless, most of the current approaches are very labor-intensive and time-ineffective because they often rely on simple keyword searches using search engines for social media (e.g., Yahoo blog search engine and forum search engines) to find user-generated contents expressing emotional distress. The search results are often rather "noisy" and the search targets are buried under a large number of other irrelevant documents, and only a few texts with genuine negative emotions can be found. For example, a news article reporting a suicide case posted on social media may match the same set of keywords as a blog written by someone who expresses a suicidal intention. Social workers and professionals often have to spend a huge amount of time to identify people who truly need help.

Techniques for text mining and affect analysis have advanced substantially in recent years. Although these techniques could help with this potentially life-saving application, little empirical research has been done. This research is intended to leverage these advanced techniques to enhance the time and cost efficiencies of these initiatives that identify people with emotional distress. We addressed the problem by designing a system called KAREN that assists social workers and professionals in searching for people with emotional distress in blogs in Chinese. Based on search keywords entered by users, the system combines search results from multiple blog search engines, and automatically analyzes and classifies the search results as showing or not showing emotional distress by combining machine learning classification (with a support vector machine and genetic algorithm) and rule-based classification (with rules obtained from experts). Two studies were conducted to evaluate the performance of the proposed system, and the results showed that (1) the classifier in the system performs better than the baseline classification models; (2) professionals can find more blog posts showing emotional distress using the

proposed system than using a regular blog search engine; and (3) professionals perceive the proposed system to be more useful than a regular blog search engine in finding people with emotional distress.

This article is organized as follows. We first review the problem of suicide and emotional distress and the characteristics of people facing these issues and their online postings. We then review relevant text mining and web mining techniques that have been used to address similar problems and might also be applied in our research. We then introduce the proposed approach, and discuss how this approach can be applied to our problem and how a system based on this approach. We then present the results of a controlled experiment and a user study conducted to evaluate the performance of the system for blogs written in Chinese. In the end, we discuss the implications and limitations of our research, and conclude the paper and suggest some future research directions.

2. Theoretical Background and Related Work

2.1 Emotional Distress and the Internet

Emotional distress and suicidal behavior have been studied under such disciplines as public health, psychology, and social sciences. In practice, questionnaires or screening instruments are often used to detect emotional distress and suicide behaviors in children and adolescents (Scouller & Smith, 2002; Feigelman & Gorman, 2008). However, many emotionally distressed and suicidal youth are reluctant to seek help, and thus it is difficult to identify them. It has been suggested that contents on the Internet, especially narratives and diaries written by youth, have great potential for the gathering of data related to youth's emotional distress and suicidal behaviors (Hessler et al., 2003; Huang et al., 2007; Cheng et al., 2015).

Analyzing user-generated online contents would be of great value to the understanding of emotional distress and prevention of suicide behaviors. In recent years, a number of studies have been published in this area. For example, a self-harm message board has been analyzed to study the role of the Internet in self-harm behaviors (Rodham et al., 2007). This idea of leveraging user-generated contents for suicide prevention is also being carried out in practice. Samaritan Befrienders Hong Kong, one of the

largest suicide prevention organizations in Hong Kong, has been running a project in which social workers monitor blogs to identify potential suicide attempters and people with emotional distress (AppleDaily, 2008). It is believed that with the help of user-generated contents on social media (e.g., blogs and social networking sites), emotional distress and suicidal behaviors could be detected earlier, and the window of opportunities to provide help can be enlarged. However, most of these detections and analyses are performed manually, which results in a very time-consuming and labor-intensive process given the sheer volume and dynamic nature of user-generated contents. In studies where automatic analysis is applied to suicide detection, the analysis usually only involves simple keyword matching (e.g., Huang et al., 2007), which is insufficient because of its low accuracy.

This research focuses on automatic detection of emotional distress in blogs, a major type of user-generated contents. Emotional distress and suicide intentions may be expressed in blogs in different ways (e.g., negative affects, suicide notes, farewell words, linguistic characteristics). Web mining and text mining techniques have achieved satisfactory performance in extracting opinions and identifying communities in blogs (Glance et al., 2005; Liu et al., 2007; Ishida, 2005; Chau & Xu, 2012; Abbasi et al., 2008; Pang & Lee, 2008; Juffinger & Lex, 2009; Kumar et al., 2010; Tang & Liu, 2010; Ceron et al., 2014). Most of these techniques have been applied to problems related to other domains such as marketing (e.g., product or movie reviews), politics (e.g., political opinions), or leisure (e.g., friend and community).

The problem of emotional distress and suicide intention detection is more complex and challenging. There are two unique characteristics of emotional distress classification. First, individual keywords may not be sufficient in revealing the overall emotions expressed in a document. It is often necessary to look at the overall context of the sentences and paragraphs and analyze the document as a whole (Aisopos et al., 2012; Zhang et al., 2009). As our goal is to find out whether the author has emotional distress, it is also important to identify whether the emotions expressed are those of the author's or someone else. Therefore, traditional classification approaches based on keyword matching without

looking at other cues such as self-referencing (Huang et al., 2007) may not be effective in identifying blogs with emotional distress. Second, unlike the expression of negative opinions, many people with emotional distress or suicidal intentions do not express their negative emotions explicitly. Human judgment is often needed to determine the actual emotions in the document. It would be desirable to incorporate these heuristics and judgment into the classification approach in the system.

Because of these challenges, we believe that effective identification and discovery of emotional distress in documents cannot be achieved without a sophisticated approach incorporating a set of advanced techniques. These techniques include sentiment and affect analysis, machine learning, domain-specific lexicons, feature selection, and rule-based classification. In the following subsections, we will review the prior literature in machine learning techniques in Section 2.2.1, domain-specific lexicons and feature selection in Section 2.2.2, and rule-based classification in Section 2.3.

2.2 Sentiment and Affect Analysis Using Machine Learning

2.2.1 Machine Learning Techniques

Machine learning has been extensively used in text-based classification and object recognition with great success in a wide range of applications, including opinion mining and sentiment analysis (e.g., the analysis of customers' opinions and illness diagnoses). Although commonly used interchangeably, opinion mining and sentiment analysis have different goals and focuses (Cambria et al., 2013; Feldman, 2013). Opinion mining is often used to collect opinions (e.g., positive, negative, and neutral) in user-generated contents for a specific subject such as consumer products and movies (e.g., Liu et al., 2007, Attardi & Simi, 2006; Yang et al., 2006; Macdonald et al., 2010). Sentiment and affect analysis focuses on categorizing emotions and affects expressed in writing into different classes such as happiness, love, attraction, sadness, hate, anger, fear, repulsion, and so on (Subasic & Huettner, 2000). For example, sentiment and affect analysis has been widely harnessed in revealing human emotions in computer-mediated communications and providing system predictions that are comparable to human judgments on distilling subjective and affective contents. The intensity of the moods of the general public during the

London bombing incident has been estimated with word frequencies and the usage of special characters in blogs (Mishne & de Rijke, 2006). The machine learning approach and lexicon-based classification on the affect intensities of web forums and blog messages also have been evaluated in previous research in the literature and the results are encouraging, showing that affects can be detected automatically (Abbasi et al., 2008).

Among all machine learning techniques, SVMs (support vector machines) are often regarded as one of the best classifiers providing good generalization capability in sentiment and affect analysis (Mullen & Collier, 2004; Saad, 2014). The SVM-based approach inherently puts a great emphasis on document-level analysis. It is a well-known and highly effective approach yielding high accuracy in sentiment and affect analysis (Mullen & Collier, 2004; Abbasi et al., 2008).

2.2.2 Feature Extraction, Domain-specific Lexicons, and Feature Selection

Most machine learning methods rely on *features* that are present in data. In machine learning research, a "feature" is a variable or a predictor in the model, similar to an independent variable in regression analysis. In the simplest implementation, every word or phrase is treated as a feature in text-based machine learning. The frequency of a word or phrase determines the value of that feature. For example, if a document collection has 5,000 unique words across all documents, each document is then represented by a vector of 5,000 features, where the value of a feature is the frequency of each word in the document of interest (Yang & Liu, 1999).

Feature extraction is the process of finding the value of each feature for every document from the raw data. For example, the value for the word "sad" in a feature vector representing a document can be found by counting how many times it appears in the document using a text analysis program. Because there might be a large number of unique words in a document collection, each document is normally represented by thousands of features without proper organization and classification. As a result, the bulky feature set often adversely affects the performance of inductive learning algorithms (Liu & Motoda, 2012). Besides, a larger feature set makes the pre-processing and training time longer.

To address the issue of large feature sets, features are often grouped into categories to reduce the total number of features, as different words can represent the same meaning or affect orientation in a document. Category-based feature extraction not only sufficiently reduces the number of features in the pre-processing step but also facilitates the later feature selection process. In addition, it has been found that category-based features can avoid the ambiguous nature of many words to greatly improve language model perplexities for training (Niesler & Woodland, 1996; Samuelsson & Reichl, 1999).

A well-developed lexicon can be used to make the feature categories more specific to a particular domain. The Linguistic Inquiry and Word Count (LIWC) lexicon (Pennebaker et al., 2007) has been used in sentiment analysis studies in the public health domain to differentiate normal individuals and those with mental problems based on their writing and linguistic styles. There is preliminary evidence that depressed individuals have a different writing style from non-depressed people (Rude et al., 2004; Pennebaker & Chung, 2011). LIWC organizes words into different categories so researchers can employ them as parameters in analysis. For example, it has been suggested that depressed and suicidal individuals tend to use significantly more self-referencing words in their writings (Rude et al., 2004; Stirman & Pennebaker, 2001; Sloan, 2005). Moreover, some other categories of words such as negations, cognitive words, and positive and negative emotional words are studied to distinguish the writing styles between mentally ill patients and normal individuals (Junghaenel et al., 2008; Gruber & Kring, 2008). Furthermore, expressive writing has been found to have a connection to mental and physical health (Pennebaker & Chung, 2011). The LIWC dictionary translated into different languages is widely used in analyses of user-generated contents including blogs and microblogs (e.g., Gill et al., 2008; De Choudhury et al., 2013; Coppersmith et al., 2014).

Because the number of features may still be large after the lexicon-based categorization of words, some feature selection techniques can be used to further reduce the number of features by finding the optimal subset of features that achieve the best classification performance. Feature selection is a crucial pre-processing step for improving the effectiveness and efficiency of the training process in machine

learning applications. Previous research has shown that feature selection may significantly improve the performance of machine learning text classifiers (Saad 2014). Since an exhaustive search over all possible feature subsets is not feasible, randomized, population-based heuristic search techniques such as genetic algorithms (GAs) can be used in feature selection (Yang & Honavar, 1998; Petricoin et al., 2002; Fang et al., 2007; Oreski & Oreski, 2014). The GA-based approach to feature subset selection, based on Darwin's natural selection theory, searches for the optimal subset according to the principle of "survival of the fittest." The algorithm starts with randomly selecting a certain number of feature subsets, which represents a population of potential solutions. Each subset is evaluated with a fitness function. A new population is then formed by selecting the subsets with a higher average fitness score. Some subsets of the new population undergo transformations such as crossover in conjunction with mutation. After multiple iterations, the GA selects the best feature subset out of all population.

2.3 Rule-based Classification with Expert Judgment

Although machine learning techniques are shown to perform well in various text classification tasks, there are also some drawbacks. First, they are entirely data-driven. If the training data set is biased, it may affect the classification performance. Also, expert judgment and experience cannot be incorporated into the model. Another issue is that machine learning techniques only treat each document as a set of features without considering the writing at the sentence or paragraph level, which may affect performance.

One way to address these issues is to use a rule-based classification approach. In rule-based classification, some rules developed by experts are used to assign score to each document. The benefit of doing this is to incorporate human judgment into the classification process. It is also possible to include sentence-level or paragraph-level analysis. While rule-based approaches have been used in sentiment analysis and emotion detection research (e.g., Hutto and Gilbert, 2014; Neviarouskaya et al., 2010; 2011; Wu et al., 2006), they have not been applied in classifying emotional distress. It would be beneficial to combine both machine learning classification and rule-based classification in order to take the advantages of both approaches.

3. Research Design

This research is intended to design, implement, and evaluate a search system that helps professionals identify people who show emotional distress in their blogs. Because of the nature of our research objective, we choose to employ the design science methodology (Hevner et al., 2004; Gregor & Hevner, 2013). Hevner et al. (2004) provides seven guidelines for conducting effective and high-quality design science research in the field of information systems (IS). It is suggested that these guidelines be followed closely to ensure that the research process and outcome are scientific. In this section, we present the design of our system, i.e., the artifact that addresses the problem described.

The system, called KAREN, which stands for Karen Automated Rating of Emotional Negativity, consists of four major components: a blog crawler, a machine learning classifier, a rule-based classifier, and result aggregation. Figure 1 presents the system architecture.

The core of our design is the classification process. Based on our review of the literature, we propose to use an *aggregation* method to combine different techniques in our classification. First, we use the SVM classifier, which has achieved the best performance in various text classification tasks (Yang & Liu, 1999; Abbasi et al., 2008). In addition, as we expect that the proportion of blogs showing emotional distress is much smaller than that of regular blogs, SVM would be a suitable technique as it is one of the classifiers that perform better when the number of positive training instances is small (Yang & Liu, 1999). Given the nature of our application, we also propose to use the lexicon defined by LIWC, which has performed satisfactorily in understanding emotions in texts, to extract words from documents into category-based features. As LIWC has 71 categories, there will be the same number of category-based features, which is not a small number. It would be good to further reduce the number of features using feature selection. We propose to use a GA-based feature selection method to do this in order to improve the classification performance of the SVM classifier.

Because of the uniqueness of the application domain as reviewed earlier, we postulate that using SVM, a machine learning classifier, alone may not be sufficient. Some expressions showing emotional

distress can only be identified when the context of the whole document is analyzed, which is not possible for SVM as it does not consider the order of words in the document. To address this problem, we propose to complement SVM with a rule-based classifier with rules obtained from experts. While it is possible to combine SVM with other machine learning classifiers such as a decision tree, we choose to complement SVM with a rule-based classifier because a rule-based classifier can perform sentence-level and paragraph-level analysis and can directly incorporate context-specific heuristics in its rules. As the SVM classifier focuses on word-level analysis and the rule-based classifier focuses on sentence-level and paragraph-level analysis, we believe that they can complement each other and obtain better performance when combined together.

When using the system, a user will first enter keywords related to emotional distress into the system, which will then be sent to various blog search engines, such as Google blog search and Yahoo blog search. The search results from these engines will be extracted and the actual contents of the blogs will be downloaded by the system to the local database. Each blog will then be analyzed by both a machine learning classifier and a rule-based classifier, and the results from the two classifiers will be aggregated into a final classification decision. Finally the search results will be presented to the user based on the classification. The workflow of a standard search session is shown in Figure 2.

The four components of the design are discussed in detail in the following subsections.

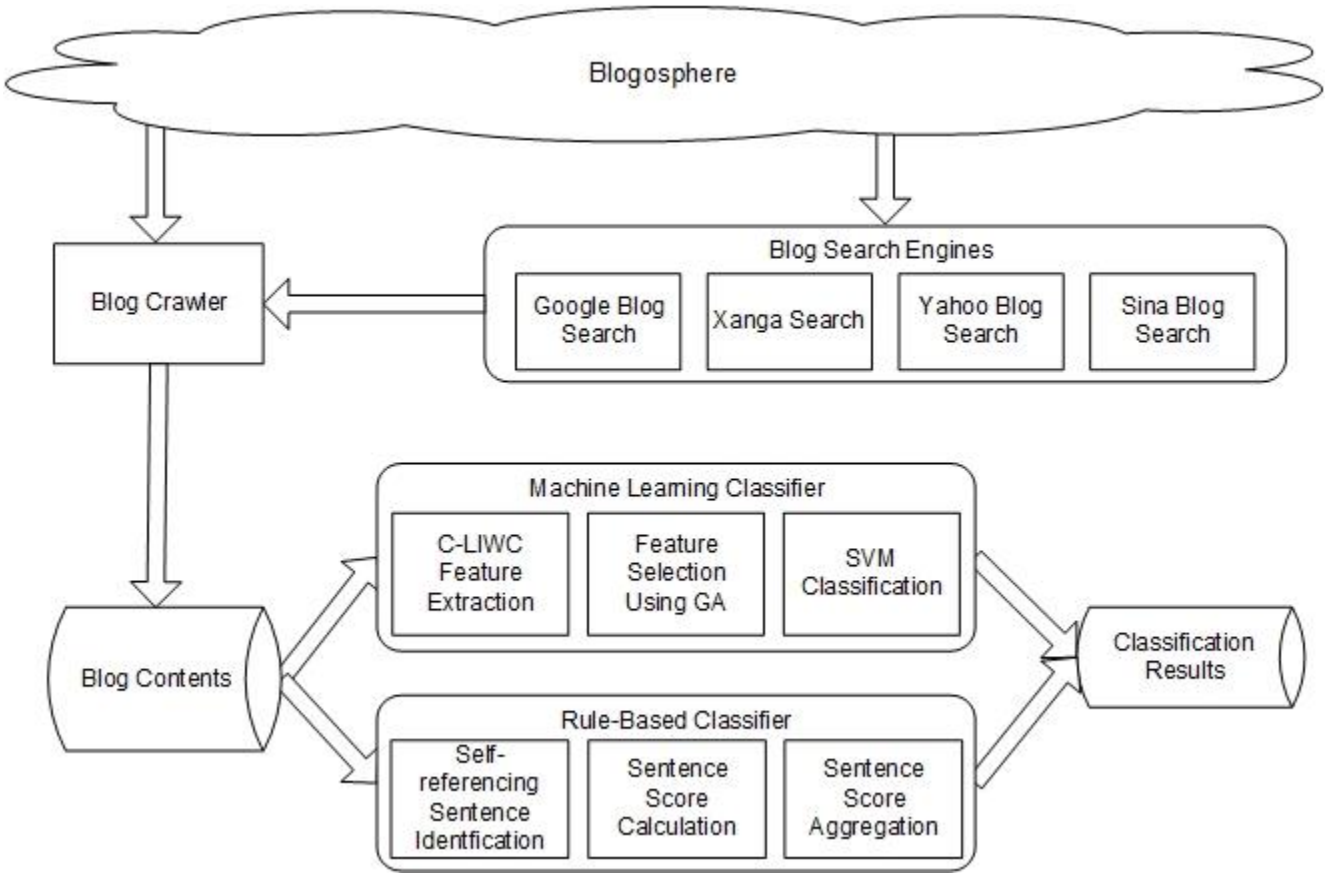

Figure 1. System Architecture

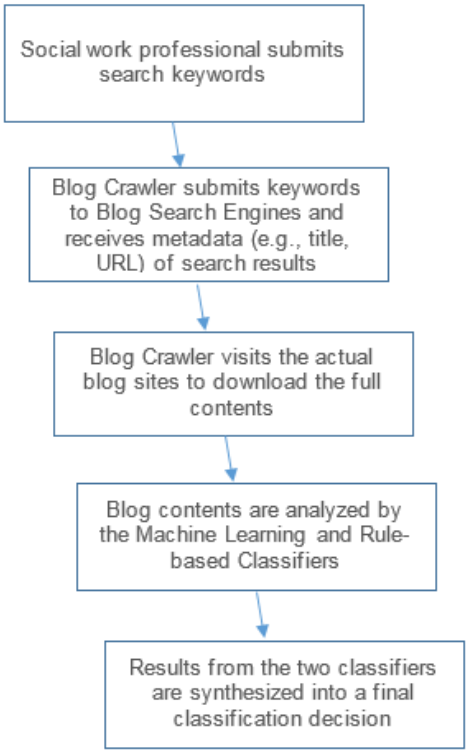

Figure 2. Workflow of a Typical Search Session

1. Blog Crawler

The first component in the proposed architecture is a blog crawler that collects blogs from different blog hosting sites. Using a meta-search approach (Chen et al., 2001), the crawler sends keywords entered by the user to blog search engines such as Google blog search and Yahoo blog search and extracts the addresses of the blogs identified. As the search engines only return the URL, title, and summary of a blog, which are not sufficient for our analysis, the crawler will also visit the hosting sites of these blogs directly to download the entire content through standard HTTP protocol.

After a blog is downloaded, our system will extract its content and perform word segmentation, i.e., tokenize the document into words, for further analysis. Simple word segmentation based on common delimiters such as spaces and punctuation marks can be employed for blogs in English. For blogs written in Chinese, which is a character-based language without explicit delimiters between words, the segmentation process is often more difficult and less accurate than for blogs written in English. In our system, we use a Chinese segmentation tool developed by the Chinese Academy of Sciences called ICTCLAS, which is a popular tool that has been used in many prior studies (Zhang et al., 2003; Zeng et al., 2011).

2. Machine Learning Classifier

The proposed architecture uses two classification models, namely a machine learning model and a rule-based model, as a classification ensemble. The models are designed to classify whether a blog is showing emotional distress based on a set of training examples. This would help professionals to identify potential emotional distress of the blog author. We use a support vector machine (SVM) as our machine learning classifier. The reason for our choice is that SVMs have been shown to be highly effective in conventional text classification and achieved the best performance among different text classifiers (Yang & Liu, 1999; Abbasi et al., 2008). We suggest that it is well suited for our application of classifying texts as whether or not showing emotional distress.

Feature Extraction

After a blog is parsed into words, each word is matched with the LIWC lexicon to determine which category it belongs to. As we are focusing on blogs written in Chinese, the Chinese version of LIWC, called C-LIWC (Huang et al., 2012), is employed. Similar to LIWC, C-LIWC provides multiple word categories such as positive or negative emotions, self-references, and causal words for text analyses on emotional and cognitive words. This approach is effective because it has been shown in many studies that people's mental health can be predicted with the words they use in writing by looking at what LIWC category the words belong to (Pennebaker, 2003). Thus, the frequency count of every word in the categories' word list is used to calculate the feature value (f_i) for each of the 71 categories in C-LIWC. The word-to-document proportion is incorporated in the calculation to reflect word importance corresponding to the document. A document is one single blog post. For each document *d*, the value f_i (for i = 1 to 71) is calculated as follows:

$$f_i = \sum_{all\ words\ w\ in\ category\ i} \frac{frequency_w}{total\ number\ of\ words\ in\ document\ d}.$$

Document length, measured by the number of words, is also added as the 72nd feature. Therefore, after this stage of processing, a vector of 72 values is created for each document *d*.

Feature Selection Using Genetic Algorithm

There are different ways of choosing which features we pass to SVM for training and performing the classification. One way is to use all the 72 features identified by LIWC and document length. However, as discussed in our literature review, it is often desirable to extract a subset of features in order to improve performance. In the proposed architecture, we use a genetic algorithm (GA) in our feature selection process. GAs have been employed for feature selection in previous research and are applicable here (Yang & Honavar, 1998; Petricoin et al., 2002; Fang et al., 2007; Oreski & Oreski, 2014). In our GA implementation, the initial population contains a fixed number of individuals (chromosomes), where each individual represents a set of a variable number of features. Each individual is represented with a binary vector of bits, where a bit value of 1 means that the corresponding feature is selected while 0 means that

the corresponding attribute is not selected. In other words, each individual in the population is a candidate solution to the feature subset selection problem. Standard GA operations like roulette wheel selection, crossover, and mutation are implemented in a standard way (Goldberg, 1989; Michalewicz, 1996). In calculating fitness value of a chromosome, the set of features represented by the chromosome is used as the input for the SVM, which will go through training and testing using 10-fold cross validation. The fitness value is calculated as the *F* value of the classification testing performance, the harmonic mean of *precision* and *recall*. All the three performance metrics have been widely used in classification and retrieval research. Readers are referred to Van Rijsbergen (1979) for more details on these measures.

3. Rule-Based Classifier

Besides the machine learning classification model, a rule-based classification model is also employed in our architecture to automatically classify a blog as showing emotional distress or not. To build our rule-based classifier, we first create a lexicon consisting of words related to emotional distress. Then, for each document to classify, we will perform the following steps.

1. For each sentence, we identify whether it is a self-referencing sentence
2. We calculate a score of emotional distress for each sentence.
3. We aggregate the scores for all sentences in a document and come up with a single score for the document.

By doing these, our rule-based classifier can classify blog content at the sentence and document levels. The sentence-level classification differentiates sentences into positive or negative emotion; as a result, the model is able to determine whether the whole document shows emotional distress from the automatically annotated sentences. In the following we discuss the details of the lexicon creation process and the three analysis steps.

Lexicon creation

Since no lexicon specifically concerning emotional distress wordings in Chinese is available, we develop our own lexicon in this model. The lexicon is constructed by manual inspection of blog contents by

professionals familiar with web discourse terminology for emotional distress. Similar lexicon creation approaches have been used in previous studies and have shown encouraging results (Abbasi & Chen, 2007; Subasic & Huettner, 2000). In this particular study, 3,147 blogs were collected from Google Blog Search, and two clinical psychologists familiar with research on emotional distress and suicide were asked to read these blog contents and extract emotional expressions and representative words of positive, negative, and neutral emotions in a macro-view. Manual lexicon creation is used since blogs contain their own terminology, which can be difficult to extract without human judgment and manual evaluation of conversation text.

Table 1: Examples and Number of Words in the Ten Lexical Groups

Group	Number of Words	Examples
Self-reference	9	自己 (self) 在下(I) 小弟(I) 本人(myself) 我(I)
Positive Emotion	34	窩心(heart-warming) 雀躍(joyful) 暢快(carefree) 驚喜(pleasant surprise) 酷愛(love)
Negative Emotion	56	心痛 (sad) 失望(disappointed) 失落(down) 沮喪(frustrated) 焦慮 (anxious)
Risk Factors	15	分手(separation) 離婚(divorce) 疾病(illness) 貧窮(poverty) 比人厄 (cheated)
Suicide Words	18	界手(self-laceration) 跳樓(jumping from a building) 燒炭(charcoal burning) 食安眠藥(taking sleeping pills) 永別(part forever)
Time	16	今朝(this morning) 每天(every day) 昨晚(last night) 聽日(tomorrow) 宜家(now)
Negation	39	唔(not) 不(no) 別 (don't) 否(negative) 沒(without) 非(not)
Leisure	184	義工(volunteer) 健身(fitness) 煲劇(watching TV drama) 動漫 (animation and comics) 旅行團(tour)
References	108	本報訊(news) 專訊(special news) 參考資料(references) 摘錄(excerpts) 撰稿(written)
Gratitude Expressions	11	共勉之(encourage each other) 加油(make effort) 鼓勵(encourage) 感恩 (appreciate) 謝天謝地(thank god)

The words in the lexicon are categorized into ten groups in the rule-based model. The ten groups are *Self-reference, Positive Emotion, Negative Emotion, Risk Factors, Suicide, Time, Negation, Leisure, References, and Gratitude Expressions*. Examples and number of words in each group are shown in Table 1. All the words are treated equally in the lexicon without individual score assignment. Different groups

of words are, however, used in different components in a sentence-level scoring process in the model (to be discussed later). Compared with C-LIWC, this lexicon is more precise and customized for the domain. This is because C-LIWC has a large coverage and contains categories and words that are not very relevant to the application domain. On the other hand, the manual lexicon contains words that have been actually used by bloggers in their online emotional expressions, which include colloquial words and domain-specific words that are not found in C-LIWC.

Self-referencing sentence identification

We want to identify self-referencing sentences as these sentences are regarded as directly reflecting the writer's cognition. Studies in psycholinguistics reveal that people who currently have depression or suicidal ideation have a distinctive linguistic style and tend to use significantly more self-referencing words (e.g., *I, me, myself*) in their writings, entailing strong self-orientation (Ramirez-Esparza et al., 2006; Rude et al., 2004; Li et al., 2014) and even withdrawal from social relationships (Stirman & Pennebaker, 2001). Although this self-referencing style is difficult to identify with human judgment, those sentences with self-referencing words are believed to provide more clues on identifying disengagement behavior and hence emotional distress. It should be noted that this is different from subjective sentence identification in some previous studies that made use of subjective words in existing knowledge and sentiment databases (Riloff & Wiebe, 2003; Zhang et al., 2009). Instead of finding the expressions of common affects like fear and anger that are normal expressions of feelings, the model is aimed at identifying emotional distress, which consists of multiple affects. Many researchers have focused on discrete affects such as fear, worry, sadness, contempt, disgust, guilt, nervousness, and anger (Abbasi et al., 2008; Subasic & Huettner, 2000). Two opposite affects, namely positive affect and negative affect, have become dominant in the literature. Although negative affect is associated with emotional distress, they are not equivalent (Crawford & Henry, 2004; Matthews et al., 1990). Emotional distress consists of multiple affects in different situations and life stressors. For instance, bereavement-related emotional distress would have affects such as sadness and nervousness (Chen et al., 1999), while diabetes-related

emotional distress would have affects such as fear and worry (Snoek et al., 2000). Also, instead of using many negative emotion words, people may talk about what has happened to them in their daily lives, which may be the causes for their emotional distress. Therefore, besides analyzing negative and positive emotion words, we also look at other words related to emotional distress such as various risk factors and suicide words, as well as words that indicate positive well-being and attitudes. More details are discussed in the following subsection.

Sentence score calculation

The procedure for calculating the emotional distress score for each sentence is shown in Figure 3. A positive value of the score means that the sentence is showing emotional distress, while a zero or negative value means otherwise. In calculating the sentence scores, we pay special attention to self-referencing sentences (sentences containing *Self-reference* words), which are more likely to be about the writer's own feelings than non-self-referencing sentences. Also, as discussed, people with emotional distress are more likely to write self-referencing sentences (Ramirez-Esparza et al., 2006; Rude et al., 2004; Stirman & Pennebaker, 2001). A self-referencing sentence's score of emotional distress is calculated based on the *Positive Emotion* words and *Negative Emotion* words that are present. Intuitively, a sentence is classified as showing emotional distress when only *Negative Emotion* words are found (Li et al., 2014; Cheng et al., 2015), and a score of 1 is first assigned. On the other hand, the sentence is considered as not having emotional distress when only *Positive Emotion* words are found, and a score of -1 is assigned. When neither *Positive Emotion* nor *Negative Emotion* words are found, the sentence is regarded in the same way as a non-self-referencing sentence. In the case where both positive and negative emotion words are found, the sentence is classified as showing negative emotion. This is to avoid missing out on possibly negative documents. Because of the nature of our application, we want to reduce the chance of not finding the documents showing emotional distress, even though doing so may result in a higher chance of classifying a normal document as showing emotional distress. *Negation* words (e.g., "no", "not", and "never") are also checked in the calculation.

Based on what we discussed, the score of each self-referencing sentence is assigned as 1 or -1 based on whether it contains any *Positive Emotion* words, *Negative Emotion* words, and *Negation* words (as shown in Lines 9-11 and 17-19 in Figure 3). We give only a score of 1 (or -1) even if the sentence contains multiple Negative Emotion (or Positive Emotion) words because we want to distinguish our approach from standard sentiment analysis methods. Therefore, instead of giving the same weight to different categories of words, we want to focus more on words that are related to emotional distress and mental well-being. For sentences showing negative emotion, the score is increased with the occurrence of words in the *Risk Factors* or *Suicide* groups (Cheng et al., 2000; Li et al., 2014). This increment is proposed because contents relating to risk factors (e.g., "divorce", "serious illness") and suicide (e.g., "suicide", "charcoal burning") provides useful information to identify emotional distress. Similarly, for sentences not showing negative emotion, the score is adjusted with the occurrence of *Leisure* words. In positive psychology, leisure is a core ingredient for overall well-being and evokes happiness (Newman et al., 2014; Zawadzki et al., 2015). In addition, if *Time* is mentioned, we will further adjust the score because the temporal connection integrates the writer's feeling with past and future events (Kuhl et al., 2015).

For non-self-referencing sentences, sentence score is not calculated using emotion words. Instead, the sentence is checked for words that reference others (*Reference*) or express thankfulness or encouragement (*Gratitude Expressions*). Under the disengagement theory, it is believed that people who reference other sources to offer opinions or convey information to others have a lower risk at depression (Stirman & Pennebaker, 2001). Giving thankful and encouraging words to others, which is shown to improve people's well-being and alleviate depression, also demonstrates a positive attitude in the writer (Bolier et al., 2013; Lyubomirsky & Layous, 2013).

Sentence score calculation

```
1. Inputs:
2.     s, a sentence
3.     lexicon, a lexicon of words divided into 10 groups
4. Output:
5.     score, the emotional distress score for sentence s
6. Procedure:
7.     score = 0
8.     if s contains (Self-reference)
9.         if s contains (Negative Emotion and not Negation)
10.            or s contains (Positive Emotion and Negation)
11.             score = 1
12.             for each (Risk Factors or Suicide) in s
13.                 if s contains (Time)
14.                     score = score + 2
15.                 else
16.                     score = score + 1
17.         else if s contains (Positive Emotion and not Negation)
18.            or s contains (Negative Emotion and Negation)
19.             score = –1
20.             for each (Leisure) in s
21.                 score = score – 1
22.    else
23.        for each (References or Gratitude expressions) in s
24.            score = score – 1
25. return score
```

Figure 3. Sentence score calculation

Sentence score aggregation

The sentence scores presented in the previous section are used to make the final decision on a document score. Since the emotional fluctuations throughout a document could be complicated, some of the scores in the middle of the document may not be meaningful and may even be confusing. The aggregation, therefore, primarily concentrates on the scores at the beginning and the end of the document.

It is believed that the summary and major theme expressed by writers generally appear at the beginning and the end of documents (Lee et al., 2002). There is difficulty, however, in defining the parameters of what constitutes the opening and the ending of a document. Static positioning does not yield significantly higher accuracy because of the reduced flexibility of the analysis. Furthermore, these parameters vary for documents by different writers who have diverse writing and organization styles. An

algorithm that dynamically defines these parameters, therefore, is crucial for improving the analysis performance.

There are many segmentation methods that have mainly been used to find sub-topics in full-length documents, e.g., by grouping sentences in blocks and partitioning content into coherent units (Hearst, 1997). Following this idea, we use the first and last blocks of self-referencing sentences in a document for the final prediction score, where a block is defined as a consecutive set of sentences with the same polarity. The other blocks were not considered as the main polarity in the document. However, because a high number of polarity changes in a document represent the inconsistency (i.e., fluctuations and unstableness) of the writer's emotion, we also use this number in the computation of the final score. Therefore, the final score is calculated as the sum of the first block's score, the last block's score, and the number of polarity transitions in a document. A document is classified as showing emotional distress if the final score is positive.

4. Result Aggregation

The results from the two classifiers are combined into a single classification result. In our context, since it is desirable not to miss any emotional distress cases, if a blog is detected as showing emotional distress by either of the classifiers, it will be classified as such. In other words, a blog will only be classified as not showing emotional distress if both classifiers give such classification.

4. Evaluation

We conducted two studies to evaluate the performance of KAREN. The first study was intended to evaluate the performance of the classifier in correctly identifying blogs with emotional distress on a static data set. In this study, the evaluation was conducted by running the proposed classifier against other benchmark approaches on computers without human subjects. The second study was a user study that focused on evaluating whether professionals could enhance their effectiveness in identifying people with

emotional distress online by using the system, like what they would do in a real usage scenario. The setup and results of the two studies are discussed in detail in the following subsections.

Study 1: Classifier performance evaluation using a static data set

Data Set

Study 1 is a controlled experiment that aimed to evaluate the performance of the classifier in the proposed system using a static data set. The data set is "static" in the sense that the data are not blogs obtained in search sessions in real time. Rather, they were downloaded from different sources and saved in our database for evaluation purpose. Although the experiment using a pre-collection of blogs is different from an actual usage scenario of our system, the use of static data can allow an easy comparison of different approaches while controlling that the data set is the same, which is a widely-adopted approach in text classification research (Yang and Liu, 1999; Pang and Lee, 2008).

To develop the static data set for evaluation, a total of 804 blogs were obtained from four different sources. We used multiple sources instead of a single source in order to increase generalizability and coverage. The first subset of data was blogs collected from the online outreaching project organized by the Hong Kong Federation of Youth Groups, one of the largest non-governmental organizations providing social services to young people in Hong Kong. These blog writers consist of individuals who were identified as possibly facing emotional difficulties. They were identified by trained volunteers working at the Hong Kong Federation of Youth Groups who searched on Yahoo's blog search engine with keywords that expressed depression or suicidal ideation. The identification process was manual and based on the judgment of these volunteers (Li et al., 2014). The content of these blogs included difficulties experienced in everyday life from home and school and documented problems with intimate relationships and friendships. This subset of examples in the experiment constitutes a corpus of 180 blogs.

The second subset of blogs was sourced from the Samaritan Befrienders Hong Kong, a non-governmental organization focusing on helping people with suicidal ideation or emotional distress. This data set encompassed examples of individuals who had been identified by trained volunteers in the

organization as possibly having emotional distress and even suicidal behavior. The volunteers searched on a wide range of blogging sites with keywords that expressed emotional distress or suicidal ideation. Site search engines and various general search engines like Google and Yahoo were used. It should be noted that the search results were almost always in the Chinese language when Chinese search keywords were submitted to these engines. This data set consisted of 239 blogs.

The third subset of the data was blogs on the web that were labeled as containing positive affects by two independent trained volunteers. The volunteers identified these blogs by browsing different blog hosting sites commonly used in Hong Kong. This data set contained 150 blogs.

The fourth subset of data comprises 235 blog posts located through the Google blog search by two trained volunteers. Some commonly used Chinese words for expressing emotional distress or suicidal ideation (such as “唔開心,” “不快樂,” “痛苦,” “絕望,” “想死,” and “自殺,” which mean “not happy,” “unhappy,” “suffering,” “despair,” “want to die,” and “suicide,” respectively) were used as the search keywords. These keywords were chosen based on findings in the literature on the writing style of people with emotional distress (Huang et al., 2007; Li et al., 2014; Pennebaker & Chung, 2011). Those blog posts included narratives and diaries containing emotions and also neutral posts such as fiction, news reports, and religious writing.

The content of all 804 blogs from the four sources was collected and further reviewed by two clinical psychologists for judgment of emotional distress. Out of these blogs, 742 were consistently rated by the two psychologists, with results in an inter-rater reliability of 0.83. The remaining 62 blogs were inconsistently rated and were discussed by the two experts to reach a consensus for each blog. As for the results, out of the 804 blogs, 274 included contents showing emotional distress and 530 included contents not showing distress. The average length of blogs is 727 characters. The sources of the data are summarized in Table 2.

Table 2: Sources of blog data used in the experiment

Data Source	Source	Search tool used	No. of blogs
#1	Hong Kong Federation of Youths Group	Yahoo blog search	180
#2	Samaritan Befrienders Hong Kong	Various	239
#3	Blog site browsing	Nil; browsing only	150
#4	Google blog searching	Google blog search	235
Total number of blogs in the sample (*n*)			**804**

Experiment Setting

In this subsection, we describe the specific parameter setting of the SVM and GA algorithms in our evaluation. Based on the results reported in previous literature (Mullen & Collier, 2004; Abbasi et al., 2008), we use a linear kernel in the SVM. It has been suggested that for a high dimensional space the linear kernel should be as good as a non-linear one (Hsu et al. 2003). The 804 posts discussed above were classified using 10-fold cross validation. In our experiment, a conventional approach using N-grams as input features to the SVM, without the rule-based model, was used as a baseline (Model 1). N-gram is a subsequence of *n* items from a given sequence, where *n* is equal to an integer value ranging from 1 to 4 in our case. Features appearing in three or fewer blogs were eliminated in order to achieve better generalization.

The second baseline model (Model 2) used C-LIWC categories as the input features. In particular, each post was parsed, and the words were checked against the C-LIWC lexicon to see which category they belonged to. The frequency of words in each category was passed to the SVM. Model 3 built on Model 2 but also used genetic algorithms (GA) for feature selection.

Model 4 used the rule-based classifier alone. Model 5 used the C-LIWC lexicon as input to the SVM and also incorporated the rule-based classifier. Two more models for comparison were also included, where feature selection was added on top of what was used in Model 5. In particular, correlation-based feature selection (CB) and recursive feature elimination (RF) was used in Model 6 and Model 7, respectively.

The main focus of the evaluation is the proposed model (Model 8) in KAREN, which incorporates all the three major components: SVM with C-LIWC categories as features, GA for feature selection, and

the rule-based classifier. In the feature selection process, the following parameter settings were used for our GA implementation: Population size was 50, and the number of generations was 200. The probability of crossover and mutation were 0.5 and 0.01, respectively. As discussed earlier, the fitness of an individual is determined with evaluating the SVM model using the training data set in each iteration.

Evaluation

Standard evaluation metrics for classification, namely precision, recall, and F-measures, were used to evaluate the performance of the classification models. Because some of the models are focused on improving the recall rate, we take the F-measure, which is balanced between precision and recall, as our main comparison metric.

The experiment results are shown in Table 3. The comparison between the two baseline models (Models 1 and 2) reveal that large reduction of the number of features from word-based features to category-based features does not necessarily lead to significant degradation of classification performance. It can be seen that the classifications with C-LIWC category-based features (Model 2) performed comparably to Model 1 in the experiment. C-LIWC categories are regarded as representative features in this domain, so the characteristics of documents can be reflected in the feature set. Therefore, a large feature set is not a practical necessity in identifying emotional distress. When GA-based feature selection is applied (Model 3), the performance performed improved.

Our results also show that the models incorporating the rule-based classifier with SVM (Models 5 to 8) consistently performed better than the baseline models using SVM alone (Models 1 to 3) or rule-based alone (Model 4). When feature selection technique is used (Models 6 to 8), the classification performance in terms of F-measure improved slightly compared with Model 5, which used all 72 features based on C-LIWC categories and document length. Among the three feature selection models, the proposed GA feature selection (Model 8) achieved the best result in terms of F-measure (0.7216), although the number of features is higher (38 vs. 17 and 20 in Models 6 and 7, respectively).

Table 3: Classification Performance

Model	No. of Features	Precision	Recall	F-measure
Model 1: SVM(N-grams)	17,560	0.5732	0.6715	**0.6185**
Model 2: SVM(C-LIWC)	72	0.7404	0.6350	**0.6837**
Model 3: SVM(C-LIWC + GA)	38	0.7542	0.6606	**0.7043**
Model 4: Rule-based	10	0.4705	0.8723	**0.6113**
Model 5: SVM(C-LIWC) + Rule-based	72	0.6171	0.8175	**0.7033**
Model 6: SVM(C-LIWC + CB) + Rule-based	17	0.6123	0.8358	**0.7068**
Model 7: SVM(C-LIWC + RF) + Rule-based	20	0.6202	0.8285	**0.7094**
Proposed Model 8: SVM(C-LIWC + GA) + Rule-based	38	0.6287	0.8467	**0.7216**

Note: SVM: Support Vector Machine; N-grams: n-grams-based features; C-LIWC: C-LIWC-based category features; CB: correlation-based feature selection; RF: recursive feature elimination; GA: Genetic algorithm.

The Effect of Training and Testing Data on Classifier Performance

In our experiment, the data were acquired from four different sources and combined into an 804-blog single data set for training the classification models. In order to test the robustness of the model, we evaluate whether a model trained using data from one or two sources would still perform well on the data obtained from other sources. As discussed in Table 2, we have four different data sources. The first two data sources (#1, #2) contain more posts showing emotional distress and the other two (#3, #4) contain more posts not showing emotional distress. Based on this, we have four settings in our robustness test. In each setting, we use one data source showing emotional distress and one not showing as the training data for our model, and the other two data sources as testing data. These combinations ensure that both the training data and testing data will not be imbalanced. The results are shown in Table 4. As can been seen, the performance in each setting is comparable to the main findings shown in Figure 3, demonstrating the generalizability of our approach.

Table 4: The Effect of Training and Testing Data

Training Data	Testing Data	Recall	Precision	F-measure
#1, #3	#2, #4	0.8634	0.6178	0.7202
#1, #4	#2, #3	0.8676	0.5960	0.7066
#2, #3	#1, #4	0.8333	0.6085	0.7034
#2, #4	#1, #3	0.8761	0.6689	0.7586

Another consideration on our experiment data is that about 34% of blogs were judged by experts as showing emotional distress. However, according to the literature, the youth prevalence rate of having emotional distress symptoms is about 9% (Leung et al., 2008). Based on this ratio, we have run another experiment as a robustness test to compare the different models using five subsets of data with a similar ratio (55 blogs showing emotional distress and 530 blogs not showing emotional distress – around 10%). The results show that in terms of the F2-measure, the proposed GA model (0.5645) performs comparably with Models 6 and 7 (0.5621 and 0.5596), which are better than the other models.

The Effect of Blog Characteristics on Classifier Performance

While our results show that the proposed model performs better, it would be interesting to study under what conditions it does so. One important factor that we have observed is the length of the blog post content. To investigate how the proposed model performs better for blog posts with different length, we divide our data set into 10 groups based on their length. As we have 804 blogs in total, each group has 80 or 81 blogs. The first group contains the 80 blogs that have the shortest contents (the lowest decile in terms of number of words), the second groups contains blogs with a length falling within the second lowest decile, and the last group contains blogs that have the longest contents (the highest decile). The average length of the blog posts in each group is shown in Table 5. We then apply the proposed model with aggregation and one with SVM only on each group, and recorded the F-measure. The results are shown in Figure 4.

As can be seen in the figure, the proposed aggregation model performs generally better than the SVM model alone when the content length is in the first seven deciles. In particular, the rule-based classifier adds the most value when the content length is within the 3rd to 7th deciles. In contrast, the aggregation model performs worse than SVM when the blog posts are long (9th and 10th deciles). We think the reason is that when the blog posts are long, SVM performs relatively better than the aggregation model because the large number of keywords in the blogs are already effective in making the classification decision. On the other hand, the rule-based approach puts more emphasis on the first block and last block of each blog post and the number of polarity transitions in the blog during the sentence score aggregation process. When a blog post is long, the first block and last block of the post would constitute only a smaller fraction of overall content, and could be less representative of the entire post. Also, because a longer post would be more likely to have a higher number of polarity transitions, it would be more likely to receive a higher final score based on our calculation. As such, the rule-based method tend to classify more long blog posts as showing emotional distress and result in more false positives (i.e., a lower precision rate).

It is also worthwhile to note that SVM alone does not perform well when the blog posts are very long (10th decile). We found that some of these long blogs contain quite a number of negative emotion words, but was classified as not showing emotional distress by our clinical psychologists. By analyzing these blogs, we found that while they did contain many negative words, there were other contents (e.g., positive emotion words or the story of another person) showing that the authors were not emotionally distressed. Given the large number of negative words in these blogs, they were still misclassified as showing emotional distress by SVM, resulting in a lower performance.

Another observation is that when the blog posts are short (1st and 2nd deciles), both the proposed aggregation model and the SVM model do not perform well. By analyzing the blog posts in these groups, we found that these documents may not contain enough informative features and therefore they could be easily misclassified by the models.

Table 5: Average Length of Blog Posts in Each Group

Decile Group	Average Length	Decile Group	Average Length
1st	47.1	6th	422.1
2nd	105.8	7th	582.7
3rd	170.4	8th	827.9
4th	230.9	9th	1249.6
5th	317.6	10th	3149.6

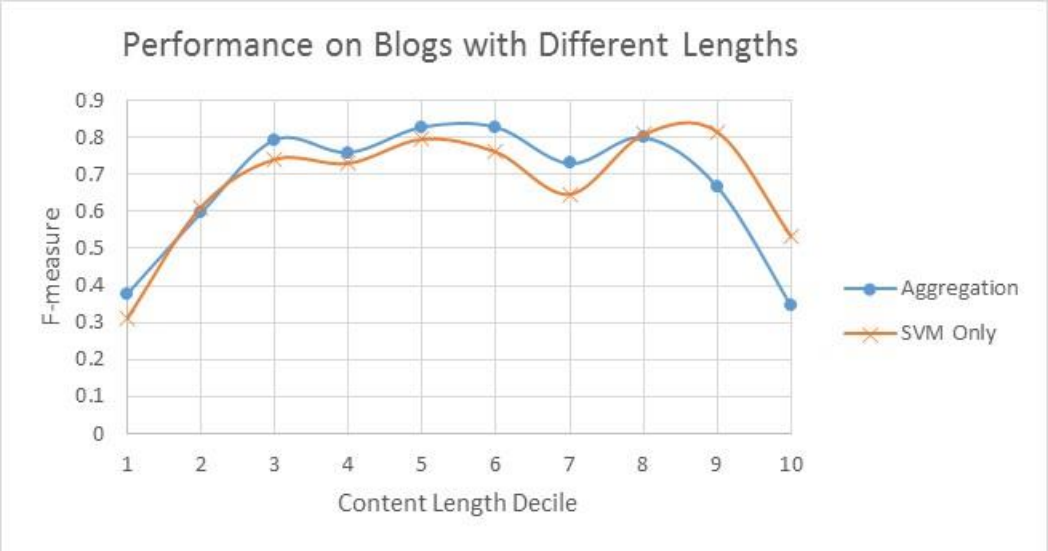

Figure 4. Performance on Blogs with Different Lengths

Besides content length, we also study the effect of several other blog characteristics, including the percentage of positive words and the percentage of negative words in each blog (calculated respectively as the number of *Positive Emotion* words or *Negative Emotion* words, as illustrated in Table 1, divided by the total number of words in a blog). Similar to the analysis on content length, we divide the data set into 10 groups corresponding to the deciles for each of these two measures. The results are shown in Figure 5 and Figure 6.

In Figure 5, we can see that the aggregation model performs better than using SVM alone in terms of F-measure when the proportions of positive emotion words is high. This is because some blogs are

actually showing emotional distress even if they contain many positive emotion words. SVM were not able to recall these blogs and identify them as showing emotional distress, while the rule-based classifier in the aggregation model could identify them correctly.

Figure 6 shows the performance of the models at different proportions of negative emotion words. We found the performance of both models are poor when the proportion of negative emotion words is low. The reason is that there are some blogs that contain very few or even no negative emotion words based on our lexicons, but are actually showing emotional distress. By analyzing the raw data, we found that some of these blogs were written in a more subtle way or used the wrong Chinese characters, so the words could not match our lexicons. These blogs would be easily missed by both models, in particular SVM as it does not have the customized lexicons or the sentence-level analysis as in the rule-based approach. As such, SVM has a high number of false negatives for these blogs, resulting in a low recall and F-measure. On the other hand, the rule-based approach, which takes other factors such as self-references and negation into account, could correctly identify some of these blogs as showing emotional distress. Therefore, the rule-based approach can successfully complement SVM, and the aggregation method achieves a better performance than SVM alone no matter the proportion of negative emotion words is high or low.

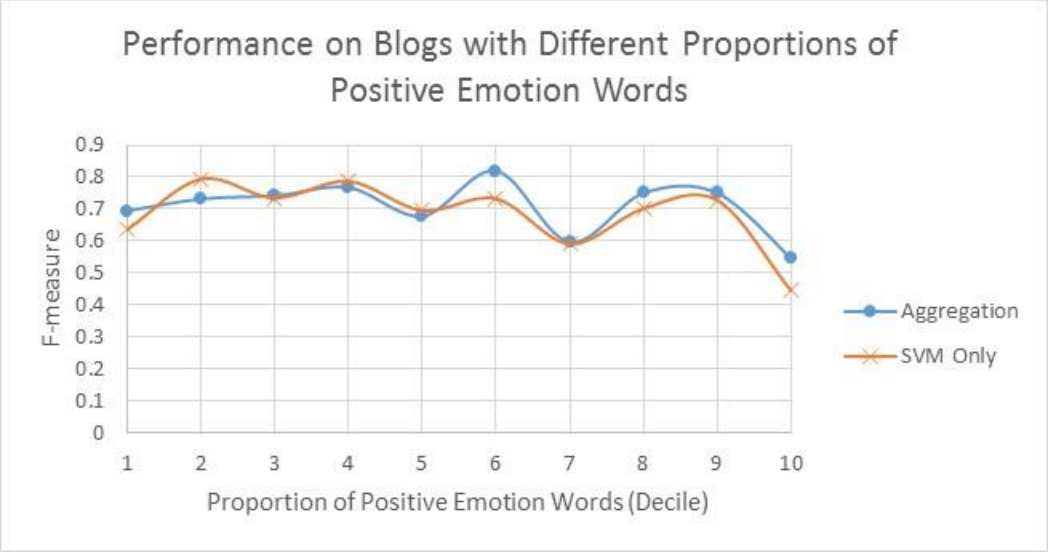

Figure 5. Performance on Blogs with Different Proportions of Positive Emotion Words

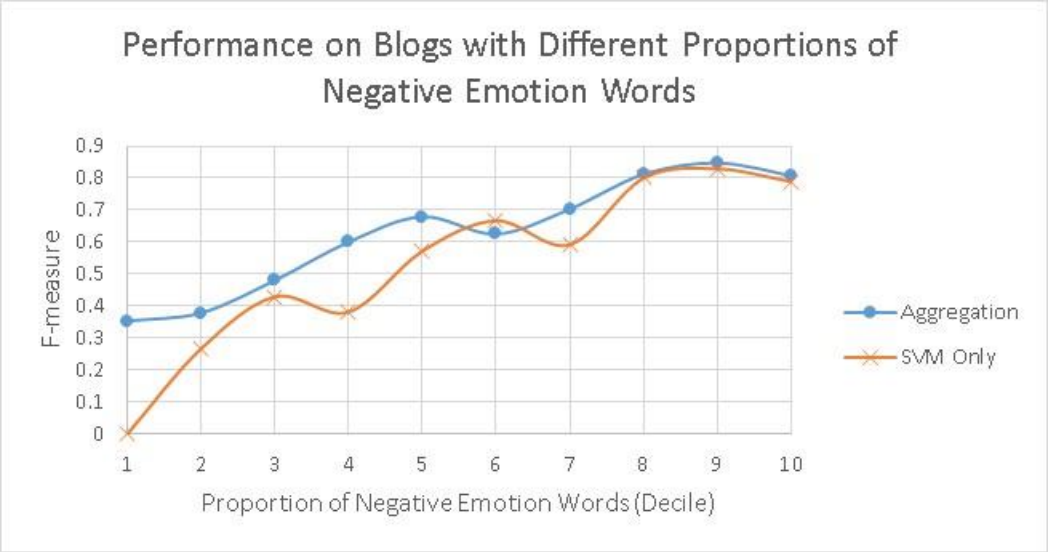

Figure 6. Performance on Blogs with Different Proportions of Negative Emotion Words

Aggregation of Classification Results

In the proposed model, a blog is classified as showing emotional distress if either of the two classifiers classified it as showing emotional distress. In other words, the results from the two classifiers are combined through a Boolean OR operation. It would be interesting to investigate if there is a better way to combine the results from the two classifiers. One way is to use the weighted scores produced by the classifiers. To test different weighting combinations, we first standardize the scores by dividing each score by the standard deviation of all scores produced by each classifier. We then calculate a weighted aggregation score for each blogs as follows:

$$Weighted\_Aggregation\_Score = (1 - w) \times SVM(C\text{-}LICW{+}GA)\_Score + w \times Rule\_Based\_Score$$

where *w* is simply a value between 0 and 1. When *w* is 0, the aggregation will be using the SVM output only. The SVM used here is the SVM using C-LIWC-based category features and Genetic Algorithms for feature selection, i.e., SVM(C-LIWC + GA) as in Model 3. A blog is classified as showing emotional distress if the weighted aggregation score is greater than or equal to 0. We adjusted the value of *w* from 0 to 1 and recorded the F-measure. The results are shown in Figure 7. The results show that the

aggregated classifier performance is consistently above the cases of $w = 1$ and $w = 0$, where there is no aggregation. We also found that the F-measure is the highest (0.7305) when the value of w is 0.7.

It should be noted that the weight of 0.7 for the rule-based classifier does not necessarily mean that the rule-based classifier is a better classifier or more important. The value could be related to the distribution of the scores for each classifier. As explained earlier, we standardized our scores by dividing them by their standard deviation. Since the raw rule-based score has a wider range, the standard deviation is higher and thus the standardized scores are much smaller in terms of magnitude than the SVM. The average of all absolute values of the standardized scores of SVM is 0.941 while that of the rule-based approach is only 0.264. A different score standardization method would possibly result in a different weighting.

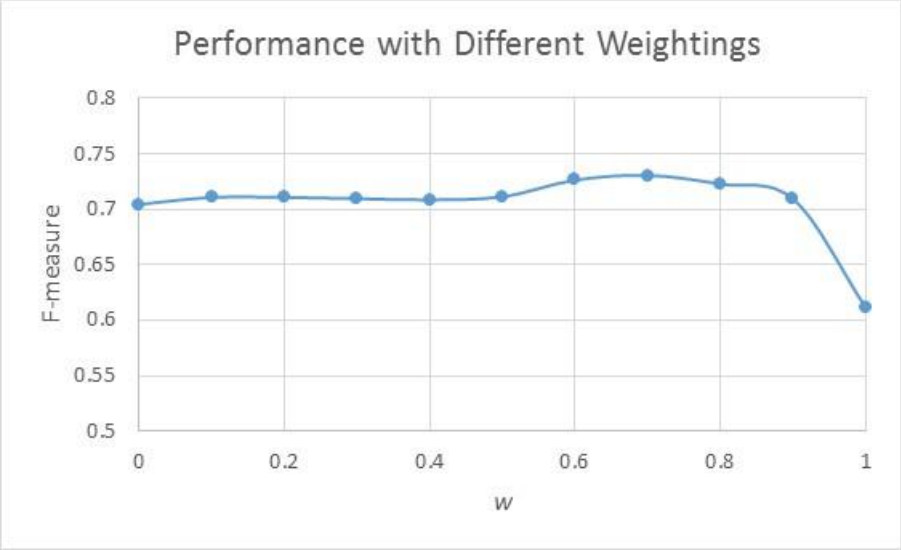

Figure 7. Performance of Classification Aggregation with Different Weightings

Comparison with Other Classifiers and Classifier Combinations

We chose to use SVM in this research as it has achieved the best performance in various text classification tasks (Yang & Liu, 1999; Abbasi et al., 2008), especially when the number of positive training instances is small (Yang & Liu, 1999). To verify if SVM is indeed suitable for our data set, we

compare it with two other popular text classifiers, namely a Naïve Bayes classifier and a decision tree classifier.

In addition, as discussed earlier, one main reason for combining SVM with a rule-based classifier is that while SVM provides good classification performance without considering word order, the rule-based approach provides sentence-level and paragraph-level analysis. We postulate that such combination will perform better than combining two similar classification approaches. We perform additional experiments to validate this.

The comparison results are shown in Table 6. As can be seen, both the Naïve Bayes classifier (0.6211) and the decision tree classifier (0.6679) perform worse than the simple SVM classifier (0.6837 as shown in Model 2 in Table 3) in terms of F-measure. The results affirm our choice of the SVM classifier in our design. Our results also show that the proposed approach combining SVM and rule-based classification, which considers sentence-level and paragraph-level analysis, achieves a higher F-measure (0.7216 as shown in Model 8 in Table 3) than an aggregation of SVM and a decision tree classifier (0.6957) and an aggregation of SVM and a Naïve Bayes classifier (0.6850). It supports our postulation that combining a machine learning classifier with a rule-based classifier performs better than combining two machine learning classifiers in this application.

Table 6: Performance of Selected Classifier Combinations

Model	Precision	Recall	F-measure
Naïve Bayes	0.5980	0.6460	**0.6211**
Decision Tree	0.6679	0.6679	**0.6679**
SVM(C-LIWC + GA) + Naïve Bayes	0.6145	0.7737	**0.6850**
SVM(C-LIWC + GA) + Decision Tree	0.6420	0.7591	**0.6957**
SVM(C-LIWC + GA) + Rule-based	0.6287	0.8467	**0.7216**

Misclassification Analysis

In addition to the quantitative analysis, we also sought to identify the causes for misclassification. Analysis of such contents is conducive to improving the classification performance and creating new categories for capturing finer details in the future. First, contents not showing emotional distress but wrongly identified (i.e., false positives) are analyzed qualitatively. The first problem is that some sensational writings, such as movie reviews and tragedy fictions, show similar writing and linguistic styles to those of depressed people (Pennebaker et al. 2003). It is difficult for the model to distinguish between these two kinds of writings using either the rule-based approach or SVM because the word usage patterns are very similar. Second, informal and colloquial Chinese expressions in some of the blog posts cannot be understood by the models if the words were not found in C-LIWC (a dictionary of formal written Chinese) in SVM or our lexicon in the rule-based approach. Third, while emotionally distressed individuals tend to use significantly more self-referencing words in their writings (Rude et al., 2001; Stirman & Pennebaker, 2001; Sloan, 2005), the same applies to people with other characteristics, such as self-conscious. However, we cannot distinguish whether those people are emotionally distressed by the self-reference category in the C-LIWC alone. Therefore, the model cannot always correctly distinguish people with emotional distress from people with other characteristics. The rule-based classifier has addressed these issue by correctly classifying some of the "marginal" posts and complement the classification of SVM to achieve a better overall performance of the aggregation approach.

Similarly, we also analyze the blogs that showed emotional distress but were wrongly identified as not showing distress (i.e., false negatives), and we found several causes. First, some of these blogs were written in a rather implicit way. For example, they may use analogy, metaphor, or sarcasm and contain few negative words, so both the SVM and the rule-based classifiers did not pick them up. Second, similar to false positives, some blogs were written in informal or colloquial Chinese. These blogs contain words that are judged as showing emotional distress by our human judges, but did not match the words in our lexicons, especially the C-LIWC dictionary which contains formal written Chinese only. Finally, as

discussed earlier, some contents are too short or contain very few negative emotion words. They may be expressing emotional distress in a single phrase and there is not enough information for the model to make the correct prediction.

It is worthwhile to note that many of the reasons for misclassification discussed are similar to those for misclassification in the traditional opinion mining and sentiment analysis literature, such as the use of implicit wordings, idiomatic expressions, or irony (Pang & Lee, 2008; Balahur et al. 2006).

From the detection and possibly life-saving prospects, it is important to boost the recall rate in order to identify more people showing emotional distress online while keeping a satisfactory precision rate. The prediction threshold was adjusted so that most recall rates shown in Table 3 are adequately high to capture those at-risk individuals. Our results show that the proposed architecture has achieved satisfactory performance. More blog posts will be selected, and false alarms should be reasonably allowed in a conservative manner. A model with a high precision rate but a low recall rate is not favorable as it might miss out on some potentially needy individuals.

Study 2: Evaluation by Professionals

A user study was designed and conducted to evaluate whether and how users can benefit from using the system for their work in a real usage scenario. The user study has two settings. The first setting aimed to evaluate the difference between the number of online posts showing emotional distress identified by the usual search method (e.g., through Google or Yahoo) and by the proposed search engine. The study also evaluated user experiences of the search process. The second setting aimed to compare the proposed search engine with classifier aggregation against one without the aggregation (i.e., using the SVM classifier only).

Comparison with Regular Blog Search Engines

In the user study, participant were asked to imagine themselves as in an Internet outreaching team hired to identify as many posts showing emotional distress as possible using the search engines. Each participant

was paid HK$200 for participating in the study and the one who correctly identified the most number of posts showing emotional distress was given an extra HK$200 as an incentive. In the first setting, participants were required to complete the searching tasks by using a regular blog search engine and KAREN separately. When using KAREN, the search results were displayed to the participants in a way similar to a standard search engine. Ten results were shown on each result page and each result contains the title and a snippet. The order of the two search engines in the user study was randomized. Participants were asked to come up with their own search queries and input them into the search engine. They then browsed the search results pages and could freely click on any of the search results to see the content of the actual online post. After the assessment, if they found the blog post to be showing emotional distress or suicidal ideation, they were required to record the URLs of the identified blog posts in an Excel file. Each search task lasted for 15 minutes, and all activities on the screen (including typing, mouse movements, and web pages visited) were recorded using a software program. After completing each search task, participants were asked to fill out a questionnaire about their experience of using the search engine.

Two main measures were used to evaluate the performance of KAREN in this study. First, we counted the number of people with emotional distress identified by the professionals in each session to measure the effectiveness of the search engines. Second, we evaluated how the professionals perceived the usefulness of the search engines. Six standard questionnaire items were used to measure perceived usefulness (Davis, 1989).

To recruit participants for the first setting of the user study, a mass email was sent to all postgraduate students in social sciences in a large university in Hong Kong. Participants were required to have previous experience in social work services. Twenty-two participants, with a mean of 4.05 years of experience in Internet outreach and online counseling services, participated in the study.

Two clinical psychologists familiar with the research domain were asked to rate the blogs found by the participants. They first rated the posts independently and then discussed the inconsistently rated

ones to reach a consensus. The results show that on average, participants were able to find significantly more individuals with emotional distress, as measured by the number of posts they found to show emotional distress, using KAREN (5.409), than the regular blog search engine that they were most familiar with, i.e., either Google or Yahoo blog search (3.864). The false positive rate of KAREN (0.148) is also much lower than that of the regular blog search engine (0.365). It shows that professionals using KAREN can identify people showing emotional distress more accurately. This would allow them to save time in their search process and to better focus their resources on those who are actually in need.

Participants also found KAREN to be more useful in completing their task, with a significantly higher perceived usefulness (4.773) than the regular blog search engine (3.939). Paired *t*-tests showed that the differences are statistically significant for both the number of posts correctly identified and the perceived helpfulness ($p < 0.05$).

Comparison with a System with SVM Only

The second setting of the user study was conducted in a very similar way to the first setting, except that participants were asked to search for online posts showing emotional distress by using the proposed search engine and by using a similar search model using SVM only (i.e., without the combination of the rule-based classifier in the classification process). Other configurations were the same as the first setting.

Nineteen participants, who have on average 3.85 years of experience in Internet outreach and online counseling services, participated in the second setting of our user study. The results show that participants were able to find more blog posts showing emotional distress using KAREN which combines SVM and rule-based classification (5.316) than using the model which uses the SVM classifier only (4.842). A paired *t*-test shows that the difference is marginally significant ($p < 0.1$). Participants also rated KAREN with a higher perceived usefulness score (4.623) than the SVM-only model (4.526), but the difference is not statistically significant. A possible reason for the insignificant difference is that the two search engines have the same user interface, and participants might not have noticed the differences in the backend algorithm.

Overall the results show that social work professionals benefit from using the proposed system. On average, the professionals were more effective in performing their tasks when using the proposed system with classifier aggregation than a system with the SVM classifier only.

5. Discussion

Contributions and Implications

This study has several important implications for the research on sentiment and affect analysis techniques, emotional distress and suicidal behavior, and the practice of social work and suicide prevention.

One major contribution of this research is the unique design that aggregates two classification techniques together with domain-specific lexicons and a GA-based feature selection component to analyze emotions expressed in user-generated contents in blogs. The proposed aggregation method achieves the best classification performance, compared with existing methods that use only one single technique or models that combine two machine learning classifiers. The results suggest that such specifically crafted rule-based classifiers may as well be needed in other domains for achieving better classification performance over traditional word-based or lexicon-based machine learning approaches. In addition, we have shown that the use of GA-based feature selection with SVM and rule-based classifier achieves very good performance in the classification task, compared with the baseline approaches. GA-based feature selection has not previously been used in this type of classification tasks, and the promising result reported here suggests that classification applications for emotion-related documents based on LIWC can benefit from the feature selection techniques. Further research would be desirable.

A second contribution is that we investigated under what conditions the aggregation method performs better. Consistent with many other studies in the literature, SVM is a suitable classifier for our textual data. Based on our analysis of the trained hyperplane of our SVM, we found that SVM is able to capture the relationship between certain keywords (mostly negative emotion words) with emotional distress in many cases. However, as discussed earlier, it is also noted that SVM does not perform well when the blog post is too long or too short, or when there are too many positive emotion words or too few

negative emotion words. This is because SVM still relies heavily on the occurrences of keywords that are good indicators of the class of the posts and does not consider the sentence- or paragraph-level context of the posts, resulting in some misclassification. On the other hand, the rules obtained from experts facilitate sentence- and paragraph-level analysis and consider the document structure and context. For example, a temporal word (e.g., "tomorrow") might not convey a special meaning when it appears alone, but would be very important in the classification process if it appears together with a suicide-related word. Such relationship has been captured in our expert rules.

As discussed earlier, we find that the rule-based classifier adds the most value when the blogs are of medium length or have very few negative emotion words. When the blog post length is medium, it can take advantage of its sentence- and paragraph-level analysis without suffering from the other problems, and thus adds the most value. These findings confirm our argument that traditional classification approaches that rely on keywords only without looking at their relationship or other cues may miss some blogs with emotional distress. More generally speaking, our findings show that a rule-based classifier will add the most value to a machine learning classifier for documents where the classification target, such as emotional distress or sentiment, are not expressed by explicit keywords.

In addition, we showed how weightings of the two classifiers can be adjusted to achieve better performance. Based on our experiment, we found that the aggregation performs the best when the value of w is 0.7. While SVM is still more useful in making the correct classification, the rule-based approach takes a complementary role and is especially useful for cases discussed above.

Our findings have several implications on the design of text classifiers. First, our results showed the limitations of keyword-based classification approach such as SVM, especially in the identification of emotional distress or other characteristics that could be quite implicit. Researchers should be cautious about such limitations in their design. Our findings also confirm that aggregating the results from different classification methods improves classification performance. The limitations of a keyword-based classification can be addressed by having a classifier with a different nature, such as a rule-based

classifier. In addition, our results show that SVM performs poorly under some conditions. Researchers need to pay attention to these conditions and consider using different classifier methods or different aggregation weighting under such conditions in order to achieve better performance.

Our research has important implications for social work practices. Emotional distress is a robust risk factor for suicidal behavior and the early detection of high-risk individuals is the key to prevent future suicidal behavior (Turecki et al., 2016). The approach proposed in this study and the system developed based on this approach are useful for social work professionals to identify bloggers with emotional distress. It is more effective and efficient than using the traditional search approaches. The system will reduce the manual efforts of the social work professionals in browsing and searching such that they can focus their attention on interacting with and providing assistance to those in need. Even though the improvement of the aggregation approach is only about 2.5% over the SVM classifier with C-LIWC and GA, this is still of practical importance in terms of the time saved and the number of true positives identified over the long run. This is especially important in situations where individuals with emotional distress and suicidal ideation need urgent help.

In addition, because the proposed approach assists professionals in identifying people at high risk more effectively, it will enable suicide prevention programs to provide more proactive and effective services. Traditional programs offering suicide hotlines or web resources are primarily reactive services which passively wait for the emotionally distressed to call or seek help in other ways (Barak, 2005; Gilat & Shahar, 2007; Luxton et al., 2011). Our proposed approach makes it possible for these programs to identify and reach out to those in need more effectively and provide timely assistance. In particular, the timely identification of youth with emotional distress and suicidal intentions and early interventions can increase the awareness among parents, schools, educators, and public health providers and help prevent tragic incidents from happening (Resnick et al., 1997), thereby reducing the possible medical, financial, and social costs associated with adolescent suicides.

Limitations

One limitation of the study is that we do not know the true psychological status of an individual who blogs about his/her emotional distress. The entire process of annotation of emotional distress relied on the textual information of the posts. It is possible that some people experiencing emotional distress never talk about their true feelings and emotions in their blogs (thus cannot be identified by our approach), while some other individuals blog about their emotional distress simply for attracting attention.

Another limitation is that the demographic characteristics of bloggers, such as gender, are not investigated. It has been found that males and females use different emotional expressions in computer-mediated communication (Thelwall et al., 2010). Although important in practice, gender differences are not incorporated into the classification process. Besides, the definition of emotional distress is complicated and subjective in nature. It is very difficult to have a clear-cut definition or classification scheme for determining which bloggers, based on their use of words and the way they write, are absolutely distressed. This has introduced possible imprecision in the machine learning process and evaluation of the classification.

6. Conclusion and Future Work

A considerable amount of studies have been carried out using machine learning in text classification applications, which often concentrate on business domains. Multiple online applications using classification-based sentiment analysis techniques have emerged to collect customers' opinions on various products and services. Research and applications on online identification of at-risk individuals to promote public health are, however, relatively rare.

Our current exploratory study demonstrates the effectiveness of the proposed approach, based on which the KAREN system was developed. It is particularly important for practical use in online detection of emotionally distressed individuals. Our evaluation studies show that the novel approach combining different techniques can facilitate the identification work, in terms of time and cost efficiencies. It is

expected that the limited and scarce resources can be shifted from the labor-intensive searching job to the implementation of intervention measures so that more people in need can receive help.

In the future, we will improve our approach and system in various aspects. The practicability and efficacy of the approach will be further evaluated. Since the prevalence rate of depression is usually low, the number of blog posts showing emotional distress is not large. The real-life application is expected to process input samples composed of a large number of normal blog posts and a relatively small number of posts showing emotional distress. In our future work, we will therefore evaluate the approach in order to show its practicability and efficacy with large datasets more representative of the real-world conditions.

Another future research direction is to analyze bloggers not only by a single post but by multiple posts. A certain amount of previous posts of bloggers, for instance, the posts in the past three months, can be analyzed to predict their emotional fluctuation. The day-to-day variation of writing styles is harnessed to predict one's health condition and frequency of visiting physicians for illnesses (Campbell & Pennebaker, 2003). It is possible to use a similar method to identify people with possible mental illness. Since decisions on whether or not one is showing emotional distress only rely on the content of the posts, there is an abundance of information such as other bloggers' comments and their interactions that can be analyzed to have a better understanding of the bloggers' thoughts. We also plan to investigate the use of social network analysis (Chau & Xu, 2012) in studying online communities who appear to be emotionally distressed.

Moreover, feature categories can be added or modified in future research. For example, pure verbal expressions omit rich information on body language, facial expression, and voice accentuation, all of which are extremely helpful in determining people's emotions. People indeed develop different symbols to express their emotion in computer-medicated communications. Emoticons, parenthetical expressions, and other commonly used symbols that convey thoughts and feelings may be incorporated in the classification approach in future work.

Lastly, our current approach only identifies blog posts showing emotional distress, but does not predict whether the author has any self-harm intentions. It would be highly valuable to further improve the proposed model such that it can also provide an estimation of such intentions, such that social work and health professionals can find these people and provide help timely.

References

Abbasi, A. & Chen, H. (2007). Affect intensity analysis of dark web forums. *Proceedings of IEEE International Conference on Intelligence and Security Informatics*, 282-288.

Abbasi, A. & Chen, H. (2008). CyberGate: A system and design framework for text analysis of computer mediated communication. *MIS Quarterly*, *32*(4), 811-837.

Abbasi, A., Chen, H., Thoms, S., & Fu, T. (2008). Affect analysis of web forums and blogs using correlation ensembles. *IEEE Transactions on Knowledge and Data Engineering*, *20*(9), 1168-1180.

Aisopos, F., Papadakis, G., Tserpes, K., and Varvarigou, T. (2012). Content vs. context for sentiment analysis: a comparative analysis over microblogs. In *Proceedings of the 23rd ACM conference on Hypertext and social media (HT '12)*. ACM, New York, NY, USA, 187-196.

AppleDaily (2008, June 29). *AppleDaily News*.

Attardi, G., & Simi, M. (2006). Blog mining through opinionated words. *Proceedings of the Fifteenth Text Retrieval Conference (TREC 2006)*.

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M, Pouliquen, B., & Belyaeva, J. (2006). Sentiment analysis in the news. *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2006).*

Barak, A. (2005). Emotional support and suicide prevention through the Internet: A field project report. *Computers in Human Behavior*, *23*, 971-984.

Bolier, L., Haverman, M., Westerhof, G. J., Riper, H., Smit, F., & Bohlmeijer, E. (2013). Positive psychology interventions: a meta-analysis of randomized controlled studies. *BMC Public Health*, 13(1), 1.

Bolon-Canedo, V., Sanchez-Marono, N., & Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. *Knowledge and Information Systems*, *34*(3), 483-519.

Borges, G., Nock, M. K., Haro Abad, J. M., et al. (2012). Twelve-month prevalence of and risk factors for suicide attempts in the World Health Organization World Mental Health Surveys, *Journal of Clinical Psychiatry, 71*, pp. 1617–1628.

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, March/April Issue, 15-21.

Campbell, R. S., & Pennebaker, J. W. (2003). The secret life of pronouns: Flexibility in writing style and physical health. *Psychological Science*, *14*, 60–65.

Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media and Society*, *16*(2), 340-358.

Chau, M. & Xu, J. (2012). Business intelligence in blogs: Understanding consumer interactions and communities. *MIS Quarterly*, *36*(4), 1189-1216.

Chen, J. H., Bierhals, A. J., Prigerson, H. G., Kasl, S. V., Mazure, C. M., & Jacobs, S. (1999). Gender differences in the effects of bereavement-related psychological distress in health outcomes. *Psychological Medicine*, 29(2), 367-380.

Chen, H., Fan, H., Chau, M., & Zeng, D. (2001). MetaSpider: Meta-searching and categorization on the web. *Journal of the American Society for Information Science and Technology*, *52*(13), 1134-1147.

Cheng, A. T., Chen, T. H., Chen, C. C., & Jenkins, R. (2000). Psychosocial and psychiatric risk factors for suicide. *The British Journal of Psychiatry*, 177(4), 360-365.

Cheng, Q., Kwok, C. L., Zhu, T., Guan, L., & Yip, P. S. F. (2015). Suicide communication on social media and its psychological mechanisms: an examination of Chinese microblog users. *International Journal of Environmental Research and Public Health*, *12*, 11506-11527.

Coppersmith, G. A., Harman, C. T., & Dredze, M. H. (2014). Measuring post-traumatic stress disorder in Twitter. *Proceedings of AAAI International Conference on Weblogs and Social Media*, Québec, Canada.

Crawford, J. R., & Henry, J. D. (2004). The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample. *British Journal of Clinical Psychology*, 43(3), 245-265.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319-340.

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Predicting postpartum changes in emotion and behavior via social media. *Proceedings of ACM CHI Conference on Human Factors in Computing Systems*, 3267–3276.

Fang, X., Sheng, O. R. L., & Chau, M. (2007). ServiceFinder: A method towards enhancing service portals. *ACM Transactions on Information Systems*, *25*(4), Article 17.

Feigelman, W. & Gorman, B. S. (2008). Assessing the effects of peer suicide on youth suicide. *Suicide and Life-Threatening Behavior*, *38*(2), 181-194.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, *56*(4), 82-89.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, *3*, 1289–1305.

Gilat, I. & Shahar, G. (2007). Emotional first aid for suicide crisis: Comparison between telephone hotline and Internet. *Journal of Psychiatry Interpersonal & Biological Processes*, *70*(1), 12-18.

Gill, A. J., French, R. M., Gergle, D., and Oberlander, J. (2008). The language of emotion in short blog texts. *Proceedings of ACM Conference on Computer-Supported Collaborative Work (CSCW)*, San Diego, California, USA.

Glance, N., Hurst, M., Nigam, K., Siegler, M., Stockton, R., & Tomokiyo, T. (2005). Analyzing Online Discussion for Marketing Intelligence. In *Proceedings of WWW 2005*, Chiba, Japan.

Goldberg, D. (1989). *Genetic algorithms in search, optimization, and machine learning*. New York: Addison-Wesley.

Gregor, S. & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly*, *37*(2), 337-356.

Gruber, J. & Kring, A. M. (2008). Narrating emotional events in schizophrenia. *Journal of Abnormal Psychology*, *117*(3), 520–533.

Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Hearst, M. A. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, *23*, 33-64.

Hessler, R.M., Downing, J., Beltz, C., Pelliccio, A., Powell, M., & Vale, W. (2003). Qualitative research on adolescent risk using e-mail: A methodological assessment. *Qualitative Sociology*, *26*(1), 111-124.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75-105.

Hong Kong Education Bureau (2016). *Report of Committee on Prevention of Students Suicides*, Hong Kong Government.

Hsu, C.-W., Chang, C.-C., and Lin. C.-J. (2003). A practical guide to support vector classification. Technical Report, Department of Computer Science, National Taiwan University. Retrieved from: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Huang, C. L., Chung, C. K., Hui, N., Lin, Y. C., Seih, Y. T., Chen, W. C., . . . Pennebaker, J. W. (2012). The development of the Chinese Linguistic Inquiry and Word Count dictionary. *Chinese Journal of Psychology*, *54*(2), 185-201.

Huang, Y., Goh, T., & Liew, C. L. (2007). Hunting suicide notes in Web 2.0 – preliminary findings. *Proceedings of the IEEE International Symposium on Multimedia – Workshops*.

Hutto, C. J. & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text, *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*.

Ishida, K. (2005). Extracting latent weblog communities - A partitioning algorithm for bipartite graph. *Proceedings of the 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, Chiba, Japan.

Joiner, Jr., T.E. (2005). *Why people die by suicide*. Cambridge, MA: Harvard University Press.

Juffinger, A. & Lex, E. (2009). Crosslanguage blog mining and trend visualisation. *Proceedings of the World Wide Web*, Madrid, Spain.

Junghaenel, D., Smyth, J. M., & Santner, L. (2008). Linguistic dimensions of psychopathology: A quantitative analysis. *Journal of Social & Clinical Psychology*, *27*(1), 36–55.

Kuhl, J., Quirin, M., & Koole, S. L. (2015). Being someone: The integrated self as a neuropsychological system. *Social and Personality Psychology Compass*, 9(3), 115-132.

Kumar, R., Novak, J., & Tomkins, A. (2010). Structure and evolution of online social networks. In P. Yu, J. Han, & C. Faloutsos (Eds.), *Link mining: Models, algorithms, and applications* (pp. 337-357).

Lee, L., Pang, B., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques, *Proceedings of EMNLP*, 79–86.

Leung, P. W., Hung, S. F., Ho, T. P., Lee, C. C., Liu, W. S., Tang, C. P., & Kwong, S. L. (2008). Prevalence of DSM-IV disorders in Chinese adolescents and the effects of an impairment criterion. *European Child & Adolescent Psychiatry*, 17(7), 452-461.

Li, T. M. H., Chau, M., Wong, P. W. C., & Yip, P. S. F. (2014). Temporal and computerized psycholinguistic analysis of the blog of a Chinese adolescent suicide. *Crisis*, *35*(3), 168-175.

Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers.

Liu, H., & Motoda, H. (2012). *Feature Selection for Knowledge Discovery and Data Mining* (vol. 454). Springer Science & Business Media, Chicago.

Liu, Y., Huang, X., An, A., & Yu, X. (2007). ARSA: A sentiment-aware model for predicting sales performance using blogs. *Proceedings of the ACM SIGIR Conference,* Amsterdam, The Netherlands.

Luxton, D. D., June, J. D., & Fairall, J. M. (2012). Social Media and Suicide: A Public Health Perspective. *American Journal of Public Health, 102(S2)*, S195-S200.

Luxton, D. D., June, J. D., & Kinn, J. T. (2011). Technology-based suicide prevention: Current applications and future directions. *Telemedicine and e-Health*, *17*(1), 50-54.

Lyubomirsky, S. & Layous, K. (2013). How do simple positive activities increase well-being?. *Current Directions in Psychological Science*, 22(1), 57-62.

Macdonald, C., Santos, R.L.T., Ounis, I., & Soboroff, I. (2010). Blog track research at TREC. *SIGIR Forum*, *44*(1), 58-75.Matthews, G., Jones, D. M., & Chamberlain, A. G. (1990). Refining the measurement of mood: The UWIST mood adjective checklist. *British Journal of Psychology*, 81(1), 17-42.

Michalewicz, Z. (1996). *Genetic algorithms + data structures = evolution programs*. Springer.

Mishne, G. & de Rijke, M. (2006). Capturing global mood levels using blog posts. *Proceedings of the AAAI Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*.

Mullen. T. & Collier, N. (2004). Sentiment analysis using support vector machines with diverse

information sources, *Proceedings of the 9th Conference on Empirical Methods in Natural Language Processing (EMNLP-04).*

Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2010). User study on AffectIM, an avatar-based Instant Messaging system employing rule-based affect sensing from text. *International Journal of Human-Computer Studies*. 68(7), 432-450.

Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2011). Affect Analysis Model: novel rule-based approach to affect sensing from text. *Natural Language Engineering*. 17(1), 95- 135.

Newman, D. B., Tay, L., & Diener, E. (2014). Leisure and subjective well-being: A model of psychological mechanisms as mediating factors. *Journal of Happiness Studies*, 15(3), 555-578.

Niesler, T. & Woodland, P. (1996). A variable-length category-based N-gram language model. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, GA, I164–I167.

Oreski, S. & Oreski, G. (2014). Genetic algorithm-based heuristic for feature selection in credit risk assessment. *Expert Systems with Applications*, 41(4), 2052-2064.

Pang, B. & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1-135.

Pennebaker, J. W. (2003). Putting stress into words: Health, linguistic, and therapeutic implications. *Behaviour Research and Therapy*, *31*(6), 539-548.

Pennebaker, J. W. and Chung, C. K. (2011). Expressive writings: Connections to physical and mental health. In H. S. Friedman (Eds.), *The Oxford handbook of health psychology.* Oxford University Press.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). The development and psychometric properties of LIWC 2007. LIWC.Net, Austin, TX, USA.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54(1),* 547-577.

Ramirez-Esparza, N., & Pennebaker, J. W. (2006). Do good stories produce good health? Exploring words, language, and culture. *Narrative Inquiry*, *10*(1), 211-219.

Riloff, E., & Wiebe, J. (2003). Learning extraction patterns for subjective expressions. *Proceedings of EMNLP*.

Rodham, K., Gavin, J., & Miles, M. (2007). I hear, I listen and I care: A qualitative investigation into the function of a self-harm message board. *Suicide and Life-Threatening Behavior*, *37*(4), 422-430.

Rude, S. S., Gortner, E. M., & Pennebaker, J. W. (2004). Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, *18*, 1121–1133.

Ruder, T. D., Hatch, G. M., Ampanozi, G., Thali, M.J., and Fischer, N. (2011). Suicide announcement on Facebook. *Crisis*, *32(5)*, pp. 280-282.

Saad, F. (2014). Baseline evaluation: An empirical study of the performance of machine learning algorithms in short snippet sentiment analysis. *Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business* 6.

Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507–2517.

Samuelsson, C., & Reichl, W. (1999). A class-based language model for large-vocabulary speech recognition extracted from part-of-speech statistics, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Mar 1999, Phoenix, AZ.

Scouller, K. M., & Smith, D. I. (2002). Prevention of youth suicide: How well informed are the potential gatekeepers of adolescents in distress? *Suicide and Life-Threatening Behavior*, *32*(1), 67-79.

Sloan, D. M. (2005). It's all about me: Self-focused attention and depressed mood. *Cognitive Therapy and Research*, *29*, 279 - 288.

Smith, A. R., Witte, T. K., Teale, N. E., King, S. L., Bender, T. W., & Joiner, T. E. (2008). Revisiting impulsivity in suicide: Implications for civil liability of third parties. *Behavioral Sciences & the*

Law, *26*(6), 779–797.Snoek, F. J., Pouwer, F., Welch, G. W., & Polonsky, W. H. (2000). Diabetes-related emotional distress in Dutch and US diabetic patients: cross-cultural validity of the problem areas in diabetes scale. *Diabetes Care*, 23(9), 1305-1309.

Stirman, S. W., & Pennebaker, J. W. (2001). Word use in poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine*, *63*, 517-522.

Subasic, P., & Huettner, A. (2000). Affect analysis of text using fuzzy semantic typing. *Proceedings of the Ninth IEEE International Conference on Fuzzy Systems, San Antonio, TX, USA*, 647-652.

Tang, L., & Liu, H. (2010). "Understanding Group Structures and Properties in Social Media," in *Link Mining: Models, Algorithms, and Applications* (eds: Yu, P., Han, J. and Faloutsos, C.), 163-185.

Thelwall, M., Wilkinson, D., & Uppal, S. (2010). Data mining emotion in social network communication: Gender differences in Myspace. *Journal of the American Society for Information Science and Technology*, *61*, 190–199.

Turecki, G. and Brent, D. A. (2016). Suicide and suicidal behavior. *The Lancet* 387.10024. 1227-1239.

Van Rijsbergen, C.J. (1979). *Information retrieval* (2nd ed.). London: Butterworths.

World Health Organization (2014). *Preventing suicide: A global imperative*. Retrieved from: http://apps.who.int/iris/bitstream/10665/131056/1/9789241564779_eng.pdf

Wu, C.-H., Chuang, Z.-J., and Lin, Y.-C. (2006). Emotion Recognition from Text Using Semantic Labels and Separable Mixture Models. *ACM Transactions on Asian Language Information Processing (TALIP)*, 5(2), pp. 165-183.

Yang, H., Si, L., & Callan, J. (2006). Knowledge transfer and opinion detection in the TREC2006 Blog Track. *Proceedings of the Fifteenth Text Retrieval Conference (TREC 2006).*

Yang, J. & Honavar, V. (1998). Feature subset selection using a genetic algorithm. In: H. Liu, & H. Motoda (Eds.) *Feature Extraction, Construction and Selection: A Data Mining Perspective*, 117-136. Kluwer.

Yang, Y. & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. *Proceedings of ICML-97, 14th International Conference on Machine Learning (Nashville, US, 1997)*, 412–420.

Yang, Y. & Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 42-49). ACM.

Yu, L. & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, *5*, 1205–1224.

Zawadzki, M. J., Smyth, J. M., & Costigan, H. J. (2015). Real-time associations between engaging in leisure and daily health and well-being. *Annals of Behavioral Medicine*, 49(4), 605-615.

Zhang, H. P., Yu, H. K., Xiong, D.Y., & Liu, Q. (2003). HMM-based Chinese lexical analyzer ICTCLAS. *Proceedings of the 2nd SIGHAN Workshop*, 184-187.

Zhang, C., Zeng, D., Li, J., Wang, F.Y., & Zuo, W. (2009). Sentiment analysis of Chinese documents: From sentence to document level. *Journal of the American Society for Information Science and Technology*, *60*(12), 2474-2487.

Zeng, D., Wei, D., Chau, M., & Wang, F. (2011). Domain-specific Chinese word segmentation using suffix tree and mutual information. *Information Systems Frontiers*, *13*(1), 115-125.