A two-stage contagious Naive Bayes classifier for detecting sociolinguistic features in text

Iena Petronella Derks[1] and Alta de Waal[1 2]

[1] Department of Statistics, University of Pretoria
[2] Center for Artificial Intelligence Research (CAIR)

1 Introduction

Online platforms allow users to masquerade themselves; making virtual interactions anonymous or misleading recipients of the interactions. It also facilitates an environment for cybercrimes, allowing users to take advantage of others and commit heinous acts. An important concern on social media usage, in particular, has to do with the security of under-age users that have access to the Internet. Children are more vulnerable to threatening situations, such as harassment [3], cyberbullying [7], and inappropriate conversations [8]. Natural language processing (NLP) techniques can be used to process and understand social media data [1]. In the area of sociolinguistics, there is evidence that links natural word use to personality and social fluctuations [5]. In NLP, the term burstiness is used to describe the tendency of word recurrence. The burstiness phenomenon is frequently exhibited in real text, in which an informative word is more likely to occur if it has already appeared in the text [2]. State-of-the-art NLP models, such as the multinomial Naive Bayes model, are often used to model text documents [4].

2 Methodology

One application area of NLP is sociolinguistics which can be defined as the relationship between social factors and linguistics. Sociolinguistics aims to isolate features to determine linguistic variation in social conditions [6]. This paper investigates classification models which can model the burstiness, or contagious effects of text as the text that we are interested in are manifestations of different social groups of people. For example, a teenager vs. an adult impersonating as a teenager will have different sentence structures. More formally defined, the purpose of this work is to:

1. *Learn the linguistic patterns among different social groups of people, and classifying unknown authors according to these patterns; and*
2. *Represent these patterns as Bayesian networks to gain an understanding of the dependency structure of words used among different social groups.*

To identify these linguistic patterns, a comparison is made between two classification techniques, namely the Naive Bayes (NB) classifier and a contagious

counterpart thereof. The NB classifier assumes that words occurring in a document are independent of each other. On the other hand, the contagious classifier captures the burstiness phenomenon. To go one step further, we investigate the dependencies between words using a Bayesian network. This allows us to understand *why* certain word patterns results in a classification.

2.1 Data Application

This paper presents a comparison between the baseline NB classifier and the proposed contagious counterpart thereof. Two data sets will be used to evaluate the performance of each method, namely the IMDB data set and the PAN 2012 data set. The IMDB data set consist of movie reviews, with binary sentiment classification. The PAN 2012 data set is originally used to identify potential predators in online conversations, with 66 927 conversations. The problem is addressed with a two-stage solution, where stage 1 is based on text classification techniques and stage 2 makes use of Bayesian networks to understand the structural dependencies among words in a document. The evaluation of stage 1 is typical classification performance, whereas the visual structural learning of the Bayesian network provides for exploratory data analysis in order to understand the conditional dependencies between words.

References

1. Chowdhury, G.G.: Natural language processing. Annual review of information science and technology **37**(1), 51–89 (2003)
2. Doyle, G., Elkan, C.: Accounting for burstiness in topic models. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 281–288. ACM (2009)
3. Kennedy, G., McCollough, A., Dixon, E., Bastidas, A., Ryan, J., Loo, C., Sahay, S.: Technology solutions to combat online harassment. In: Proceedings of the first workshop on abusive language online. pp. 73–77 (2017)
4. Madsen, R.E., Kauchak, D., Elkan, C.: Modeling word burstiness using the dirichlet distribution. In: Proceedings of the 22nd international conference on Machine learning. pp. 545–552. ACM (2005)
5. Pennebaker, J.W., Mehl, M.R., Niederhoffer, K.G.: Psychological aspects of natural language use: Our words, our selves. Annual review of psychology **54**(1), 547–577 (2003)
6. Spolsky, B., Widdowson, H., et al.: Sociolinguistics, vol. 1. Oxford University Press (1998)
7. Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., Daelemans, W., Hoste, V.: Automatic detection and prevention of cyberbullying. In: International Conference on Human and Social Analytics (HUSO 2015). pp. 13–18. IARIA (2015)
8. Yenala, H., Jhanwar, A., Chinnakotla, M.K., Goyal, J.: Deep learning for detecting inappropriate content in text. International Journal of Data Science and Analytics **6**(4), 273–286 (2018)