Biomedical Informatics Insights

ORIGINAL RESEARCH

Topic Categorisation of Statements in Suicide Notes with Integrated Rules and Machine Learning

Aleksandar Kovačević[1], Azad Dehghan[2], John A. Keane[2] and Goran Nenadic[2]

[1]Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia. [2]School of Computer Science, University of Manchester, Manchester, UK. Corresponding author email: g.nenadic@manchester.ac.uk

Abstract: We describe and evaluate an automated approach used as part of the i2b2 2011 challenge to identify and categorise statements in suicide notes into one of 15 topics, including *Love*, *Guilt*, *Thankfulness*, *Hopelessness* and *Instructions*. The approach combines a set of lexico-syntactic rules with a set of models derived by machine learning from a training dataset. The machine learning models rely on named entities, lexical, lexico-semantic and presentation features, as well as the rules that are applicable to a given statement. On a testing set of 300 suicide notes, the approach showed the overall best micro F-measure of up to 53.36%. The best precision achieved was 67.17% when only rules are used, whereas best recall of 50.57% was with integrated rules and machine learning. While some topics (eg, *Sorrow*, *Anger*, *Blame*) prove challenging, the performance for relatively frequent (eg, *Love*) and well-scoped categories (eg, *Thankfulness*) was comparatively higher (precision between 68% and 79%), suggesting that automated text mining approaches can be effective in topic categorisation of suicide notes.

Keywords: text mining, text classification, suicide notes, sentiment mining

Biomedical Informatics Insights 2012:5 (Suppl. 1) 115–124

doi: 10.4137/BII.S8978

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

Introduction

Automated processing and categorisation of subjective and affective statements (eg, in blogs, tweets, suicide notes) have been both a challenging and hot topic in the humanities and text mining communities in the last decade, in particular with the development of Web 2.0 technologies. Several methods have been developed to automatically identify main messages, sentiments and opinions presented in such environments, across different domains and communities.[1–4]

The need for quantitative and computational processing of suicide notes in particular has been highlighted as a way to identify any risks of (repeat) attempts as presented in Web 2.0 sources.[5,6] The aim of the i2b2 Medical NLP challenge 2011 (Track II) was to classify statements in de-identified suicide notes into 15 different topics (see the challenge description[7] for detailed description of the task). These topics included eleven emotions (*Hopelessness, Love, Guilt, Thankfulness, Anger, Sorrow, Hopefulness, Happiness Peacefulness, Fear, Pride, Forgiveness*) and four other categories such as *Instructions, Information, Blame* and *Abuse*. The task was to categorise each line (roughly corresponding to a sentence) into one or more of these topics, or to tag them as referring to none of these topics.

Previous work on computational analysis of suicide notes has been concerned with content analysis (eg, distribution of positive, negative and emotion words).[6,8,9] Various discriminative features (eg, emotional concepts, part-of-speech tags (POS), readability scores, etc.) have been considered.[6,8] A comparative study between a set of automatic classification algorithms and human counterparts to distinguish genuine suicide notes from simulated ones showed promising results, as nine out of ten machine classification algorithms outperformed the human counterparts (a team of 11 mental health professionals) in distinguishing genuine notes from elicited ones.[8]

Our approach to the task was a hybrid method that integrates rule-based and machine learning (ML) predictions into a topic-categorisation module. Rule-based predictions combined lexical and syntactic patterns with common expressions empirically associated with a given category. The machine learning module consists of a set of classifiers built using a set of features to provide sentence-level predictions. The two prediction modules were combined using three different approaches, which corresponded to the three runs submitted to the challenge. Our best run combined predictions based on rules and all ML scores, resulting in an (micro) F-measure of 53.36%, with 50.47% recall and 56.61% precision. The following sections describe the method in more detail and provide discussions of the results.

Methodology

An analysis of a set of 300 suicide notes that had been provided by the organisers of the i2b2 2011 challenge[7] as the training data revealed that most lines consisted of single sentences, but that there were cases where several sentences appeared in the same line. We further noted that lines that had several topic categories attached to them were likely to have either multi-focal sentences (ie, sentences that contain several statements, either about related or unrelated issues) or indeed several separate sentences. Therefore, the general idea underlying our approach was to determine topic categories for each of the sentences in a note and then integrate sentence-level predictions at the line level.

The system developed for the topic-categorisation consists of four major modules: (A) pre-processing, (B) rule-based predictions, (C) ML-driven predictions and (D) result integration module. Figure 1 provides a detailed system architecture diagram.

A. Pre-processing notes. Before sentence splitting, each note was first split into lines, as the final system output was at the line level and a simple way was needed to identify sentences belonging to a particular line. Each line was then filtered to remove a set of symbols (eg, ^, {, *) that caused problems in the further steps in our workflow (eg, in parsing). Given that suicide notes have a number of typos, we performed spelling correction using Google spell checking API[10] and then applied the Stanford CoreNLP tools[11–13] for tokenisation, sentence splitting, part-of-speech tagging, lemmatisation, shallow parsing and recognition of common named entities (see below). We did not split up multi-focal sentences into separate units.

B. Rule-based prediction module. We relied on lexico-semantic processing that has been tailored for each of the topic categories by considering (1) category-specific lexical and syntactic patterns (eg, *'inform <person>'* for *Instructions*), and (2) common "frozen" expressions empirically

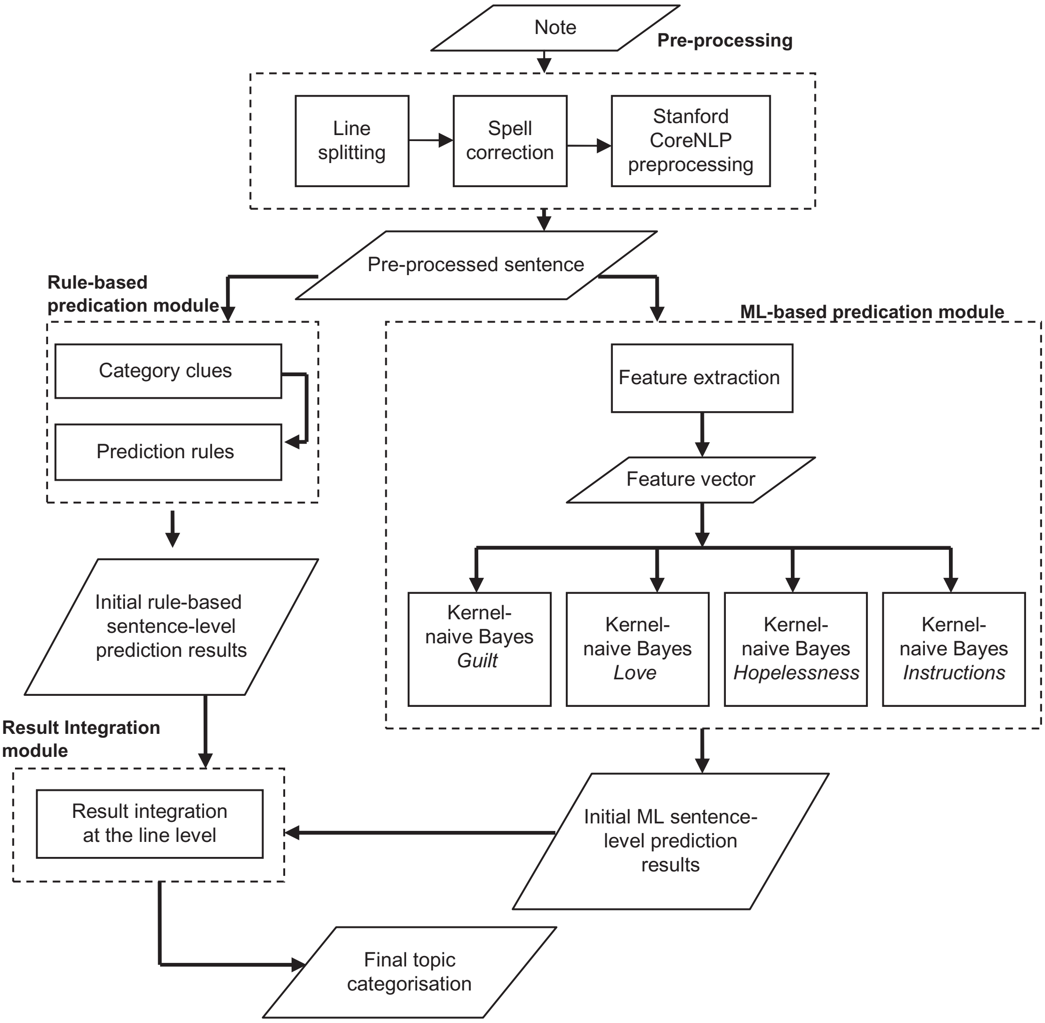
Note
Pre-processing
Line splitting
Spell correction
Stanford CoreNLP preprocessing
Pre-processed sentence
Rule-based predication module
Category clues
Prediction rules
ML-based predication module
Feature extraction
Feature vector
Kernel-naive Bayes *Guilt*
Kernel-naive Bayes *Love*
Kernel-naive Bayes *Hopelessness*
Kernel-naive Bayes *Instructions*
Initial rule-based sentence-level prediction results
Result Integration module
Result integration at the line level
Initial ML sentence-level prediction results
Final topic categorisation

Figure 1. System architecture.

associated with a given category (eg, '*I love you*' for the *Love* category). All rules (see Table 1 for details) were specified at the literal/lemma level and relied on associated category-specific dictionaries. These dictionaries include common clue terms or expressions (represented as regular expressions to account for variations) associated with particular syntactic patterns and/or expressions indicative of a specific category. These terms were primarily extracted from the training data through manual inspection of a subset of notes, and further extended with synonyms and related words from various online resources, including WordNet[14] and Thesaurus.[15] For example, strong clues for identification of sentences belonging to the *Information* class were prepositional phrases (PPs), specifically those expressing relations in space: for instance, a common indicator was a mention of a PP with prepositions *in* or *on* followed by a noun phrase whose head described a physical space (eg, *box, drawer, bag, pocket, car, trunk*, etc). The corresponding semantic sets (eg, for physical space) were defined for each category as a specific dictionary that contained associated nominal or verbal clues. For example, *Instructions* sentences often contained the word *please* followed by a verb in its infinitive form optionally followed by a proper noun or pronoun; associated verbs were defined by the *Instructions*-specific dictionary.

Another approach built into rules was to predict topic categories based on specific, typically "frozen" common laymen expressions. For example, *Hopelessness* statements were strongly indicated by expressions such as *I can' t take it anymore* or *I can' t stand it*. The rules were developed using the General Architecture for Text Engineering (GATE)[16] and the Java Annotation Pattern Engine (JAPE) grammar

Table 1. Number of rules and dictionary entries per category.

Topic category	Number of rules	Number of dictionary entries
Instructions	25	61
Hopelessness	21	40
Love	8	0
Guilt	9	14
Information	21	30
Blame	15	0
Thankfulness	6	13
Hopefulness	9	0
Sorrow	6	0
Anger	4	0
Happiness_ peacefulness	5	0
Fear	4	0
Pride	4	0
Forgiveness	1	0
Abuse	2	11

formalism. For each rule, the training dataset was used to assess its precision and recall. Precision values were used to rank rules within a given category. Categories associated with the rules that have fired for a given sentence were considered as initial rule-based predictions for that sentence.

C. ML-based predication module. The main objective of this module was to provide sentence-level predictions using a set of ML models with a pre-defined set of features. Given that the training data set had only five categories that had enough data in order to train ML models (*Guilt, Hopelessness, Instructions, Information* and *Love*), we created five ML models for these categories only. The models were designed to recognise sentences from a particular category and to classify all the other sentences in one opposing class (referred as *Other*). The features used were clustered into four groups:

1. **Lexical features** included a set of most significant uni-, bi- and tri-grams extracted for each of the five classes from the training dataset. Significance was measured using the *likelihood* measure and we have selected the top 500 features for each category. In addition, the features included the lemma of the first finite verb of the sentence, along with the tense assigned to the sentence (determined by a hand-crafted set of rules that relied on part of speech tags and shallow phrases). We also included the presence of negation cues as returned by *NegEx*.[17]
2. **Named-entity features** were used to indicate the presence of common named entities, such as personal names, addresses, dates, times etc. We used the Stanford CoreNLP suite to identify the following named-entity types: *person, location, date, time, money, organization, ideology, nationality, religion, title* and *misc*. As the notes were anonymised by mapping some of these entity type mentions to a fixed set of values (eg, '*Jane*, *John*, *3333 Burnett Ave.*', etc.), the extraction was also done by simple string matching. Additionally, task-specific semantic classes, for example those referring to a family member, financial information, disease etc. were considered. A careful analysis of sentences from the training corpus (grouped by categories) revealed that sentences from particular groups contained key terms that distinguished them from other categories: for example, sentences from the *Hopelessness* category often contained disease mentions, whereas *Instructions* and *Information* sentences contain financial terms ('*cash*', '*cheque*', '*business*' etc.). We also noticed that a number of sentences contained terms indicating sentiment ('*bad*', '*appalling*', etc.), family members ('*mom*', '*dad*', '*sister*' etc.), endearments ('*baby*', *honey*' *etc.*), body parts ('*arm*', '*chest*', etc.), ache (with many synonyms for pain) and pills ('*pill*', '*tablet*', '*capsule*' etc.). Based on our findings, the eight semantic lexicons (corresponding to the mentioned semantic types) were created manually (except for diseases) in a two step process. As a first step we collected all terms by searching through the training data and then in the second step we used the Internet and WordNet to find synonyms for the collected words. We did not perform any statistical tests in order to see if particular semantic classes of words could be useful for separating emotion labels; we rather left it to the ML models and feature selection methods to decide which ones were useful. The disease lexicon was automatically collected from the UMLS[18] and the Disease onotology.[19]
3. **Rule-based features** represented a set of binary features provided by the rule-based module (module (B) above). Each rule was represented by a corresponding feature that indicated if that

rule has fired in the sentence. We experimented by using the rules precision as feature values, but binary values (fired/not fired) showed better performance on the training data.

4. **Presentation features** included attributes that described some aspects of how the statement (in a given sentence) has been made. Previous work[8] indicated that such features could be useful for the classification of suicide notes. We considered readability of the sentence, the gender of the person who wrote the note (determined at the note level and then propagated to all sentences) and three features that represented the (local) line number in which the sentence appeared, the number of lines and the number of sentences in the note. In order to calculate the readability, we used the Flesch and Kincaid readability scores[20,21] as returned by the *Flesh* tool.[22] In order to extract the gender feature, a set of hand-crafted rules was used to identify greetings (eg, "signed John/Jane", "love John/Jane" etc.). In cases where there were no greetings found, we used the Koppel-Argamon algorithm[23,24] as implemented in *Lingua-EN-Gender*-1.0.[25]

Each of the five classifiers was built in RapidMiner[26] using the *Naive Bayes* model with *kernel density estimation*. Feature selection was performed by a genetic algorithm integrated into *RapidMiner* and with the *Fast Correlation-Based Filter* method.[27,28] In order to estimate the performance of the selected features, the five-fold cross-validation method was applied as standard.

D. Result integration module. The role of this module was to integrate the predictions made by the rule- and ML-based models in order to produce the final output for a given line. Note that both modules provided sentence-level predictions only, which were then combined. Therefore, the system could provide multi-label annotation both at the sentence and line level. We created three different workflows (runs) for the result integration as follows.

In **Run 1**, the goal was to optimise recall. For a given line, we therefore collected all topic categories returned by all the rules that fired for any of its sentences and all the predictions returned by applying the ML models to them. Based on empirical evidence from the training data, we decided not to use the ML model built for the *Information* category, given its relatively poor performance (ie, no improvements) compared to other "large" categories (see Table 4). Similarly, we decided not to use the rule-based features in the ML models, again based on the results of experimentation on the training data (the best recall was achieved when the rule-based features were omitted; data not shown).

In **Run 2**, we considered only predictions returned by the rule-based module. The goal for this run was to optimise precision. Final predictions for a given line included all categories returned by the rules from each sentence in that line.

In **Run 3** we used all the initial results from the rule-base module (as in Run 2) and a single best prediction from the four ML models (again, the *Information* category was omitted). In this run, the ML models included the rule-based features (as opposed to Run 1). Overall, we aimed at optimising the F-measure with this run. Predicted labels of the ML models were ranked by *prediction confidence* as provided by *RapidMiner*, and the label with the highest confidence was used. These values where comparable since all of the ML models were based on the same approach (*Naive Bayes*).

Results and Discussion

The task was evaluated on a test dataset containing another set of 300 suicide notes as provided by the organisers. The "gold" annotation topic categories were manually provided for each line by three annotators. The organisers estimated the inter-annotation agreement as 0.546 (using Krippendorff's alpha coefficient).[7]

The system performance was primarily estimated using the *micro*-averaged F-measure, which averages the results across all annotations (line-level). The test results of our system are given in Table 2. Run 1 gave the best results, with the highest F-measure (53.36%) and the highest recall (50.47%). As expected, the best precision (67.17%) was achieved in Run 2, with all predictions coming from the rule-based module. Run 3 was an attempt to compromise between the first two runs, as reflected by the results (but it failed to get the best F-score).

Category-specific results are given in Table 3. We note that the "large" categories (such as *Instructions, Hopelessness, Love* and *Guilt*) have reasonably high and comparable performance, with *Love* consistently showing the best results (F-measure of 67.34%). The exception is the *Information* category (F-measure of

Table 2. Micro-averaged results on the test data.

	P-micro	R-micro	F-micro
Run 1	0.5661	**0.5047**	**0.5336**
Run 2	**0.6717**	0.3797	0.4851
Run 3	0.5900	0.4764	0.5271

Abbreviations: P, precision; R, recall; F, F-measure.

29.73%), probably due to the very broad scope of this topic. The results for mid- and low-frequency categories relied on rules only, and typically showed poor performance, with notable exceptions of *Thankfulness* (F-measure of 72.53%) and *Happiness_ peacefulness* (F-measure of 53.85%). Still, the rules (Run 2) overall provided relatively high precision (67.17%).

Run 3 attempted to optimise F-measure, but the drop in recall was significant probably due to (1) excluding the less confident predictions from the ML models, and (2) using the ML models with rule-based features, which proved to increase precision but have the reverse effect on recall (data not shown). Table 3 also shows the *macro*-averaged results (averaged over topic categories), which were significantly lower than the micro-averaged ones, given that there were categories (eg, *Sorrow* and *Abuse*) with no correct predictions.

When compared to the results on the training dataset (see Table 4), there are drops in the overall micro F-measure of between 6.38 and 7.86 percentage points. There were differences in the performance drops for specific categories: while *Love* performed mostly consistently (drop of 3%–5%), performance for the *Information* category dropped between 14 and 19 percentage points, indicating again the wider scope of this category that has not been captured by rules or ML approaches (likely due to lexical variability and limitations of our topic dictionaries). There were also significant drops in performance for *Guilt*, in particular in the runs that included ML-based predictions, indicating again that the models have not generalised well (see Table 5 for FP and FN examples).

While the rules (Run 2) did not fail for some of the "large" categories (*Hopelessness, Love* and *Guilt*), there were significant drops for *Instructions* (a large category) and *Information* (a wide scope) when compared to the training data. As expected, the rules developed for the mid- and low-frequency categories in principle did not show consistent performance. Notable exceptions are *Thankfulness* (one of the "easiest" categories to predict) and *Happiness_ peacefulness*, both of which provided even better performance on the test dataset than on the training data.

We also note that the overall drop in precision for Run 2 (rules only) between the two datasets was

Table 3. Per-category performance on the test data.

Topic category	Frequency	Run 1			Run 2			Run 3		
		P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Instructions	382	0.5509	0.6649	0.6026	0.7076	0.4372	0.5405	0.5692	0.6571	0.6100
Hopelessness	229	0.5775	0.5371	0.5566	0.7154	0.3843	0.5000	0.5950	0.5197	0.5548
Love	201	**0.6802**	**0.6667**	**0.6734**	**0.7943**	**0.5572**	**0.6550**	**0.7832**	**0.5572**	**0.6512**
Guilt	117	0.4857	0.4359	0.4595	0.6230	0.3248	0.4270	0.5294	0.3846	0.4455
Information	104	0.5000	0.2115	0.2973	0.5000	0.2115	0.2973	0.5000	0.2115	0.2973
Blame	45	0.2381	0.1111	0.1515	0.2381	0.1111	0.1515	0.2381	0.1111	0.1515
Thankfulness	45	**0.7174**	**0.7333**	**0.7253**	**0.7174**	**0.7333**	**0.7253**	**0.7174**	**0.7333**	**0.7253**
Hopefulness	38	0.3077	0.1053	0.1569	0.3077	0.1053	0.1569	0.3077	0.1053	0.1569
Sorrow	34	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Anger	26	0.5000	0.0385	0.0714	0.5000	0.0385	0.0714	0.5000	0.0385	0.0714
Happiness_ peacefulness	16	**0.7000**	**0.4375**	**0.5385**	**0.7000**	**0.4375**	**0.5385**	**0.7000**	**0.4375**	**0.5385**
Fear	13	0.3636	0.3077	0.3333	0.3636	0.3077	0.3333	0.3636	0.3077	0.3333
Pride	9	0.6667	0.2222	0.3333	0.6667	0.2222	0.3333	0.6667	0.2222	0.3333
Forgiveness	8	**1.0000**	0.1250	0.2222	1.0000	0.1250	0.2222	1.0000	0.1250	0.2222
Abuse	5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Overall—macro F		0.4859	0.3064	0.3414	0.5223	0.2664	0.3301	0.4980	0.2940	0.3394

Note: Frequency represents the number of lines in the test dataset.
Abbreviations: P, precision; R, recall; F, F-measure.

Table 4. Per category performance on the training data.

Topic category	Frequency	Run 1			Run 2			Run 3		
		P	***R***	***F***	***P***	***R***	***F***	***P***	***R***	***F***
Instructions	820	0.6093	**0.7476**	0.6714	0.8120	0.5634	0.6652	**0.7347**	0.6280	0.6772
Hopelessness	455	0.6465	0.5626	0.6016	**0.8203**	0.3912	0.5298	0.6639	0.5253	0.5865
Love	296	0.6916	0.7196	**0.7053**	0.7946	**0.6014**	**0.6846**	0.7208	**0.6892**	**0.7047**
Guilt	208	0.5349	0.5529	0.5437	0.6333	0.3654	0.4634	0.5954	0.4952	0.5407
Information	295	**0.7521**	0.3085	0.4375	0.7561	0.3153	0.4450	0.4256	0.5525	0.4808
Blame	107	0.8108	0.2804	0.4167	0.8108	0.2804	0.4167	0.8108	0.2804	0.4167
Thankfulness	94	**0.8871**	**0.5851**	**0.7051**	**0.8750**	**0.5957**	**0.7089**	**0.8871**	**0.5851**	**0.7051**
Hopefulness	47	0.4595	0.3617	0.4048	0.4595	0.3617	0.4048	0.4595	0.3617	0.4048
Sorrow	51	0.4800	0.2353	0.3158	0.5000	0.2157	0.3014	0.4800	0.2353	0.3158
Anger	69	0.8571	0.0870	0.1579	0.8571	0.0870	0.1579	0.8571	0.0870	0.1579
Happiness_ peacefulness	25	**0.7273**	0.3200	0.4444	0.7273	0.3200	0.4444	0.7273	0.3200	0.4444
Fear	25	0.5000	0.4400	0.4681	0.5000	0.4400	0.4681	0.5000	0.4400	0.4681
Pride	15	0.5000	0.3333	0.4000	0.5000	0.3333	0.4000	0.5000	0.3333	0.4000
Forgiveness	6	**1.0000**	**0.6667**	**0.8000**	**1.0000**	**0.6667**	**0.8000**	**1.0000**	**0.6667**	**0.8000**
Abuse	9	0.5000	0.4444	0.4706	0.5000	0.4444	0.4706	0.5000	0.4444	0.4706
Overall—macro F		0.6066	0.4372	0.4923	0.7031	0.3988	0.4907	0.6003	0.4371	0.4944
Overall—micro F		0.6339	0.5686	0.5995	0.7727	0.44370	0.5637	0.6477	0.5432	0.5909

Note: Frequency represents the number of lines in the training dataset.
Abbreviations: P, precision; R, recall; F, F-measure.

significant and even larger than (expected) drop in recall, indicating some confusion between categories (eg, between *Instructions* and *Information*; see Table 6). In many cases, the difference between an *Instruction* and *Information* is very subtle and requires sophisticated processing (eg, '*you will find my body*'). *Information* additionally showed a high degree of lexical variability, which was difficult to "capture" with rules or with the ML models. *Instructions* did show more syntactic constraints, which resulted in reasonable performance overall.

Another example where the rule-based approach showed a significant drop in precision (from 81% to 24%) was the *Blame* category (see Table 7 for examples). An inherent limitation of our rule-based approach was reliance on topic-specific dictionaries mainly derived from the dataset. Our manual analysis for *Blame* did not come up with any specific lexical constraints, which made the rules less productive. In addition, a number of FP cases were due to confusion with *Guilt* (see tables 5 and 7 for some examples) as with *Information* and *Instructions*, the differences can be very subtle.

Tables 3 and 4 show that our approach could profile the *Thankfulness* and *Love* categories relatively well, whereas *Sorrow* and *Anger*, as well as *Abuse* proved to be challenging, with virtually no or very few correct predictions in the test dataset. In addition to the training data and examples being scarce for these categories (very few rules and basically no category-specific dictionary, see Table 1), it also seems that wider and deeper affective processing

Table 5. Examples of FPs and FNs for *Guilt.*

Example sentence	Predicted topic	Correct topic
Mary I'm sorry but it had to be done as I cant go on any longer.	*Guilt*	*Hopelessness*
To my beloved children, Please forgive me for taking this step.	*Guilt*	No annotation
I can not believe I have been so bad a husband as to merit this.	*Guilt*	*Blame*
I am no good to My family or myself either.	No annotation	*Guilt*
I caused you so much unhappiness and worry.	*Blame*	*Guilt*
I have sinned and must pay the penalty.	No annotation	*Guilt*
Love Jane Please forgive me all of you.	*Forgiveness*	*Guilt*

Table 6. Examples of confusion between *Instructions* and *Information.*

Example sentence	Predicted topic	Correct topic
I do not think that it will be necessary to phone him as I have written about the necessary things, Do n't hesitate to call him if necessary, reversing the charges	*Information*	*Instructions*
I think now that they will find my body up Burnet Ave. to the side of the road not to far.	*Instructions*	*Information*
This letter gives him authority to turn over to you and Mr. John Johnson, my attotney, free access to my room and personal effects.	*Instructions*	*Information*

is needed to identify the subtle lexical expression of grief, sadness, disappointment, anger etc. (see Table 8 for some examples). Of course, the task proved to be challenging even for human annotators (Krippendorff's alpha coefficient of 0.546), with many gold standard annotations that could be considered as questionable or at least inconsistent. This is particularly the case with muti-focal sentences, where many labels seems to be missing (for example, '*My mind seems to have goen a blank, Forgive me. I love you all. so much.*' is not labelled as *Love*; '*(signed) John My wisfe is Mary Jane Johnson 3333 Burnet Ave. Cincinnati, Ohio OH*-636-2051 *Call her first*' was annotated only as *Instructions*, but not as *Information*).

In the current approach, we did not try to split individual multi-focal sentences apart and process the parts individually (of course, all sentences in a given line were processed separately). Instead, we hypothesised that we could collect the results from each of the separate ML models and all of the triggered rules at the sentence level, and thus produce multi-label annotations (both at the sentence and consequently at the line level). For example, the sentence '*Wonderful woman, I love you but can't take this any longer*.' triggered two rules (one for *Love* and one for *Hopelessness*); the ML models for those two classes also gave positive predictions, while the other two ML models predicted the *Other* label. This resulted in the final prediction for the sentence consisted of both *Love* and *Hopelessness* labels. Still, future work may explore if splitting multi-focal sentences would provide better precision, given that some weak evidence in separate parts of the multifocal sentence could be combined by an ML model to provide (incorrect) higher confidence and thus result in an FP. However, the experiments on both the training and testing data have shown that there was no "over-generation" of labels. The rules were built to have high precision, so in most cases only one rule fired per sentence and cases with more then two fired rules were very rare. An analysis of the ML results revealed that in the majority of cases only one of the four ML models predicted their respective categories for a given sentence. Cases where more than one ML predictions were made seem to be related to multi-focal sentences, and our best results were achieved with all ML predictions taken into account (run 1).

Conclusion

Identification of topics expressed in suicide notes proved to be a challenging task for both manual and automated analyses. Our approach to the prediction of

Table 7. Example FPs and FNs for the *Blame* category.

Example sentence	Predicted topic	Correct topic(s)
Perhaps so, I told you two years ago that they were driving me that way	No annotation	*Blame*
Mom, you should have known what was about to happen after I told you my troubles now I will get my rest	*Guilt*	*Blame*
I hope the people who made me do this Go to Hell Jane	*Anger*	*Blame, Anger*
I might have killed you if you had been around me in the last 10 months	*Blame*	*Anger*
Dear John I 'm all too sorry I have caused you all the trouble I have	*Guilt*	*Blame*
Jane—Forgive me for all the misery I have caused you	*Forgiveness*	*Blame*

Table 8. Example of FPs and FNs for the *Sorrow* and *Anger* categories.

Example sentence	Predicted topic	Correct topic(s)
Dear Mom, I 'm sorry for all the Bad things I 've done in the past.	*Guilt*	*Sorrow*
I am ill and heart broken	*Hopelessness*	*Sorrow*
I grieve that I could not have had the juy of being close to our babies, but that is no one 's fault	No annotation	*Sorrow*
Jane dear I 'm sorry that I have been making you unhappy—I 'm all twisted up inside	*Sorrow*	*Guilt, Hopelessness*
You left me and did not say anything So darling this is your divorce my darling wife	No annotation	*Anger*
Hopeing that Father inlow will be very happy for everything now	No annotation	*Anger*
At one thirty in the morning I was awaking by a Nurse you know how I hate them damn things	*Anger*	No annotation

topic categories relied on combining hand-crafted rules (which included both lexical, syntactic and lexico-semantic components) and various features used in the ML models (which included lexical, lexico-semantic and presentation features, and named entities and rules that were linked to corresponding sentences). The results showed reasonable performance for frequent and relatively well-scoped topics (eg, *Thankfulness, Love, Instructions*), whereas infrequent and non-focused categories (eg, *Sorrow, Anger, Blame, Information*) proved to be challenging. Future work will need to be informed by a detailed error analysis and in particular further investigations in prediction confusions between various topic categories. The effects of particular features (eg, presentation, named entities, etc.) on performance will also need to be further explored. Still, the current approach not only indicates the limits of the component technologies, but also demonstrates the potentials of combining or selecting different approaches for different topic categories.

Acknowledgement

This work was partially supported by a PhD scholarship from the School of Computer Science, University of Manchester and EPSRC (to AD) and the Serbian Ministry of Education and Science (projects III47003, III44006, to AK, GN).

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. Bollen J, Mao H, Zeng XJ. Twitter mood predicts the stock market. *Journal of Computational Science*. 2011;2(1):1–8.
2. Vukadinovic-Greetham D, Hurling R, Osborne G, Linley A. Social networks and positive and negative affect. 7th International Conference on Applications of Social Networks Analysis. *ASNA*. 2010.
3. Wu S, Tan C, Kleinberg J, Macy M. Does Bad News Go Away Faster? Proc. 5th International AAAI Conference on Weblogs and Social Media 2011.
4. Golder S, Macy MW. Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength across Diverse Cultures. *Science*. 333(6051):1878–81, doi: 10.1126/science.1202775.
5. Hjelmeland H, Knize BL. Why we need qualitative research in suicidology. *Suicide and Life-Threatening Behavior*. 2010:40(1).
6. Pestian JP, Matykiewicz P, Grupp-Phelan J. Using natural language processing to classify suicide notes. In Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing. *BioNLP*. 08:96–97, ACL.
7. Pestian JP, Matykiewicz P, Linn GM, Wiebe J, Cohen K, Brew C, et al. Sentiment Analysis of Suicide Notes: A Shared Task. *J Biomed Informatics Insight*. 2012;5 (Suppl. 1):3–16.
8. Pestian JP, Nasrallah H, Matykiewicz P, Bennett A, Leenaars A. Sentiment Suicide Note Classification Using Natural Language Processing: A Content Analysis. *J Biomed Informatics Insight*. 2010;(3):19–28.
9. Mohammad S, Yang T. Tracking Sentiment in Mail: How Genders Differ on Emotional Axes, Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA), ACL HLT 2011, http://aclweb.org/anthology-new/W/W11/W11-17.pdf
10. http://code.google.com/p/google-api-spelling-java/.
11. Toutanova K, Klein D, Manning C, Singer Y. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. Proceedings of the HLT-NAACL;Edmonton, Canada. 2003;173–80.
12. Finkel JR, Grenager T, Manning C. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the ACL; Ann Arbor, USA. 2005: 363–70.
13. Klein D, Manning C. Accurate Unlexicalized Parsing. Proceedings of the 41st Annual Meeting of the ACL; Sapporo, Japan. 2003:423–30.

14. http://wordnetweb.princeton.edu/perl/webwn.
15. http://thesaurus.com/.
16. http://www.gate.ac.uk.
17. Chapman W, Bridewell W, Hanbury P, Cooper G, Buchanan B. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform*. 2001;34(5):301–10.
18. http://www.nlm.nih.gov/research/umls/.
19. http://do-wiki.nubic.northwestern.edu/index.php/Main_Page.
20. Flesch R. A new readability yardstick. *Journal of Applied Psychology*. 1948;32(3):221–33.
21. Kincaid J, Fisburne R, Rogers R, Chissom B. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. Memphis, TN: Naval Technical Training, US Naval Air Station; 1975.
22. http://flesh.sourceforge.net/.
23. Koppel M, Argamon S, Shimoni A. Automatically categorizing written texts by author gender. *Lit Linguistic Computing*. 2002;17(4):401–12.
24. Argamon S, Koppel M, Fine J, Shimoni A. Gender, Genre, and Writing Style in Formal Written Texts. *Text*. 2003;23(3):321–46.
25. http://search.cpan.org/~eekim/Lingua-EN-Gender-1.0/Lingua/EN/Gender.pm.
26. Mierswa I, Wurst M, Klinkenberg R, Scholz M, Euler T. YALE. Rapid Prototyping for Complex Data Mining Tasks. Proceedings of the 12th ACM SIGKDD International Conference on KDD; Philadelphia, USA. 2006: 935–40.
27. Yu L, Liu H. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. Proceedings of the Twentieth International Conference on Machine Leaning; Washington, DC, USA. 2003:856–63.
28. http://www.public.asu.edu/~huanliu/FCBF/FCBFsoftware.html.