The Role of Features and Context on Suicide Ideation Detection

Yufei Wang*
School of ITEE
The University of Queensland
Brisbane, Australia
Yufei.Wang1@uq.net.au

Stephen Wan and Cécile Paris
CSIRO Data61
Sydney, Australia
firstname.lastname@data61.csiro.au

Abstract

There is a growing body of work studying suicide ideation, expressions of intentions to kill oneself, on social media. We explore the problem of detecting such ideation on Twitter, focusing on the impact of a set of features drawn from the literature and on the role of discussion context for this task. Our experiments show a significant improvement upon the previously published results for the O'Dea et al. (2015) dataset on suicide ideation. Interestingly, we found that stylistic features helped while social media metadata features did not. Furthermore, discussion context was useful. To further understand the contributions of these different features and of discussion context, we present a discussion of our experiments in varying the feature representations, and examining their effects on suicide ideation detection on Twitter.

1 Introduction

According to World Health Organisation, a suicide occurs every 40 seconds worldwide (WHO, 2014). Suicidal death has destructive effect on both family (Cerel et al., 2008) and community (Levine, 2008) level. Tragically, many suicide cases can be prevented (Bailey et al., 2011). As social media platforms, such as Twitter[1], are often used as channels to discuss mental health topics, there is a need for new technologies to deliver online mental health support (Daine et al., 2013). Such services may be particular important for the youth, well represented on social media, for whom suicide is the second leading cause of death (WHO, 2014).

Consequently, there is a growing body of work that studies suicide ideation, expressions of intentions to kill oneself, on platforms such as Twitter. For example, O'Dea et al. (2015) describe a data set of Twitter posts that has been annotated by mental health and social media experts for (i) the presence of suicide ideation, and (ii) the level of severity of the ideation. In that text classification work, lexical features alone were used. However, intuitively, one might expect that information, such as the discussion context, might each provide valuable information to detect cases of suicide ideation.

For example, information from the surrounding discussion context, perhaps by friends, might indicate the presence of genuine suicide ideation. Two examples, Post A and Post B and their respective replies, are shown below.[2]

Post-A: *Okay goodbye, im going to kill myself tomorrow @ the retreat thing.*
Reply-A: *@ANON No plz dont.*

Post-B: *Listening to ultra live stream rn in ANON's car da gonna kill myself*
Reply-B: *@ANON I was watching it at work!!*

Although both cases contain the key phrase "kill myself", the replies indicate that Post-A is a more concerning post than Post-B, as the respondent answers sympathetically and supportively. However, the reply to Post-B focuses on the topic of the "live stream", seemingly dismissing the phrase "kill myself" as a colloquialism.

In this paper, we describe our exploration of these different feature sets for suicide ideation detection. We perform this study using the data set of O'Dea et al. (2015) as it contains annotations of suicide ideation and also of the severity of that ideation. That is, it also includes cases of non-genuine suicide ideation (based on uses of

*This work was performed while Yufei Wang was on a CSIRO Student Vacation Scholarship, 2015-2016

[1] www.twitter.com

[2] Examples have been modified to remove Twitter handles.

Yufei Wang, Stephen Wan and Cecile Paris . 2016. The Role of Features and Context on Suicide Ideation Detection. In *Proceedings of Australasian Language Technology Association Workshop*, pages 105−113.

the word "suicide" for metaphorical or humorous purposes). In addition, the data set also includes metadata for each Twitter post and the discussion context following each annotated post.

Our contributions are as follows.

1. We improve on the results published in O'Dea et al. (2015);
2. We describe a unified feature set drawn from the literature of mental health and suicide ideation analytics; and
3. We present a novel analysis on the impact of discussion level features for suicide ideation detection on Twitter.

Interestingly, we find that the literature-inspired feature sets only marginally improved upon the classification results. Specifically, for this work, stylistic features helped but social media features did not. Furthermore, discussion context was useful but only provided a small gain in performance. This is a surprising outcome, and so we investigate the roles of these features and of the discussion context further.

In the remainder of this paper, we describe the O'Dea et al. (2015) dataset and the previously published results in Section 2. We survey the related work from which our feature set was inspired in Section 3. Section 4 outlines the stylistic and social media metadata features used in this work, as well as providing an analysis about the contributions of these feature types. We examine the role of discussion context in Section 5. Finally, we present concluding remarks in Section 6.

2 The O'Dea et al. (2015) Dataset and Classification Results

In this work, we base our study of features relevant in suicide ideation detection on an existing Twitter dataset that contains judgements on the severity of the suicide ideation and also a rich collection of supplementary data for the post in question, such as the following discussion and the Twitter metadata (O'Dea et al., 2015). In this section, we will briefly describe the dataset, along with the machine learning features and algorithm used to obtain published performance results.

2.1 The Dataset

Twitter data was collected by O'Dea et al. (2015) using queries based on words relating to general English words about suicide ideation (Jashinsky et al., 2014), such as: suicidal; suicide; kill myself; my suicide note; never wake up; better off dead; suicide plan; tired of living; die alone; go to sleep forever.

Attribute	SI	PC	SC	All
Num. Twitter posts	534	1029	258	1821
Num. Unique words	2545	3016	694	4750
Avg. Num. words	17.5	14.9	10.9	15.1

Table 1: Descriptive summary statistics about each class label.

Of these, 2000 Twitter posts occurring between February and April 2014 were randomly sampled and annotated using three categories of severity listed here from least to most severe: "Safe to Ignore"(SI), "Possibly Concerning"(PC) and "Strongly Concerning"(SC) according to their suicide risk (O'Dea et al., 2015). Table 1 presents summary statistics about each class.

2.2 Prior classification results

The best performing system found by O'Dea et al. (2015) was a Support Vector Machine (SVM) (Joachims, 1999) with a feature set of unigrams weighted by TF-IDF scores. For these features, casing was ignored. To focus on the impact of using different feature types, we continue using SVM as the classifier and TF-IDF for lower-cased unigram features.

We successfully replicated the previous result reported by O'Dea et al. (2015), built using the Python Scikit-learn package[3]. We achieved a 10-fold cross-validation accuracy of 66% that is slightly better than the reported result of 63% in O'Dea et al. (2015).

We suspect this difference is due to variations in the text preprocessing. We thus experimented with different text preprocessing variants for n-gram lexical features. These are as follows:

- **N-gram** We extended the feature set to include uni-, bi- and tri-gram, where longer n-grams potentially captures phasal information.
- **Text Preprocessing** We tokenised the text using the Twokenize tool from Carnegie Melon University (CMU), which provides a

[3]http://scikit-learn.org/stable/index.html

Features	Accuracy	Macro-F1 (p-value)
Baseline	66.4%	58.6 (-)
1-3 NGrams	66.0%	57.7 (p = 0.275)
CMU	66.6%	59.0 (p = 0.432)

Table 2: Accuracy and macro-F1 scores for different variants of our baseline.

treatment of social media conventions such as emoji.[4].

We summarise thse results in Table 2 Given our multi-class scenario, a more informative metric than accuracy is the macro-F1 score, which we present here (scaled to lie from 0 to 100) and use in the remainder of this paper. For this experiment and in the remainder of this paper, we consistently report on 10-fold cross-validation results, using the same fold splits each time. For significance tests, we use the Wilcoxon Signed Ranks (Wilcoxon, 1945) test. Following the evaluation procedure of the 2016 CL Psych shared task, (Milne et al., 2016), we use macro-F1 as it gives "more weight to infrequent yet more critical labels", noting that the shared task and the classification task described in this paper shared much in common, albeit for different data sets. In this paper, significant results are in **bold** font.

We found that using a larger n-gram size did not help, decreasing the macro-F1 score to 57.7. We suspect this is due to the short nature of Twitter. Using the CMU tool provided a small improvement in macro-F1 (59.0), which we attribute to Twokenise's more comprehensive treatment of social media text conventions.

We note that character n-grams have also been explored in the literature, as a means to abstract beyond the noisy nature of social media. This has been experimented in the past by Coppersmith et al. (2016) and Malmasi et al. (2016). We focus on unigram features here to allow a straightforward comparison with the previously published results for the dataset.

In the remainder of this paper, as our baseline, we use our re-implementation of the O'Dea et al. (2015) classifier, using the Twokenise tool to create unigram features.

[4]https://github.com/myleott/ark-twokenize-py

3 Features used in Suicide-related Research

3.1 A Survey

One recent focus of computational linguistics research community has been on natural language processing tools to facilitate mental health research. This has been coordinated as shared tasks in the 2011 i2b2 Medical NLP Challenge[5] as well as the recent 2015 and 2016 shared tasks in the Computational Linguistics and Clinical Psychology (CL Psych) series (Coppersmith et al. (2015b) and Milne et al. (2016), respectively).

In this short survey, we focus on related work that examines different facets of text studied that help to characterise mental illness, with a particular focus on work on detecting suicide ideation. We can characterise features used as being: (i) **stylistic**, or (ii) **social media metadata**:

The **stylistic** features for analysing suicide-related text often uses features from the Linguistics Inquirer Word Count (LIWC) (Tausczik and Pennebaker, 2010). LIWC provides features such as articles, auxiliary verbs, conjunctions, adverbs, personal pronouns, prepositions, functional words, assent, negation, certainty and quantifier and have been used by Coppersmith et al. (2014) and De Choudhury et al. (2013) to study mental health signals in Twitter. Coppersmith et al. (2015a) employ the features to characterise mental illness, such Attention Deficit/Hyperactivity Disorder (ADHD) and Seasonal Affective Disorder (SAD).

These have also been applied to other data sources besides Twitter. For analyses of text on suicide ideation, Matykiewicz et al. (2009), uses LIWC to study suicide notes of suicide completers. Kumar et al. (2015) look at Reddit discussions following a celebrity suicide. Cohan et al. (2016) use the features to categorise mental health forum data in the 2016 CL Psych shared task.

In addition to LIWC, other stylistic features are possible. For example, Pestian et al. (2010) examines the use of readability metrices, such as the Flesch and Kincaid readability scores. Liakata et al. (2012) describe the role of features such as grammatical subject and object, grammatical triples, and negation in detecting emotion in the i2b2 dataset.

Social media metadata features have also pre-

[5]https://www.i2b2.org/NLP/Coreference/Call.php

viously been explored in the analysis of mental health related content. For example, metadata such as the time of post has previously been studied by Huang et al. (2015) and De Choudhury et al. (2013). Interestingly, De Choudhury et al. (2013) link time of posting to an insomnia index.

De Choudhury et al. (2013) also examines Twitter discussions, looking at the proportion of reply posts and the fraction of retweets as features. Related features are possible with other data sources besides Twitter. For example, Cohan et al. (2016) examine the role of discussion thread length for forum data.

A more complex set of features derived from the social media platform are network-related features. Colombo et al. (2016) perform social network analysis and examine the friend vs follower distributions in their analysis of Twitter networks and suicide ideation.

4 Evaluating Literature-Inspired Features

In this section, we describe our literature-inspired feature set covering (i) stylistic features and (ii) social media features. Our focus is on Twitter data which differs from other text given its short length, its informality in style, spelling and grammaticality. Consequently, instead of LIWC, we use a range of tools that are optimised for Twitter analytics, such as the CMU preprocessing tools, which provides Part-of-Speech tags for Twitter, and our own Twitter specific versions of the stylistic features listed above.

4.1 Stylistic Features

Following related work in examining stylistic linguistic features in analysing the language of mental health discussions (for example, Kumar et al. (2015) and Coppersmith et al. (2015a)), we examine a set of features that capture the linguistics attributes associated with the style of writing, such as orthography or words that have a strong syntactic element like pronouns. Similar features have been proven successful in sentiment analysis domain (for example, Mohammad et al. (2013) look at part-of-speech features and Brody and Diakopoulos (2011) examines orthographic features).

The features we explored are as follows:

- **Generic Text Attributes** The number of *chars*, *tokens* in the Twitter message.
- **Orthographic** This feature group includes the number of *all-upper-letter word*, *all-lower-letter word*, *words starting with upper letter*, *words containing continuously repeated letters* and *ratio of all uppercase to all lowercase* words in one tweet.
- **Sympathy Response Words** The number of words associated with a sympathetic response. We use the following categories:
  - please: *please, pls, plz*
  - no: *no, not, none, nope*
- **Punctuation** The number of *question marks*, *exclamation marks* and *colons* in the tweet.
- **Personal Pronoun** Three Boolean features to indicate the presence of 1st, 2nd and 3rd person pronouns. We define these as:
  - 1st: *I, me, myself, im, I'm*
  - 2nd: *u, you, yourself*
  - 3rd: *she, he, hers, his, her, him, herself, himself*
- **Question Words** The number of question words, such as: *why, what, whats, what's, when, where*, and *how*.
- **Time References** The number of time references, searching keywords including: *tomorrow*, *today*, *yesterday*, *now*, and the names of days (including abbreviations).
- **Auxiliary Verbs** The number of auxiliary and modal verbs, including: *am, is, are, do, does, have, has, going, gonna, was, were, did, had, gone, shall, can, may, might, could, would, should, will, must*.
- **Part-of-Speech (POS) features** The counts for POS tags provided by the CMU Twitter NLP tool (Gimpel et al., 2011).

4.2 Social Media Features

The Twitter Application Programming Interface (API)[6] provides additional metadata in addition to the message content. Some of these features capture elements of the social environment of the Twitter user posting the message, such as the size of their Twitter community (through the follower

[6]For full documentation, please view the Twitter Developer documentation: http://dev.twitter.com

and followee counts), and the level of conversational interaction for the current discussion, as given by the number of replies or retweets (Boyd et al., 2010).

The features we examined and our intuitions for using them were as follows:

- **Number of replies** The number of replies could indicate if the content was concerning enough to evoke one or more responses.
- **The timestamp of the post** Tweets posted at certain hours, for example late in the night, may be potentially more concerning.
- **Account features** These features capture the extent to which the Twitter user has personalised their Twitter account. The degree of personalisation could indicate the presence of spam accounts. We use 5 types of features: (i) whether the author has changed the default profile, (ii) whether author uses the default image, (iii) whether the author has provided a personal web URL; (iv) the number of followers; and (v) the number of friends (where both parties follow each other).
- **Tweet Special Elements** The count of special elements in a tweet, including: *retweet* flags, *favourite* flags, *hashtags*, *URLs present*, *user mentions*. This could indicate the style of communication.
- **Message Truncation** If the message is truncated, this could indicate that the content has been copied or reposted, potentially indicating that the content did not originate with the author.

4.3 Feature Normalization

So far, we introduced features with different units and scaling. In a linear model, such as the SVM, features with larger scale will be assigned higher weight during training stage. To avoid this, we normalised each feature independently by removing mean and scaling them to unit variance, as shown in following equation:

$$X_{norm} = \frac{X - \mu}{\sigma}$$

4.4 Results

In Table 3, we present 10-fold cross validation results for the dataset using the baseline features, as well as variants of the classifier that combine the stylistic and social media metadata features outlined above with the baseline features. The results show that performance is relatively unchanged when using social media features and stylistic features seem to help marginally. However, these results are not statistically significant.

Model	Macro-F1 (P-value)
Baseline (1-gram TFIDF)	58.6 (-)
+ Stylistic	60.2 (p = 0.084)
+ Social Media	58.5 (p = 1.000)

Table 3: Classification performance for different feature types.

Features	Macro-F1 (P-Value)
All	38.7 (-)
All - Style. Ling.	27.7 ($p = 0.002$)
All - POS	36.6 ($p = 0.010$)
All - Social Media	38.7 ($p = 1.000$)

Table 4: Metadata Features Performance

The lack of improvement was surprising, given the prevalence of these features in the literature. We thus performed a feature ablation study for social media features and the stylistic linguistic features. To gain insights on the contribution of these features types, this study was done without unigram features.

The results are presented in Table 4. The lower overall score indicates that the baseline classifier heavily relies on the unigram features, indicating that this is a strongly lexical task. We note that stylistic features capture textual cues, such as auxiliary verbs and pronouns, that may overlap somewhat with the unigram features. This is why we see so little benefit when they are added to the unigram features, as shown earlier in Table 3.

Removing POS features, as a subcategory of the stylistic features, only drops performance marginally, We infer that features to do with content, such as pronouns and sympathetic features are thus more useful cues in detecting suicide ideation.

Again, we find that social media features do not contribute greatly. One reason why this result may differ from related work is the nature of the data set, which may differ substantially from other data studied in related work. For example, it may be the case that timestamps do not matter for this Twitter dataset, which was collected under

different conditions than the work of De Choudhury et al. (2013), where Twitter content is much more strongly aligned to suicide attempts.

In addition, although the number of replies was useful in related work, in this data set most posts only had a single response, as shown in Figure 1. Furthermore Figure 2 shows that there is little difference in the length of discussion across different class labels.

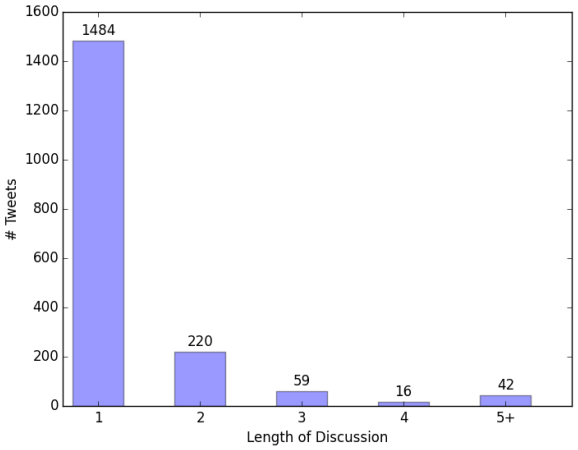

Figure 1: Distribution of Discussion Length

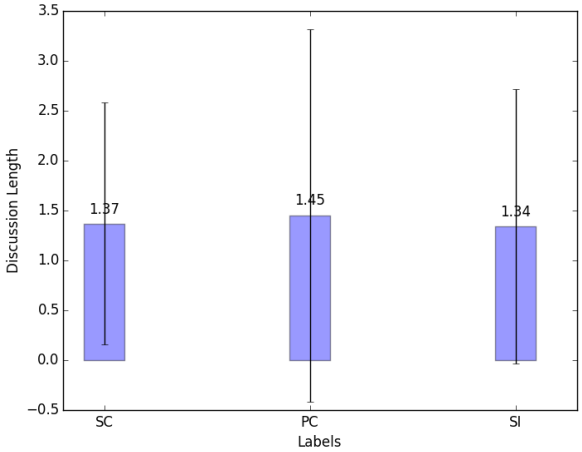

Figure 2: Averaged discussion length for each class label.

5 Discussion Context

One facet of the O'Dea et al. (2015) dataset is that it contains the responses to the annotated post. Although in a real-world intervention system that classifies a newly created Twitter post, responses may not be available, it may still be useful to gauge their role in suicide ideation detection.

Our motivation here in examining the responses is that these could lead to alternative methods for labelled data acquisition. For example, if responses turn out to be strongly correlated with the level of concerns for suicide ideation, perhaps by virtue of containing sympathetic content, we can explore methods that capitalise on this. Our aim here is to understand the feasibility of data acquisition approaches based on responses.

In exploring the role of the text in responses for suicide ideation detection, our work is similar to the recent 2016 CL Psych shared task, where forum discussions were the main source data. As a result, many participants explored the discussion as extra text context from which to derive features. For example, Malmasi et al. (2016) used the discussion structure to look at the posts preceding and following the discussion post in question. Pink et al. (2016) look at concatenations of discussion reply chains as a source of features. We used a similar approach in this work, except that we focus on the much shorter Twitter discussions.

We incorporate information about the discussion context by examining the responses to the Twitter post in question, or the "triggering post". When using the additional context of discussion responses, the feature representation of the triggering post can be augmented with feature representations based on the text of the responses. Given the results of the preceding section, we focus on unigram features for responses.

The two methods we explored are:

- **Merge Text** In the simplest approach, the text of original Twitter post and all responses are merged together into one text. Unigram features are extracted from this combined text. The length of this feature $|V|$ where V is the vocabulary size.

- **Split Text** In this representation, we keep the text of the triggering post and the text of the responses separate, resulting in two sets of unigram features. The size of this feature vector is $2|V|$.

5.1 Results

In Table 5, we present the results for the discussion features, showing that performance increases when maintaining some discussion structure (using the split text variant). Indeed, by collapsing the discussion, the triggering post and the responses, into a single text block, which one might want to do for the purposes of simplifying the model, the results are negatively affected.

Model	Macro-F1 (P-value)
Baseline (1-gram TFIDF)	58.6 (-)
+ Disc. (Merge Text)	57.1 (p = 0.375)
+ Disc. (Split Text)	60.7 (p = 0.084)
Disc. Split Text + Stylistic	**61.7 (p = 0.010)**
All	62.3 (p = 0.193)

Table 5: Classification performance for different feature types. All means "Disc. Split Text + Stylistic + Social media"

If we combine this with the stylistic features for the triggering post and for the responses, the gains are culminative with performance increasing to 61.7 (+3.1 macro-F1 points), a significant improvement above the baseline ($\alpha = 0.05$). We also conduct experiment including both stylistic features and social media features with results shown in All in Table 5. As we expected, by incorporating social media features, we only gain mild 0.6% F1 score improvement which is not statistical significant ($\alpha = 0.193 > 0.05$) compared with "Disc. Split Text + Stylistic".

5.2 The Role of the First Response

We observed a statistically significant positive improvement of 3.1 macro-F1 points. Although this is a positive improvement, it is slight. This is a surprise given the motivating example above. In particular, we expected that content-based features from the responses would help more in labelling the triggering post.

We performed subsequent experiments to see whether additional features that capture more of the discussion structure would help. For the results reported in Table 5, responses were treated as single amalgamated unit. However, one might expect that it is the first response that potentially sheds the most light as to whether there is a severe suicide ideation in the triggering post, since the subsequent responses may contain divergent topics.

Approach	Macro-F1 (P-Value)
Baseline	58.6 (-)
All Responses	60.7 (p = 0.084)
First Response	60.5 (p = 0.105)

Table 6: Investigating the role of the first response

We investigated this by creating variants of the system that would use just the first response, compared to the system described above, which uses all responses. The results are presented in Table 6. We observe that the performance is almost identical, if not marginally worse. We believe that this is because, while the first response may indicate the severity of the ideation, sympathetic responses tend to be shorter. Thus segmenting the discussion after the first response means that the feature representations is less rich.

	Resp.	SC	PC	SI	All
Chars	FR	55.8	62.0	69.1	63.2
	OR	69.8	57.7	69.2	62.1
Words	FR	10.8	12.0	13.4	12.2
	OR	13.2	10.5	12.4	11.31

Table 7: Average lengths of the first response (FR) vs. other responses (OR) in terms of characters and words.

Class	Words
SC	you, i, **don't**, to, **no**, it, **do**, me, that, **please**
PC	you, i, to, that, it, the, me, a, and, **don't**
SI	you, i, to, the, a, it, that, is, and, me

Table 8: Top 10 most frequent words in the first response (ordered by rank).

To explore this negative result further, we checked to see if indeed the first responses were shorter. Table 7 presents the average length of the first responses (compared to other responses) in terms of characters and words. Interestingly, for the SC class, the length of the first response is indeed shorter than the other responses. Furthermore, this is not the case for the other class labels.

This shorter length was associated with sympathetic responses. Table 8 provides a summary view of these responses by showing the top 10 words for the first response for each class label, with sympathetic terms bolded (terms that correspond to responses like "no, don't do it please"). The SC case has more of these words in its top 10 list, compared to the other class labels.

As the SVM was not able to utilise this information, we checked to see if a partially heuristic approach would work. We implemented a variant of the suicide ideation detection system that would first check the length of the first response. If this was less than a certain threshold, it would be deemed to be of the SC class. Otherwise, we

Thres.	6	8	10	12	14
F1	38	33	29	27	25

Table 9: A partially heuristic approach based on the length of the first response (in words). *thresh.* stands for threshold.

used the trained model (with the combination of the discussion split text features and the stylistic feature set, see Table 5) select the best class label.

In Table 9, we present the cross-validation results for this heuristic approach. The results show that this manual heuristic does not perform well. Thus, we are unable to beat the simpler model that simply treats the entire set of responses as single text. Unfortunately, given that we were not able to detect any stronger boost in performance, we conclude that basing an alternative mechanism for automatic data acquisition on the use of responses is not feasible.

6 Conclusion

In this paper, we explored a range of literature-inspired features that for the task of detecting suicide ideation on Twitter. We focused on stylistic linguistic and social media metadata features for use in addition to unigram features, finding that it was the stylistic features that helped for our dataset. We described a number of further investigations on the role of discussion context for this classification task, finding that discussion context helps. Furthermore, both discussion context and stylistic features can be combined to achieve a significant improvement in performance, compared with the previously published performance on this dataset. We also explored the contributions of different feature types and variations in representing the discussion context. We found that a simple representation that does not make a distinction between the first and following responses worked best. From these results, we conclude that unigram features still represent a strong baseline, reflecting perhaps that suicide ideation detection is a task that is heavily influenced by lexical cues.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable feedback. This work is supported by CSIRO Undergraduate Vacation Scholarships.

References

Rahn K. Bailey, Tejas C. Patel, Jaymie Avenido, Milapkumar Patel, Mohammad Jaleel, Narviar C. Barker, Jahanzeb Ali Khan, Shahid All, and Shagufta Jabeen. 2011. Suicide: Current trends. *Journal of the National Medical Association*, 103(7):614 – 617.

Danah Boyd, Scott Golder, and Gilad Lotan. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pages 1–10. IEEE.

Samuel Brody and Nicholas Diakopoulos. 2011. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!! using word lengthening to detect sentiment in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 562–570, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Julie Cerel, John R. Jordan, and Paul R. Duberstein. 2008. The impact of suicide on the family. *Crisis*, 29(1):38–44. PMID: 18389644.

Arman Cohan, Sydney Young, and Nazli Goharian. 2016. Triaging mental health forum posts. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 143–147, San Diego, CA, USA, June. Association for Computational Linguistics.

Gualtiero B. Colombo, Pete Burnap, Andrei Hodorog, and Jonathan Scourfield. 2016. Analysing the connectivity and communication of suicidal users on twitter. *Computer Communications*, 73(Pt B):291–300, jan.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015a. From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses. *the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015b. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, Denver, Colorado, June 5. Association for Computational Linguistics.

Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. 2016. Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 106–117, San Diego, CA, USA, June. Association for Computational Linguistics.

Kate Daine, Keith Hawton, Vinod Singaravelu, Anne Stewart, Sue Simkin, and Paul Montgomery. 2013. 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, oct.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 2:128–137.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47. Association for Computational Linguistics.

Xiaolei Huang, Xin Li, Tianli Liu, David Chiu, Tingshao Zhu, and Lei Zhang. 2015. Topic model for identifying suicidal ideation in chinese microblog. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, PACLIC 29, Shanghai, China, October 30 - November 1, 2015*.

Jared Jashinsky, Scott H Burton, Carl L Hanson, Josh West, Christophe Giraud-Carrier, Michael D Barnes, and Trenton Argyle. 2014. Tracking suicide risk factors through twitter in the us. *Crisis*.

Thorsten Joachims. 1999. Making large-scale support vector machine learning practical. pages 169–184.

Mrinal Kumar, Mark Dredze, Glen Coppersmith, and Munmun De CHoudhury. 2015. Detecting Changes in Suicide Content Manifested in Social Media Following Celebrity Suicides. *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 85–94.

Heidi Levine. 2008. Suicide and its impact on campus. *New Directions for Student Services*, 2008(121):63–76.

Maria Liakata, Jee-Hyub Kim, Shyamasree Saha, Janna Hastings, and Dietrich Rebholz-Schuhmann. 2012. Three Hybrid Classifiers for the Detection of Emotions in Suicide Notes. *Biomedical Informatics Insights*, 5(1):175–184.

Shervin Malmasi, Marcos Zampieri, and Mark Dras. 2016. Predicting Post Severity in Mental Health Forums. pages 133–137.

Pawel Matykiewicz, W Duch, and John P Pestian. 2009. Clustering semantic spaces of suicide notes and newsgroups articles. *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, (June):179–184.

David N. Milne, Glen Pink, Ben Hachey, and Rafael A. Calvo. 2016. Clpsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 118–127, San Diego, CA, USA, June. Association for Computational Linguistics.

Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.

Bridianne O'Dea, Stephen Wan, Philip J Batterham, Alison L Calear, Cecile Paris, and Helen Christensen. 2015. Detecting suicidality on twitter. *Internet Interventions*, 2(2):183–188.

John Pestian, Henry Nasrallah, Pawel Matykiewicz, Aurora Bennett, and Antoon Leenaars. 2010. Suicide Note Classification Using Natural Language Processing: A Content Analysis. *Biomedical informatics insights*, 2010(3):19–28.

Glen Pink, Will Radford, and Ben Hachey. 2016. Classification of mental health forum posts. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 180–182, San Diego, CA, USA, June. Association for Computational Linguistics.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods.

WHO. 2014. Preventing suicide: A global imperative. Technical report, World Health Organisation, Geneva, Switzerland.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.