Classification of Helpful Comments on Online Suicide Watch Forums

Ramakanth Kavuluru*
Div. of Biomedical Informatics
University of Kentucky
Lexington, Kentucky

María Ramos-Morales
Dept. of Computer Science
University of Kentucky
Lexington, Kentucky

Tara Holaday
College of Social Work
University of Kentucky
Lexington, Kentucky

Amanda G. Williams
Psychological Sciences Dept.
Western Kentucky University
Bowling Green, Kentucky

Laura Haye
College of Social Work
University of Kentucky
Lexington, Kentucky

Julie Cerel
College of Social Work
University of Kentucky
Lexington, Kentucky

{rvkavu2, maria.e.ramos, t.holaday, laurahaye, julie.cerel}@uky.edu, amanda.williams569@topper.wku.edu

ABSTRACT

Among social media websites, Reddit has emerged as a widely used online message board for focused mental health topics including depression, addiction, and suicide watch (SW). In particular, the SW community/subreddit has nearly 40,000 subscribers and 13 human moderators who monitor for abusive comments among other things. Given comments on posts from users expressing suicidal thoughts can be written from any part of the world at any time, moderating in a timely manner can be tedious. Furthermore, Reddit's default comment ranking does not involve aspects that relate to the "helpfulness" of a comment from a suicide prevention (SP) perspective. Being able to automatically identify and score helpful comments from such a perspective can assist moderators, help SW posters to have immediate feedback on the SP relevance of a comment, and also provide insights to SP researchers for dealing with online aspects of SP. In this paper, we report what we believe is the first effort in automatic identification of helpful comments on online posts in SW forums with the SW subreddit as the use-case. We use a dataset of 3000 real SW comments and obtain SP researcher judgments regarding their helpfulness in the contexts of the corresponding original posts. We conduct supervised learning experiments with content based features including n-grams, word psychometric scores, and discourse relation graphs and report encouraging F-scores ($\approx 80-90\%$) for the helpful comment classes. Our results indicate that machine learning approaches can offer complementary moderating functionality for SW posts. Furthermore, we realize assessing the helpfulness of comments on mental health related online posts is a nuanced topic and needs further attention from the SP research community.

*corresponding author

BCB'16, October 2–5, 2016, Seattle, WA, USA.
 ISBN 978-1-4503-4225-4/16/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2975167.2975170.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*; I.5.2 [**Pattern Recognition**]: Design Methodology —*Classifier design and evaluation, Feature evaluation and selection*

General Terms

Algorithms, Experimentation

Keywords

text classification, suicide prevention

1. INTRODUCTION

In this age of technology and the Internet, people are increasingly using online forums such as Reddit, Quora, and Stack Overflow to discuss popular topics, ask questions to obtain answers, and share personal stories to obtain feedback. Reddit in particular focuses on entertainment, news, and various other topics including music, sports, gaming, and food. Reddit users (or redditors) can submit URLs or text posts and can receive comments, up-votes, and down-votes. Comments can also be upvoted/downvoted just like posts. Reddit's ranking algorithm sorts both posts and the corresponding comments based on the time of submission and the difference between the numbers of up-votes and down-votes. The sorted list of posts is displayed as a bulletin board on Reddit. As such, redditors' voting activity and time of submission automatically decide the exposure of a post/comment to other redditors and external visitors given they are more likely to read only the first few posts in the ranked list. Due to this feature Reddit calls itself "the front page of the Internet".

Reddit also supports focused communities or forums that facilitate discussion on particular topics. These are called subreddits and SuicideWatch (SW) is one such subreddit where people post about their issues (or of someone they know) related to having suicidal thoughts to receive feedback and support from the community. In communities with such a sensitive topic, human moderators make sure that comments left for a post are not abusive in any way. A moderator may remove links and comments from the subreddit

if they find them objectionable or off topic, ban spammers or other abusive users from submitting to the subreddit, and add other users as moderators. The SW subreddit currently has 13 moderators carrying out such tasks.

Posts and comments can be made at any time of the day from any part of the world. Hence, with a small number of moderators it might be tedious to promptly review and moderate all comments. An objectionable comment can slip through the cracks and could already be seen by the original poster before it is caught by one of the moderators. Besides this, as SW subreddit grows it is increasingly important to also identify comments that are useful or helpful to the original poster from a suicide prevention (SP) perspective. This is important because the default ranking algorithm used by Reddit is common to all of Reddit and its communities and is based on metadata (time of submission, up-votes, down-votes) that do not have any explicit link to the content of the post. So communities such as SW might need custom comment ranking based on the relevance of particular comments to specific posts from an SP perspective.

The main goal of the work we report here is to *automatically identify helpful comments to posts in the SW subreddit* using supervised machine learning and natural language processing techniques. The approaches we use also inherently assign probability estimates to comments for their utility from a prevention perspective. These scores may also be used in future as part of the ranking algorithm for SW subreddit comments. To this end:

1. We used a dataset of 3000 SW subreddit comments, which were first hand-labeled with the help of three domain experts (students who are working on SP research) from University of Kentucky and Western Kentucky University.

2. We trained supervised models for text classification on this dataset with three types of features: n-grams, word psychometric scores [32], and discourse relation graphs [31]. We achieve F-scores in the 80–90% range with tight 95% confidence intervals based on repeated experiments with different subsets of the dataset.

Our overall objective is not to replace the human moderators but instead to prioritize or suggest useful comments to moderators and guide them in devising better ranking algorithms by also incorporating the content based usefulness/helpfulness score as a ranking component. The probability estimates of comment usefulness may also be used to color code comments in a certain way to convey the degree of helpfulness (this can be useful given only human moderators can actually delete comments). Furthermore, we believe that large scale linguistic analysis of helpful comments could yield meaningful insights into suitable discourse structures for writing helpful responses to posts on SW forums.

Although we specifically deal with the Reddit SW community, we believe our approach can be used for any other online forum that offers community based feedback to posts on suicide, self-harm, or other mental health related themes. Suicide is the 10th most common cause of death in the U.S. As of 2014, it is the second leading cause for the 10–34 age group and fourth leading cause among people in the 35–54 age group (`http://www.cdc.gov/violenceprevention/suicide/statistics/`). Also, 90% of US young adults use at least one social networking site and overall 65% of American adults are on at least one such site [25]. In this context, we believe our effort offers a meaningful exploration of computational methods for analyzing comments to SW posts.

2. RELATED WORK

Mining social media data (e.g., Twitter, Facebook, and Reddit) for health related information has deservedly garnered significant attention in the recent past. On one hand there are general efforts in identifying health related information in social data [33] and on the other we have results in specific subdomains such as pharmacovigilance [28], disease surveillance [4], and mental health and substance abuse monitoring [3, 7, 16]. Please see a brief survey by Paul et al. [23, Section 2] for more recent results.

With regards to suicide related research involving social media, it has been established that online social networks can play both positive (SP oriented) and negative (enabling suicidal ideation) roles [19]. Won et al. [35] conducted an interesting study to show that social media variables based on blog posts are significantly correlated with nation-wide suicide numbers in Korea. A similar result was later reported on the US population by Jashinsky et al. [15] based on Twitter data. Recently, Burnap et al. [2] built text classifiers that identify tweets discussing suicide ideation. They also studied the follower-friend networks of tweeters posting tweets that convey suicidal ideation [5].

De Choudhury et al. [8] tackled the highly consequential problem of identifying shifts to suicidal ideation from mental health discourse on Reddit. Specifically, they model the transition to ideation in terms of users posting in other mental health subreddits during a time period followed by a post in the SW subreddit in a subsequent time period. They are able to achieve an accuracy of 77.5% in classifying users who shifted to suicide ideation. In contrast to these efforts, we focus on the complementary task of identifying helpful comments to posts in the SW subreddit. To our knowledge, our attempt is the first in addressing this particular problem. The overall goal is both to classify helpful comments and also interpret the results to glean insights into effective communication strategies for online responses to social media posts from users expressing suicidal thoughts.

3. DATA COLLECTION & ANNOTATION

Using Reddit's API we collected 11,730 SW posts and 36,563 corresponding comments during the two year period from April 2013 to April 2015. We used REDCap [13], a web application for building and managing online surveys and databases, to have three raters annotate 3000 randomly selected comments from this dataset. Our raters were students working in SP research from the University of Kentucky and Western Kentucky University. They annotated the comments as either 'not helpful', 'helpful generic', or 'helpful specific' from an SP perspective. They rated the comments in the context of their original posts using a set of guidelines that we created (included in the appendix of the paper) based on a well known book on suicidology [1] and inputs from a licensed psychologist and established researcher in the area (Dr. Cerel, a co-author of this paper). The idea was to start with these basic guidelines and build predictive models for identifying helpful comments and derive linguistic traits that are associated with such comments.

The rationale for the three classes is to first identify comments that are not helpful in general. In case the comment

was deemed helpful, we would like to know whether the comment was specifically referring to or addressing topics discussed in the original post (the 'helpful specific' class). If not, the comment could be helpful but only as a generic message (e.g., to call some SP hotline) that is not tailored for a particular post. If we are able to identify comments that are helpful and also specific to the original post, we might be able to conduct large scale analysis of linguistic traits or features of such posts given comment length alone is generally not a reliable predictor of helpfulness based on our experiments. This may enable SP researchers in designing effective communication strategies for handling social media based disclosure of suicidal intentions and thoughts. As discussed in the appendix, there could be cases where a generic looking message might be the most appropriate and specific depending on what is revealed by the original poster.

We measured agreement between the three annotators based on two different scenarios. In the first one, we collapsed the generic and specific helpful class annotations into a single 'helpful' category thus making it a binary task while in the second scenario we maintained the original three class annotations. For these situations, based on the Fleiss' κ [10] for inter-rater agreement involving more than two raters, we obtain the agreements shown in Table 1; the qualitative assessment of the agreement is in the final column [18]. From the table, we see that agreement is nearly more than

Table 1: Inter-rater agreement assessment

	Agr. Freq.	Agr. %	Fleiss' κ	Assessment
Two class	1617	53.88%	0.35	Fair
Three class	983	32.76%	0.25	Fair

half of the dataset for the two class scenario and around a third for the three class version. When we examined agreements between pairs of annotators (instead of all three), two class agreement is around 70% and three class agreement is around 50% for all three possible rater pairings, which confirms that agreement goes down when the number of raters is increased. It could also be the case that assessing helpfulness of comments is a non-trivial and inherently subjective task that deserves additional attention from SP researchers. Nevertheless, given the agreement is only deemed 'fair' when considering all three annotators, instead of considering the full dataset of 3000 comments, we considered two subsets of relatively higher quality.

The first dataset is what we call the exact-2-class dataset which is essentially where all three annotators agree in the two class scenario (from Table 1). We also obtain a second larger subset of the full dataset called the majority-3-class dataset obtained by considering only those instances where there is a majority vote at the three class level (given a majority is not always guaranteed with three classes). It is straightforward to see that the exact-2-class dataset is a subset of the majority-3-class one. The details with class counts for both datasets are shown in Table 2. In both datasets, helpful examples constitute the majority classes (72% in exact-2-class and 62% in majority-3-class) and among helpful comments the specific class has roughly three times as many instances as the corresponding generic class.

Table 2: Details of final datasets considered

	hlp. spec.	hlp. gen.	¬helpful	total
Exact-2-class	907	259	451	1617
Majority-3-class	1175	440	995	2610

We did not experiment with the dataset where all three raters agreed at the three class level (2nd row of Table 1) given it results in a much smaller dataset and might contain examples that are very straightforward to classify. On the other hand, we still had to resort to the subsets in Table 2 because of the low agreement if we were to consider the full dataset of 3000 comments. By imposing a majority vote in the 3-class situation (majority-3-class) or exact match after collapsing to the 2-class scenario (exact-2-class), we are considering datasets that have higher quality (in terms of agreement) than the full dataset. Next, we discuss the supervised machine learning experiments conducted and corresponding results obtained.

4. CLASSIFICATION EXPERIMENTS

On both datasets in Table 2, we conducted supervised learning experiments for both binary (helpful vs ¬helpful) and multi class scenarios (helpful specific, helpful generic, and ¬helpful)) with three different types of features:

1. unigrams and bigrams from comments and original posts,
2. aggregate word psychometric scores from comments and original posts using the linguistic inquiry and word count (LIWC) program, and
3. frequent subgraphs of the rhetorical structure theory (RST) relation graphs extracted from comments and posts

Next, we briefly discuss each of these features separately.

4.1 N-gram features

These are straightforward tokens (unigrams) and adjacent token sequences of length two (bigrams) extracted from the comment text. Given we also have a 'helpful specific' class, we also included the unigrams and bigram features from the original post as a separate set of features. We also included the n-grams that occur in both the comment and original post as separate features to further capture, albeit in a naive manner, how well the comment seemed to address aspects discussed by the original poster. N-gram features provide a strong baseline in many text classification experiments [34] and our results here also indicate the same. Given the length of the comment could also play a role, we also included it (as number of tokens) as a separate numeric feature.

4.2 LIWC aggregate scores

LIWC (`http://liwc.wpengine.com/`) is a licensed software program that analyzes free text documents and returns scores for various psychological and other types of dimensions based on narrative content. Employing peer reviewed linguistic research [32] on psychometrics of word usage, LIWC aggregates scores for different dimensions based

on specific dictionaries with words that are pre-assigned (by linguistic experts) scores for each dimension. A few dimensions for which scoring is supported in LIWC are shown in Figure 1. As shown in the figure, LIWC computes numerical scores for both positive and negative emotions, personal concerns, and cognitive processes discussed in the content.

The intuition is that given LIWC attempts to capture psychometrics of word usage and prior knowledge that suicidal tendencies and other mental health conditions have a strong psychological component, we thought that including these features could help our case in predicting useful comments. LIWC features have been used in prior efforts [21] as meta-features of the content of a textual narrative. We used these features from both posts and comments in our models.

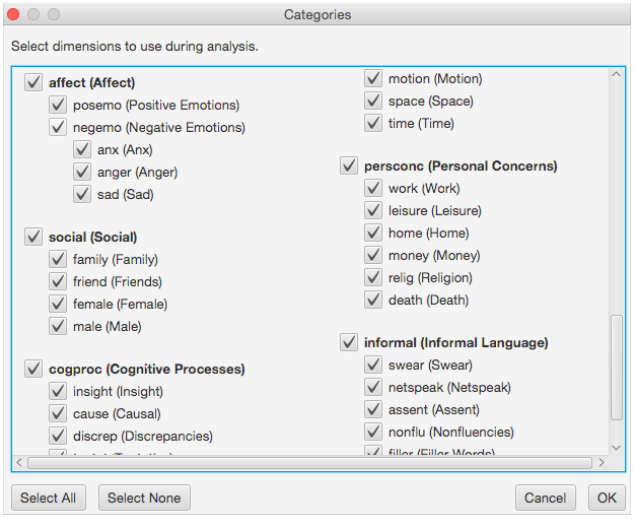

Figure 1: Screenshot of some dimensions supported by LIWC

4.3 RST relation subgraphs

RST [31] provides an explanation of the coherence of texts and is intended to describe texts from the perspective of its organization or discourse. It describes relations between segments of text from a sentence, paragraph or document, in the form of a directed graph. There are many available RST relations, but we only focused on 18 of those relations, specifically: attribution, background, cause, comparison, condition, contrast, elaboration, enablement, evaluation, explanation, joint, manner-means, summary, temporal, topic-change, topic-comment, same-unit, and textual-organization.

According to RST, a text narrative is composed of these relations connecting various units of the narrative where a unit is a coherent textual segment. Each manifestation of a relation has so called 'nucleus' and 'satellite' components. What a nucleus or satellite means depends on the specific RST relation. For example, for the 'evidence' RST relation, the nucleus is a claim and the satellite is a textual unit that contains information that is supposed to increase the reader's belief in the stated claim. In Figure 2, we show the RST graph for a sample comment from our dataset. The full comment is presented in the figure caption and can also be obtained from the leaves in order from left to right. For a few relations (e.g., 'contrast' and 'joint' in Figure 2), there is no distinction between a nucleus and a satellite given both components are equally important and do not have an explicit dependency such as in the 'condition' or 'attribution' relations. Recent applications of RST in natural language processing include text summarization, sentiment analysis, and machine translation. For a detailed review of applications of RST please see the article by Taboada and Mann [30].

We used Scala based RST parser discussed in efforts by Surdeanu et al. [29] and Jansen et al. [14] to get the RST relations of our posts and comments. The parser detects the 18 RST relations introduced earlier. We kept track of the numbers of each type of relation for each post and comment. RST relations for documents are represented as directed graphs as explained earlier. We stored the RST graph for all posts and comments in a database and obtained frequent subgraphs (of size at least two edges) for all comments in our labeled datasets using gSpan [36], a frequent graph pattern mining program, with minimum frequency set to 10% (appearing in at least 10% of full comment dataset), which resulted in 621 subgraphs. The frequent subgraphs were used as features in our models just like n-grams. We used these subgraphs instead of simple counts of RST relations because the subgraphs inherently capture finer aspects of discourse.

4.4 Results and Discussion

We used these features in a linearSVC model, an implementation of linear support vector machine (SVM) in the Python scikit-learn framework [24]. We converted the SVM scores to probability estimates using the sigmoid function. Our results for the binary classification scenario for the two datasets are shown in Table 3. We considered all possible combinations of features and also experimented with ensemble models (final three rows) including model averaging, stacking, and voting, all of which are built upon hundred base linearSVC classifiers. For stacking, the second stage classifier is also a linearSVC model. The performance measures reported are averages determined from experiments conducted with hundred distinct 80%-20% proportional train-test splits of the corresponding datasets. Besides average precision (P) and recall (R), we also show the 95% confidence interval around the mean F-score where all are computed assuming the helpful class as the positive class.

From the first two rows of Table 3, it is evident that there is statistically significant performance improvement when adding n-grams from the original post in addition to comment n-grams for both datasets as observed by non overlapping confidence intervals. From the 3rd row we see that LIWC scores alone achieve performance that is comparable to using n-grams for the smaller dataset. This shows that word psychometric scores provide a strong signal for classification. RST subgraphs (row 4) alone also do well but result in 5% F-score loss over the n-grams model. It is interesting to note that both LIWC and RST features seem to offer substantially higher recall than other approaches. Using LIWC scores and n-grams provides the best average F-score for the binary classification scenario, although using all features (row 7) is very close to this value and also improves further for the majority-3-class dataset. However, these performance differences are not statistically significant given the confidence intervals overlap. Considering F-scores, the single SVM model with all features outperforms the ensemble approaches we used. While stacking seems to produce better precision overall, this is at a major loss in recall.

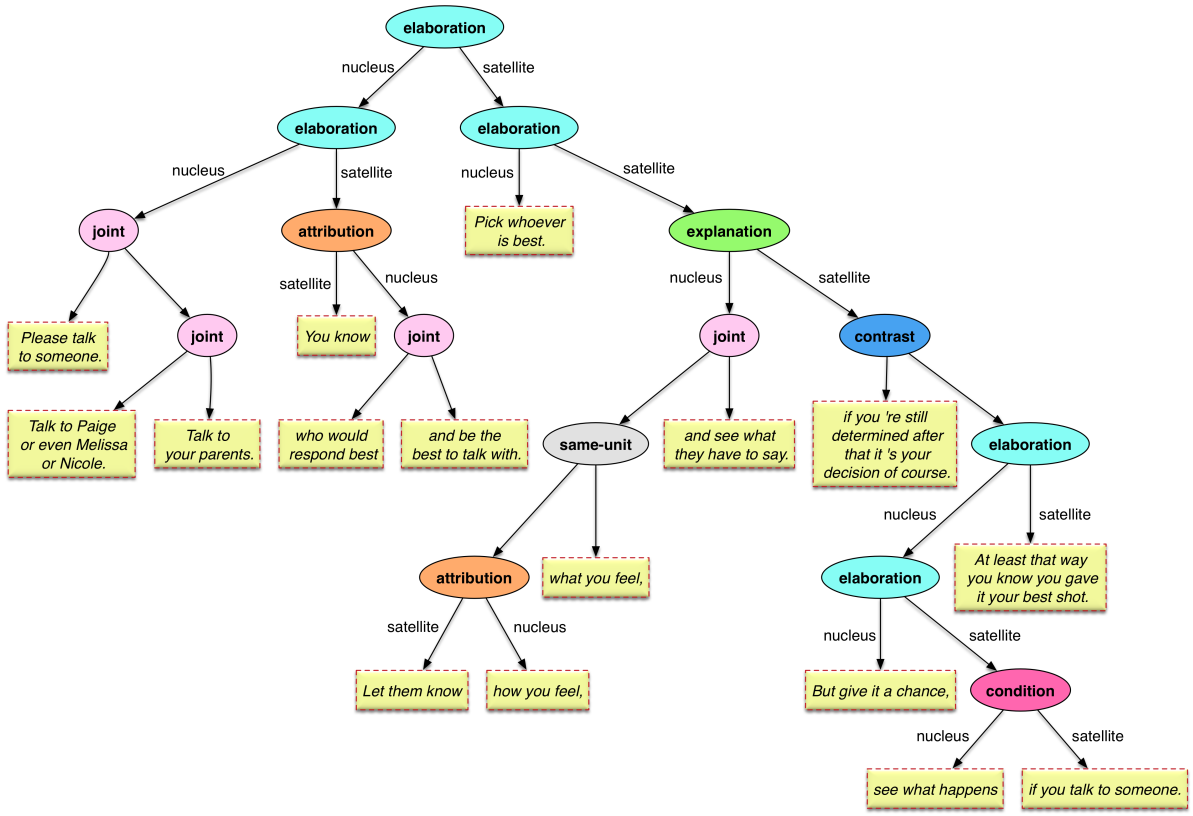

Figure 2: RST relation graph for an example useful comment: "Please talk to someone. Talk to Paige or even Melissa or Nicole. Talk to your parents. You know who would respond best and be the best to talk with. Pick whoever is best. Let them know how you feel, what you feel, and see what they have to say. if you're still determined after that it's your decision of course. But give it a chance, see what happens if you talk to someone. At least that way you know you gave it your best shot"

Table 3: Average binary classification performances over hundred distinct 80%-20% train-test splits

Features	exact-2-class			majority-3-class		
	P	R	F-score 95% CI	P	R	F-score 95% CI
comment n-grams	81.63	95.41	87.98 ± 0.20	73.10	89.98	80.65 ± 0.23
comment & post n-grams	82.40	96.31	88.80 ± 0.20	73.83	90.07	81.14 ± 0.19
LIWC scores	79.66	97.06	87.50 ± 0.18	62.00	**99.29**	76.34 ± 0.06
RST subgraphs	72.22	**99.92**	83.84 ± 0.01	62.07	98.97	76.23 ± 0.08
n-grams & RST subgraphs	82.56	95.56	88.58 ± 0.22	74.22	89.54	81.16 ± 0.21
n-grams & LIWC scores	83.14	96.04	**89.11** ± 0.20	73.81	90.04	81.11 ± 0.20
ALL	83.37	95.65	89.07 ± 0.20	74.29	89.63	**81.23** ± 0.20
model averaging	80.74	97.14	88.17 ± 0.19	72.06	91.99	80.80 ± 0.16
stacking	**84.30**	89.54	86.82 ± 0.19	**75.19**	79.82	77.41 ± 0.18
voting	80.78	97.03	88.15 ± 0.19	72.19	91.79	80.81 ± 0.19

Our results for the multiclass scenario are shown in Table 4. We assess the performance based on 100 runs on different splits of the datasets. Given there are three classes, we use macro averages for precision, recall, and F-score commonly used for text classification [20, Chapter 13] when the class distributions are not heavily skewed. So the values shown in Table 4 are mean values (over 100 runs) of the macro measures obtained in each run. We also used the average F-scores of just the two helpful classes (F_M^U) akin to how in sentiment analysis assessments are often done based

Table 4: Average multiclass classification performances over hundred distinct 80%-20% train-test splits

Features	exact-2-class				majority-3-class			
	P_M	R_M	F_M 95% CI	F_M^U 95% CI	P_M	R_M	F_M 95% CI	F_M^U 95% CI
comment n-grams	65.88	57.82	57.16 ± 0.50	51.23 ± 0.62	60.47	54.86	54.15 ± 0.41	47.67 ± 0.52
comment & post n-grams	66.82	59.08	58.71 ± 0.51	52.98 ± 0.66	62.46	56.51	56.21± 0.39	50.39 ± 0.48
LIWC scores	61.77	53.87	52.89 ± 0.50	46.95 ± 0.58	57.06	51.22	49.52 ± 0.38	43.46 ± 0.47
RST subgraphs	61.40	46.77	42.52 ± 0.32	37.35 ± 0.19	56.29	44.57	40.61 ± 0.28	33.27 ± 0.24
n-grams & RST subgraphs	65.00	58.56	58.28 ± 0.56	52.91 ± 0.69	62.33	56.51	56.32 ± 0.43	50.72 ± 0.54
n-grams & LIWC scores	67.86	**60.71**	**61.13** ± 0.50	55.72 ± 0.65	63.55	**57.82**	**57.70** ± 0.36	52.20 ± 0.47
ALL	67.03	60.59	61.12 ± 0.46	**56.17** ± 0.65	62.52	57.54	57.56 ± 0.41	**52.41**± 0.50
model averaging	68.97	57.75	57.41 ± 0.49	51.89 ± 0.61	65.37	56.39	55.80 ± 0.37	47.79 ± 0.49
voting	**69.27**	57.49	57.01 ± 0.48	51.36 ± 0.59	**65.61**	56.12	55.33 ± 0.38	49.09 ± 0.50

on average F-score for the positive and negative classes (ignoring the neutral class). Even in the multiclass scenario, adding n-grams from original posts provided statistically significant improvements over the case when n-grams from comments are used. This is not unexpected given our annotators look at the original post before assigning the class label to each comment. LIWC and RST features do not do as well for this scenario compared with the top scores. LIWC scores with n-grams provide the best macro F-score and recall for both datasets but using all features provides the best macro F-score computed over the two helpful classes. However, the performance differences between these two feature combinations are not statistically significant.

The ensemble approaches underperform like in the binary case except for a slight bump in the precision. At this point, it is not clear to us why ensemble models that typically improve over stand alone models did not work for this problem. It is possible that dataset size is not large enough to obtain diverse classifiers that can work together in a complementary fashion. Overall, the performance takes a major loss compared with the binary case (Table 3) across all models for the three class scenario. Looking at the precision and recall values from Tables 3 and 4, it is also interesting to see that recall is always higher than precision in the former, while it is the converse in the latter. From the confusion matrices, this manifested with having many more false positives than false negatives. Analyzing the confusion matrix of results from our best method for the three class scenario revealed that among 'helpful generic' comments around 50% were misclassified as 'helpful specific' and 20% were misclassified as 'not helpful'. That is, only 30% of the generic comments were correctly classified. This explains the higher performance measures when the helpful classes are merged into one category in the binary classification scenario. Also, in the same matrix, 20% of the 'not helpful' comments were misclassified as 'helpful specific' and 10% of 'helpful specific' comments were misclassified as 'not helpful'.

Given our motivation is also to look for linguistic traits, we analyzed the top RST subgraph features that offered most predictive power based on feature selection approaches. Here are a few examples from the top 25 subgraph features, where the first item is essentially five subgraphs with different varying relations at the leaf node.

- elaboration ⇒ elaboration ⇒ elaboration, attribution, contrast, explanation, joint
- joint ⇒ joint ⇒ elaboration
- elaboration ⇐ joint ⇒ joint

Elaboration is the most common relation and features in more than half of all comments and it is not surprising that it features as a node in many subgraphs. However, these results make it clear that attribution, contrast, and explanation are also important relations in helpful comments. For example, in Figure 2) the comment author is trying to convey the two contrasting actions available to the poster (see blue colored contrast relation node). This correlates well with recent findings that guided contrasting can be an effective means to cope with mental health problems such as depression [11].

5. CONCLUDING REMARKS

Suicide is a preventable leading cause of death in the US. According to the world health organization, the statistics of suicide related deaths are even more alarming in low and middle income countries accounting for 75% of all such deaths (`http://www.who.int/mental_health/prevention/suicide/suicideprevent/en/`). The increasing availability and affordability of personal computers and internet access and the subsequent growth of user engagement in online social networks points to a fast growing landscape of online mental health discourse. Motivated by the anonymity offered by the so called 'throwaway' online accounts, which are used for some posts, disinhibited users are increasingly sharing intimate information about their mental health concerns which underlie most suicide ideation scenarios. Although this can lead to unwanted consequences [19], through advances in computational social science, these developments offer new opportunities to understand suicide ideation.

In this paper, we reported results of an interdisciplinary project to automatically identify helpful comments to SW posts from an SP perspective. Based on annotations by

three different raters, we curated a dataset of Reddit SW posts and corresponding comments that have hand-labeled judgments indicating their helpfulness. We conducted text classification experiments using lexical, psychometric, and discourse related features and demonstrated that machine learned models offer a strong potential in the binary task of identifying helpful comments ($\approx 90\%$ F-score). From our literature review and consulting with our collaborators in SP research, we observed that guidelines were not available for effective communication when responding to suicide and self-harm related online posts. Our preliminary results show that presence of attribution, contrast, and explanation discourse relations as part of an elaborated response seemed to associate well with helpful comments. To our knowledge, this is the first effort to use an explicit SP lens to assess comment helpfulness.

Next, we identify some future research directions based on limitations of our current approach. Our results for the three class scenario are not very encouraging. This could be due to a couple of reasons: 1. The labeling task is inherently difficult in the three class scenario (lower κ from Table 2) 2. Our methods were not suitable for discriminating between subtle variations in the two helpful classes. Building a larger dataset by potentially involving post-annotation disagreement resolution meetings might be helpful. Refining the annotation guidelines might be essential and so also is involving annotators who have lived experience of losing a loved one due to suicide related incidents. Asking experts to also annotate specific segments of the comment that helped them reach their judgment could be helpful in teasing out differences in annotation practices. On the other hand, employing more sophisticated approaches especially convolutional neural networks can lead to performance gains based on our prior success in using them for biomedical text classification tasks [27]. Instead of using all LIWC dimensions, only using those specific to mental health domain might improve model performance as was the case in prior efforts [6]. We also propose to use promising new features extracted from posts and comments including

- dependency relations [9] involving psychometric words identified through LIWC,
- aggregate sentiment scores computed with large sentiment lexicons that were shown to achieve state-of-the-art results in SemEval tasks [17] and in general sentiment assessment of tweets [12],
- biomedical named entities from posts and comments from the Unified Medical Language System [22], and
- tree substitution grammar fragments that were shown to be useful for text classification [26].

Acknowledgments

We thank anonymous reviewers for their helpful comments that improved the organization of this paper.

This effort was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences through Grant UL1TR000117 and the Kentucky Lung Cancer Research Program through Grant PO2-415-1400004000-1. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

6. REFERENCES

[1] A. L. Berman, M. M. Silverman, and B. M. Bongar. *Comprehensive textbook of suicidology*. Guilford Press, 2000.

[2] P. Burnap, W. Colombo, and J. Scourfield. Machine classification and analysis of suicide-related communication on twitter. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 75–84. ACM, 2015.

[3] D. Cameron, G. A. Smith, R. Daniulaityte, A. P. Sheth, D. Dave, L. Chen, G. Anand, R. Carlson, K. Z. Watkins, and R. Falck. Predose: A semantic web platform for drug abuse epidemiology using social media. *Journal of biomedical informatics*, 46(6):985–997, 2013.

[4] L. E. Charles-Smith, T. L. Reynolds, M. A. Cameron, M. Conway, E. H. Lau, J. M. Olsen, J. A. Pavlin, M. Shigematsu, L. C. Streichert, K. J. Suda, et al. Using social media for actionable disease surveillance and outbreak management: A systematic literature review. *PloS one*, 10(10):e0139701, 2015.

[5] G. B. Colombo, P. Burnap, A. Hodorog, and J. Scourfield. Analysing the connectivity and communication of suicidal users on twitter. *Computer Communications*, 73:291–300, 2016.

[6] M. De Choudhury and S. De. Mental health discourse on Reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the Eighth AAAI International Conference on Weblogs and Social Media (ICWSM)*, pages 71–80. Citeseer, 2014.

[7] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz. Predicting depression via social media. In *Proceedings of the Seventh AAAI International Conference on Weblogs and Social Media (ICWSM)*, pages 128–137, 2013.

[8] M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2098–2110. ACM, 2016.

[9] M. de Marneffe, B. MacCartney, and C. Manning. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 449–454, 2006.

[10] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[11] A. Fritzsche, B. Schlier, G. Oettingen, and T. M. Lincoln. Mental contrasting with implementation intentions increases goal-attainment in individuals with mild to moderate depression. *Cognitive Therapy and Research*, pages 1–8, 2016.

[12] S. Han and R. Kavuluru. On assessing the sentiment of general tweets. In *Canadian Conference on Artificial Intelligence*, pages 181–195. Springer, 2015.

[13] P. A. Harris, R. Taylor, R. Thielke, J. Payne, N. Gonzalez, and J. G. Conde. Research electronic data capture (REDCap) – a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of biomedical informatics*, 42(2):377–381, 2009.

[14] P. Jansen, M. Surdeanu, and P. Clark. Discourse complements lexical semantics for non-factoid answer reranking. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, Volume 1: Long Papers*, pages 977–986, 2014.

[15] J. Jashinsky, S. H. Burton, C. L. Hanson, J. West, C. Giraud-Carrier, M. D. Barnes, and T. Argyle. Tracking suicide risk factors through twitter in the us. *Crisis*, 35(1):51–59, 2014.

[16] R. Kavuluru and A. Sabbir. Toward automated e-cigarette surveillance: Spotting e-cigarette proponents on Twitter. *J. of biomedical informatics*, 61:19–26, 2016.

[17] S. Kiritchenko, X. Zhu, and S. M. Mohammad. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, pages 723–762, 2014.

[18] J. Landis and G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[19] D. D. Luxton, J. D. June, and J. M. Fairall. Social media and suicide: a public health perspective. *American Journal of Public Health*, 102(S2):S195–S200, 2012.

[20] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[21] E. Momeni, K. Tao, B. Haslhofer, and G.-J. Houben. Identification of useful user comments in social media: A case study on flickr commons. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '13, pages 1–10. ACM, 2013.

[22] National Library of Medicine. Unified Medical Language System Reference Manual. `http://www.ncbi.nlm.nih.gov/books/NBK9676/`.

[23] M. Paul, A. Sarker, J. Brownstein, A. Nikfarjam, M. Scotch, K. Smith, and G. Gonzalez. Social media mining for public health monitoring and surveillance. In *Pacific Symposium on Biocomputing.*, volume 21, pages 468–479, 2015.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[25] Pew Research Center. Social media usage: 2005–2015. `http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/`.

[26] M. Post and S. Bergsma. Explicit and implicit syntactic features for text classification. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, Volume 2: Short Papers*, pages 866–872, 2013.

[27] A. Rios and R. Kavuluru. Convolutional neural networks for biomedical text classification: application in indexing biomedical articles. In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 258–267. ACM, 2015.

[28] A. Sarker, R. Ginn, A. Nikfarjam, K. O'Connor, K. Smith, S. Jayaraman, T. Upadhaya, and G. Gonzalez. Utilizing social media data for pharmacovigilance: a review. *Journal of biomedical informatics*, 54:202–212, 2015.

[29] M. Surdeanu, T. Hicks, and M. A. Valenzuela-Escárcega. Two practical rhetorical structure theory parsers. In *Proceedings of NAACL-HLT*, pages 1–5, 2015.

[30] M. Taboada and W. C. Mann. Applications of rhetorical structure theory. *Discourse studies*, 8(4):567–588, 2006.

[31] M. Taboada and W. C. Mann. Rhetorical structure theory: Looking back and moving ahead. *Discourse studies*, 8(3):423–459, 2006.

[32] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[33] S. Tuarob, C. S. Tucker, M. Salathe, and N. Ram. An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages. *Journal of biomedical informatics*, 49:255–268, 2014.

[34] S. Wang and C. D. Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 90–94. Association for Computational Linguistics, 2012.

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

[36] X. Yan and J. Han. gSpan: Graph-based substructure pattern mining. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002)*, pages 721–724, 2002.

Appendix: Annotation Guidelines

We developed the following brief guidelines loosely based on general guidance indicated in a suicidology resource [1] and inputs from our collaborator Dr. Cerel.

Helpful generic

- A helpful comment will strive to help the original poster (OP) in an appropriate way to potentially prevent suicide or self-harm.
- It might provide information on suicide hotlines or general advice to seek help including suggestions to consider seeing a psychiatrist, talking to family members.
- If the OP asks a question, the comment does not answer the question or any specific aspects of the main post, but nevertheless includes generic advice.
- This may also include comments that express empathy toward the OP's situation. However, in this case, the comment must also indicate in some way that the situation will improve, giving a sense of hope to the OP.
- When the OP does not provide information about specific details of what he/she is going through but simply expresses intentions to end his/her life, even comments that are essentially questions asking for further information could be construed as useful for our purposes.

Helpful specific

- This category essentially has the same criteria as the previous 'helpful genric' category, and in addition addresses specific issues mentioned by the OP.
- The comment may provide emotional support and understanding while also specifically addressing aspects or issues mentioned by the OP.
- Sometimes advice to call 911 or family members can be considered as specific to the original post if the post actually contains details that need such a response. For example, people posting about having already consumed or overdosed on some medication.

Not helpful

A comment is not helpful

- if it has a clear judgmental tone.
- if it is just short and superficial reassurance without conveying enough information,
- if it is a stereotypical response or is conveying a 'get over it' or 'suck it up' response. For example, if the commenter is saying that the problem the OP has is not unique to him/her and others like him/her also go through it. Stereotypical assumptions are not helpful, such as assuming that the OP's problem is related to his/her gender or race. However, depending on the situation, if the comment actually goes on further to encourage and give hope, it could be construed as useful.
- if the comment author responds defensively to a potentially polarizing view expressed by the OP going through suicidal thoughts. As in, the comment author is taking the post's text personally and is reacting to it defensively rather than keeping suicide prevention as the main motivation for the comment.