Libertas Academica
FREEDOM TO RESEARCH

ORIGINAL RESEARCH

# A Hybrid Approach to Sentiment Sentence Classification in Suicide Notes

Sunghwan Sohn[1,*], Manabu Torii[2,*], Dingcheng Li[1], Kavishwar Wagholikar[1], Stephen Wu[1] and Hongfang Liu[1]

[1]Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN. [2]ISIS Center, Georgetown University Medical Center. *Equal contribution. Corresponding author email: sohn.sunghwan@mayo.edu

**Abstract:** This paper describes the sentiment classification system developed by the Mayo Clinic team for the 2011 I2B2/VA/Cincinnati Natural Language Processing (NLP) Challenge. The sentiment classification task is to assign any pertinent emotion to each sentence in suicide notes. We have implemented three systems that have been trained on suicide notes provided by the I2B2 challenge organizer—a machine learning system, a rule-based system, and a system consisting of a combination of both. Our machine learning system was trained on re-annotated data in which apparently inconsistent emotion assignment was adjusted. Then, the machine learning methods by RIPPER and multinomial Naïve Bayes classifiers, manual pattern matching rules, and the combination of the two systems were tested to determine the emotions within sentences. The combination of the machine learning and rule-based system performed best and produced a micro-average F-score of 0.5640.

**Keywords:** sentiment classification, suicidal emotion, natural language processing, machine learning

This article is available from http://www.la-press.com.

## Introduction

The I2B2/VA/Cincinnati 2011 Natural Language Processing (NLP) Challenge has two tracks; track 1 is co-reference resolution in clinical notes, and track 2 is sentiment classification in suicide notes. This paper describes Mayo Clinic team's work on track 2. In sentiment classification, the data come from actual notes written by people who have committed suicide. The sentences that indicate patient sentiments are assigned to one or more emotions. The sentiment classification task is to identify emotional sentences in the suicide notes and assign the corresponding suicidal emotions to them.

Suicide is a leading cause of death among teenagers and adults under 35.[1] It is the second leading cause of death in the United States for 25 to 34 year olds and the third for 15 to 25 year olds.[2] Suicide notes contain rich emotional expressions from suicide victims and provide a fundamental basis for accessing the suicidal mind. There have been some attempts to study suicide notes using computational methods. Pestian et al investigated suicide note classification using NLP and various machine learning (ML) techniques to distinguish actual suicide notes from crafted ones.[3] Matykiewicz et al developed a system using the unsupervised ML technique to calculate a real-time index for the likelihood assessment of repeated suicide attempts and to differentiate real suicide notes from other notes in newsgroups.[4] Huang et al used a simple keyword search method to find emotional content in blog data and identify bloggers at risk of suicide.[5] The above studies attempt a binary classification to distinguish actual suicide notes from others at a document level. However, deep sentiment analysis in suicide notes has not yet been explored much with computational approaches using advanced ML and NLP techniques. Track 2 of the 2011 I2B2/VA/Cincinnati NLP challenge inaugurates this more complicated task of sentiment analysis to classify 15 emotions at the sentence level within actual suicide notes.

Mayo Clinic team implemented a machine learning system based on RIPPER[6] and multinomial Naïve Bayes[7] using Weka machine learning suite[8] and also a hand-coded rule-based system consisting of regular expressions to identify key phrases and phrasal patterns in text. We also investigated the combination of both methods to augment the classification performance.

## Data

The training set consists of 600 actual suicide notes and the gold standard annotation of the emotional sentences. Each emotional sentence was manually labeled with one or more emotions from a set of 15 different emotions. Table 1 shows the emotion list and brief annotation guidelines provided by the I2B2 organizer.

Distribution of assigned emotion types is highly skewed. Figure 1 shows the statistics in the training set. The eight most frequent emotions cover 90% of the training data; the remaining emotions rarely occur. The challenge organizer provided 300 suicide notes as a test set that has a similar distribution of the emotional classes.

## Method
### Data preprocessing

Noticing that certain types of named entities (eg, person name) play a key role in the current task, we implemented a simple NER program based on regular expressions. Detected phrases were converted into special tokens before n-grams were extracted from text, eg, converting "Mrs. J. J. John" to "__NAME__". Entity types considered were, from the most to least frequent types detected in the training corpus, person name, salutation (eg, "Dear Jane," and "Darling : "), date and time (eg, "01-01 9 AM" and "January 1st, 2001"),

**Table 1.** Emotions and their annotation guidelines.

| Emotion | Annotators guideline |
| --- | --- |
| Abuse | Was abused verbally, physically, mentally ... |
| Anger | Is angry with someone ... |
| Blame | Is blaming someone ... |
| Fear | Is afraid of something ... |
| Guilt | Feels guilt ... |
| Hopelessness | Feels hopeless ... |
| Sorrow | Feels sorrow ... |
| Forgiveness | Is forgiving someone ... |
| Happiness_peacefulness | Is feeling happy or peaceful ... |
| Hopefulness | Has hope for future ... |
| Love | Feels love for someone ... |
| Pride | Feels pride ... |
| Thankfulness | Is thanking someone ... |
| Instructions | Giving directions on what to do next |
| Information | Giving practical information where things stand |
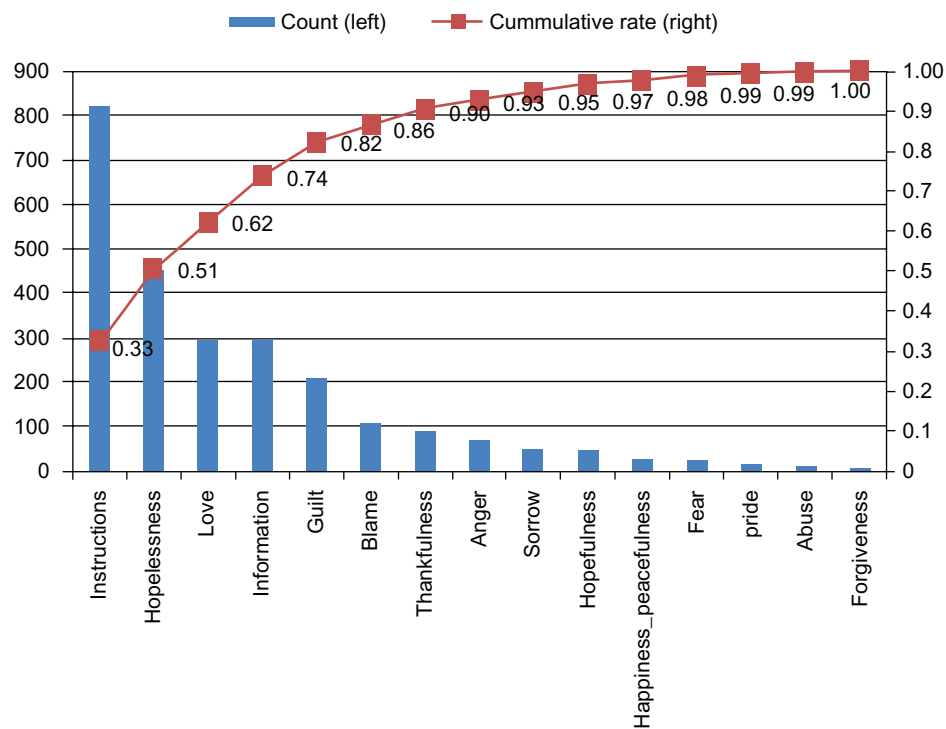
**Figure 1.** Statistics of 600 suicide notes in the training set.

money (eg, "Two hundred dollars" and "$80.00"), address (eg, "3333 Burnett Ave, Cincinnati, Ohio"), car (eg, "the car", "my truck", and "Toyota"), location (eg, "Children's Hospital"), and phone number (eg, "636 2051"). These phrases types must be place holders inserted during anonymization.

## Systems development

Three systems developed by the Mayo Clinic team are described below.

**System 1** (ML emphasis) is based on Weka implementations[8] of multinomial Naïve Bayes (MNB)[7] and RIPPER.[6] In this system, just as our other systems, we assumed that sentences in notes (classification units) are independent of each other and also that assignment of one emotion type is independent of that of the other types. With these assumptions, detection of a particular emotion was tackled as a binary classification and thus the challenge task was viewed as a set of 15 independent problems.

In training and applying a system, each sentence was regarded as a set of n-gram tokens (contiguous n tokens), and represented as a binary vector indicating the presence/absence of pre-selected n-grams

(features) in it. We included 1-, 2-, and 3-grams as feature candidates, and information gain[12] was used to select informative features for the detection of each emotion. Among several machine learning algorithms we tested using Weka, the MNB classifier produced the best overall performance in cross-validation on the training corpus. Yet, we noticed that rule induction methods, such as RIPPER, generally performed better than MNB for the two emotions, love and thankfulness. Therefore, in the final configuration, MNB was used[a] for six emotions (blame, guilt, hopelessness, information, instructions, and thankfulness), and RIPPER was used for two emotions (love and thankfulness). One emotion, thankfulness, was predicted based on the union of both algorithms' output: if either algorithm predicts the presence of the emotion, then that prediction was used. The remaining eight emotions were not identified by our ML methods. Thus, we used a simple phrase pattern matching for those emotions but only for obvious cases, for example, "forgiveness", we used the pattern matching "I forgive {you, him, her, …}."

---

[a]Inclusion of 2- and 3-grams and binary vector representation, not term frequency vector, does not conform to the multinomial model, but the performance was improved with them and so we used them anyway.

Besides the overall architecture described above, we employed additional techniques:

**Token normalization**—All letters were lower-cased, and digits were converted to the same number '9'. Provided data was already tokenized, but we further process it to make the input text consistent, eg, "can n't" and "cannot" were converted to more prevalent "ca n't".

**Classifier Ensemble**—For MNB, five models were trained using different thresholds for information gain (0.01, 0.005, 0.001, 0.0005, and 0.0002), and the highest predicted value from these models was used during model application. For RIPPER, nine models were trained using three pruning data sizes (1/3, 1/5, and 1/10) each with three different random seeds, and majority voting was used to determine the final prediction.

**Corpus re-annotation**—We noticed irregular sentence splitting and annotation inconsistency in the training corpus. Hoping to mitigate these challenges, we manually split sentences and re-annotated emotions in the training corpus. The system performance was improved in cross-validation when the system was trained and tested on the re-annotated corpus. The performance, however, was degraded when it was trained on the re-annotated corpus but evaluated on the original corpus. In our second attempt, we did not modify sentence split, but reviewed false positives and negatives by the system observed in cross-validation on the training corpus and re-annotated them if the gold standard annotations seem to be inconsistent, similar to the re-annotation process by Patrick et al in 2010 I2B2/VA Challenge.[13] Using this second corpus, we could generally observe a slight improvement (~1% F-score) in performance in 10-fold cross-validation on the training corpus.

**System 2** (Rule emphasis) implements manually developed pattern matching rules, using Perl regular expressions. This is a simple string matching rules and does not employ any syntactic information. Each note was processed using the GENIA tagger[9–11] to obtain part-of-speech (POS) tags and normalized forms of words. For each emotion, keywords and their description patterns were manually compiled from the training data to generate rules. The normalized word forms, POS tags, and named-entity templates were also utilized to generalize patterns. We also use synonyms in WordNet to expand patterns. For example, one of the rules for emotion "love" is: normalized sentence = ~ /I ([^ ]+ ){0,3} love (you|him|her|them|my|NAME_NE)/i.

For each sentence, we applied every pattern-matching rule we developed and assigned any matching emotion to that sentence. This heuristic allows more than one emotion for a given sentence. Figure 2 presents a brief process of the rule-based system.

It is difficult to generalize some emotions, such as instructions and information, according to rules, and we observed worse performances using the rules than ML. For those two emotions, we used machine learning results from System 1 instead of rules.

**System 3** (the Union of System 1 and System 2) implements the combined results from System 1 and 2 results. Any emotion found by Systems 1 and 2 are merged to produce the final outputs. This system is intended to improve recall performance, assuming Systems 1 and 2 have enough mutually exclusive outputs.

## Post processing

In all three systems, we applied an additional process to check input lines against those in the training corpus. There were certain phrases repeated across different notes, such as salutations, eg, "Dear __NAME__" and common brief expressions, eg, "I love you". If an input line to be processed is found among training instances, the emotion assigned in the annotated instances should be generally applicable to this input line. This mechanism was motivated to avoid silly errors by automated systems, particularly machine learning models. For example, opening salutations, such as "Dear Jane", were falsely assigned with instructions by a machine learning model in some runs of cross-validation tests, possibly due to the frequent occurrences of person names in instructions. With this mechanism, "dear __NAME__", a normalized form of "Dear Jane", was compared against all such normalized instances in the training data, and false assignment of instructions could be avoided after reviewing emotions assigned to the found training

```
Preprocess the sentence (word/named entity normalization, POS tagging)
For each preprocessed sentence
      For each emotion rules
         Apply each rule
              If match found, assign the given emotion to the sentence
              Else apply next rule
      Continue until consume all emotion rules
```

**Figure 2.** Summary of pattern matching rules.

instances, specifically, by confirming more than two-thirds of the found training instances were not assigned with instructions.

## Evaluation measurement

For evaluation of system outputs, precision, recall, and F-score are used:

$$Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN}$$

$$F\text{-}score = \frac{2*(precision*recall)}{(precision+recall)}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

## Results

The three systems described in the Method section were evaluated with Table 2 showing the results on the test set. Overall, System 3 (Union) produced the highest number of TPs, while System 1 (ML emphasis) and System 2 (Rule emphasis) produced the same number of TPs. The ML System produced the lowest number of FPs followed by the Rule and Union Systems. The evaluation results on the test set are summarized in Table 3. The Union System achieved the highest micro average F-score of 0.5640, but it is very close to the F-score for the ML System (0.5636).

The ML System produced the highest precision of 0.6112, followed by the Rule System (0.5906) and the Union System (0.5709). While the Union System produced the lowest precision, it achieved the highest recall of 0.5574. The highest micro average F-score of the Union System is due to the relatively high recall level, and in spite of relatively low precision.

Rare emotion types (ie, emotion types rarely assigned in the training corpus) were either not identified at all or identified very poorly. For instance, sorrow, forgiveness, and abuse did not have any TP cases in all three systems. Blame, hopefulness, anger, happiness_peacefulness, fear, and pride produced very low F-scores.

Table 4 shows the statistics of emotions, emotional sentences and also total sentences in the test set. The gold standard has an average of 1.16 emotions per emotional sentence, while our best system (System 3) has an average of 1.26 emotions per emotional sentence. The ratio of emotions and the total number of sentences (column Emot/Tsent) are very close between the gold standard and our best system (0.61 vs. 0.60).

## Discussion

Our system of sentiment classification combines ML algorithms and regular expression pattern rules. It is built upon techniques like named entity recognition,

**Table 2.** Result statistics on the test set.

| Emotion | System 1 (ML) | | | System 2 (Rule) | | | System 3 (Union) | | |
|---|---|---|---|---|---|---|---|---|---|
| | TP | FN | FP | TP | FN | FP | TP | FN | FP |
| Instructions (382) | 254 | 128 | 382 | 254 | 128 | 135 | 254 | 128 | 135 |
| Hopelessness (229) | 130 | 99 | 79 | 123 | 106 | 85 | 143 | 86 | 119 |
| Love (201) | 133 | 68 | 53 | 135 | 66 | 54 | 143 | 58 | 74 |
| Guilt (117) | 58 | 59 | 49 | 49 | 68 | 42 | 64 | 53 | 57 |
| Information (104) | 51 | 53 | 79 | 51 | 53 | 79 | 51 | 53 | 79 |
| Thankfulness (45) | 31 | 14 | 20 | 38 | 7 | 34 | 39 | 6 | 34 |
| Blame (45) | 1 | 44 | 3 | 4 | 41 | 4 | 4 | 41 | 7 |
| Hopefulness (38) | 0 | 38 | 0 | 1 | 37 | 2 | 1 | 37 | 2 |
| Sorrow (34) | 0 | 34 | 1 | 0 | 34 | 12 | 0 | 34 | 12 |
| Anger (26) | 1 | 25 | 0 | 3 | 23 | 3 | 3 | 23 | 3 |
| Happiness_peacefulness (16) | 3 | 13 | 0 | 3 | 13 | 6 | 3 | 13 | 6 |
| Fear (13) | 2 | 11 | 4 | 3 | 10 | 5 | 3 | 10 | 5 |
| Pride (9) | 1 | 8 | 0 | 1 | 8 | 0 | 1 | 8 | 0 |
| Forgiveness (8) | 0 | 8 | 0 | 0 | 8 | 0 | 0 | 8 | 0 |
| Abuse (5) | 0 | 5 | 0 | 0 | 5 | 0 | 0 | 5 | 0 |
| **Overall (1272)** | **665** | **607** | **423** | **665** | **607** | **461** | **709** | **563** | **533** |

**Note:** In Emotion column, number in () is the number of the given emotion in the gold standard.

**Table 3.** Evaluation results on the test set.

| Emotion | System 1 (ML) | | | System 2 (Rule) | | | System 3 (Union) | | |
|---|---|---|---|---|---|---|---|---|---|
| | Prec | Rec | F-sco | Prec | Rec | F-sco | Prec | Rec | F-sco |
| Instructions | 0.653 | 0.665 | 0.659 | 0.653 | 0.665 | 0.659 | 0.653 | 0.665 | 0.659 |
| Hopelessness | 0.622 | 0.568 | 0.594 | 0.591 | 0.537 | 0.563 | 0.546 | 0.624 | 0.582 |
| Love | 0.715 | 0.662 | 0.687 | 0.714 | 0.672 | 0.692 | 0.659 | 0.711 | 0.684 |
| Guilt | 0.542 | 0.496 | 0.518 | 0.538 | 0.419 | 0.471 | 0.529 | 0.547 | 0.538 |
| Information | 0.392 | 0.490 | 0.436 | 0.392 | 0.490 | 0.436 | 0.392 | 0.490 | 0.436 |
| Thankfulness | 0.608 | 0.689 | 0.646 | 0.528 | 0.844 | 0.650 | 0.534 | 0.867 | 0.661 |
| Blame | 0.250 | 0.022 | 0.041 | 0.500 | 0.089 | 0151 | 0.364 | 0.089 | 0.143 |
| Hopefulness | 0 | 0 | 0 | 0.333 | 0.026 | 0.049 | 0.333 | 0.026 | 0.049 |
| Sorrow | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Anger | 1.000 | 0.038 | 0.074 | 0.500 | 0.115 | 0.188 | 0.500 | 0.115 | 0.188 |
| Happiness_peacefulness | 1.000 | 0.188 | 0.316 | 0.333 | 0.188 | 0.240 | 0.333 | 0.188 | 0.240 |
| Fear | 0.333 | 0.154 | 0.211 | 0.375 | 0.231 | 0.286 | 0.375 | 0.231 | 0.286 |
| Pride | 1.000 | 0.111 | 0.200 | 1.000 | 0.111 | 0.200 | 1.000 | 0.111 | 0.200 |
| Forgiveness | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Abuse | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Overall (micro-avg†)** | **0.6112** | **0.5228** | **0.5636** | **0.5906** | **0.5228** | **0.5546** | **0.5709** | **0.5574** | **0.5640** |

**Note:** †Micro averaged – ie, obtained by using a global count of each emotion and averaging these sums.
**Abbreviations:** Prec, precision; Rec, recall; F-sco, F-score.

word normalization, POS tagging, and synonyms expansion through Wordnet. These NLP techniques helped to better generalize data and to improve the performance of ML and rule-based systems.

As for the ML approach, the major advantage we observed was that it generalized well across the problem space, when provided with sufficient training data. Our ML system produced better performance on the final test set (F-score of 0.564) than what it achieved in a 10-fold cross-validation on the training corpus (F-score of 0.552). Meanwhile, some emotions have a small number of instances (eg, abuse and forgiveness) and they posed a great challenge to the ML system. Apart from the inherent challenges in identifying human emotions, irregularity of data formatting and annotation posed additional challenges. Our other ML exploration includes stemming of words for token normalization, application of SMOTE[14] and also

support vector machine (SVM) with unequal error costs for the class imbalance problem, application of a sequential labeling method to exploit cross-line dependency, and incorporation of POS and other syntactic information as features. These attempts either did not result in improved performance or could not be explored fully within the limited time frame during the challenge event.

Some emotions are expressed with relatively explicit indication keywords and simple patterns (eg, love and thankfulness). For these kinds of emotions pattern matching rules seem to be effective. Meanwhile, the other emotions are expressed in various ways and are difficult to generalize phrasal patterns by manual pattern matching rules (eg, hopelessness, guilt, instructions, and information). Although a few description patterns can be found in these emotions (ie, instructions has a description pattern in that the sentence starts with a base verb form; information contains address, phone number, money information, etc.), they are not always correct and do not cover many cases. In addition, the shortcoming of our rule-based system is that it was over-fit to the training data. Therefore the classification performance on the test data was much lower than that of the training data (F-score of 0.628 vs. 0.555 in the training and test sets, respectively).

The Union system produced the highest recall because it merged both systems' outputs; however,

**Table 4.** Emotion statistics in the test set.

| | Emot | Esent | Emot/Esent | Emot/Tsent |
|---|---|---|---|---|
| Gold standard | 1272 | 1098 | 1.16 | 0.61 |
| System 1 (ML) | 1088 | 889 | 1.22 | 0.52 |
| System 2 (rules) | 1126 | 907 | 1.24 | 0.54 |
| System 3 (union) | 1242 | 984 | 1.26 | 0.60 |

**Notes:** Emot is the number of emotions, Esent is the number of sentences that contain emotion, Tsent (=2086) is the total number of sentences in the test set.

its precision was lower than the others. Although the Union system produced the highest micro average F-score, a simple union of two system outputs did not achieve much gain compared to the ML system.

Besides the challenges pertaining to sentiment analysis, insights we gained from the current exercise include the irregularity and heterogeneous nature of real-life data broadly gathered for clinical NLP applications. We believe these problems are not able to resort only to data, format, and/or annotation standardization because it must be the reality that incoming data are always not well-formatted. NLP systems need to accommodate unexpected vocabulary, formatting and inconsistent annotation to some extent. That being said, regular formatting of text and consistent annotation would be highly desired. Notably, in System 1 (ML emphasis), our re-annotation of the training corpus improved emotion classification by ~1% in micro average F-score on both training and test sets.

In the current corpus, we observed that the exact same sentence can have different emotions assigned to it in the gold standard. For example, the following cases are inconsistently assigned a given emotion (ie, sometimes assigned a given emotion, sometimes not):

> *"Thanks _NAME_." <e = "thankfulness">*
> *"I love you." <e = "love">*

In addition, similar sentences often disagree with their emotion class. We assume that this is because the feeling of emotion could be subjective, and in addition, it could be affected by the context of the whole document. This fact might cause inconsistency in emotion annotation. The relatively low F-score of this task (mean = 0.4875 of all participated teams in this I2B2 challenge) might also reflect this intrinsic difficulty in emotion assignment. To partially incorporate nearby contexts, we tried to use one previous sentence as well as the current one in ML training, but this did not improve the classification performance.

For some emotions (eg, sorrow, blame, and anger), it seems necessary to understand the overall contextual meaning of the sentence rather than using simple indication keywords. The ML trained without syntactic/semantic features and string pattern matching rules is prone to fail in correctly identifying those emotions.

Even further, some emotions seem to be annotated based on document level understanding rather than handling individual sentences separately. Those emotions are hard to classify correctly unless the system understands the overall context and feeling of the person. The current system trained and tested without considering deep syntactic/semantic aspects would face the difficulty in correctly identifying them.

## Conclusion

We investigated techniques to classify sentiment in the sentences of suicide notes in I2B2/VA/Cincinnati 2011 NLP Challenge. Both ML algorithms with a voting ensemble method and pattern matching rules as well as the combination of both were tested. Our best system (union of the ML and rule-base system) achieved a micro average F-score of 0.5640, which ranked 3rd in this challenge. The rule-based system performed better in rare emotions and emotions expressed by clear indication keywords than the ML system, while the ML system performed better in emotions expressed in more complicated ways. Though the union of the ML and rule-based results produced the highest recall, it degraded the precision level and so did not achieve a significant gain in overall F-score compared to the ML system alone.

Possible improvements have been observed in several aspects. Based on our observations, the sentence emotion could be affected by nearby or entire document emotion(s). Thus, utilizing this information as a feature might improve the overall classification performance. It would be also helpful to build an ontology to define key concepts in the domain along with synonymous words representing the same concept, since the amount of data would be rather limited and it is difficult to observe potential emotional keywords in the training data. Rich and deep syntactic level features (eg, subject-object relation or the imperative form) might also help improve a system further.

## Acknowledgements

## Disclosures

Author(s) have provided signed confirmations to the publisher of their compliance with all applicable legal

and ethical obligations in respect to declaration of conflicts of interest, funding, authorship and contributorship, and compliance with ethical requirements in respect to treatment of human and animal test subjects. If this article contains identifiable human subject(s) author(s) were required to supply signed patient consent prior to publication. Author(s) have confirmed that the published article is unique and not under consideration nor published by any other publication and that they have consent to reproduce any copyrighted material. The peer reviewers declared no conflicts of interest.

## References

1. O'Connor R, Sheehy N. Understanding suicidal behaviour: Leicester: BPS Books; 2000.
2. Sussman M, Jones S, Wilson T, Kann L. The Youth Risk Behavior Surveillance System: updating policy and program applications. *J Sch Health*. 2002; 72(1):13–7.
3. Pestian J, Nasrallah H, Matykiewicz P, Bennett A, Leenaars A. Suicide Note Classification Using Natural Language Processing: A Content Analysis. *Biomedical Informatics Insights*. 2010;3:19–28.
4. Matykiewicz P, Duch W, Pestian J. Clustering semantic spaces of suicide notes and newsgroups articles. BioNLP '09 Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing; 2009.
5. Huang Y-P, Goh T, Li C. Hunting Suicide Notes in Web 2.0 – Preliminary Findings. Ninth IEEE International Symposium on Multimedia 2007—Workshops; 2007:517–21.
6. Cohen WW. Fast effective rule induction. In: *Proc of the Twelfth International Conference on Machine Learning*; 1995: MORGAN KAUFMANN PUBLISHERS, INC; 1995:115–23.
7. McCallum A, Nigam K. A comparison of event models for naive bayes text classification. In: *Proc of the AAAI-98 Workshop on Learning for Text Categorization*, Citeseer; 1998:41–8.
8. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*. 2009;11(1):10–8.
9. Kulick S, Bies A, Liberman M, et al. Integrated Annotation for Biomedical Information Extraction. HLT/NAACL 2004 Workshop: Biolink; 2004: 61–8.
10. Tsuruoka Y, Tateishi Y, Kim J-D, et al. Developing a Robust Part-of-Speech Tagger for Biomedical Text. Advances in Informatics—10th Panhellenic Conference on Informatics, LNCS 3746; 2005:382–92.
11. Tsuruoka Y, Tsujii JI. Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data. *Proceedings of HLT/EMNLP*. 2005:467–74.
12. Forman G. An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research*. 2003;3: 1289–305.
13. Patrick JD, Nguyen DHM, Wang Y, Li M. A knowledge discovery and reuse pipeline for information extraction in clinical notes. *Journal of the American Medical Informatics Association*. 2011;18(5):574–9.
14. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*. 2002;16(1):321–57.