Machine Classification and Analysis of Suicide-Related Communication on Twitter

Pete Burnap
School of Computer Science
and Informatics
Cardiff University
burnapp@cardiff.ac.uk

Gualtiero Colombo
School of Computer Science
and Informatics
Cardiff University
colombog@cardiff.ac.uk

Jonathan Scourfield
School of Social Sciences
Cardiff University
scourfield@cardiff.ac.uk

ABSTRACT

The World Wide Web, and online social networks in particular, have increased connectivity between people such that information can spread to millions of people in a matter of minutes. This form of online collective contagion has provided many benefits to society, such as providing reassurance and emergency management in the immediate aftermath of natural disasters. However, it also poses a potential risk to vulnerable Web users who receive this information and could subsequently come to harm. One example of this would be the spread of suicidal ideation in online social networks, about which concerns have been raised. In this paper we report the results of a number of machine classifiers built with the aim of classifying text relating to suicide on Twitter. The classifier distinguishes between the more worrying content, such as suicidal ideation, and other suicide-related topics such as reporting of a suicide, memorial, campaigning and support. It also aims to identify flippant references to suicide. We built a set of baseline classifiers using lexical, structural, emotive and psychological features extracted from Twitter posts. We then improved on the baseline classifiers by building an ensemble classifier using the Rotation Forest algorithm and a Maximum Probability voting classification decision method, based on the outcome of base classifiers. This achieved an F-measure of 0.728 overall (for 7 classes, including suicidal ideation) and 0.69 for the suicidal ideation class. We summarise the results by reflecting on the most significant predictive principle components of the suicidal ideation class to provide insight into the language used on Twitter to express suicidal ideation.

Categories and Subject Descriptors

[**Computer Science**]: artificial intelligence, text analysis, Web-based interaction, human safety

1. INTRODUCTION

It is recognised that media reporting about suicide cases has been associated with suicidal behaviour [29] and con-

cerns have been raised about how media communication may have an influence on suicidal ideation and cause a contagion effect between vulnerable subjects [14]. With the advent of open and massively popular social networking and microblogging Web sites, such as Facebook, Tumblr and Twitter (frequently referred to as social media), attention has focussed on how these new modes of communication may become a new, highly interconnected forum for collective communication and, like news media reporting, lead to contagion of suicidal ideation.

Social science and medical research have investigated the impact that communication on the topic of suicide via the World Wide Web may have on vulnerable subjects, with particular attention to the younger generation [9]. [2] conducted a qualitative study by interviewing young adults who engage in suicidal behaviours and use websites dedicated to these themes. [31, 5] also conducted online searches for Web resources containing suicide-related terms and describing suicide methods. They presented a qualitative analysis of the resources they discovered and concluded that, although neutral and anti-suicide Web sites occurred most frequently, pro-suicide forums and Web sites encouraging suicidal behaviour were also present and available, suggesting that more prevention plans specifically focused on Web resources are required. Building on this, [19] have reviewed online suicide intervention and prevention literature, concluding that there is a lack of published evidence about online prevention strategies and more attention is required to develop and evaluate online preventative approaches. [33] also studied the impact of Facebook suicide notes on suicidal behaviour, reporting that it was not yet clear to what extent suicide notes on online social media actually induce copycat suicides. They note that suicide and social media effects deserve further evaluation and research.

Other studies have focused on the written communication of suicide on the Web via bulletin boards [18], newsgroups [24], chat rooms [4], and web forums [22]. These are mostly qualitative analyses and where quantitative data are used in web-related suicide studies, they tend to rely solely on human classification, which is difficult to implement at scale. Computational methods have only been used in a small number of suicide communication studies.

Some studies report a positive correlation between suicide rates and the volume of social media posts that may be related to suicidal ideation and intent [37, 20]. There is also a developing body of literature on the topic of identifying suicidal language on Twitter [15, 35], but very few attempts to use machine classification to automatically identify suici-

dal language and differentiate between this and other forms of suicide-related communication, such as awareness raising and reporting of suicides. The differentiation is a requirement for the purposes of analysing the characteristics of suicidal ideation on social media. [10, 8] study depression and other emotional states expressed via social media. Suicidal language is likely to include emotive content and possible signs of depression but we do not suggest depression and suicidal ideation are synonymous in this paper. Two very recent papers presented the results of Twitter studies aiming to classify 'risky' language [1] and levels of 'distress' [16] - both reporting classification performance that has potential for improvement (around 60-64%). An important step in providing support to suicidal social media users is to understand how suicidal ideation is communicated. Recent studies have shown that people are more likely to seek support from non-professional resources such as social media, rather than risk social stigmatisation by seeking formal treatment [16].

Thus, our study aims to contribute to the literature on understanding communications on the topic of suicide in social media by (i) creating a new human-annotated dataset to help identify features of suicidal ideation, (ii) creating a set of benchmark experimental results for machine learning approaches to the classification of suicidal ideation, and (iii) developing a machine classifier capable of distinguishing between worrying language such as suicidal ideation, and flippant references to suicide, awareness raising about suicide and reports of suicide. This last contribution is especially relevant to quantify actual volumes of worrying language on social media for the purposes of understanding risk to human safety, as opposed to all references to suicide. The research presented in this paper comprises an analysis of data collected from the microblogging website Twitter, the text of which has been classified into one of seven suicide-releated categories by a crowdsourced team of human annotators. We then use a range of machine learning classification methods to identify suicidal ideation in tweets and analyse the predictive features of suicidal ideation to help explain the language used by perceived suicidal social media users.

2. RELATED WORK

The Durkheim project is aiming to mine social media data to identify markers of harmful behaviour[1]. The project will study a group of US war veterans who will opt-in to share their Twitter, Facebook and LinkedIn posts over time. There are so far no publicly available results from this study but the group has recently published the results of a suicide prediction task, using text from the clinical notes of US war veterans to identify text-based statistically significant *signals* of suicidality, with around 60% accuracy [30]. They found clinical notes of people who had died through suicide frequently recorded behaviours indicative of fear, agitiation and delusion.

Written text has also been analysed in a number of recent studies that have analysed suicide notes and developed machine classifiers to identify topics and emotions expressed by people who have taken their lives [17, 34, 27, 39, 23, 12]). Many of these papers attempt to classify text at a sentence level, which would suggest short strings much like those that would be posted to social media. However, suicide notes are written by people who have accepted suicide and then go on to harm themselves, whereas the current research is particularly interested in identifying suicidal thinking or ideation prior to self-harm, which may differ from the language used in suicide notes. Additionally, handwritten notes, even at sentence level, are not constrained by string length. Twitter posts are limited to 140 characters, which forces authors to use short, informal language that may differ from the way they would normally express feelings on paper. Finally, social media data are noisy, contain a broad range of topics, and language use varies over time. These features arguably make the task of classifying suicidal ideation more complex than it would be in a discrete recording of pre-suicide thoughts and feelings in a suicide note.

A small number of studies have investigated the communication of suicidal ideation on social media. However, they are mainly focused on a comparison with national death rates. For example, in Korea [37] and the US [20] research has attempted to identify a positive correlation between the frequency of suicide-related posts on social media and the number of recorded suicide cases. Suicide related posts were identified using a set of keywords relating to general concepts such as suicide and depression [37] or relating to suicide methods [20].

[15] analysed the Twitter posts of a person who had recently died through suicide. They studied the posts sent in the twenty-four hours prior to death, finding an increase in positive emotions (though not statistically significant) and a change in focus from the self to others as the time of death approached. As this was only a single person study, and given the fact the person had attempted to make the posts rhyme (thereby perhaps using different language to achieve this), the authors propose larger studies of a wider range of Twitter posts. They used the Linguistic Inquiry and Word Count (LIWC) software to identify parts of speech, emotional words and cognitive processes among other concepts [26]. LIWC was also used in [16] as a sampling technique to identify 'sad' Twitter posts that were subsequently classified using a machine learning into levels of distress on an ordinal scale, with around 64% accuracy in the best-case.

Also studying linguistic features of suicidal ideation, [35] used an online panel of young (early 20s) Twitter users to examine the association between suicide-related tweets and suicidal behaviour. They identified that particular phrases such as "want to commit suicide" were strongly associated with lifetime suicide attempts, the most powerful predictor of future suicide. They also noted that other phrases that suggest suicidal intent, such as "want to die", were less strongly associated. The variation here could suggest the flippant use of such phrases on social media when having a bad day - hence the additional challenges posed to classification of suicidal ideation on social media. Finally, [1] used machine learning to classify 'risky' and 'non risky' Tweets, as defined by human annotators, with an accuracy of around 60%. They created word lists to represent a number of topics and emotions related to suicide, finding references to insults, hurt and bullying in the 'risky' category.

3. DATA

3.1 Data Collection and Annotation

Rather than manually developing a word list to represent suicidal language, we generated a lexicon of terms by collect-

---

[1]http://www.durkheimproject.org/news/durkheim-project-will-analyze-opt-in-data/

ing anonymised data from known suicide Web forums, blogs and microblogs, and asking human annotators to identify whether it contained references to suicidal ideation. First we collected user posts from four known Web sites identified by experts in the field [31, 5] as being used to discuss suicidal themes for support and prevention. The selected Web sites either had dedicated sections[2],[3] or are specifically designed for suicidal discussions[4],[5]. Then we collected data from microblogging site Tumblr [6] - specifically, content containing self-classified suicidal ideation (i.e. text posts 'tagged' with the word 'suicide').

For each of the resulting Web sites we then collected an equal number of 200 posts, retrieved in chronological order, with a total of 800 text posts. These posts, and 1000 posts randomly selected from the Tubmlr sample, were subsequently human annotated using the crowd-sourcing online service Crowdflower[7]. To avoid difficulties in the annotation of long pieces of text we discarded posts having a length greater than the five percent longer than the average post length for each of the websites considered. Human annotators were asked to identify content containing suicidal ideation using a binary criteria by answering the question 'Is this person suicidal?'.

We then applied the Term Frequency/Inverse Document frequency (TF.IDF) method to the corpus of annotated documents in order to identify terms that appear frequently in the suicidal ideation class but appear with less frequency in the non-ideation class. This process identifies terms that can be used to distinguish between the two classes. In the TF.IDF process we considered n-grams of 1 to 5 words in length, and ranked the top 500 terms. These terms were further analysed by two experienced suicide researchers to remove terms not specifically related to suicide, as well as duplicate keywords. This resulted in a final list of 62 keywords and phrases that suggested possible suicide intent. Illustrative examples are *asleep and never wake*, *don't want to exist* and *kill myself*. These search terms were then used to collect data from Twitter via the Twitter Streaming Application Programming Interface (API).

Twitter data were collected for a continuous period of six weeks from 1st February 2014 using the suicide-related search terms, resulting in a dataset of over four million posts. In parallel we monitored traditional media over the same period to identify the names of reported suicide cases in England. We then retrieved a second data set from Twitter using the name and surname of the deceased as search keywords. Here, the underlying idea was to collect different types of posts with a connection to suicide other than those more directly expressing suicidal ideation (which was the aim of the first dataset collection). All names were removed from the text before analysis.

Following the data collection we produced a random sample of 1000 tweets from both datasets, with 80% of posts from the collection of suicide related search terms, and the remaining from the 'names' dataset. The human annotation task was repeated using the same crowdsourcing service.

This time human annotators were asked to classify data into either one or more of the six suicide related categories listed below, or into the seventh category representing tweets that cannot be classified into any of them. This coding frame was developed with expert researchers in suicide studies to capture the best representation of how people generally communicate on the topic of suicide.

Table 1: Types of Suicidal communication with relative % proportion in dataset

class	description	% of dataset
c1	Evidence of possible suicidal intent	13
c2	Campaigning (i.e. petitions etc.)	5
c3	Flippant reference to suicide	30
c4	Information or support	6
c5	Memorial or condolence	5
c6	Reporting of suicide (not bombing)	15
c7	None of the above	26

Text annotation can be a subjective task, so to limit the amount of subjectivity we required at least 4 human annotations per tweet as per the convention in related research [36]. CrowdFlower provides an agreement score for each annotated unit, which is based on the majority vote of the trusted workers [21]. Because the crowdsourcing service continues to recruit workers until the task is complete, there is no guarantee that all workers will annotate the same set of units. Therefore we cannot calculate traditional inter-rater reliability (IRR) scores, such as Krippendorf's Alpha or Cohen's Kappa to determine agreement between all annotators. However, CrowdFlower has been shown to produce an agreement score that compares well to these classic measures [21]. Based on the output from our annotator task we can determine agreement on each unit. The purpose of the experiments performed in this paper are to establish the accuracy of a machine classifier when assigning tweets to a particular class of suicidal communication, and thus it is the agreement score for the unit of analysis (each tweet), and not the overall human agreement for all units that is important for validation. We removed all tweets with less than 75% agreement - again, following established methods from related research [36], and discarded any where less than three out of four annotators (75%) agreed on the dominant class for each tweet. Annotators were allowed to select multiple class labels and the majority choice was taken. The distribution of tweets to classes from c1-c7 is shown in Table 1. Note that the dominant class was flippant and improper use of suicide-related phrases and expressions, with actual suicidal intent or thinking being in the minority (about 13% of the total). The fact that four people unknown to each other and without being influenced by each other's annotations could agree to this level would suggest that it is possible for human annotators to agree on what constitutes the language of suicidal ideation, and what is simply a flippant reference to suicide. The resulting dataset of 816 Tweets was subsequently used to train a machine learning classifier (details are provided in the next section), which is only slightly below the dataset sizes of other similar analyses of emotive content on social media (e.g. [25, 36, 3, 7]).

3.2 Feature Preparation

We used the text of the tweets in order to train and test

a number of machine classifiers to identify suicidal ideation and differentiate between this and other types of suicide-related communication, including flippant references to suicide. Three features sets were derived from the text as follows:

- Features representing *lexical characteristics* of the sentences used, such as the Parts of Speech (POS), and other *language structural features*, such as the most frequently used *words* and *phrases*. These are standard features used in most text mining tasks. References to self and others are also captured with POS – these terms have been identified in previous research as being evident within suicidal communication;

- Features representing *sentiment, affective and emotional features and levels* of the terms used within the text. These were incorporated because of the particularly emotive nature of the task. Emotions such as fear, anger and general aggressiveness are particularly prominent in suicidal communication [1]

- Features representing ideosyncratic language expressed in short, informal text such as social media posts within a limited number of characters. These were extracted from the annotated Tumblr posts we collected to try and incorporate the language used on social media that may not be identified using standard text mining features.

3.2.1 Feature Set 1

For the first set of features, and part of the second set, we derived features used in [34], published within the special issue on *sentiment analysis of suicide notes* [28]. We will refer to this set of features as $Set1$. More specifically, the $Set1$ feature set included the following:

- *Parts of Speech.* We used to the Stanford Part-Of-Speech (POS) Tagger[8] to assign each word in a Tweet a POS label. Examples are nouns (broken down into singular, plural, proper), verbs (specifying tenses such as present, past and present participle), 1st vs 3rd person references, adjective and adverbs (comparative, superlative), pronouns (personal, possessive), as well as other tags representing conjunctions, determiners, cardinal numbers, symbols, and interjections. For each of POS we considered the frequency of each in a Tweet as a feature.

- *Other Structural Features.* For this we considered the inclusion of negations in the sentence (total number), the specific use of a first person pronoun (either singular or plural), and external communication features such as the inclusion of a URL in a tweet or a mention symbol (indicating a retweet or reply).

- *General Lexical Domains.* These features represent general lexical categories such as home, religion, psychology, sociology, etc. These were extracted using WordNet Domains labels,[9]

- *Affective Lexical Domains.* These are a set of categories specifically related to domains representing 'affective' concepts. These include concepts representing moods, situations eliciting emotions, or emotional responses such as joy, anger, grief, sadness, enthusiasm, surprise, love, hate, and happiness; but even more specific sub-categories such as amicability, belligerence, bad-temper, unrest, and trepidation; and opposites such as positive-negative concern, negative fear, positive-negative suspense, self-esteem, self consciousness, self-pity, and self-deprecation. These are very appropriate for the specific language we are investigating in this study.

- *Sentiment Score.* Using SentiWordNet[10] each words is assigned a score between zero and one for both positivity and negativity. The sum all words in a Tweet were used as features.

- *Words.* The most frequently used *words* and *n-grams* in terms of (first 100) unigrams, bigrams and trigrams contained in the training set.

- *Keyword list.* We also included each of the 62 keywords derived from the Web form text that were used for the pre-filtering search (e.g. 'asleep and never wake', 'don't want to try anymore', 'end it all', 'isn't worth living', 'my life is pointless', 'kill myself', 'to live any more', 'want to end it', 'want to disappear', 'want to die', etc..). Each of the search terms were included as individual features together with one global binary feature representing the inclusion of any of them in a Tweet.

3.2.2 Feature Set 2

Given the psychological and emotional expressiveness of suicidal ideation, we then explored a second set of features by using the Linguistic Inquiry and Word Count *LIWC* text analysis software [26] to extract more specific labels representing affective emotions and feelings within the text. We refer to these features as $Set2$. These include a more extensive breakdown of categories that may be more suitable for the particular language of emotional distress that we would expect to be present in suicidal ideation. Examples are related to death, health, money, religion, occupation, and achievement, senses (e.g. feeling, hearing, seeing), and three other groups of terms related to 'cognitive mechanisms', 'affect', and 'social words'. These can be further broken down into labels representing family, friends, humans; anxiety, anger, sadness and positive and negative emotions; and terms related to certainty, inhibition, insight, causal, inclusivity and exclusivity. A subset of these features (sadness) were used in [16], but we have incorporated a wider range of the feature set to enable us to distinguish between distress and other forms of suicide-related communication (e.g. grief, support and reporting).

3.2.3 Feature Set 3

Next, due to the noisy nature of social media, where short, informal spelling and grammar are often used, we developed a set of regular expression (RegEx) and pattern matching rules from our collection of suicide-related posts collected

---

from social networking website Tumblr. We refer to these features as *Set*3. These were annotated as part of the human annotation process conducted earlier and introduce language from short informal text related to the six suicide related categories to assist the classifier. Examples of these expressions for each class (numbered 1-6 here) include:

1: '.+((\cutting |\depres|\sui)|\these|\bad|\sad).+ (\thoughts| \feel).+' to represent phrases such as 'suicidal / cutting / bad / these . . . thoughts / feelings'; '.+\wan\w.+d[ie].+' for expressions as 'want/wanted/wanting to die'; '.+\end.+ (\all|\it|\life).+' for sentences with 'end/ending it all' and 'end my life'; and '.+ (can.+|don.+|\take).+(\go|\live|\anymo| \cop|\alive).+' covering a wide range of phases including 'can't take anymore', 'can't/don't want to live/cope anymore', 'don't want to be alive', 'can't take it anymore', and 'can't go on'. In addition, we added a list of individual words and n-grams including 'trigger warning', 'tw', 'eating disorder', 'death', 'selfharm' and 'self harm', 'anxiety', and 'pain'.

2: '.+(\need|\ask|\call|\offer).+\help.+' related to phrases as 'call/offer for/of help' and individual terms as 'shut' (e.g. website shut down) and 'stop' (e.g. bullying).

3: '.+(\kill\hat\throw)' for phrases including 'kill/killing /hate myself', '.+(\f**k.+' for swearwords such as 'f**k/ f**king', '.+ (\boy\girl).+(\friend)' for expressions with 'boy-friend' and 'girlfriend', and '.+(\ just)\.+(\like).+' covering expression including 'just' . . . like'. In addition, some words related to general topics such as 'work' and 'school' have also been included since they are representing contexts more favourable to flippant language rather than genuine expression of distress and suicidal intent.

4: '.+(\talk|\speak).+\to.+(\one|\some|\any).+' related to phr-ases as 'talk / speak to someone/somebody' and words such as 'web', 'blog', 'health' , and 'advice'.

5: '.+miss.+(\you|\her|\him).+' related to phrases such as 'miss/missing you/her/him' and '.+(\kill|\die|\comm).+(day| month|year).+' to represent specific time references.

6: '.+(\took|\take).+\own.+\life.+' covering expressions including 'took/taken his/her own life' and words related to suicide methods such as 'hanged', 'hanging' and 'overdose'.

Note that the regular expressions included in the third class representing flippancy were also identified within those related to the first suicidal class (and vice versa). However, we decided to associate RegExs to only one of the two classes according to the nature of the annotated tweets, for example phrases as 'hate myself' or 'kill myself' were frequently associated with flippant posts whereas terms such as 'wanted to die' and 'want to end it' were more likely to be included in tweets containing evidence of suicidal thinking.

3.2.4 Data-driven Features

We built a fourth feature set that we will refer to as the *combined* set, incorporating the union of all of the features in the three previous groups. Given the large number of features associated with each tweet, and potential for co-linearity between features in the *combined* set, we applied *Principal Component Analysis (PCA)* as a dimension reduction procedure to convert the set of all possibly correlated variables within the *combined* set into a new set of linearly uncorrelated features (called principal components).

The text of the tweets was also incorporated as a feature set for all experiments. We transformed each Tweet into a word vector using ngrams of size 1 to 5, and retained between 100 and 2000 words (in increments of 100, 300, 500, 1000, 1500 and 2000). The optimum performance was 1-3grams with 500 words retained, and we only present these results in this paper.

4. MACHINE CLASSIFICATION METHOD

4.1 Baseline Experiments

We first conducted baseline experiments using the Weka machine learning libraries[11]. We used the four derived features sets with the most popular classifiers from the special issue on classification of suicidal topics in [27]. These were *Support Vector Machine (SVM)*, *Rule Based (we used Decision Trees (DT))*, and *Naive Bayes (NB)*.

Support Vector Machines (SVM) have been shown to work very well with short informal text [25, 38], including promising results when classifying other mental health issues [11]. Feature vectors are plotted in high-dimensional space and hyperplanes (lines that separate the data points) are used to try and find the optimum way to divide the space such that the data points belonging to the different human assigned classes are separated. Multiple hyperplanes can be used and the optimal hyperplane will be the line that maximizes the separation between classes. Rule-based approaches are able to iteratively identify the feature from a set of training data that maximises information gain in a classification exercise - or put another way, it quantifies the significance of how using one feature as a rule to classify a tweet as suicidal ideation, reduces the uncertainty as to which class it belongs to. Performing this step multiple times creates a hierarchical and incremental set of rules that can be used to make classification decisions. We used a J48 decision tree (C4.5) to perform rule-based experiments. Finally, given the prevalence of individual words or short combinations of words that would be associated with suicidal ideation, it is logical to incorporate probabilistic classifiers into the experiments as they make classification decisions based on the likelihood of feature occurrence. Specific terms and phrases prevalent in each class can be identified and learned by the classifier. We implemented a Naive Bayes algorithm as a probabilistic approach.

4.2 Ensemble Experiments

The individual baseline experiments produced a set of results that achieved a reasonable performance but clearly required refining (see Table 2). This could suggest that the sample was not large enough to allow the classifier to learn a suitable set of predictive features. It could also suggest the features themselves were either not adequate to represent the latent meaning that human annotators identified when assigning each tweet to a class, or the features were not being suitably utilised during the learning phase. Both sample size and feature set limitations led us to incorporate an *ensemble* classification approach, which enabled us to combine the base classifiers and different methods of feature sampling during the learning phase. There are two very popular ensemble approaches. One is Boosting [13] (e.g. AdaBoost), which aims to 'boost' the performance of a classifier by iteratively adding a new classifier to the ensemble where each new classifier is trained on data for which the previous iteration performed poorly. An advantage of this is that, for

---

[11]http://www.cs.waikato.ac.nz/ml/weka/

smaller samples, the more difficult to classify instances can be focussed on to improve classifier performance. However, this approach has also been reported to reduce classifier accuracy by forcing new classifiers to focus on difficult data points at the sacrifice of other data. The second popular method is Bagging [6], which takes a bootstrap sample of data points and trains a classifier on each sample, averaging out the probabilities for each class across all classifiers in the ensemble.

In [32] the authors propose an ensemble approach known as Rotation Forest (RF), which splits the feature set into a number of smaller sets before sampling from each set and running Principal Component Analysis (PCA) on each subset, creating a number of different principal components for each subset of features, and subsequently building a number of classifiers using these. This approach showed a performance improvement over Bagging and Boosting and provided a logical choice of method to refine our baseline classifiers, given the 1444 features all measuring properties of the text, possible colinearity between features, and variance of features in the training data. We hypothesised that splitting the features into a number of subsets and deriving a range of principal components from these, rather than deriving principal components from all features at once, would reduce the number of false negative results by using a wider range of principal components. We therefore repeated the experiments from the baseline phase with a RF ensemble classifier.

Ensemble meta classifiers can incorporate a number of combined baseline classifiers. We experimented with incorporating all the classifiers used in the baseline experiments to determine how the principles of RF could improve these. As the initial results showed varying performance between classifiers - for example, the NB produced the lowest numbers of false positives using $Set1$ and $Set3$, but SVM produced the lowest false negatives in both cases - we chose to incorporate a second metaclassifier within the RF that used a voting principle as a mechanism to assign the label with maximum probability across all base classifiers to new instances. SVM, J48 Decision Tree and Naive Bayes classifiers were integrated within the RF classifier as an ensemble, with the classifier producing the maximum probability for new instances being selected for each classification decision. We ran two experiments with the RF approach - one with all three baseline classifiers and another with just NB and SVM classifiers. Table 3 shows the notable difference in performance when using DT to classify suicidal ideation, thus it was dropped and the ensemble approach performed much better. We have only reported the results of the NB and SVM combination.

5. RESULTS AND EVALUATION

We used a 10-fold cross validation approach in the evaluation of our classification experiments. This approach iteratively trains the classifier on 90% of the training data and tests on the remaining 10%. After 10 iterations, the results are calculated by taking the mean accuracy across all models. The results are provided in this section at two levels. Tables 2 and 3 present the results for each of the baseline classifiers - Naive Bayes (NB), J48 Decision Tree (DT), and Support Vector Machine (SVM). Each row represents the results using a different set of features. The final column in the table provides the results of the Rotation Forest (RF)

ensemble classifier. Table 2 provides the weighted average results across all classes, while Table 3 provides the results of the key class of interest - suicidal ideation. Evaluation followed standard classification measures of *Precision* measuring false positives, *Recall* measuring false negatives, and *F-measure* a harmonized mean. In the Tables we represent the best scores in bold, and the best precision and recall for each feature set in italic.

Table 2: Machine Classification Results: All classes

Feature \ Classifier		NB	DT	SVM	RF
$Set1$	P	*0.694*	0.635	0.692	0.672
	R	0.681	0.641	*0.689*	0.667
	F	0.681	0.637	*0.682*	0.664
$Set2$	P	0.683	0.620	0.698	*0.703*
	R	0.667	0.622	0.696	*0.702*
	F	0.667	0.620	0.689	*0.696*
$Set3$	P	0.694	0.638	0.690	*0.708*
	R	0.679	0.642	0.686	*0.707*
	F	0.680	0.636	0.680	*0.702*
Combined	P	0.674	0.622	0.695	**0.732**
	R	0.659	0.617	0.689	**0.729**
	F	0.658	0.617	0.690	**0.728**
PCA (combined)	P	0.607	0.552	0.594	*0.647*
	R	0.561	0.547	0.586	*0.591*
	F	0.563	0.549	0.581	*0.591*

Table 3: Machine Classification Results: Suicidal Ideation

Feature \ Classifier		NB	DT	SVM	RF
$Set1$	P	0.514	0.464	**0.657**	0.587
	R	*0.731*	0.410	0.564	0.474
	F	0.603	0.435	*0.607*	0.525
$Set2$	P	0.491	0.397	*0.652*	0.589
	R	*0.705*	0.372	0.577	0.423
	F	0.579	0.384	*0.612*	0.493
$Set3$	P	0.505	0.530	*0.647*	0.614
	R	*0.705*	0.449	0.564	0.449
	F	0.588	0.486	*0.603*	0.519
Combined	P	0.496	0.447	0.551	*0.644*
	R	0.718	0.487	0.692	**0.744**
	F	0.586	0.466	0.614	**0.690**
PCA (combined)	P	0.400	*0.446*	0.441	0.438
	R	0.590	0.526	0.385	*0.628*
	F	0.477	0.482	0.411	*0.516*

The three baseline models perform similarly across all classes for feature set 1,2 and 3, with SVM slightly outperforming NB in most cases, and DT performing least well (see Table 2). In two out of 3 cases NB achieved the best precision score and SVM the best recall in all three - leading us to test an ensemble approach. It is interesting to note that combining all feature sets led to only a 0.001 improvement in precision and actually reduced recall by 0.07 when compared to *Set 2*. Furthermore, applying a dimension reduction method - principle component analysis - led to a further reduction in performance when applied to all features (see bottom three rows of Tables 2 and 3). However, when

the training data was split into smaller samples, with principle components derived for each sample - thus broadening the diversity of components while retaining complexity - we saw a performance increase, going from a maximum performance of P=0.695 and R=0.689 to P=0.732 and R=0.729 across all classes when applying the RF approach combined with a Maximum Likelihood voting metaclassifier.

When digging deeper into the key class of interest - the suicidal ideation class - we see a reduced performance for all base classifiers (see Table 3). The confusion matrix for the best performing classification model (see Table 4) shows that this is largely due to confusion between c1 (suicidal ideation) and c3 (flippant reference to suicide). This was always going to be a challenge given the subjective nature of the task and the difficulty human annotators found in agreeing on this. Sarcasm and irony are notable text classification challenges that are yet to be resolved. This is primarily due to the same language often being used in serious and flippant cases. However, the SVM baseline classifier still achieved a Precision performance of 0.657, which was in fact the best performance - even better than the RF classifier. Indeed, the baseline SVM generally outperformed the other base classifiers, and the RF ensemble, for the individual sets of features. This is in line with other existing research in this area, though we have achieved a higher performance. Yet when combining all features, and applying principle component analysis to smaller subsets of training data, the RF model performed significantly better than any other classification model for the suicidal ideation class. The maximum Recall was 0.744, which is only a slight improvement of 0.013 over the NB baseline using *Set 1*, but the maximum F-measure was 0.69 as compared to 0.61. These results suggest that the ensemble of multiple base classifiers with a maximum probability meta classifier offers a promising way forward for the multi-class classification of suicidal communication and ideation in 'noisy' short informal text, such as social media posts. The 'none of the above' confusion also suggests there may be other latent topics not present in our set of class labels. Identifying these may be a useful task for future research. Table 5 provides P, R and F results for the best performing classifier across all classes for comparison.

Table 4: Confusion matrix for the best performing classification model

class	c1	c2	c3	c4	c5	c6	c7
c1	*58*	0	15	0	0	0	5
c2	0	*18*	1	4	0	4	1
c3	11	0	*143*	0	1	5	17
c4	0	4	5	*18*	0	2	6
c5	1	1	1	0	*31*	1	1
c6	0	6	9	7	2	*76*	4
c7	20	0	23	0	2	4	*94*

6. DISCUSSION

In this section we analyse the components produced by running the Principle Component Analysis (PCA) method on the *combined* set that resulted in the best set of results, as shown in Tables 2 to 5. The application of PCA reduced the features set from 1444 to 255 attributes in terms of main

Table 5: Precision, Recall, and F-measure for the best performing classification model

class	P	R	F
c1	0.644	0.744	0.690
c2	0.621	0.643	0.631
c3	0.726	0.808	0.765
c4	0.621	0.514	0.563
c5	0.861	0.861	0.861
c6	0.826	0.731	0.776
c7	0.734	0.657	0.694

components. For the seven suicide related classes we show in Tables 6 and 7 the most representative principal components and briefly discuss what each class represents in terms of the features in the component and the particular language used in it.

Note that while the distribution of the components per class mirrors the total number of annotation per class (therefore penalising the classes less represented in our data set such as 'memorials') in Tables 6, 7 and in the related discussion we are giving priority to the most representative class of posts containing evidence of possible suicidal intent. We can observe the following characteristics of the features included for each class component:

c1: Many of the features that appear dominant in the suicidal ideation class are those related to phrases and expressions identified in the suicide literature as being significantly associated within the language of suicide. In particular, beside a limited number of uni/bi/tri-grams generated directly from the training set, the terms derived from a number of suicide related Web sites were fundamental in classifying suicidal ideation in Twitter data. As were the regular expression features derived from Tumblr posts. Examples like 'end it all now' and 'want to be dead' and regex including expression of 'depressive/suicidal/self harming' ...'thoughts /feelings' appear strongly related to suicidal ideation and are clearly discriminating for this specific class. Other terms (such as 'killing myself' and the regex containing 'die' ... 'my sleep') become effective for classification when used besides other attributes such as lexical features that express surprise, exaggeration and emphasis (e.g. adverbs ('really'), predeterminers (e.g. 'such' 'rather')), and words mapped to specific 'affective' domains such as 'alarm' and 'misery'. Note that some other concepts and terms appear with a negative correlation as expressions of opposite affective states, such as 'security' and 'admiration'.

c2: For the class representing campaigning and petitions we can observe more general concepts, again expressed by regular expressions and language clues (word-lists in our terminology), such as 'support/help', 'blog' as well as more specific terms (e.g 'safety plea') and expressions ('put an end to this'). Some of the Wordnet domain features require further examination as they appear confusing at first - for example 'racing' is picking up on the words 'run' and 'running' that are related to campaigns.

c3: As the confusion matrix in Table 4 shows, the class concerning a 'flippant' use of suicidal language is the one presenting the major difficulties in classification, since it includes many of the same linguistic features of suicidal ideation. However, the principal components derived for this class identify certain attributes that are the opposite

type of sentiment from emotional distress. These include affective states such as 'levity', 'gaiety', 'jollity', and 'cheerfulness', as well as popular conversational topics, such as casual remarks about the weather. The confusion occurs where phrases such as 'kill myself' are used frivolously.

c4: The class representing posts related to information and support (and prevention) appear mostly represented by specific words (often unigrams and 'tags') directly linked to the support services (e.g. #police, #officers, internet and suicide) and/or topicality (such as sexual references ('#lgtb'), and the domains of self-harm and #suicide).

c5: For the class concerning memorial messages, as may be expected, direct mentions of the name of the deceased appear highly influential as well as 'time' references (e.g. 'a month ago', 'a year since') in association with terms such as 'killed' and 'died' (well captured by one of our regular expressions). In addition labels and tags as 'rip' and terms expressing 'love' 'and 'affection' are also part of the components associated with this class. Again, we see some Wordnet domains appearing - 'mathematics' and 'agriculture' are related to specific words such as 'add' and 'grow'.

c6: The class concerning news reports related to suicide presents features such as words representing sources of information (e.g. #bbc news), types of news (research study or statistical report), and direct mentions of the name of the deceased (as well as general concepts related to the particular case, such as in the one here reported of the 'TV' domain). Note that the last three classes of memorial, information/support, and news reporting all share the common characteristics of including URL links within the tweets which, consequently, does not result in an effective feature for discrimination between these different classes.

c7: Finally, the class of posts annotated as not related to any of the previous classes exhibits attributes such as general phrases related to self doubt (such as 'what's wrong with me and 'hate myself') and emotional states (such as 'jitteriness' and 'admiration'). These are phrases that could appear in tweets relating to emotional distress but are also clearly evident in general everyday 'chatter'.

7. CONCLUSION

In this paper we developed a number of machine classification models built with the aim of classifying text relating to communications around suicide on Twitter. The classifier distinguishes between the more worrying content, such as suicidal ideation, and other suicide-related topics such as reporting of a suicide, memorial, campaigning and support. We built a set of baseline classifiers using lexical, structural, emotive and psychological features extracted from Twitter posts. We then improved on the baseline classifiers by building an ensemble classifier using the Rotation Forest algorithm, achieving an F-measure of 0.728 overall (for 7 classes, including suicidal ideation) and 0.69 for the suicidal ideation class.

We summarised and attempted to explain the results by reflecting on the most significant predictive principle components of each class to provide insight into the language used on Twitter around suicide-related communication. From this analysis we observed that *word-lists and regular expressions (regex)* extracted from online suicide-related discussion fora and other microblogging Web sites appear capable of capturing relevant language 'clues', both in terms of single words, n-grams (word-lists) and more complex patterns.

Table 6: Principal components per class

c1 - Evidence of possible suicidal intent
0.185word_list1_end it all_521+0.185end it all+0.179it all now +0.179all now+0.175it all
0.149word_list1_want to be dead_554-0.133_-0.129i think +0.125word_list1_to commit suicide_547+0.114really
0.149word_list1_want to be dead_554+0.145wn_affect11_alarm _496-0.123number of adverb superlative_211-0.121word_list7 _relationship_780+0.118regEx_class6_+.+\report.+_701
0.153thinking about killing+0.153about killing myself +0.153about killing+0.147so im+0.147wn_affect11_misery_314
0.119number of predeterminers_206+0.117regEx_class1_+.+ ((\cutting|\depres|\sui)|\these|\bad|\sad).+(\thoughts|\feel) .+_667+0.115wn_domain_astrology_160-0.106bombing
0.231regEx_class1_+.+(\bdie).+(\bmy).+\bsleep.+0.177word _list_want to be dead_554-0.155wn_domain_dentistry_113 -0.146wn_affect11_security_277-0.129wn_affect11_admiration

c2 - Campaigning (i.e. petitions etc.)
0.25 word_list2_support_746-0.134wn_domain_racing_84 +0.119regEx_class2_+.+blog.+_683+0.113wn_domain_jewellery
0.189safety+0.188plea+0.188safety plea+0.188plea over
0.187end to+0.187word_list_ put an end to this_540 +0.187an end to+0.187an end+0.152r i

c3 - Flippant reference to suicide
0.112wn_domain_meteorology_166+0.11 to live+0.107 wn_affect1_jollity_333+0.107wn_affect11_levity_327 +0.107wn_affect11_levity-gaiety_378
0.14 word_list_ want to be here anymore_575-0.13number of existentials (there)_196+0.126wn_affect11_cheerfulness_459 -0.111so-0.111really
0.162wn_affect11_jollity_333+0.162wn_affect11_levity_327 +0.162wn_affect11_levity-gaiety_378+0.128or +0.113wn_domain_meteorology_166
-0.159myself-0.144regEx_class3_total_662-0.136regEx_class3_+. +(\to).+(\kill|\disapp).+_672-0.125to kill myself-0.125to kill

c4 - Information or support
0.152and anxiety self-harm+0.152challenge+0.152 challengesexps to#lgbt+0.152young people#mylgbthealth
0.175#police #officers in+0.175#suicide preventiontoday +0.175#suicide prevention+0.175officers trained+0.175#police
0.21 internet & suicide+0.21 between internet & +0.21 & suicide http+0.21 & suicide+0.21 internet

c5 - Memorial or condolence
0.155regEx_class5_+.+(\kill|\die|\comm).+(day|month| year.+_692+0.138wn_domain_mathematics_117 +0.13 wn_domain_agriculture_104-0.12wn_domain_tax_126 -0.116number of interjections_215
0.125wn_affect11_love_324+0.125love+0.112rip *name_ replaced*+0.11 rip *name_replaced*+0.107rip

c6- Reporting news of someone's suicide (not bombing)
0.178bbc news+0.15 number+0.15 deaths by+0.15 deaths by suicide+0.15 number of+0.15 by suicide from
0.158research-0.123off-0.107self+0.106to study link+0.1 see_626
0.129regEx_class6+.+friend.+_690+0.12 friend_608-0.114regEx _class2_+.+blog.+_683-0.101adverb_599+0.101killed
+0.144self+0.121wn_domain_tv_184+ 0.101*name_replaced*13+0.101*name_replaced*+0.093dead

These appear particularly effective for the suicidal ideation class, expressing emotional distress. *Lexical and grammar features* such as POSs appear mostly ineffective and scarcely present in the principal components (only some mentions as predeterminers, existential clauses and superlatives that, however, also relate to more specific 'affective' language features than only pure lexical ones). *Affective lexical domains*, appear instead very relevant (such as those represented by the WordNet library of 'cognitive synonyms') and able to well represent the affective and emotional states associated to this particular type of language.

Concepts and labels representing broader semantic domains (also derived form the WordNet library) are, on the contrary, not effective. In fact, although they appear rather numerous as attributes within the principle components they reveal to be, on close inspection, for the majority of cases irrelevant and mostly generated by a 'confusion' and 'misrepresentation' of words (such as sentences like 'my reason crashed' associated to the 'motor-racing' domain, and 'suicide watch' associated to 'numismatic').

Sentiment Scores generated by software tools for sentiment analysis appear also ineffective and either scarcely or not at all included within the principal features of each class. Note that this is true for both basic tools that only provide a binary representation of positive and negative score values (SentiWordNet) as well as more sophisticated text analysis software that generate sentiment scores over a larger range of labels representing emotional states (LIWC).

Table 7: Principal components per class

c7-None of the above
0.15 dont want +0.149word_list_ dont want to be here_518 +0.149regEx_class7_+.+(\don.+).+(\wan.+).+here.+_707
0.136to live+0.133live+0.124hate myself +0.113hate myself for+0.113myself for
0.213regEx_class1_+.+(\die).+(\my).+\sleep.+_677 +0.2 wn_affect11_jitteriness_335+0.147wn_affect11_admiration _501+0.132wn_domain_mythology_135+ 0.207regEx_class1_+.+(\die).+(\my).+\sleep.+_677
+0.133wn_domain_town_planning_79-0.12wn_domain_painting _121+0.12bombing

A classifier for suicide-related language could potentially make an important contribution to suicide prevention. Monitoring individual social media accounts via keywords that suggest possible suicidal ideation is controversial territory, as shown by the recent withdrawal of the Samaritans Radar app in the UK [12] but there is nonetheless potential for such a lexicon to contribute to prevention in some way, as long as acceptability to social media users is thoroughly investigated. The 'real-time' identification of aggregate levels of suicide-related communication at scale in online social networks, which could be facilitated by the ensemble classifier produced in this research, is one possible approach. This could potentially aid the identification of emerging suicide clusters and the concentration of suicidal communication following particular events such as a celebrity suicide. Our classifier goes beyond the recognition of suicidal language insofar as it also aids identification of other kinds of communication, in recognition that social media platforms can be used for multiple purposes, including the reporting of news and marshalling of campaigns. Monitoring of suicide news reporting in social media is another potential avenue where text mining and machine classification techniques could be applied. The identification of flippant use of suicidal language could be especially useful. The methods needs further development, ideally with a larger sample of social media postings, and application to platforms other than Twitter. Finally, we note that it is important to retain collaboration with domain experts in suicidology throughout the experimental and interpretation phases of future research to improve classification accuracy by incorporating prior knowledge of the characteristics of suicidal language - especially given the significance of the affective features in this paper.

8. ACKNOWLEDGMENTS

9. REFERENCES

[1] A. Abboute, Y. Boudjeriou, G. Entringer, J. Aze, S. Bringay, and P. Poncelet. Mining twitter for suicide prevention. In *Natural Language Processing and Information Systems*, volume 8455 of *Lecture Notes in Computer Science*, pages 250–253. Springer, 2014.

[2] D. Baker and S. Fortune. Understanding self-harm and suicide websites. *Crisis: The Journal of Crisis Intervention and Suicide Prevention*, 29(3):118–122, 2008.

[3] L. Barbosa and J. Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 36–44. Association for Computational Linguistics, 2010.

[4] K. Becker and M. H. Schmidt. When kids seek help on-line: Internet chat rooms and suicide. *reclaiming children and youth*, 13(4):229–230, 2005.

[5] L. Biddle, J. Donovan, K. Hawton, N. Kapur, and D. Gunnell. Suicide and the internet. *Bmj*, 336(7648):800–802, 2008.

[6] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[7] P. Burnap, O. F. Rana, N. Avis, M. Williams, W. Housley, A. Edwards, J. Morgan, and L. Sloan. Detecting tension in online communities with computational twitter analysis. *Technological Forecasting and Social Change*, 2013.

[8] M. D. C. S. Counts and M. Gamon. Not all moods re created equal! a exploring human emotional states in social media. 2012.

[9] K. Daine, K. Hawton, V. Singaravelu, A. Stewart, S. Simkin, and P. Montgomery. The power of the web: a systematic review of studies of the influence of the internet on self-harm and suicide in young people. *PloS one*, 8(10):e77555, 2013.

[10] M. De Choudhury, S. Counts, E. J. Horvitz, and A. Hoff. Characterizing and predicting postpartum

---

depression from shared facebook data. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 626–638, New York, NY, USA, 2014. ACM.

[11] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz. Predicting depression via social media. In *ICWSM*, 2013.

[12] B. Desmet and V. Hoste. Emotion detection in suicide notes. *Expert Systems with Applications*, 40(16):6351–6358, 2013.

[13] Y. Freund and R. E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer, 1995.

[14] M. Gould, P. Jamieson, and D. Romer. Media contagion and suicide among the young. *American Behavioral Scientist*, 46(9):1269–1284, 2003.

[15] J. F. Gunn and D. Lester. Twitter postings and suicide: An analysis of the postings of a fatal suicide in the 24 hours prior to death. *Present tense*, 27(16):42, 2012.

[16] C. Homan, R. Johar, T. Liu, M. Lytle, V. Silenzio, and C. Ovesdotter Alm. Toward macro-insights for suicide prevention: Analyzing fine-grained distress at scale. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology*, pages 107–117, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[17] Y.-P. Huang, T. Goh, and C. L. Liew. Hunting suicide notes in web 2.0-preliminary findings. In *Multimedia Workshops, 2007. ISMW'07. Ninth IEEE International Symposium on*, pages 517–521. IEEE, 2007.

[18] A. Ikunaga, S. R. Nath, and K. A. Skinner. Internet suicide in japan: A qualitative content analysis of a suicide bulletin board. *Transcultural psychiatry*, page 1363461513487308, 2013.

[19] N. Jacob, J. Scourfield, and R. Evans. Suicide prevention via the internet: A descriptive review. *Crisis: The Journal of Crisis Intervention and Suicide Prevention*, 35(4):261, 2014.

[20] J. Jashinsky, S. H. Burton, C. L. Hanson, J. West, C. Giraud-Carrier, M. D. Barnes, and T. Argyle. Tracking suicide risk factors through twitter in the us. 2013.

[21] V. Kolhatkar, H. Zinsmeister, and G. Hirst. Interpreting anaphoric shell nouns using antecedents of cataphoric shell nouns as training data. In *EMNLP*, pages 300–310, 2013.

[22] M. T. Lehrman, C. O. Alm, and R. A. Proaño. Detecting distressed and non-distressed affect states in short forum texts. In *Proceedings of the Second Workshop on Language in Social Media*, pages 9–18. Association for Computational Linguistics, 2012.

[23] M. Liakata, J.-H. Kim, S. Saha, J. Hastings, and D. Rebholz-Schuhmann. Three hybrid classifiers for the detection of emotions in suicide notes. *Biomedical informatics insights*, 5(Suppl 1):175, 2012.

[24] P. Matykiewicz, W. Duch, and J. Pestian. Clustering semantic spaces of suicide notes and newsgroups articles. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*,

pages 179–184. Association for Computational Linguistics, 2009.

[25] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, 2010.

[26] J. Pennebaker, M. Francis, and R. Booth. *Linguistic Inquiry and Word Count: A computerized text analysis program.* 2001.

[27] J. Pestian, H. Nasrallah, P. Matykiewicz, A. Bennett, and A. Leenaars. Suicide note classification using natural language processing: A content analysis. *Biomedical informatics insights*, 2010(3):19, 2010.

[28] J. P. Pestian, P. Matykiewicz, M. Linn-Gust, B. South, O. Uzuner, J. Wiebe, K. B. Cohen, J. Hurdle, and C. Brew. Sentiment analysis of suicide notes: A shared task. *Biomedical informatics insights*, 5(Suppl 1):3, 2012.

[29] J. Pirkis and R. W. Blood. Suicide and the media. *Crisis: The Journal of Crisis Intervention and Suicide Prevention*, 22(4):155–162, 2001.

[30] C. Poulin, B. Shiner, P. Thompson, L. Vepstas, Y. Young-Xu, B. Goertzel, B. Watts, L. Flashman, and T. McAllister. Predicting the risk of suicide by analyzing the text of clinical notes. *PloS one*, 9(1):e85733, 2014.

[31] P. R. Recupero, S. E. Harms, and J. M. Noble. Googling suicide: surfing for suicide information on the internet. *Journal of Clinical Psychiatry*, 2008.

[32] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso. Rotation forest: A new classifier ensemble method. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(10):1619–1630, 2006.

[33] T. D. Ruder, G. M. Hatch, G. Ampanozi, M. J. Thali, and N. Fischer. Suicide announcement on facebook. *Crisis: The Journal of Crisis Intervention and Suicide Prevention*, 32(5):280–282, 2011.

[34] I. Spasić, P. Burnap, M. Greenwood, and M. Arribas-Ayllon. A naïve bayes approach to classifying topics in suicide notes. *Biomedical informatics insights*, 5(Suppl 1):87, 2012.

[35] H. Sueki. The association of suicide-related twitter use with suicidal behaviour: A cross-sectional study of young internet users in japan. *Journal of affective disorders*, 2014.

[36] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

[37] H.-H. Won, W. Myung, G.-Y. Song, W.-H. Lee, J.-W. Kim, B. J. Carroll, and D. K. Kim. Predicting national suicide numbers with social media data. *PloS one*, 8(4):e61809, 2013.

[38] C. Yang, K. H. Lin, and H.-H. Chen. Emotion classification using web blog corpora. In *Web Intelligence, IEEE/WIC/ACM International Conference on*, pages 275–278. IEEE, 2007.

[39] H. Yang, A. Willis, A. De Roeck, and B. Nuseibeh. A hybrid model for automatic emotion recognition in suicide notes. *Biomedical informatics insights*, 5(Suppl 1):17, 2012.