Sentiment Analysis: Detection of stress in workplace using email data

Adwin Singh
Middle Tennessee State University – Computer Science Department
Advisor – Dr. Sal Barbosa

Introduction

- Sentiment Analysis is an area of study in natural language processing which deals with extracting opinions or sentiments from natural language.
- As per Ekman's model, there are 6 basic set of emotions anger, sadness, fear, joy, anger and disgust.
- In this research I am proposing a framework to perform sentiment analysis on email dataset to detect emotions from email text and identify stress
- In business workspace, email still remains the most commonly used mode of communication and statistics indicates a total 281 billion emails are exchanged per day.

Aims

- Identify the optimum text preprocessing techniques to clean the email dataset.
- Develop method to label emails based on the presence of emotions in the email text for supervised learning approach
- Develop a supervised classification model to classify emails for the six basic emotions and detect stress

Background

- Previous research done on email dataset for sentiment classification showed text features such as TF-IDF improved the accuracy of the classification
- In sentiment analysis lexicon based methods has generated a lot of interest in recent times

Abstract

Identifying stress using email especially in a workplace would be helpful in addressing possible cases before it becomes chronic. Emotion analysis from text documents has gained lot of attention in recent years. But research in emotion analysis has been difficult due to lack of annotated datasets. Emotion analysis on large Email data a ubiquitous means of social networking and communication, has not been studied thoroughly. This paper proposes a new process to analyze email text to extract emotions of the writer to identify indicators of stress. This process includes a weakly supervised labelling approach to generate a labeled dataset, a new lexicon to identify stress based on the email corpus and classification model to classify emails with indicator of stress and experiment with the labelled dataset. Initial classification results show that the process is able to classify emails indicating stress with good accuracy and recall rate.

Methods

Text Cleanup
Developed in python to clean email text. For email set consider the first email in the chain.

Data Labelling
Method to label data set using distant supervision using emotional intensity of words in NRC emotional intensity lexicon

Feature Generation
Features - Bag of words, TF-IDF, word count of emotion words in text
Output – Numpy array of features for every email text

Classification Models
- Support Vector Machine
- Logistic Regression
- Naïve Bayes

Metrics

- **Accuracy –** Rate of correct predictions
- **Precision –** Ratio of correct positive predictions to the total predicted positives
- **Recall –** Ratio of correct positive predictions to the total positive samples

Results

	Logistic Regression	Naive Bayes	SVM
Avg Accuracy	81%	78%	86%
Avg Precision	51%	40%	71%
Avg Recall	25%	36%	54%

Conclusions

Results indicate that Support Vector machine outperformed other classification models. As per [Liu and Lee, 2018], they also found SVM classification performed better for email classification. But there is a lot of scope for improving the overall precision and recall. Average precision of 71% and recall of 54% across all 6 basic emotions is not very efficient. Labelling method also needs improvement which could also improve the overall classification as emails can be mislabeled. The method used to label dataset needs to be tested on other data sets.

Future Work

- Identify new features to improve the classification model.
- Data labelling method needs to be tested on a different human labelled dataset to check its accuracy.
- Further research on identifying different intensity of emotions and how to combine them to identify indicators of stress.

References

[Davcheva et al., 2019] Davcheva, E., Adam, M., and Benlian, A. (2019).User dynamics in mental health forums - a sentiment analysis perspective. Internationale Tagung Wirtschaftsinformatik, 14.

[Erosynidis et al., 2017] Erosynidis, D., Symeonidis, S., and Arampatzis, A. (2017). A comparison of pre-processing techniques for twitter sentiment analysis. 21st International Conference on Theory and Practice of Digital Libraries (TPDL 2017).

[Han et al., 2018] Han, H., Zhang, J., Yang, J., Shen, Y., and Zhang, Y. (2018). Generate domain-specic sentiment lexicon for review sentiment analysis. Multimedia Tools and Applications, 77:1-16.

[Hassan Yousef et al., 2014] Hassan Yousef, A., Medhat, W., and Mohamed, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5.

[Hutto and Gilbert, 2015] Hutto, C. and Gilbert, E. (2015). Vader: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014.

[Khan et al., 2014] Khan, F., Bashir, S., and Qamar, U. (2014). Tom: Twitter opinion mining framework using hybrid classication scheme. Decision Support Systems.

[Klimt and Yang, 2004] Klimt, B. and Yang, Y. (2004). The enron corpus: A new dataset for email classication research. pages 217-226.

[Liu, 2010] Liu, B. (2010). Sentiment analysis and subjectivity.

[Liu and Lee, 2015] Liu, S. and Lee, I. (2015). A hybrid sentiment analysis framework for large email data. 2015 10th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), 10(5):324-330.

[Liu and Lee, 2018] Liu, S. and Lee, I. (2018). Ecient and ective email sentiment analysis through k means labeling and support vector machine classication. Cybernetics and Systems, 49:181-199.

[Mohammad and Yang, 2011] Mohammad, S. and Yang, T. (2011). Tracking sentiment in mail: How genders dier on emotional axes. Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011), pages 70-79.

[Pillai et al., 2018] Pillai, R. G., Thelwall, M., and Orasan, C. (2018). Detection of stress and relaxation magnitudes for tweets. WWW '18: Companion Proceedings of the The Web Conference 2018, 41:1677-1684.

[Tang et al., 2014] Tang, G., Pei, J., and Luk, W.-S. (2014). Email mining: Tasks, common techniques, and tools. Knowledge and Information Systems, 41:1-31.

[Tausczik and Pennebaker, 2010] Tausczik, Y. R. and Pennebaker, J. W.(2010). The psychological meaning of words: Liwc and computerized textanalysis methods. Journal of Language and Social Psychology, 29(1):24-54.

[The Radicati Group, 2018] The Radicati Group, I. (2018). Email statistics report, 2018-2022. Email Statistics Report, 2018-2022.