MIS Quarterly

RESEARCH NOTE

FINDING PEOPLE WITH EMOTIONAL DISTRESS IN ONLINE SOCIAL MEDIA: A DESIGN COMBINING MACHINE LEARNING AND RULE-BASED CLASSIFICATION[1]

Michael Chau
Faculty of Business and Economics, The University of Hong Kong,
Pokfulam, HONG KONG {mchau@business.hku.hk}

Tim M. H. Li
Department of Social Work and Social Administration, The University of Hong Kong,
Pokfulam, HONG KONG {tim.mh.li@connect.hku.hk}

Paul W. C. Wong
Department of Social Work and Social Administration, The University of Hong Kong,
Pokfulam, HONG KONG {paulw@hku.hk}

Jennifer J. Xu
Computer Information Systems, Bentley University,
Waltham, MA 02452 U.S.A. {jxu@bentley.edu}

Paul S. F. Yip
HKJC Center for Suicide Research and Prevention, Faculty of Social Sciences, and
Department of Social Work and Social Administration, The University of Hong Kong,
Pokfulam, HONG KONG {sfpyip@hku.hk}

Hsinchun Chen
Department of Management Information Systems, The University of Arizona,
Tucson, AZ 85721 U.S.A. {hchen@eller.arizona.edu}

Many people face problems of emotional distress. Early detection of high-risk individuals is the key to prevent suicidal behavior. There is increasing evidence that the Internet and social media provide clues of people's emotional distress. In particular, some people leave messages showing emotional distress or even suicide notes on the Internet. Identifying emotionally distressed people and examining their posts 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 system performance in searching and analyzing blogs written by people who might be emotionally 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 approaches.

[1]Sumit Sarker was the accepting senior editor for this paper. Manish Agrawal served as the associate editor.

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

Introduction

Emotional distress is a prevailing complex social and public-health problem in modern societies. About 9.2% of people worldwide have had suicidal ideation at least once in their lifetime, 2% have had that in the past 12 months (Borges et al. 2010), and around 804,000 individuals take their own lives every year (World Health Organization 2014). Emotional distress is a robust risk factor for suicidal behavior, and the early detection of high-risk individuals is the key to prevent suicidal behavior (Turecki et al. 2016). 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 suicide had expressed their intentions on social media (Hong Kong Education Bureau 2016). It has been suggested that content on the Internet, especially narratives and diaries written online, have great potential for understanding people's emotional distress and suicidal behaviors (Cheng et al. 2015; Hessler et al. 2003; Huang et al. 2007). In view of this, some nongovernmental organizations (NGOs) have started to actively search for these distressed and negative self-expressions in social media to identify potentially severely depressed people in order to provide help and follow-up services. However, 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 content expressing emotional distress (e.g., Huang et al. 2007). The search results are often rather "noisy" and the search targets are buried under a large number of irrelevant documents, and only a few texts showing 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 entry written by someone who expresses suicidal intention. Social workers and professionals often have to spend a large amount of time to identify people who truly need help.

Techniques for text mining and affect analysis have advanced substantially in recent years (Liu 2012; Pang and Lee 2008). Web-mining and text-mining techniques have achieved satisfactory performance in extracting opinions and identifying communities in blogs (Abbasi et al. 2008; Chau and Xu 2012; Ceron et al. 2014; Glance et al. 2005; Ishida 2005; Juffinger and Lex 2009; Kumar et al. 2010; Liu et al. 2007; Pang and Lee, 2008; Tang and Liu 2010). Many 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., friends and community). Although these techniques could help with this potentially life-saving application, little empirical research has been conducted.

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.

Theoretical Background and Related Work

Sentiment and Affect Analysis Using Machine Learning

Sentiment and Affect Analysis

Machine learning has been extensively used in text-based classification and object recognition with great success in a wide range of applications, including sentiment and affect analysis (Cambria et al. 2013; Feldman 2013). 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 and Huettner 2000). For example, the intensity of the general public's moods during a bombing incident in London was estimated with word frequencies and the usage of special characters in blogs (Mishne and de Rijke 2006).

The affect intensities of web forums and blog messages were also evaluated in previous research, and the results were encouraging, showing that affects could be detected automatically (Abbasi et al. 2008). Among various 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 and Collier 2004; Saad 2014). The SVM-based approach inherently emphasizes document-level analysis. It is a well-known and highly effective approach yielding high accuracy in sentiment and affect analysis (Abbasi et al. 2008; Mullen and Collier 2004).

Lexicon-Based Feature Extraction

Most machine learning methods rely on features, which are variables or predictors, that are present in the data. A well-developed lexicon can be used to make the features extracted more specific to a particular domain. For instance, the linguistic inquiry and word count (LIWC) lexicon (Pennebaker et al. 2007) organizes words into different categories so that researchers can employ them as features for analysis. It has been suggested that this kind of category-based features can avoid the ambiguous nature of many words to greatly improve language-model perplexities (Niesler and Woodland 1996; Samuelsson and Reichl 1999). LIWC has been used in sentiment analysis studies in public health. For example, preliminary evidence suggests that depressed individuals have a different writing style from that of non-depressed people (Rude et al. 2004; Pennebaker and Chung, 2011). Depressed and suicidal individuals tend to use significantly more self-referencing words in their writing (Rude et al. 2004; Sloan 2005; Stirman and Pennebaker 2001). Other categories of words, such as negations, cognitive words, and positive and negative emotional words, also are used to identify the writing styles in mentally ill patients (Gruber and Kring 2008; Junghaenel et al. 2008). The LIWC lexicon translated into different languages is widely used in analyses of user-generated content, including blogs and microblogs (e.g., Coppersmith et al. 2014; De Choudhury et al. 2013; Gill et al. 2008).

Feature Selection Techniques

Feature selection techniques can be used to reduce the number of features by finding the optimal subset of features that achieve the best classification performance. Feature selection is a crucial preprocessing 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 (Fang et al. 2007; Oreski and Oreski 2014; Yang and Honavar 1998). 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 populations.

Rule-Based Classification with Expert Judgment

Although machine learning techniques are shown to perform well in various text classification tasks, some drawbacks exist. First, they are entirely data driven. If the training data set is biased, it may affect the classification performance. Second, expert judgment and experience cannot be incorporated into the model. Third, 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, where some rules developed by experts are used to assign a 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. Combining both machine learning and rule-based classification to take advantage of both approaches may be beneficial.

System Design

This research aims 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 followed the design science methodology (Gregor and Hevner 2013; Hevner et al. 2004). In this section, we present the design of our system (i.e., the

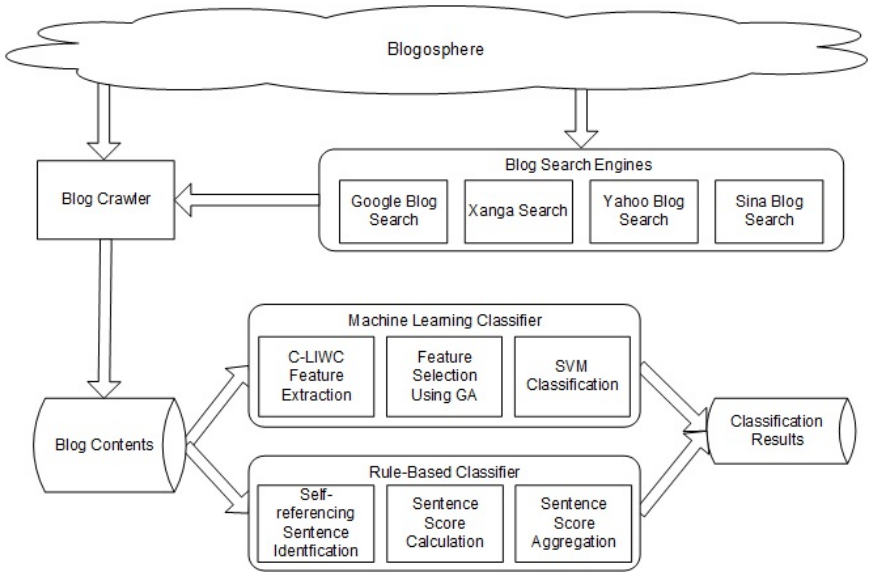

Figure 1. System Architecture

artifact that addresses the classification problem described in the previous sections). 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 (Abbasi et al. 2008; Yang and Liu 1999). 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 and 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, further reducing the number of features using feature selection would be beneficial. We propose to use a GA-based feature selection method to improve the 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 as it can perform sentence-level and paragraph-level analysis and 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.

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 content 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 typical search session is shown in Figure 2.

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

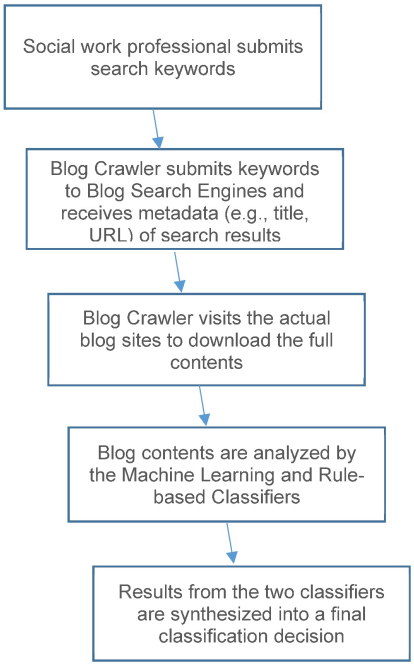

Figure 2. Workflow of a Typical Search Session

Blog Crawler

The first component in the proposed architecture is a blog crawler that collects blogs from different blog-hosting sites. Using a metasearch 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 insufficient 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, a popular tool that has been used in many prior studies (Zeng et al. 2011; Zhang et al. 2003).

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 shows emotional distress based on a set of training examples. This would help professionals identify potential emotional distress of the blog author. We use an SVM as our machine learning classifier as SVMs have been shown to be highly effective in conventional text classification and achieved the best performance among different text classifiers (Abbasi et al. 2008; Yang and Liu 1999). 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 many studies show that people's mental health can be predicted with the words they use in writing by observing 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 blog post. For each document d, the value f_i (for i = 1 to 71) is calculated as follows:

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

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, extracting a subset of features to improve performance is often desirable. In the proposed architecture, we use a generic algorithm (GA) in our feature selection process. GAs have been employed for feature selection in previous research and are applicable here (Fang et al. 2007; Oreski and Oreski 2014; Yang and Honavar 1998). 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 such as roulette-wheel selection, crossover, and mutation are implemented in a standard way (Goldberg 1989; Michalewicz 1996). In calculating the 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 tenfold 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.

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. We identify whether each sentence 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.

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

Lexicon Creation

Since no lexicon specifically concerning emotional distress words in Chinese is available, we develop our own lexicon in this model. The lexicon is constructed by the manual inspection of blog content 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 and Chen 2007; Subasic and Huettner 2000). In this particular study, 3,147 blogs were collected from Google blog search, and two clinical psychologists familiar with emotional distress and suicide research were asked to read these blog content 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 the manual evaluation of conversation text.

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. The

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 (heartwarming), 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)

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 as 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 actually been 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 they directly reflect 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 writing, entailing strong self-orientation (Li et al. 2014; Ramirez-Esparza et al. 2006; Rude et al. 2004) and even withdrawal from social relationships (Stirman and Pennebaker 2001). Although this self-referencing style is difficult to identify with human judgment, 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 and Wiebe, 2003; Zhang et al. 2009).

Sentence-Score Calculation

Instead of finding expressions of common affects such as fear and anger, the model is aimed at identifying emotional distress that consists of multiple affects. Many researchers have studied discrete affects such as fear, worry, sadness, contempt, disgust, guilt, nervousness, and anger (Abbasi et al. 2008; Subasic and Huettner, 2000). The identification of two opposite affects—namely, positive and negative—has become dominant in the literature. Although the negative affect is associated with emotional distress, these two terms are not equivalent (Crawford and 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 in their daily lives, which may be the cause 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.

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 shows 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

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

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 and Pennebaker 2001). A self-referencing sentence's score of emotional distress is calculated based on the positive emotion and negative emotion words present. Intuitively, a sentence is classified as showing emotional distress when only negative emotion words are found (Cheng et al. 2015; Li et al. 2014), 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 overlooking possibly negative documents. Because of the nature of our application, we want to reduce the chance of not finding 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, negative emotion, 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 or positive emotion words, respectively, because we want to distinguish our approach from standard sentiment analysis methods. Therefore, instead of giving the same weight to different word categories, we want to focus more on words 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 content 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, the sentence score is not calculated using emotion words. Instead, the sentence is checked for words that reference others (references) or express thankfulness or encouragement (gratitude expressions). Under the disengagement theory, people who reference other sources to offer opinions or convey information to others have a lower risk of depression (Stirman and 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 and Layous 2013).

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, 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). However, defining the parameters of what constitutes the opening and the ending of a document is difficult. Static positioning does not yield significantly higher accuracy because of the reduced analysis flexibility. 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.

Several segmentation methods have been used to find subtopics 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 computing 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, which is classified as showing emotional distress if the final score is positive.

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.

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 as 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 aims to evaluate the performance of the classifier in the proposed system using a static data set, 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. Although the experiment using a precollection 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 (Pang and Lee 2008; Yang and Liu 1999).

To develop the static data set for evaluation, a total of 804 blogs were obtained from four different sources. We used multiple sources 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 nongovernmental 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 nongovernmental 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 such as 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 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, 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 and 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 emotional distress judgment. Out of these blogs, 742 were consistently rated by the two psychologists, which 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 content showing emotional distress, and 530 included content not showing distress. The average length of blogs is 727 characters. The sources of the data are summarized in Table 2.

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 (Abbasi et al. 2008; Mullen and Collier 2004), 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 nonlinear one (Hsu et al. 2003). The 804 posts discussed above were classified using tenfold 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). An 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 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 com-

Table 2. Sources of Blog Data Used in the Experiment.

Data Source	Source	Search Tool Used	No. of Blogs
#1	Hong Kong Federation of Youth Groups	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**

parison 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) were used in Model 6 and Model 7, respectively.

The 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: the population size was 50, the number of generations was 200, and the probability of crossover and mutation were 0.5 and 0.01, respectively. As discussed earlier, the fitness of an individual is determined by evaluating the SVM model using the training data set in each iteration.

Experiment Results

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 a 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 the 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 was applied (Model 3), the performance improved.

Our results also show that the models incorporating the rule-based classifier with SVM (Models 5 to 8) performed better than the baseline models using SVM alone (Models 1 to 3) or rule-based alone (Model 4). When the feature selection technique was used (Models 6 to 8), the classification performance in terms of F-measure improved slightly compared to 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 versus 17 and 20 in Models 6 and 7, respectively).

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. 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 data obtained from other sources. As presented in Table 2, the first two data sources (number 1 and 2) contain more posts showing emotional distress, and the other two (number 3 and 4) contain more posts not showing emotional distress. Accordingly, we create four settings in our robustness test. In each setting, we use one data source showing emotional distress and one otherwise as the training data for our model and the other two data sources as testing data. These combinations ensure that both the training and testing data are not heavily imbalanced. The results are shown in Table 4. As can be seen, the performance in each setting is comparable to the main findings shown in Figure 3, demonstrating the generalizability of our approach.

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

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.

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

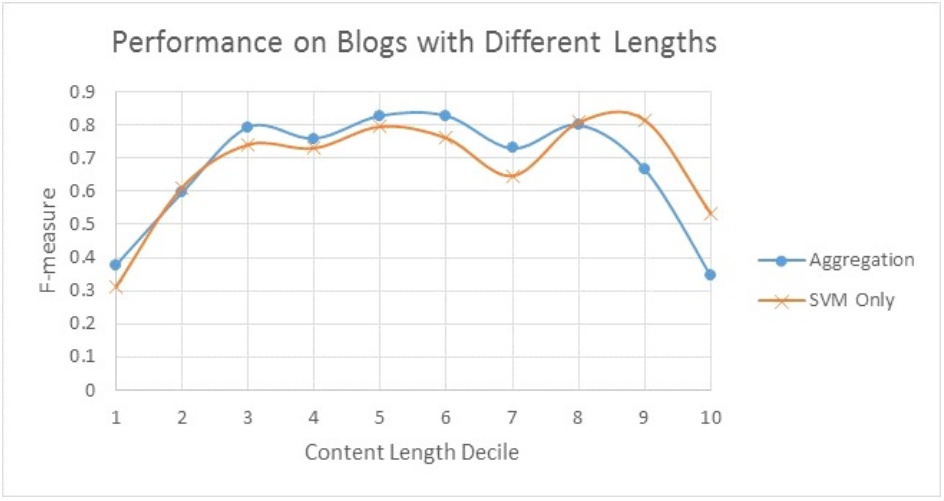

Figure 4. Performance on Blogs with Different Lengths

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

run another experiment as a robustness test to compare the different models using five data subsets 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) 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 lengths, we divide our data set into ten 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 content (the lowest decile in terms of word count), the second groups contains blogs with a length falling within the second lowest decile, and the last group contains blogs that have the longest content (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 record the F-measure. The results are shown in Figure 4.

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 third to seventh deciles. In contrast, the aggregation model performs worse than SVM when the blog posts are long (ninth and tenth deciles). We think the reason for SVM's relatively better performance when classifying long blogs is that the large number of keywords in the blogs already makes the classification decision rather effective. The rule-based approach, however, emphasizes the first and last blocks 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 and last blocks represent only a smaller portion of the entire post and could be less representative of the overall content. Also, because a longer post is more likely to have a higher number of polarity transitions, receiving a higher final score based on our calculation would be more likely. As such, the rule-based method tends to classify longer blog posts as showing emotional distress, producing more false positive results (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 (tenth decile). We found that some of these long blogs contain a number of negative emotion words but were classified as not showing emotional distress by our clinical psychologists. By analyzing these blogs, we found that while they did contain many negative affect words, other content (e.g., positive emotion words or stories of another person) showed that the authors were not emotionally distressed. Given the large number of negative words in these blogs, SVM still misclassified these posts as showing emotional distress, resulting in lower performance.

Another observation is that when a blog post is short (first and second deciles), neither the proposed aggregation model nor the SVM model performs well. Our analysis of these short posts shows that many of them do not contain enough informative features and therefore are easily misclassified by both models.

Besides content length, we also study the effect of several other blog characteristics, including the percentages of positive and 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 word count in a blog). Similar to the analysis on content length, we divide the data set into ten groups corresponding to the deciles for each of these two measures. The results are shown in Figures 5 and 6.

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

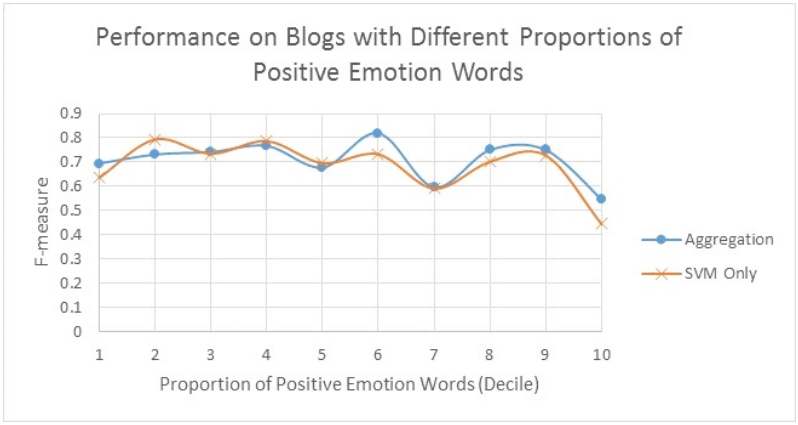

Figure 5. Performance on Blogs with Different Proportions of Positive Emotion Words

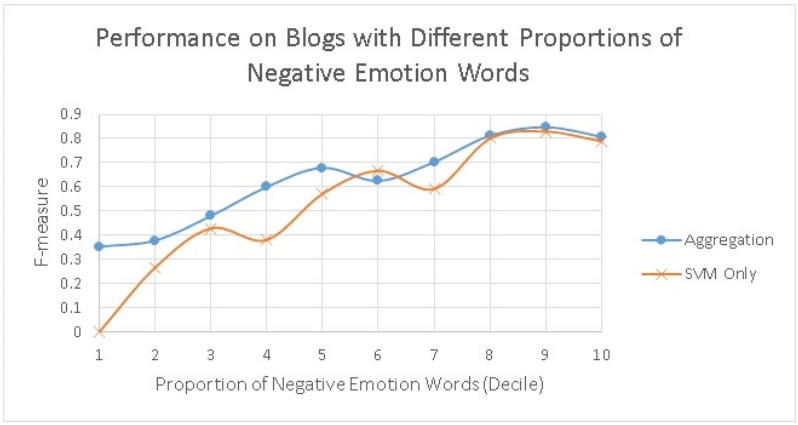

Figure 6. Performance on Blogs with Different Proportions of Negative Emotion Words

some blogs actually show emotional distress even if they contain many positive emotion words. SVM was 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 that the performance of both models is poor when there is a low percentage of negative emotion words as some blogs contain very few or even no negative emotion words based on our lexicons but actually show emotional distress. By perusing these blog texts, we found that some of these blogs were written more subtly, and one can only capture intimations of emotional distress by reading between the lines. Some other blogs used illegal Chinese characters that do not match any characters in our lexicons. In either case, these blogs would be easily missed by both models, especially by SVM as it lacks the customized lexicons or the sentence-level analysis as in the rule-based approach. As a result, SVM has a higher number of false negatives for these blogs, causing low recall and F-measure. On the other hand, the rule-based approach, which considers other factors such as self-references and negation, is more capable of identifying 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 regardless of the low or high proportion of negative emotion words.

Misclassification Analysis

In addition to the quantitative analysis, we also sought to identify the causes for misclassification by performing qualitative analysis, which is conducive to improving the classification performance and creating new categories to capture finer details in the future. Our qualitative analysis of the content that do not show emotional distress but are wrongly identified (i.e., false positives) reveals several problems. The first problem is that some sensational writing, such as movie reviews and tragedy fictions, show similar writing and linguistic styles to those of depressed people (Pennebaker et al. 2003). Distinguishing between these two kinds of writing using either the rule-based approach or SVM is difficult for the model because the word-usage patterns are very similar. Second, neither SVM, which uses C-LIWC (a dictionary of formal written Chinese), nor the rule-based approach, which uses our own lexicons, can understand informal and colloquial Chinese expressions in some of the blog posts if the words are not found in these dictionaries. Third, while emotionally distressed individuals tend to use significantly more self-referencing words in their writing (Rude et al. 2001; Sloan 2005; Stirman and Pennebaker 2001), the same applies to people with other characteristics, such as self-consciousness. However, we cannot determine 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 this issue by correctly classifying some of the marginal posts and complemented the classification of SVM to achieve a better overall performance of the aggregation approach.

Similarly, we also analyze blogs that showed emotional distress but were wrongly identified as not showing distress (i.e., false negatives), and we have identified several causes. First, some of these blogs were written rather implicitly. 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 red-flag them. 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 content are too short or contain very few negative emotion words. They may express emotional distress in a single phrase, and there is not enough information for the model to make a correct prediction. It is worthwhile to note that many of the reasons for misclassification discussed are similar to those for misclassification in the traditional sentiment analysis literature, such as the use of implicit words, idiomatic expressions, or irony (Balahur et al. 2006; Pang and Lee 2008).

From the detection and possibly life-saving prospects, it is important to boost the recall rate 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 overlook some potentially needy individuals.

In addition to the analyses reported, we also investigate how the aggregation of classification results impacts the proposed method's performance and how the proposed method compares with different classifier combinations. Please refer to the Appendix for a more detailed discussion.

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 aims 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 that by the proposed search engine. The study also evaluated user experiences of the search process. The second setting aims 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, participants 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 highest 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. With KAREN, the search results were displayed to the participants in a way similar to a standard search engine. Ten results were shown on each results page, and each result contained the title and a snippet. The order of the two search engines in the user study was randomized. Participants were

asked to devise 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 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. A total of 22 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 with which they were most familiar (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). This 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 very similarly 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.

A total of 19 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 the 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 with 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, possibly because the two search engines have the same user interface, and participants might not have noticed the differences in the back-end 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.

Discussion and Conclusion

This study demonstrates the effectiveness of the KAREN system for practical use in the online detection of emotionally distressed individuals. This study has several important implications for 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 blog content. The proposed aggregation method achieves the best classification performance compared to existing methods that use only one 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 a rule-based classifier achieves satisfactory performance in the classification task compared to 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.

The second contribution is that we investigated the conditions under which 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 can capture the relationships between certain keywords (mostly negative emotion words) and emotional distress in many cases. However, as discussed earlier, SVM does not perform well when a 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 a 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's length is medium, it can take advantage of its sentence- and paragraph-level analysis without suffering from 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, is not expressed explicitly using negative keywords.

Our findings have several implications for the design of text classifiers. First, our results show the limitations of keyword-based classification approaches such as SVM, especially in the identification of emotional distress or other characteristics that could be implicit. Researchers should be cautious about such limitations in their design. Our findings also confirm that aggregating results from different classification methods improves classification performance. The limitations of 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 to achieve better performance.

Our research has important implications for social work practices. 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. The system will reduce manual efforts of 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. It is expected that the limited 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.

However, this study has some limitations. First, we did 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 (and thus cannot be identified by our approach), while other individuals blog about their emotional distress simply to seek attention. In addition, the demographic characteristics of bloggers, such as gender, were not investigated. It has been found that males and females use different emotional expressions in computer-mediated communication (Thelwall et al. 2010). In this study, gender differences were not considered in the classification process. The definition of emotional distress is complicated and subjective in nature; this may introduce possible imprecision in the machine learning process and evaluation of the classification.

In the future, we will improve our approach and system in various aspects. First, the practicability and efficacy of the approach will be further evaluated. 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. Therefore, we will evaluate

the approach to show its practicability and efficacy with large data sets more representative of real-world conditions. Second, we will analyze bloggers not with a single post but with 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 (Campbell and Pennebaker 2003). There is also an abundance of information, such as other bloggers' comments and interactions, that can be analyzed to better understand the bloggers' thoughts (Chau and Xu 2012). Third, emoticons, parenthetical expressions, and other commonly used symbols that convey thoughts and feelings can be incorporated in the classification approach in future work. We believe that these future works would be highly valuable in further improving the proposed model.

Acknowledgments

This project is supported in part by a grant from the General Research Fund of the Hong Kong Research Grants Council (project number 742012B) and a grant from the Azalea (1972) Endowment Fund. We thank the senior editor, associate editor, and the reviewers for their invaluable comments and suggestions throughout the review process. We are grateful to the Hong Kong Federation of Youth Groups and the Samaritan Befrienders Hong Kong for providing the data used in this study. We also thank Broderick Koo and Ben Ng for program development, Angie Shum, Tom Li, and Chris Wong for data evaluation and analysis, the staff at the HKU-HKJC Centre for Suicide Research and Prevention for their contribution and useful suggestions, and all the participants who took part in our evaluation studies.

References

Abbasi, A., and Chen, H. 2007. "Affect Intensity Analysis of Dark Web Forums," in *Proceedings of IEEE International Conference on Intelligence and Security Informatics*, New Brunswick, NJ, May 23-24, pp. 282-288.

Abbasi, A., and Chen, H. 2008. "CyberGate: A System and Design Framework for Text Analysis of Computer-Mediated Communication," *MIS Quarterly* (32:4), pp. 811-837.

Abbasi, A., Chen, H., Thoms, S., and Fu, T. 2008. "Affect Analysis of Web Forums and Blogs Using Correlation Ensembles," *IEEE Transactions on Knowledge and Data Engineering* (20:9), pp. 1168-1180.

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M., Pouliquen, B., and Belyaeva, J. 2006. "Sentiment Analysis in the News," in *Proceedings of Workshop on Intelligent Analysis and Processing of Web News Content*, IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 2216-2220.

Bolier, L., Haverman, M., Westerhof, G. J., Riper, H., Smit, F., and Bohlmeijer, E. 2013. "Positive Psychology Interventions: A Meta-Analysis of Randomized Controlled Studies," *BMC Public Health* (13:1) (https://www.ncbi.nlm.nih.gov/pubmed/23390882).

Borges, G., Nock, M. K., Haro Abad, J. M., Hwang, I., Sampson, N. A., Alonso, J., Andrade, L. H., Angermeyer, M. C., Beautrais, A., Bromet, E., Bruffaerts, R., de Girolama, G., Florescu, S., Gureje, O., Hu, C., Karam, E. G., Kovess-Masfety, V., Lee, S., Levinson, D., Medina-Mora, M. E., Ormel, J., Posada-Villa, J., Sagar, R., Tomov, T., Uda, H., Williams, D. R., and Kessler, R. C. 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., and Havasi, C. 2013. "New Avenues in Opinion Mining and Sentiment Analysis," *IEEE Intelligent Systems* , March/April, pp. 15-21.

Campbell, R. S., and Pennebaker, J. W. 2003. The Secret Life of Pronouns: Flexibility in Writing Style and Physical Health," *Psychological Science* (14), pp. 60-65.

Ceron, A., Curini, L., Iacus, S. M., and Porro, G. 2014. "Every Wweet 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), pp. 340-358.

Chau, M., and Xu, J. 2012. "Business Intelligence in Blogs: Understanding Consumer Interactions and Communities," *MIS Quarterly* (36:4), pp. 1189-1216.

Chen, H., Fan, H., Chau, M., and Zeng, D. 2001. "MetaSpider: Meta-Ssearching and Categorization on the Web," *Journal of the American Society for Information Science and Technology* (52:13), pp. 1134-1147.

Chen, J. H., Bierhals, A. J., Prigerson, H. G., Kasl, S. V., Mazure, C. M., and Jacobs, S. 1999. "Gender Differences in the Effects of Bereavement-Related Psychological Distress in Health Outcomes," *Psychological Medicine* (29:2), pp. 367-380.

Cheng, A. T., Chen, T. H., Chen, C. C., and Jenkins, R. 2000. "Psychosocial and Psychiatric Risk Factors for Suicide," *The British Journal of Psychiatry* (177:4), pp. 360-365.

Cheng, Q., Kwok, C. L., Zhu, T., Guan, L., and Yip, P. S. F. 2015. "Suicide Communication on Social Media and its Psychological Mechanisms: an Examination of Chinese Microblog Users," I*nternational Journal of Environmental Research and Public Health* (12), pp. 11506-11527.

Coppersmith, G. A., Harman, C. T., and Dredze, M. H. 2014. "Measuring Post-Traumatic Stress Disorder in Twitter," *Proceedings of 8th International AAAI Conference on Weblogs and Social Media*, Ann Arbor, MI, pp. 579-582.

Crawford, J. R., and 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), pp. 245-265.

Davis, F. D. 1989. "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Quarterly* (13:3), pp. 319-340.

De Choudhury, M., Counts, S., and Horvitz, E. 2013. "Predicting Postpartum Changes in Emotion and Behavior Via Social Media," in *Proceedings of ACM CHI Conference on Human*

Factors in Computing Systems, New York: ACM, pp. 3267-3276.

Fang, X., Sheng, O. R. L., and Chau, M. 2007. "ServiceFinder: A Mmethod towards Enhancing Service Portals," *ACM Transactions on Information Systems* (25:4), Article 17.

Feldman, R. 2013. "Techniques and Applications for Sentiment Analysis," *Communications of the ACM* (56:4), pp. 82-89.

Gill, A. J., French, R. M., Gergle, D., and Oberlander, J. 2008. "The Language of Emotion in Short Blog Texts," in *Proceedings of ACM Conference on Computer-Supported Collaborative Work* San Diego, CA.

Glance, N., Hurst, M., Nigam, K., Siegler, M., Stockton, R., and Tomokiyo, T. 2005. "Analyzing Online Discussion for Marketing Intelligence," in *Proceedings of 2nd Annual Workshop on the Weblogging Ecosystem*, Chiba, Japan.

Goldberg, D. 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning*, Boston: Addison-Wesley.

Gregor, S., and Hevner, A. R. 2013. "Positioning and Presenting Design Science Research for Maximum Impact," *MIS Quarterly* (37:2), pp. 337-356.

Gruber, J., and Kring, A. M. 2008. "Narrating Emotional Events in Schizophrenia," *Journal of Abnormal Psychology* (117:3), pp. 520-533.

Hearst, M. A. 1997. "TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages," *Computational Linguistics* (23), pp. 33-64.

Hessler, R. M., Downing, J., Beltz, C., Pelliccio, A., Powell, M., and Vale, W. 2003. "Qualitative Research on Adolescent Risk Using e-Mail: A Methodological Assessment," *Qualitative Sociology* (26:1), pp. 111-124.

Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. "Design Science in Information Systems Research," *MIS Quarterly* (28:1), pp. 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 (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., Lam, B. C. P., Chen, W.-C., Bond, M. H., and Pennebaker, J. W. 2012. "The Development of the Chinese Linguistic Inquiry and Word Count Dictionary," *Chinese Journal of Psychology* (54:2), pp. 185-201.

Huang, Y., Goh, T., and Liew, C. L. 2007. "Hunting Suicide Notes in Web 2.0: Preliminary Findings," in *Proceedings of the 9th IEEE International Symposium on Multimedia—Workshops*.

Hutto, C. J., and Gilbert, E. 2014. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," in *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*.

Ishida, K. 2005. "Extracting Latent Weblog Communities: A Partitioning Algorithm for Bipartite Graph," in *Proceedings of the 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, Chiba, Japan.

Juffinger, A., and Lex, E. 2009. "Crosslanguage Blog Mining and Ttrend Visualisation," in *Proceedings of the 18th International Conference on World Wide Web*, Madrid, Spain.

Junghaenel, D., Smyth, J. M., and Santner, L. 2008. "Linguistic Dimensions of Psychopathology: A Quantitative Analysis," *Journal of Social and Clinical Psychology* (27:1), pp. 36-55.

Kuhl, J., Quirin, M., and Koole, S. L. 2015. "Being Someone: The Integrated Self as a Neuropsychological System," *Social and Personality Psychology Compass* (9:3), pp. 115-132.

Kumar, R., Novak, J., and Tomkins, A. 2010. "Structure and Evolution of Online Social Networks," in *Link Mining: Models, Algorithms, and Applications*, P. Yu, J. Han, and C. Faloutsos (eds.), New York: Springer, pp. 337-357.

Lee, L., Pang, B., and Vaithyanathan, S. 2002. "Thumbs Up? Sentiment Classification Using Machine Learning Techniques," in *Proceedings of 2002 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA: Association for Computational Linguistics, pp. 79-86.

Leung, P. W., Hung, S. F., Ho, T. P., Lee, C. C., Liu, W. S., Tang, C. P., and Kwong, S. L. 2008. "Prevalence of DSM-IV Disorders in Chinese Adolescents and the Effects of an Impairment Criterion," *European Child and Adolescent Psychiatry* (17:7), pp. 452-461.

Li, T. M. H., Chau, M., Wong, P. W. C., and Yip, P. S. F. 2014. "Temporal and Computerized Psycholinguistic Analysis of the Blog of a Chinese Adolescent Suicide," *Crisis* (35:3), pp. 168-175.

Liu, B. 2012. *Sentiment Analysis and Opinion Mining*. San Rafael, CA: Morgan and Claypool Publishers

Liu, Y., Huang, X., An, A., and Yu, X. 2007. "ARSA: A Sentiment-Aware Model for Predicting Sales Performance Using Blogs," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information*, Amsterdam, The Netherlands.

Lyubomirsky, S., and Layous, K. 2013. "How Do Simple Positive Activities Increase Well-Being?." *Current Directions in Psychological Science* (22:1), pp. 57-62.

Matthews, G., Jones, D. M., and Chamberlain, A. G. 1990. "Refining the Measurement of Mood: The UWIST Mood Adjective Checklist," *British Journal of Psychology* (81:1), pp. 17-42.=

Michalewicz, Z. 1996. *Genetic Algorithms + Data Structures = Evolution Programs*," Heidelberg, Germany: Springer-Verlag.

Mishne, G., and de Rijke, M. 2006. "Capturing Global Mood Levels Using Blog Posts," in *Proceedings of the AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, pp. 145-152.

Mullen. T., and Collier, N. 2004. "Sentiment Analysis Using Support Vector Machines with Diverse Information Sources," in *Proceedings of the 9th Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA: Association for Computational Linguistics, pp. 412-418.

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), pp. 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), pp. 95-135.

Newman, D. B., Tay, L., and Diener, E. 2014. "Leisure and Subjective Well-Neing: A Model of Psychological Mechanisms as Mediating Factors," *Journal of Happiness Studies* (15:3), pp. 555-578.

Niesler, T., and Woodland, P. 1996. "A Variable-Length Category-Based n-Gram Language Model," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, GA, pp. I164-I167.

Oreski, S., and Oreski, G. 2014. "Genetic Algorithm-Based Heuristic for Feature Selection in Credit Risk Assessment," *Expert Systems with Applications* (41:4), pp. 2052-2064.

Pang, B., and Lee, L. 2008. "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval* (2:1-2), pp. 1-135.

Pennebaker, J. W. 2003. "Putting Stress Into Words: Health, Linguistic, and Therapeutic Implications," *Behaviour Research and Therapy* (31:6), pp. 539-548.

Pennebaker, J. W., and Chung, C. K. 2011. "Expressive Writings: Connections to Physical and Mental Health," in *The Oxford Handbook of Health Psychology*, H. S. Friedman (ed.), Oxford, UK: Oxford University Press.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., and Booth, R. J. 2007. "The Development and Psychometric Properties of LIWC 2007," LIWC.Net, Austin, TX.

Pennebaker, J. W., Mehl, M. R., and Niederhoffer, K. G. 2003. "Psychological Aspects of Natural Language Use: Our Words, Our Selves," *Annual Review of Psychology* (54:1), pp. 547-577.

Ramirez-Esparza, N., and Pennebaker, J. W. 2006. "Do Good Stories Produce Good Health? Exploring Words, Language, and Culture," *Narrative Inquiry* (10:1), pp. 211-219.

Riloff, E., and Wiebe, J. 2003. "Learning Extraction Patterns for Subjective Expressions," in *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA: Association for Computational Linguistics,pp. 105-112.

Rude, S. S., Gortner, E. M., and Pennebaker, J. W. 2004. "Language Use of Depressed and Depression-Vulnerable College Students," *Cognition and Emotion* (18), pp. 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," in *Proceedings of the 14th International Conference on Knowledge Technologies and Data-Driven Business*, New York: ACM.

Samuelsson, C., and 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*, Phoenix, AZ.

Sloan, D. M. 2005. "It's All About Me: Self-Focused Attention and Depressed Mood," *Cognitive Therapy and Research* (29), pp. 279-288.

Snoek, F. J., Pouwer, F., Welch, G. W., and Polonsky, W. H. 2000. "Diabetes-Related Emotional Distress in Dutch and U.S. Diabetic Patients: Cross-Cultural Validity of the Problem Areas in Diabetes Scale," *Diabetes Care* (23:9), pp. 1304-1309.

Stirman, S. W., and Pennebaker, J. W. 2001. "Word Use in Poetry of Suicidal and Nonsuicidal Poets," *Psychosomatic Medicine* (53), pp. 517-522.

Subasic, P., and Huettner, A. 2000. "Affect Analysis of Text Using Fuzzy Semantic Typing," in *Proceedings of the 9th IEEE International Conference on Fuzzy Systems*, San Antonio, TX, pp. 647-652.

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

Thelwall, M., Wilkinson, D., and 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), pp. 190-199.

Turecki, G., and Brent, D. A. 2016. "Suicide and Suicidal Behaviour," *The Lancet* (387:10024), pp. 1227-1239.

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

World Health Organization. 2014. *Preventing Suicide: A Global Imperative* (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* (5:2), pp. 165-183.

Yang, J., and Honavar, V. 1998. "Feature Subset Selection Using a Genetic Algorithm," in *Feature Extraction, Construction and Selection: A Data Mining Perspective*, H. Liu and H. Motoda (eds.), Boston: Kluwer Academic Publishers, pp. 117-136.

Yang, Y., and 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*, New York: ACM, pp. 42-49.

Zawadzki, M. J., Smyth, J. M., and Costigan, H. J. 2015. "Real-Time Associations between Engaging in Leisure and Daily Health and Well-Being," *Annals of Behavioral Medicine* (49:4), pp. 605-615.

Zeng, D., Wei, D., Chau, M., and Wang, F. 2011. "Domain-Sspecific Chinese Word Segmentation Using Suffix Tree and Mutual Information," *Information Systems Frontiers* (13:1), pp. 115-125.

Zhang, C., Zeng, D., Li, J., Wang, F. Y., and 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), pp. 2474-2487.

Zhang, H. P., Yu, H. K., Xiong, D. Y., and Liu, Q. 2003. "HMM-Based Chinese Lexical Analyzer ICTCLAS," in *Proceedings of the Second Annual SIGHAN Workshop on Chinese Language Processing*, Hoboken, NJ: IEEE Press, pp. 184-187.

About the Authors

Michael Chau is an associate professor in the Faculty of Business and Economics and the Warden of Lee Chi Hung Hall at the University of Hong Kong. His research focuses on the cross-disciplinary intersection of information systems, computer science, business analytics, and information science, with an emphasis on the applications of data, text, and web mining in various business, education, and social domains. He is particularly interested in text mining and data analytics research. His research has resulted in over 130 publications in high-quality journals and conferences. He is the recipient of the HKU Outstanding Young Research Award (2014) and Knowledge Exchange Award (2013, 2016), and is highly ranked

in several research productivity studies. He served as a program co-chair of the International Conference on Information Systems in 2013 and is the founding co-chair of the Pacific Asia Workshop on Intelligence and Security Informatics series. Michael served as an associate editor for *MIS Quarterly* from 2014-2017. He received his Ph.D. in Management Information Systems from the University of Arizona and his B.Sc. in Computer Science and Information Systems from the University of Hong Kong.

Tim M. H. Li is passionate about working on projects that drive innovation and technology forward and help people to better lives. His research interests lie in digital humanities including web-based interventions, the internet and suicide, youth social withdrawal (*hikikomori*), and computerized handwriting assessment. He has over 3 years of postdoctoral research experience managing multi-disciplinary teams, coordinating large-scale projects, and publishing academic articles in top-tier international journals. Tim received a Ph.D. in Youth Studies from the Department of Social Work and Social Administration at the University of Hong Kong, as well as an M.Sc. and a B.Eng. degree in Computer Science from the Department of Computer Science at the University of Hong Kong. He is a senior engineer specialized in machine learning, natural language processing, and data visualization.

Paul W. C. Wong holds a D.Psyc. (Clinical) and is currently an associate professor in the Department of Social Work and Social Administration at the University of Hong Kong. Paul has been involved in suicide prevention research and mental health promotion and practice in Hong Kong since 2003. His recent research projects include social withdrawal behaviour (a.k.a. *hikikomori*), using animals as part of psychological and educational interventions, youth positive development interventions for local and ethnic minority young people in Hong Kong, and helping care givers of children with developmental delays. He is currently an Honorary Fellow of the HKU-HKJC Centre for Suicide Research and Prevention, HK Police College, and Council Member of the Hong Kong Psychological Society; he was National Representative of the International Association for Suicide Prevention (2010-2015). He is also the Director of the B.Soc.Sci (Counselling) and Deputy Director of M.Soc.Sci. (Counselling) programs. He has published about 80 academic articles on a number of suicide-related, and mental-health-related issues, and he is the developer of an award-winning website (www.depression.edu.hk) and the author of *The Belated Dialogues between the Suicides and Their Families*.

Jennifer J. Xu is an associate professor of Computer Information Systems at Bentley University. She received her Ph.D. in Management Information Systems from the University of Arizona. Her research interests include business intelligence and analytics, machine learning, data science, FinTech, social network analysis, human–computer interaction, and enterprise systems. She has published more than 60 articles in Information Systems journals, books, and conference proceedings. She is currently serving on the editorial boards of *Journal of the Association for Information Systems*, *Communications of the AIS*, and *Journal of Security Informatics*.

Paul S. F. Yip is the Chair Professor of Population Health at the Department of Social Work and Social Administration and the director of the Centre for Suicide Research and Prevention at the University of Hong Kong. He is the principal investigator of an online crisis support for youth and has developed a number of innovative suicide prevention projects. He served as the chairman of the Committee of Preventing Students' Suicide, as a member of the steering committee on Population Policy, and as an associate member of the Central Policy Unit of the Hong Kong SAR Government. He is a recipient of a medal of honor from the Hong Kong SAR Government, 2017; the Stengel Research award in 2012; Outstanding Supervisor and Researcher of the University of Hong Kong in 2011 and 2009, respectively. He has published more than 400 research papers relating to population health and suicide prevention.

Hsinchun Chen is the University of Arizona Regents' Professor and Thomas R. Brown Chair Professor in Management and Technology. He is also a Fellow of ACM, IEEE, and AAAS. Hsinchun served as the lead program director of the Smart and Connected (SCH) Program at the NSF (2014-2015), a multi-year multi-agency health IT research program of in the United States. He is author/editor of 20 books, 300 journal articles, and 200 refereed conference articles covering digital library, data/text/web mining, business analytics, security informatics, and health informatics. His overall h-index is 96 (32,000 citations for 900 papers according to Google Scholar), among the highest in MIS and top 50 in Computer Science. Hsinchun founded the Artificial Intelligence Lab at The University of Arizona in 1989; the Lab has received more than $40 million in research funding from NSF, NIH, NLM, DOD, DOJ, CIA, DHS, and other agencies (100 grants, 50 from the NSF). He has served as editor-in-chief, senior editor or associate editor for several major ACM/IEEE journals, *MIS Quarterly*, and *Decision Support Systems*, and for several Springer journals. He has also served as a conference or program chair of major conferences in the field. Hsinchun is also a successful IT entrepreneur. His COPLINK/i2 system for security analytics was commercialized in 2000 and acquired by IBM as its leading government analytics product in 2011. He is internationally renowned for leading research and development in the health analytics (data and text mining; health big data; DiabeticLink and SilverLink) and security informatics (counter terrorism and cyber security analytics; security big data; COPLINK, Dark Web, Hacker Web, and AZSecure) communities.

Appendix

Aggregation of Classification Results

In the proposed model, a blog is classified as showing emotional distress if at least one of the two classifiers classified it as such. In other words, the results from the two classifiers are combined through a Boolean OR operation. Investigating whether a better way exists to combine the results from the two classifiers would be interesting. 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 blog 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 use the SVM output only. The SVM used here is the SVM using C-LIWC–based category features and GAs 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, displayed in Figure A1, 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 better 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.

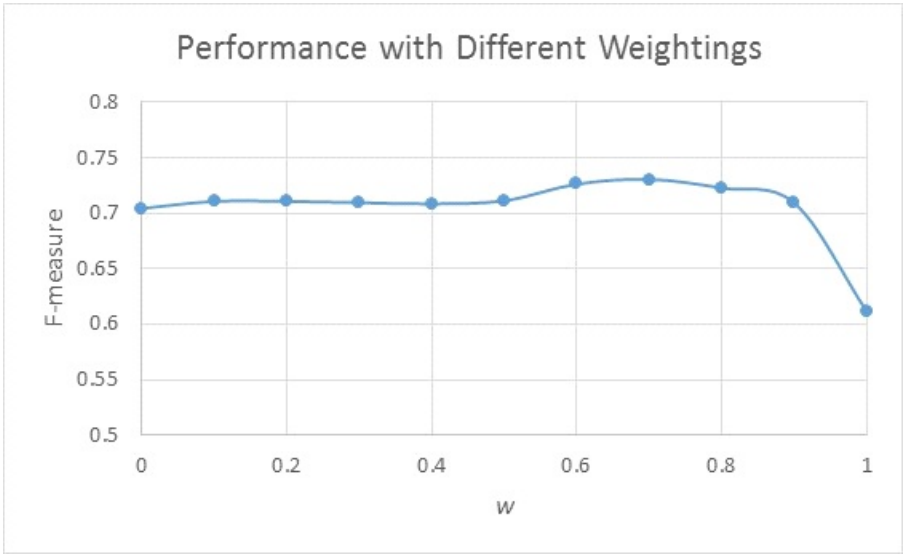

Figure A1. Performance of Classification Aggregation with Different Weightings

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, especially when the number of positive training instances is small. To verify if SVM is indeed suitable for our data set, we compare it with two other popular text classifiers, namely, a Naive 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 a combination will perform better than combining two similar classification approaches. We perform additional experiments to validate this.

The comparison results are shown in Table A1. As can be seen, both the Naive 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 support 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 Naive Bayes classifier (0.6850). This 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 A1. Performance of Selected Classifier Combinations

Model	Precision	Recall	F-measure
Naive Bayes	0.5980	0.6460	**0.6211**
Decision Tree	0.6679	0.6679	**0.6679**
SVM(C-LIWC + GA) + Naive 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**