Automatic Identification of Suicide Notes from Linguistic and Sentiment Features

Annika Marie Schoene and Nina Dethlefs
University of Hull, UK
amschoene@googlemail.com

Abstract

Psychological studies have shown that our state of mind can manifest itself in the linguistic features we use to communicate. Recent statistics in suicide prevention show that young people are increasingly posting their last words online. In this paper, we investigate whether it is possible to automatically identify suicide notes and discern them from other types of online discourse based on analysis of sentiments and linguistic features. Using supervised learning, we show that our model achieves an accuracy of 86.6%, outperforming previous work on a similar task by over 4%.

1 Introduction

The World Health Organisation outline in a recent report that suicide is the second leading cause of death for people aged 15-29 worldwide (2014). The total number of suicides per year is around 800,000 people. In recent years, there has been a trend recognized that especially young people tend to publish their suicide notes or express their suicidal feelings online (Desmet and Hoste, 2013).

Research in psychology has long recognised that our drive or motivation can affect the way in which we communicate, leading to the assumption that our spoken and written language represents those shifting psychological states (Osgood, 1960). This argument was elaborated by Cummings and Renshaw (1979), who suggest that there is a shift in people's linguistic expression due to the aroused cognitive state suicidal individuals experience.

Facebook has recently developed an online feature which relies on users reporting other users if they feel that they are at risk of committing suicide (Morese, 2016). New features such as the Facebook feature are undoubtedly important in suicide prevention as suicide is not only a result of mental health issues, but of various sociocultural factors and especially individual crisis (Worldwide, 2016). Therefore it has been argued by Desmet and Hoste (2013) that there is a need for automatic procedures that can spot suicidal messages and allow stakeholders to quickly react to online suicidal behaviour or incitement. This paper aims to investigate the linguistic features in discourse that are representative of a suicidal state of mind and automatically identify them based on supervised classification.

2 Related Work

Traditionally, the linguistic analysis of suicide notes has been conducted in the field of forensic linguistics in order to provide evidence for the genuineness of suicide notes in settings such as police investigations, court cases or coroner inquiries, where expert evidence is given by professionals, such as forensic linguists (Coulthard and Johnson, 2007).

As argued above, there is great impact potential in the automatic identification of suicide notes, e.g. on social media sites, in order to prevent such cases. Previous work in this direction by Jones and Benell (2007) developed a supervised classification model based on linguistic features that can differentiate genuine from forged suicide notes. The authors found that structural features such as nouns, adjectives or average sentence length were reliable predictors of genuine notes and report an overall classification accuracy of 82%.

An alternative direction was taken in research by Pestian et al. (2010) and Pestian et al. (2012), who investigate the impact of sentiment features on the identification of suicide notes. The authors

Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), pages 128–133,
Berlin, Germany, August 11, 2016.

focus particularly on those emotion features that have been shown to play a role in the clinical assessment of a person (Pestian et al., 2010).

Most work to date has focused on the identification of genuine suicide notes against forged ones. Also, different types of features have been shown to be useful in this task. In this paper, we explore the impact of combining these features into a model that can differentiate suicide notes from other types of discourse, such as depressive notes or love letters–which share several linguistic features with genuine suicide notes.

3 Data Collection and Annotation

3.1 Corpora

We use three datasets for our analysis: a corpus of genuine suicide notes and two corpora for comparison. The latter two were collected from public posts made to the Experience Project website.[1]

- **Genuine Suicide Notes (GSN)**: this corpus contains genuine suicide notes which we collected from various sources, including newspaper articles and already existing corpora from other academic resources, e.g. Shneidman and Farberow (1957), Leenaars (1988) and Etkind (1997). Only notes of which there was a full copy available were included.
- **Love / happiness (LH)**: this corpus comprises 142 posts from the Experience Project's public groups 'I Think Being In Love Is One Of The Best Feelings Ever' and 'I Smile When I Think Of You'. We chose this topic as it could have interesting linguistic similarities with suicide notes in its use of cognitive verbs. However, there are also important differences as emotions are expected to be largely positive. Posts were collected randomly with an equal number of men and women to a keep a demographic balance.
- **Depression / loneliness (DL)**: the DL corpus was collected as it may be close in the emotions and language usage to the GSN corpus and could therefore demonstrate clear differences in how depressed and suicidal people communicate. This corpus was collected randomly from the Experience Project's group 'I Fight Depression And Loneliness Everyday'.

[1] http://www.experienceproject.com/

All corpora were collected from the public domain, but nevertheless anonymised in order to protect the privacy of the author as well as the privacy of those referenced in the communication. The other two corpora were chosen as both differ significantly in topic, purpose and arguably emotions. Similar research in this area has been conducted by Bak et al. (2014), who investigated how self-disclosure is used in twitter conversations, where self-disclosure is used as a means of gathering social support as well as "to improve and maintain relationships". Although it could be argued that suicide notes are a form of self-disclosure the purpose of a suicide note is different to the one mentioned by Bak et al (2014). The purpose of a suicide note is manifold and can range from statements of their current feelings, apologies or instructions, but not all suicide notes are written in order to comfort the survivors (Wertheimer, 2001). Therefore the level or type of self- disclosure may be another interesting feature to be included into a further analysis.

3.2 Features

All three corpora were annotated manually and on clause level by one author including the following features based on previous work discussed in Section 2.

Sentiment Features In terms of sentiment features, we annotated the following 12 emotions on a clause level: *fear, guilt, hopelessness, sorrow, information, instruction, forgiveness (fg), happiness/peacefulness (hp), hopefulness, pride, love* and *thankfulness*. Feature values were the number of occurrences of each emotion in a note, e.g. `sorrow=2`. These emotions are based on the work of Pestian et al. (2012), who uses 'abuse', 'anger' and 'blame' in addition. However, there were too few examples of these in our data, so that we excluded them from our analysis. Furthermore Yang et al. (2012) showed that assigning the emotions to a positive, neutral and negative group can improve classification accuracy. We therefore include grouping of emotions as well, again representing them by their number of occurrence, e.g. `positive=4`. The concepts 'information' and 'instruction' were assigned to the neutral group.

Some clauses can contain more than one emotion, so annotation features were not always mutually exclusive. In such cases, we chose to annotate the most prominent emotion. For example, in the

Category	GSN	LH	DL
No. of tokens in corpus	20,534	10,051	17,161
No. of notes in corpus	142	142	142
Ave. no. of words in note	141	71	121
No. of clauses in corpus	1,305	787	1,135
Ave. clause length	15	12	15

Table 1: Quantitative comparison of corpora collected in terms of number of words of each corpus, number of documents notes, average number of words in each note, clauses in each corpus and average clause length.

clause "i know that i will die dont be mean with me please" [sic] both *instruction* and *forgiveness* are possible, but only the first emotion was annotated as it appeared to be the prominent one.

Linguistic Features In terms of linguistic features, we used Python's Natural Language Toolkit (NLTK) (Loper and Bird, 2002) to extract POS tag information, the most frequent lexical items, 2-grams and 4-grams. In addition, we used the LIWC tool [2] to extract note length, cognitive processes, tenses (past, present, future), average sentence length, relativity, negation, signs (e.g. +, &), adverbs, adjectives, and verbs. Finally, we manually annotated the feature 'endearment', which referred to words such as 'Dear' at the beginning of a note or post. Work by Gregory (1999) previously established a significant influence of 'endearment'.

Corpora Statistics Table 1 shows a quantitative comparison of our three corpora showing the number of words of each corpus, number of documents notes, average number of words in each note, clauses in each corpus and average clause length. We can see that while each corpus contained exactly the same number of documents/notes, other statistics such as the number of words or clauses vary substantially across corpora.

Previous work by Gregory (1999) can perhaps help shed some light on these differences. Gregory (1999) found that suicide notes are often greater in length due to the fact that the suicidal individual wants to convey as much information as possible. This is due to the note writer's feeling that they will not have time to convey this information at a later point (Gregory, 1999). In our corpora, we can see this tendency clearly reflected in the overall lengths of notes. In addition, the corpora differ noticeably in the average length per note. The notes in the GSN corpus are almost double in length compared to the LH corpus.

It can be seen in Table 1 that more similarities are found between the suicidal GSN corpus and the depressive DL corpus that with the love corpus LH. A possible explanation for this is work by Alvarez (1971) who explains that it is known in a clinical setting that there is a similarity between the state of mind of a suicidal person, and a person who experiences depression. When comparing the LH corpus to the other two corpora it is clear that although the number of tokens in the corpus is smaller, the sentence length is almost as high as the one of the GSN and DL corpora. It could be argued that this phenomenon may be due to a higher amount of adjectives used in a sentence, which will be tested at a later point. In addition to this, it has been argued that people who communicate under stress tend to break their communication down into shorter units (Osgood, 1959), thus perhaps pointing to a higher stress level of the suicidal individuals. The research however suggested that there is no significant difference in the overall length per unit when comparing suicide notes to regular letters to friends and simulated suicide notes (Osgood, 1959).

4 Classification Experiments

We use the WEKA toolkit (Hall et al., 2009) for our supervised learning experiments. Table 2 shows an overview of the models compared: a logistic tree regressor (LMT), a J48 decision tree classifier, a Naive Bayes classifier, and a simple majority baseline (Zero-R). All models were trained using 10-fold cross-validation in order to minimize variability in results (Alpaydin, 2012). The results are shown in Table 2, with the first box in the table including both sentiment and linguistic features, the second box only including sentiment features and the last box including only linguistic features. As can be seen, the best performance is achieved by a combination of sentiment and linguistic features by an LMT tree regressor with an overall accuracy of 86.61%. The following regression equation was learnt for a suicide note:

[2] http://liwc.wpengine.com/compare-versions/

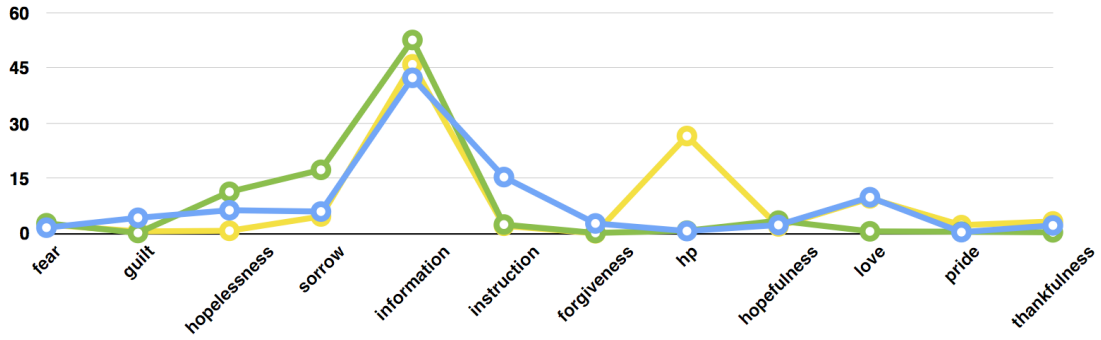

Figure 1: Different sentiments found in the GSN corpus (blue), the LH corpus (yellow) and the DL corpus (green). Emotions are in line with those expected by previous psychological studies. Emotion *hp* refers to happiness/peacefulness.

Classifier	ACC	PRE	REC	F-Score
LMT	**86.61**	**0.86**	**0.86**	**0.86**
J48	78.87	0.78	0.78	0.78
Naive Bayes	74.17	0.76	0.74	0.74
Zero-R	32.86	0.21	0.32	0.23
LMT	**78.63**	**0.79**	**0.78**	**0.78**
J48	71.36	0.71	0.71	0.71
Naive Bayes	69.01	0.72	0.69	0.68
Zero-R	32.86	0.21	0.32	0.23
LMT	**75.35**	**0.75**	**0.75**	**0.75**
J48	67.60	0.67	0.67	0.67
Naive Bayes	65.96	0.67	0.66	0.65
Zero-R	32.86	0.21	0.32	0.23

Table 2: Classification accuracy, precision, recall and F-Score metrics for different Weka classifiers. The first set of results includes *all features*, the second set is based on *sentiment features* only, and the last set of based on *linguistic features* only.

$$1.98 + [fear] * -0.55 + [guilt] * 0.76 + [sorrow] * -0.13 + [instruction] * 0.66 + [fg] * 2.28 + [endearment] * 2.95 + [signs] * 1.51 + [cognitive] * -0.11 + [relativity] * -0.05 + [negations] * -0.1 + [adverb] * -0.11 + [noun] * 0.01$$

Our results exceed previously reported results on the (slightly different) task of classifying genuine suicide notes against forged ones by Jones and Benell (2007), who achieved 82%.

4.1 Discussion of Sentiment Features

Apart from the overall classification accuracy, we were interested in the contribution of the individual sentiment and linguistic feature sets. To this end, we conducted a sentiment analysis in order to identify which emotions are present in the three corpora and which proved to be most significant (Figure 1). 'Information' is the most frequent in all three corpora. This may be due to the fact that the clauses labelled as 'information' are mainly descriptive and inform the reader of things such as where a specific item is placed or give instructions (Yang et al., 2012). Examples are "I know it is going to hard with William and Sister." (information) or "Please see that Charles gets a Mickey Mouse Watch for his birthday." (instruction).

Furthermore, the results of the **GSN corpus** correspond to the findings of Lester and Leenaars (1988), who argue that there is a high likelihood that a person leaves instructions behind for the survivors. Also, Foster (2003) found that 60% of people convey their love for those who they leave behind in a suicide note, which would explain why the emotional concept of 'love' is so prominent. A further observation is that certain emotions occur with a higher percentage in the GSN corpus and less or not at all in the other two. This can be explained by the higher degree of confusion that Leenaars (1988) found in the emotions in suicide notes compared to other types of discourse. Our LMT model confirms this—5 different emotions are used in the regression equation, more than for the other two datasets (see below).

In relation to the **LH corpus**, Ben-Ze'ev (2004) argued that the emotions 'happiness' and 'love' are closely related to each other because sharing activities with a loved one can generate happiness on both sides. Therefore it is not surpris-

ing that besides the feature 'information', 'love' and 'happiness' are the two most predictive emotions in the LH corpus. Since people who wrote notes in the LH corpus are happily in love, the need for expressing negative emotions is reduced in this group. The LMT model identified the presence of 'happiness/peacefulness' and the absence of 'hopelessness' as the most important predictors.

Regarding the **DH corpus**, primary emotional concepts are 'hopelessness', 'sorrow' as well as 'anxiety'. These match the emotions that the Mental Health Foundation describe on their website[3] as typical feelings people experience when suffering from depression. We can argue that overall the emotions identified in the individual corpora match those that we expected based on previous research and psychological studies. Based on the LMT model, the presence of 'sorrow' was the most important predictor with 'hopelessness' and 'fear' also playing a role.

4.2 Discussion of Linguistic Features

Linguistic features which improved the classification accuracy substantially were the length of a note, number of verbs and nouns as well as the features endearment, and relativity.

Gregory (1999) argues that suicide notes are greater in length due to the fact that the author wants to convey as much information as possible, due to their feeling that they will not have time to convey this information at a later point. This proved to be true for the three corpora analysed as the average length of the GSN corpus (144.6 words) was substantially higher than the other two (LH= 70.78 words, DL= 120.85).

Gregory (1999) further found that suicidal individuals use more nouns and verbs in their notes. This was confirmed by Jones and Benell (2007), who explain that a person who is going to commit suicide is under a higher drive and therefore more likely to refer to a large amount of objects (nouns). Our LMT model identified the number of nouns and verbs as a significant predictor.

Previous work by Ogilvie et al. (1966) identified a high frequency of emotional endearment in genuine suicide notes, which was confirmed by our analysis. Interestingly, in our LMT model the feature 'endearment' is important both for suicide notes (in its presence) and for depressed notes (in its absence), thereby representing one of the most important contrasts between these (in many ways similar) datasets.

A further predictor identified by our LMT model was 'signs', e.g. the use of '+' or '&' instead of 'and'. Previous research by Wang et al. (2012) also identified this tendency, but Wang et al. (2012) excluded the feature, applying automatic spelling correction to increase accuracy. We argue that the feature might be important in relation to Osgood and Walker's argument (1959) that spelling or punctuation errors can be a direct result of the drive that suicidal people experience. This is particularly noteworthy since the feature harldy occurs in the LH and DL corpora.

Finally, 'relativity' refers to references to space, motion and time in a note. Handelman and Lester (2007) found fewer references made to inclusive space made in suicide notes. Again, we confirm this with the lowest relativity in the GSN corpus and the higher in the LH corpus.

5 Conclusion and Future Work

The automatic identification of suicide notes is an important research direction due to its potential for suicide prevention. In this paper, we have demonstrated that using a combination of sentiment analysis and linguistic features, it is possible to learn a model of the emotions and linguistic features that are representative of suicide notes, and tell them apart from other types of discourse, such as depressive notes or love notes. Our study can be seen as an initial investigation, which comes with some limitations and could lead to a number of future research directions.

A potential limitation of our study is that the notes included in our GSN corpus were written at various points in time, which means that some of the notes are as old as 60 years. The posts collected from the Experience Project are all drawn from an online community, so that a comparison with online suicide notes would be appropriate to investigate whether language change affects the linguistic features characteristic of recent notes.

[3] https://www.mentalhealth.org.uk/a-to-z/d/depression

References

E. Alpaydin. 2012. *Introduction to Machine Learning*. MIT Press, Cambridge, Massachusetts, second edition.

A. Alvarez. 1971. *The savage God: A study of suicide*. Norton, New York.

J. Bak, C. Y. Lin, and A. H. Oh. 2014. Self-disclosure topic model for classifying and analyzing Twitter conversations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar.

A. Ben-Ze'ev. 2004. *Love Online Emotions on the Internet*. Cambridge University Press, Cambridge.

M. Coulthard and A. Johnson. 2007. *An Introduction to FORENSIC LINGUISTICS: Language in Evidence*. Routledge, Abington.

H. Cummings and S. Renshaw. 1979. SLCA - 3: A meta theoretical approach to the study of language. *Human Communication Research*, 5:291–300.

B. Desmet and V. Hoste. 2013. Emotion detection in suicide notes. *Expert Systems with Applications*, 40:6351–6358.

M. Etkind. 1997. *A Collection of Suicide Notes*. The Berkeley Publishing Group, New York.

T. Foster. 2003. Suicide note themes and suicide prevention. *International Journal of Psychiatry in Medicine*, 33:323–331.

A. Gregory. 1999. The decision to die: The psychology of the suicide note. In D.Canter and L.Alison, editors, *Interviewing and deception*. Aldershot, Ashgate, UK.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November.

L. Handelman and D. Lester. 2007. The content of suicide notes from attempters and completers. *Crisis*, 28:102–104.

N.J. Jones and C. Benell. 2007. The development and validation of statistical prediction rules for discriminating between genuine and simulated suicide notes. *Archives of Suicide Research*, 11:219–233.

A. Leenaars. 1988. *Suicide Notes*. Human Sciences Press, New York.

D. Lester and A.A. Leenaars. 1988. The moral justification of suicide in suicide notes. *Psychological Reports*, 63:106.

Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70.

F. Morese. 2016. Facebook adds new suicide prevention tool in the UK. http://www.bbc.co.uk/newsbeat/article/35608276/facebook-adds-new-suicide-prevention-tool-in-the-uk.

D. Ogilvie, P. Stone, and E. Shneidman. 1966. Some characteristics of genuine vs. simulated suicide notes. In D. C. Dunphy, D. M. Ogilvie, M. S. Smith, and P. J. Stone, editors, *The general inquirer: A computer approach to content analysis*. MIT Press, Cambridge, MA.

World Health Organisation. 2014. First WHO Suicide Report. `http://www.who.int/mental_health/suicide-prevention/en/`.

E.G. Osgood, C.E. and Walker. 1959. Motivation and language behaviour: A content analysis of suicide notes. *Journal of Abnormal Psychology*, 59:58–67.

C. Osgood. 1960. The cross-cultural generality of visual- verbal synesthetic tendencies. *Behavioural Sciences*, 5:146–169.

J. Pestian, H. Nasrallah, P. Matykiewicz, A. Bennet, and A. Leenaars. 2010. Suicide Note Classification Using Natural Language Processing: A Content Analysis. *Biomedical Informatics Insights*, 3:19–28.

J. Pestian, P. Matykiewicz, and M. Linn-Gust. 2012. What's in a note: Construction of a suicide note corpus. *Biomedical Informatics Insights*, 5:1–6.

E. Shneidman and N. Farberow. 1957. *Clues to suicide*. McGraw-Hill Book Company Inc., New York.

W. Wang, L. Chen, M. Tan, S. Wang, and A. Sheth. 2012. Discovering Fine- grained Sentiment in Suicide Notes. *Biomedical Informatics Insights*, 1:137–145.

A. Wertheimer. 2001. *A Special Scar: The Experiences of People Bereaved by Suicide*. Routledge, London, 2nd edition.

Befrienders Worldwide. 2016. Suicide statistics. http://www.befrienders.org/suicide-statistics.

H. Yang, A. Willis, A. De Roeck, and B. Nuesibeh. 2012. A hybrid model for automatic emotion recognition in suicide notes. *Biomedical Informatics Insights*, 5:17–30.