Biomedical Informatics Insights

ORIGINAL RESEARCH

Sentiment Analysis of Suicide Notes: A Shared Task

John P. Pestian[1], Pawel Matykiewicz[1], Michelle Linn-Gust[2], Brett South[3], Ozlem Uzuner[4], Jan Wiebe[5], K. Bretonnel Cohen[6], John Hurdle[7], Christopher Brew[8]

[1]Cincinnati Children's Hospital Medical Center, University of Cincinnati, Cincinnati OH. [2]American Association of Suicidology, Washington DC. [3]United States Veteran's Administration, Salt Lake City, UT. [4]University at Albany, SUNY, Albany, NY. [5]University of Pittsburgh, Pittsburgh, PA. [6]University of Colorado, Denver, CO. [7]University of Utah, Salt Lake City, UT. [8]The Ohio state University, Columbus, OH. Corresponding author email: john.pestian@ecchmc.org

Abstract: This paper reports on a shared task involving the assignment of emotions to suicide notes. Two features distinguished this task from previous shared tasks in the biomedical domain. One is that it resulted in the corpus of fully anonymized clinical text and annotated suicide notes. This resource is permanently available and will (we hope) facilitate future research. The other key feature of the task is that it required categorization with respect to a large set of labels. The number of participants was larger than in any previous biomedical challenge task. We describe the data production process and the evaluation measures, and give a preliminary analysis of the results. Many systems performed at levels approaching the inter-coder agreement, suggesting that human-like performance on this task is within the reach of currently available technologies.

Keywords: Sentiment analysis, suicide, suicide notes, natural language processing, computational linguistics, shared task, challenge 2011

Biomedical Informatics Insights 2012:5 (Suppl. 1) 3–16

doi: 10.4137/BII.S9042

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

Introduction

In this paper we describe the 2011 challenge to classify the emotions found in notes left behind by those who have died by suicide. A total of 106 scientists who comprised 24 teams responded to the call for participation. The results were presented at the Fifth i2b2/VA/Cincinnati Shared-Task and Workshop: Challenges in Natural Language Processing for Clinical Data in Washington, DC, on October 21–22, 2011, as an American Medical Informatics Association Workshop. The following sections provide the background, methods and results for this initiative.

Background Content of Notes

All age groups leave suicide notes behind between 10% and 43% of the time. What is in a suicide note? Menniger suggested that "the wish to die, the wish to kill and the wish to be killed must be present for suicide to occur,"[1] but there is a paucity of research exploring the presence of these motives in suicide notes. Brevard, Lester and Yang analyzed notes to determine if Menniger's concepts were present. Without controlling for gender, they reported more evidence for the wish to be killed in suicide notes of completers (those who successfully complete suicide) than the notes of non-completers.[2] Leenaars, et al revisited Menninger's triad and compared 22 suicide to 22 parasuicide notes that were carefully matched. They concluded that the notes from completers were more likely to have content reflecting anger or revenge, less likely to have escape as a motive, and, although it was not statistically significant, there was a tendency to show self-blame or self-punishment. In another study of 224 suicide notes from 154 subjects, note-leavers were characterized as young females, of non-widowed marital status, with no history of previous suicide attempts, no previous psychiatric illness, and with religious beliefs. Suicide notes written by young people were longer, rich in emotions, and often begging for forgiveness. Another study noted that statements found significantly and more frequently in genuine notes included: the experience of adult trauma, expressions of ambivalence; feelings of love, hate and helplessness, constricted perceptions, loss and self-punishment. One important and consistent finding is the need to control for differences in age and gender Leenaars et al.[3]

Using suicide notes for clinical purposes

At least 15% of first attempters try again, most often successfully dying by suicide. "Determining the likelihood of a repeated attempt is an important role of a medical facility's psychiatric intake unit and notoriously difficult because of a patient's denial, intent for secondary gain, ambivalence, memory gaps, and impulsivity."[4] One indicator of the severity and intent is simply the presence of a suicide note. Analysis has shown that patients presenting at an emergency department with non-fatal self-harm and a suicide note suggests that these patients were likely to be at increased risk for completing suicide at a later date.[5] Evidence of a suicide note may illuminate true intentions, but the lack of one does not squelch questions like: without a note is the patient substantially less severe, how many patients died by suicide without leaving a note behind, or is there a difference between the notes of completers and attempters? Valente's matched notes from 25 completers and attempters found differences in thematic content like fear, hopelessness and distress. On the other hand, Leenaars found no significant difference between thematic groups.[3,6]

These studies, however, were unable to take advantage of advanced Natural Language Processing (NLP) and machine learning methods. Recently, Handleman incorporated basic NLP methods like word-counts and a rough approximation of a semantic relationship between a specific word and a concept. For example, the concept of *time* was semantically represented by the words *day* or *hour*. The univariate analysis using only word count found no difference between notes, which is contrary to our previous results. When gender was controlled, some semantics differences like positive emotions, time, religion, and social references emerge.[7] Our interpretation of this gap between conclusions suggest these notes offer opportunity to explain some of the variation in suicide susceptibility, but require sophisticated NLP for a fuller understanding. Like Handelman, our initial attempts to understand the linguistic characteristics of these notes was to review the differences between linguistic characteristics like word count, parts of speech and emotional annotation. We found significant difference between the linguistic and emotional characteristics of the notes. Linguistic differences

(completer/simulated): word count 120/66 $P = 0.007$, verbs 25/13 $P = 0.012$, nouns 28/12 $P = 0.0001$, and prepositions 20/10 $P = 0.005$. Emotional differences: completers gave away their possessions 20% of the time, simulated, never did.[8]

Corpus Preparation

The corpus used for this shared task contain the notes that were written by 1319 people before they died by suicide. They were collected between the years of 1950 and 2011 by Dr. Edwin Shneidman and Cincinnati Children's Hospital Medical Center. The database construction began in 2009 and is approved by the CCHMC IRB (#2009-0664). Each note was scanned into the Suicide Note Module (SNM) of our clinical decision support framework called CHRISTINE. The notes were scanned to the SNM and then transcribed to a text-based version by a professional transcriptionist. Each note was then reviewed for errors by three separate reviewers. Their instructions were to correct transcription errors but leave errors like spelling, grammar and so forth alone.

Anonymization

To assure privacy, the notes were anonymized. To retain their value for machine learning purposes, personal identification information was replaced with like values that obscure the identity of the individual.[9] All female names were replaced with "Jane," all male names were replaced with "John," and all surnames were replaced with "Johnson." Dates were randomly shifted within the same year. For example, Nov 18, 2010, may have been changed to May 12, 2010. All addresses were changed to 3333 Burnet Ave., Cincinnati, OH, 45229, the address of Cincinnati Children's Hospital Medical Center main campus.

Annotators

It is the role of an annotator to review a note and select which words, phrases or sentences represent a particular emotion. Recruiting the most appropriate annotators led us to consider "vested volunteers," or volunteers who had an emotional connection to the topic. This emotion connection is what makes this approach different than crowd-sourcing[10] where there is no known emotional connection. In our case, these vested volunteers are routinely called *survivors of suicide loss* and they are generally active in a number of suicide communities. Approximately 1,500 members of several online communities were notified via e-mail or indirectly via Facebook suicide bereavement resource pages. Of those communities, two groups included Karyl Chastain Beal's online support groups *Families and Friends of Suicides* and *Parents of Suicides*, and the *Suicide Awareness Voices of Education*, directed by Daniel Reidenberg, PsyD. were most active. The notification included information about the study, its funding source and what would be expected of a participant. Respondants were vetted in two stages. The first stage included insuring that the inclusion criteria (21 years of age, English as a primary language, willingness to read and annotate 50 suicide notes) were met. The second stage included a review of the e-mail that potential participants were asked to send. In the email, respondents were asked to describe their relationship to the person lost to suicide, the time since the loss, and whether or not the bereaved person had been diagnosed with any mental illness. Demographic information about the vested volunteers is described below. Once fully vetted, they were given access to the training site. They also were reminded that they could opt out of the study at any time if they had any

Table 1. Example of a note annotation for different span with corresponding Krippendorff's α and the majority rule.

		I	hate	you	I	love	you	α
	a_1		hate			love		
Token	a_2	anger, hate	anger, hate		love	love		≈0.570
	a_3	anger, blame	anger, blame	anger, blame	love	love	love	
	a_1		hate			love		
Sentence	a_2		anger, hate			love		≈0.577
	a_3		anger, blame			love		
Majority	m		anger, hate			love		

difficulties and they were given several options for support. Training consisted of an online review and annotation of 10 suicide notes. If the annotator agreed with the gold-standard at least 50% of the time, they were asked to annotate 50 more notes.

Emotional assignment

Each note in the shared task's training and test set was annotated at least three times. Annotators were asked to identify the following emotions: abuse, anger, blame, fear, guilt, hopelessness, sorrow, forgiveness, happiness, peacefulness, hopefulness, love, pride, thankfulness, instructions, and information. A special web-based tool was used to collect, monitor and arbitrate the annotation. The tool collects annotation at the token and sentence level. It also allows for different concepts to be assigned to the same token. This makes it impossible to use simple k inter-annotator agreement coefficient.[11] Instead, Krippendoff's α[12] with Dice's coincidence index[13] was used. Artstein and Poesio[14] provided excellent explanation of the differences and applicability of variety of agreement measures. There is no need to repeat their discourse, however, it is worth explaining how it applies to the suicide note annotation task.

Table 1 shows an example of a single note annotation done by three different coders. At a glance, one can see that the agreement measure has to accommodate multiple coders (a_1, a_2, a_3), missing data, and multi-level agreement ("anger, hate" and "anger, blame" where $d_{Dice} = 1/2$ vs. "hate" and "anger, hate" where $d_{Dice} = 1/3$). Krippendoff's α accommodates all these needs and enables calculations for different spans. Despite that annotators were asked to annotate sentences, they usually annotated clauses and in some cases phrases. For this shared task, the annotation at the token level was merged to create sentence level labels. This is only an approximation to what happens in suicide notes. Many notes do not have typical English grammar structure so none of the known text segmentation tools would work well with this unique corpora. Nevertheless, this crude approximation yields similar inter-annotator agreement (see Table 2). Finally, a single gold standard was created from these three sets of sentence level annotations. There was no reason to adopt any *a priori* preference for one annotator over another, so the democratic principle of assigning a majority annotation was used (see Table 1). This remedy is somewhat similar to the Delphi method, but not as formal.[15] The majority annotation consists of those codes assigned to the document by two or more of the annotators. There are, however, several possible problems with this approach. For example, it could be that majority of the annotation will be empty. The arbitration phase focused on notes with the

Table 2. Annotator characteristics.

Response to call	
Annotators	
Direct contact	1500
Indirect contact	Unknown
Not eligible	10
Completed training	169
Withdrew	17
Respondents who fully completed the task	64
Gender and age	
Males	10%
Females	90%
Average age (SD)	47.3 (11.2)
Age range	23–70
Education level	
High school degree	26
Associates degree	13
Bachelors	23
Masters	34
Professional (PhD/MD/JD)	4
Connection to suicide	
Survivor of a loss to suicide	70
Mental health professional	18
Other	12
Time since loss	
0–0 years	27
3–3 years	25
6–60 years	14
11–15 years	13
16 years or more	12
Relationship to the lost	
Child	31
Sibling	23
Spouse or partner	15
Other relative	9
Parent	8
Friend	5
Performance	
Number of notes annotated at least once	1278
Number of notes annotated at least twice	1225
Number of notes annotated at least three times	1004
Mean (SD) annotation time per note	4.4 min (1.3 min)
Token inter-annotation agreement	0.535
Sentence inter-annotation agreement	0.546

lowest inter-annotator agreement where this situation could occur. Annotators were asked to re-review the conflicting notes, however, not all of them completed the final stage of the annotation process. There were ≈37% of sentences that had a concept assigned by only one annotator.

Evaluation

Micro- and macro-averaging

Although we rank systems for purposes of determining the top three performers on the basis of micro-averaged F_1, we report a variety of performance data, including the micro-average, and macro-average. Jackson and Moulinier comment (for general text classification) that: "No agreement has been reached … on whether one should prefer micro-or macro-averages in reporting results. Macro-averaging may be preferred if a classification system is required to perform consistently across all classes regardless of how densely populated these are. On the other hand, micro-averaging may be preferred if the density of a class reflects its importance in the end-user system"[16] p160–161. For the present biomedical application, we are more interested in a system's ability to reflect the intent. We, therefore, emphasize the micro-average.

Systems comparison

A simple table showing micro-averaged F_1 scores show the relationship between systems' outputs and the gold stan-dard but does not give insight how the individual submissions differ from each other. Even the z-test on two proportions does not do good job of comparing the system outputs.[17] It is conceivable that two systems may produce the same F_1 scores but err on different sentences. It may be possible to create ensamble classifier[18] from different systems if they specialize in different areas of automation. In order to diagnose this problem, we used hierarchical clustering with minimum variance aggregation technique to create a dendrogram that will cluster similar system outputs in the same branches.[19] The distance between submissions was calculated using inverse F_1 score ($d = 1\ F_1$).

The data

It is our goal to be fully open-access with data from all shared tasks. The nature of these data, however, requires special consideration. We required each team to complete a Data Use Agreement (DUA). In this DUA, teams were required to keep the data confidential and only use it for this task. Other research using the data is encouraged, but an approved Institutional Review Board protocol is required to access the data first.

Results

The results are described below. First a description of the annotators and their overall performance is provided. Then a description of the teams and their locations as described. More about the teams' performance is described in the workshop's proceedings. After this, each team's performance is listed. '

Annotators

The characteristics of the annotators are described in Table 2.

Figure 1. Geographic location of participants.

Table 3. Characteristics of the data.

Description	Total	Average	St. dev	Min	Max
Word count	146739	102.399	112.178	3	888.000
Swear	105	0.073	0.48	0	7.690
Family	2029	1.416	2.24	0	17.650
Friend	305	0.213	0.794	0	12.500
Positive emotion	7869	5.491	5.096	0	42.860
Negative emotion	3017	2.105	2.834	0	33.330
Anxiety	356	0.248	0.788	0	9.090
Anger	650	0.453	1.132	0	10.000
Sad	814.4	0.568	1.309	0	16.670
Cognitive process	19512.39	13.616	6.380	0	66.670
Biology	4267	2.977	3.324	0	25.000
Sexual	1453	1.01	2.044	0	25.000
Ingestion	172	0.12	0.496	0	5.560
Religion	917	0.64	1.845	0	27.270
Death	971	0.677	1.858	0	33.330

Participants

A total of 35 teams enrolled in the shared task. The geographic locations of these teams are shown in Figure 1. A total of 24 teams ultimately submitted results. There were a total of 106 participants on these teams. Team size ranged from 1 to 10. The averages size was 3.66 (SD = 1.86).

Characteristics of the data

Selected characteristics of the data are found in Table 3. This table provides and overview of the data using Linguistic Inquiry and Word Count, 2007. This software contains within it a default set of word categories and a default dictionary that defines which words should be counted in the target text files.[20]

Ranking

The ranking by each team is listed in Table 4. It provides each team's F_1 (micro-average), precision and recall. The highest score 0.6139 was achieved by *Open University* team. The scores range between 0.6139 and 0.29669 suggesting that different methods were used to achieve same goal.

It is interesting to look at relationship between different systems. Figure 2 provides a visual representation of the clustered results including the gold standard reference. It shows that the two most similar systems are *richardw* and *nrciit* and the F_1 between them is 0.7636. The F_1 between all systems (excluding the gold standard) ranges between 0.21 and 0.76 with the mean ≈0.522. This means that systems took fairly different approaches in solving the task, ie, each system makes errors on different sentences. In fact, there are only 118 sentence/label combinations that were false negatives across all systems and three sentence/label combinations

Table 4. Team ranking using micro-average F_1, precision and recall.

Team	F_1	Precision	Recall
Open university	0.61390	0.58210	0.64937
MSRA	0.58990	0.55915	0.62421
Mayo	0.56404	0.57085	0.55739
Nrciit	0.55216	0.55725	0.54717
Oslo	0.54356	0.60580	0.49292
Limsi	0.53831	0.53810	0.53852
Swatmrc	0.53429	0.57890	0.49607
UMAN	0.53367	0.56614	0.50472
Cardiff	0.53339	0.54962	0.51808
LT3	0.53307	0.54374	0.52280
UTD	0.51589	0.55089	0.48506
OHSU	0.50985	0.53351	0.48821
Wolverine	0.50315	0.45334	0.56525
TPAVACOE	0.50234	0.49922	0.50550
CLiPS	0.50183	0.51889	0.48585
SIP	0.49727	0.67429	0.39387
SRI & UC Davis	0.48003	0.49831	0.46305
DIEGO-ASU	0.47506	0.41791	0.55031
Ebi	0.45636	0.60077	0.36792
Duluth	0.45269	0.45985	0.44575
Columbia	0.43017	0.42125	0.43947
Pxs697	0.40288	0.37192	0.43947
Lassa	0.38194	0.35089	0.41903
Saeed	0.37927	0.37059	0.38836
SNAPS	0.35294	0.58684	0.25236
Senti6	0.29669	0.30532	0.28852

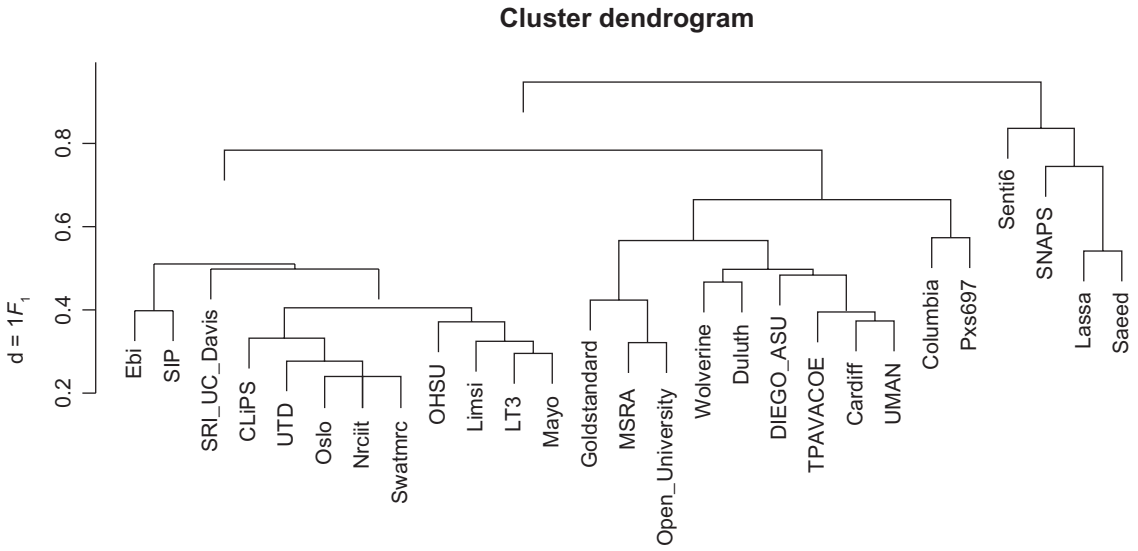

Figure 2. Comparison of different systems' outputs using distance $d = 1\ F_1$ and hierarchical clustering with minimum variance condition.

that were false positives across all systems. When we remove these 121 sentence/label combinations from the test data, the F_1 increases, for all systems on average, by 0.0223. Examples of these combinations are in the Table 5. Appendix 1 provides a listing of all systems.

On the other hand, if we would look at errors made by at least one system there were 5539 total combinations of sentence/label that were assigned by at least one system but were not present in the test gold standard and there were 1234 total combinations of sentence/label that were not assigned by at least one system but were present in the test gold standard. This leaves 38 sentence/label combinations that every system got right.

Even though there were frequent errors committed by individual classifiers, there were very few of the same errors committed by all systems. This suggests that appropriate ensemble of sentence classifiers might perform much better than a single instance classifier or even better than an ensemble of human experts. These findings make it more difficult to prove that there is a connection between the IAA that is calculated for human behavior and the F_1 that is calculated for machine learning output.

Discussion

Observations on running the task and the evaluation

Evaluations like the Challenge 2011 usually provide a laboratory of learning for the managers as well as the participants. In our case a few observations resonate. First, without the vested-volunteers it is unlikely we would have been able to conduct this challenge. Their courage was admirable, even when it led to churning

Table 5. Examples of sentence/label combinations that were misclassified by all systems.

Error type	Text ID	Sentence	Annotator	System
False negative	200909031138 4664	"Goodbye my dear wife Jane."	love	none
False negative	200809091809 2119	"I ask God alone to judge my action."	guilt	none
False negative	200812181837 2227	"I hope something is done to John Johnson, for I do not wish to die in vain."	anger	none
False positive	200908201415 0445	"respectfully Mary P.S. I love you BABY."	none	love
False positive	200812181838 1506	"Dearest Jane I am about to commit suicide. x Please notify police that I am in the deserted garage at the top of Terrace in Cincinnati near the rose bowl."	none	instructions
False positive	200809091735 1923	"John: I can't take your cruel unkind treatment any longer."	none	hopelessness

such deep emotional waters. Next, we relearned that emotional data remain a challenge. In our previous Shared Task, an inter-annotator agreement of 0.61 was achieved using radiology data.[9] Here we were able to attain a 0.546, which given the variation in data and annotators is appropriate. We conjecture that part of this difference is due to psychological phenomenology. That is, each annotator has a psychological perspective that he/she brings to emotionally-charged data and this phenomenology causes a natural variation.[21] Whether our use of vest-volunteers biased the interoperation, we are not sure. Preliminary analysis, suggests that these volunteers identify a smaller set of labels than mental health professionals. Finally, we wonder the what, if any bias traditional macro and micro F score introduce to this analysis. This question is apropos when dealing with multilabel-multiclass problems. Measures like micro and macro precision, recall, f1, hamming loss, ranked loss, 11-point average, break-even point, and alpha-evaluation are exploring this issue but consensus has yet to emerge.[22–26] The relation between inter-annotator agreement and automated system performance is not clear. The belief is that low IAA results in weak language models[27] but this connection was never formally established.

Acknowledgements

This research and all the related manuscripts were partially supported by National Institutes of Health, National Library of Medicine, under grant R13LM01074301, Shared Task 2010 Analysis of Suicide Notes for Subjective Information. Suicide Loss *Survivors* are those who have lost a loved one to suicide. We would like to acknowledge the roughly 160 suicide loss survivor volunteers who annotated the notes. Without them this research could not be possible. Their desire to help is inspiring and we will always be grateful to each and everyone of them.

We would like to acknowledge the efforts of Karyl Chastain Beal's online support groups Families and Friends of Suicides and Parents of Suicides and the Suicide Awareness Voices of Education, a non-profit organization directed by Danial Reidenberg, PsyD.

Finally, we acknowledge the extraordinary work of Edwin S. Shneidman, PhD and Antoon A. Leenaars, PhD who have had an everlasting impact on the field of suicide research.

Disclosures

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

References

1. Menninger K. *Man Against Himself*. Harcourt Brace, 1938.
2. Brevard A, Lester D, Yang B. A comparison of suicide notes written by suicide completers and suicide attempters. *Crisis*. 1990;11:7–11.
3. Leenaars AA, Lester D, Wenckstern S, Rudzinski D, Breward A. A comparison of suicide notes written by suicide notes and parasuicide notes. *Death Studies*. 1992;16.
4. Freedenthal S. Challenges in assessing intent to die: can suicide attempters be trusted? *Omega (Westport)*. 2007;55(1):57–70.
5. Barr W, Thomas J, Leitner M. Self-harm or attempted suicide? do suicide notes help us decide the level of intent in those who survive? *Accid Emerg Nurs*. 2007;15(3):122–7.
6. Valente SM. Comparison of suicide attempters and completers. *Med Law*. 2004;23(4):693–714.
7. Handelman LD, Lesterd D. The content of suicide notes from attempters and completers. *Crisis*. 2007;28(2):102–4.
8. Pestian JP, Matykiewicz P, Grupp-Phelan J, Arszman-Lavanier S, Combs J, Kowatch R. Using natural language processing to classify suicide notes. Chicago, IL, October 2008. American Medical Informatics Association.
9. Pestian JP, Brew C, Matykiewicz P, et al. A shared task involving multi-label classification of clinical free text. In ACL, editor, *Proceedings of ACL BioNLP*, Prague, June 2007. Association of Computational Linguistics.
10. Howe J. The rise of crowdsourcing. *Wired Magazine*. 2006;14(6):1–4.
11. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. Apr 1960;20(1):37–46.
12. Krippendorff K. *Content Analysis: An Introduction to its Methodology*. Sage Publications, Beverly Hills, CA, 1980.
13. Dice LR. Measures of the amount of ecologic association between species. *Ecology*. July 1945;26(3):297–302.
14. Artstein R, Poesio M. Inter-Coder agreement for computational linguistics. *Computational Linguistics*, 2008;34(4):555–96.
15. Dalkey NC, Rand Corporation. *The Delphi Method: An Experimental Study of Group Opinion*. Defense Technical Information Center, 1969.
16. Jackson P, Moulinier I. *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization*. John Benjamins Publishing Co, 2002.
17. Uzuner O, Sibanda TC, Luo Y, Szolovits P. A de-identifier for medical discharge summaries. *Artificial Intelligence in Medicine*. Jan 2008;42(1):13–35.

18. Kuncheva L, Whitaker C. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*. May 2003;51(2):181–207.
19. Gan G, Ma C, Wu J. *Data Clustering: Theory, Algorithms, and Applications*. SIAM, Society for Industrial and Applied Mathematics, May 2007. ISBN 0898716233.
20. Pennebaker JW, Chung CK, Ireland M, Gonzales A, Booth RJ. The development and psychometric properties of liwc 2007. *Austin, TX, LIWC. Net*, 2007.
21. Pestian JP, Matykiewicz P, Leenaars AA, et al. Distinguishing between completer and simulated suicide notes: A comparison of machine learning methods. In: *Association of Computational Linguistics*, 2008—In Review.
22. Min-Ling Zhang, Zhi-Hua Zhou. ML-KNN: a lazy learning approach to multi-label learning. *Pattern Recognition*. Jul 2007;40(7):2038–48.
23. Andre Elisseeff, Jason Weston. A kernel method for multi-labelled classification. In: *In Advances in Neural Information Processing Systems* 14. 2001;14:681–7.
24. Matthew R. Boutell, Jiebo Luo, Xipeng Shen, Christopher M. Brown. Learning multi-label scene classification. *Pattern Recognition*. September 2004;37(9):1757–71.
25. Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*. Mar 2002;34(1):1–47.
26. Tsoumakas G, Katakis, Taniar D. Multi label classification: An overview. *International Journal of Data Warehousing and Mining*. 2007;3(3):1–13.
27. Savova SP Ogren, Chute C. Constructing evaluation corpora for automated clinical named entity recognition. In Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'*08), Marrakech, Morocco, May 2008. European Language Resources Association (ELRA).

Appendix

Appendix 1. System description.

Team name	System name	Feature engineering	Feature selection	Number of features in the model	Data matrix sparseness
Cardiff	TopClass	Stanford POS tagger, WordNet lexical domains, emotive lexicons, internally assembled lexicons, manually identified patterns	frequency, mutual information, principal component analysis	245	N/A
CLiPS Research Center	GoldDigger	Multi-label training sentences re-annotated into single-label instances. Token unigrams (incl. function words and punctuation).	None	6,941 (# of tokens in training)	N/A
Columbia	Columbia	Lexical, syntactic, and machine-learned features	No	30	Very sparse
DEIGOASU	Emotion Finder	Clause level polarity features, unigrams and WordNet Affect emotion categories, Syntactical features (eg, sentence offset in the note)	Semi automated: the clause level and syntactic features manually selected and a greedy algorithm developed for selecting the rest of the features for each category	14,300	0.0025
Duluth	Duluth-1	Manual inspected combined with use of Ngram Statistics Package	Manual selection, looking for features uniquely associated with a particular emotion (based on intuition and Ngram Statistics Package output)	Approximately 1–30 rules per emotion, mainly consisting of unigram and bigram expressions	N/A
European Bioinformatics Institute	ebi	Word unigrams and bigrams, POS, negation, grammatical relations (subject, verb, and object)	Using frequency as threshold	Unigram (1,379), Bigram (8,391), POS (6), GR (775), verb (550)	
LIMSI	LIMSI	SVM classifiers and manually-defined transducers	None	160,272	N/A
LT3, University College Ghent, Belgium	LT3	MBSP shallow parser (lemma, POS), token tri-grams (highly frequent in positive instances), Senti- WordNet and Wiebe Subjectivity clues scores	Experimental: manual compilation of 17 feature sets, experiments to determine best feature set per label	5975 average (min 1747, max 6699)	0.00270 average (min 0.00189, max 0.00426)

Feature weighting	Learning algorithm	Manual rules	Estimation technique	Micro-average F_1 score
None	naive Bayes	Java regular expressions	cross validation	0.533
None	One- vs.-all SVMs trained on emotion- labeled and unlabeled instances, returning probability estimates per instance, per class. Two experimentally determined probability thresholds: one for emotion labels & one for the no-emotion class	None	10-fold CV	0.5018
Using ridge estimator	Logistic regression with ridge estimator	No	MLE	0.43
TF-IDF for unigrams	SVM with polynomial kernel	Intuitive lexical and emotional clues were manually translated to rules using regular expression and sentiment analysis of the clauses	2-fold cross validation	0.47
Rules for each emotion checked in order of frequency of emotion in training data, at most 2 emotions assigned	Human intuition	Perl regular expressions	N/A	0.45
None	SVM, CRF, SVM + CRF	Yes	9-fold cross validation	0.456
Combination of Binary and frequency weighting	LIBLINEAR SVM classifiers (one per emotion class) using following features: POS tags, General Inquirer, Heuristics, Unigrams, Bi-grams, Dependency Graphs, Affective Norms of English Words (ANEW)	Cascade of UNITEX transducers (one per emotion class)	10-fold cross validation	0.5383
None	Binary SVM, one classifier per label	None	50 bootstrap resampling rounds (3000 train, 1633 test)	0.5331

(*Continued*)

Appendix 1. (*Continued*)

Team name	System name	Feature engineering	Feature selection	Number of features in the model	Data matrix sparseness
Microsoft Research Asia	eHuatuo	Spanning 1–4 grams and general 1–4 garms	Positive frequency is divided by negative frequency by leveraging Live-journal weblog information	14428 selected features from spanning 1–4 grams	N/A
National Research Council Canada	NRC	Word unigrams and bigrams, thesaurus matches, character 4-grams, document length, various sentence-level patterns	None	71061	608448/ (71061 * 4633) = 0.00185
Oslo	Oslo	Stems and bigrams from PorterStemmer; part-of-speech from TreeTagger; dependency patterns from MaltParser; first synsets from WordNet	No constraints	Mean = 28289.3; std. dev. = 18924.7	Mean = 0.0017; std. dev. = 0.0008
SRI, UC Davis		Stanford Core-NLP generated POS tags, addressing features, unigrams & bigrams, LIWC (original and customized), emotion sequence and sentence position	Regularization in Log-Linear Model	On the order of thousands (comparable to text classification problems)	Very sparse (comparable to text classification problems)
UMAN		NLTK for significant uni-, bi- and tri-grams (likelihood measure), Stanford CoreNLP for NLP and NER, hand-crafted semantic lexicons, Flesh tool (for readability scores), Lingua-EN-Gender-1.013 (for gender feature) and manually written rules for sentence tense and some NER classes	genetic algorithm, Fast Correlation-Based Filter method and top 500 uni-, bi- and tri-grams	1690	0.013

Feature weighting	Learning algorithm	Manual rules	Estimation technique	Micro-average F_1 score
The confidence score from SVM	SVM classifier and pattern matching	No	10 fold cross validation	0.5899
Feature vectors normalized to unit length	Binary SVM; one-classifier-per-label	None	10-fold cross validation	0.5522
N/A	Six binary linear one- vs.-all cost-sensitive SVM classifiers	None	10-fold cross-validated grid search over all permutations of feature types and cost factors	0.54356
Frequency counts	Log-linear model, tuned with L-BFGS, followed by single step self training	None	5-fold cross validation	0.49
None	Nave Bayes with kernel density estimation	1. Frozen/common layman expressions 2. lexico-syntactic patterns using GATE/JAPE grammar	5-fold cross validation	0.5336