We then tested the hypothesis that MLAs could distinguish between completer and simulated notes as well as MHPs. Copies of the notes were given to five MHPs who classified them as either written by a completer or an simulator. MLA feature space was defined by matrix of selected characteristics from four sources: words, parts of speech, concepts, and readability indexes. Collinearity was eliminated by removing highly correlated features. The final feature space included: specific words (such as ”love”, ”life”, ”no”), specific parts of speech (such as, personal pronouns, verbs) Kincaid readability index and emotional concepts (such as anger, and hopelessness). We then tested the following algorithms’ ability to distinguish between completer and simulator notes: *decision trees* - J48, C4.5, LMT, DecisionStump, M5P; *classification rules* - JRip, M5, OneR, PART; *function models* - SMO, logistic builds, multinomial logistic regression, linear regression; *lazy learners* and *meta learners*[5].

3 Results

A significant difference was found 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. This difference justified testing the classification hypothesis. Emotionally, completers gave away their possessions 20% of the time, simulators, never did. Mental health experts accurately classified the notes 71% of the time. The MLAs were accurate 60-79% of the time with SMO giving the highest results when the word count, part-of-speech, and readability vectors were included. Performance weakened when the emotional vector was included, yet the emotional vector was the primary source of data for the MHPs.

4 Conclusion

Machine learning methods for classifying suicide and non-suicide notes are promising. Future efforts to represent the thoughts of the suicidal patient will require larger sample sizes, inclusion of attempters response to open-ended questions, biological and clinical characteristics.

5 Acknowledgements

We acknowledge Drs. AA Leenaars, ES Shneidman, the divisions of Biomedical Informatics, Emergency Medicine and Psychiatry at Cincinnati Children’s Hospital Medical Center, University of Cincinnati and Ohio Third Frontier program for their generous support of this work.

References:

[1] Jeffrey A Bridge, Tina R Goldstein, and David A Brent. Adolescent suicide and suicidal behavior. *J Child Psychol Psychiatry*, 47(3-4):372–394, 2006.

[2] P M Lewinsohn, P Rohde, and J R Seeley. Psychosocial risk factors for future adolescent suicide attempts. *J Consult Clin Psychol*, 62(2):297–305, 1994.

[3] E S Shneidman. Suicide as psychache. *J Nerv Ment Dis*, 181(3):145–147, 1993.

[4] ES Shneidman and NL Farberow. *Clues to Suicide*. McGraw Hill Paperbacks, 1957.

[5] I.H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools ad Techniques*. Morgan Kaufman, 2nd edition, 2005.

Suicide notes provide material for natural language processing. Previous approaches have examined suicide notes using content analysis [26], sentiment analysis [17, 29], and emotion detection [22]. In the age of cyberspace, suicide notes are now also written in the form of web blogs and can be identified as carrying the potential risk of suicide [14]. 2.3. Online User Content. Cash et al.Â Suicide notes are useful materials for training a classifier. The current dataset of suicide notes is quite small. Automatic detection on online user content will be a promising way for suicide detection and prevention. Our proposed method investigated a better solution with effective feature engineering on a bigger social dataset than the previous work. Recently, Natural Language Processing (NLP) strategies have been used with Electronic Health Records to increase information extraction from free text notes as well as structured fields concerning suicidality and this allows access to much larger cohorts than previously possible. This paper presents two novel NLP approaches â€“ a rule-based approach to classify the presence of suicide ideation and a hybrid machine learning and rule-based approach to identify suicide attempts in a psychiatric clinical database. Good performance of the two classifiers in the evaluation study suggest they can be us... Natural language processing to classify suicide notes.pdf. pattern recognition in narrative.pdf. Pennebaker&Chung_Al-Qaeda.pdf. Pers Soc Psychol Bull-2001-McAdams-474-85.pdf. Predicting military suicide.pdf. Revisions. Version ID.Â This page is currently connected to collaborative file editing. All edits made will be visible to contributors with write permission in real time. Changes will be stored but not published until you click the "Save" button. Natural language processing identified phrases from the notes associated with the suicide attempt outcome. We enriched this group of phrases with a clinically focused list of terms representing known risk and protective factors for suicide attempt in adolescents. We then applied the random forest machine learning algorithm to develop a classification model.Â Using NLP to codify pre-admission electronic health record (EHR) notes to detect suicidal behavior is a promising approach [28]. EHR clinician notes are likely to capture important correlates of suicidal behavior in aggregate over time since mental health clinicians are trained in biopsychosocial mental health evaluation, including risk assessment [29]. However, the 1990s saw mathematicians and information technology specialists join the effort by actively employing the methods of mathematical statistics, computational linguistics and natural language processing (NLP) in particular for quick processing of large massifs of textual data.Â Despite a pressing need to investigate suicide notes, being small, they do not offer opportunities to look at all the features of speech production of suicidal individuals.Â 17. Pestian J. P., Matykiewicz P., Grupp-Phelan J. Using natural language processing to classify suicide notes // Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing (BioNLP '08) / Association for Computational Linguistics, Stroudsburg, PA, USA, 2008. Ð . 96-97.