Nikolaj Tollenaar

Research and Data Centre (WODC), Den Haag, the Netherlands
Utrecht university, Utrecht, the Netherlands

Multilabel text classification of police registrations for cyber- and digitized crime: a comparison of classical machine learning with deep learning

1 Background

Crime victim surveys report substantial rates of online victimization, while police-recorded cyber and digitized crimes are rare. Such crime is not properly recorded in the police registration in the Netherlands, but also internationally [1]. Free text fields in the police registration do however contain information about cybercrime. For operational and or police policy purpose, typically text queries applying keywords are used to obtain additional quantified information on cybercrime. However, this is an inefficient method for information retrieval, as you have to pre-specify keywords and leads to a significant portion of false positives.

2 Study objectives

The aim of this study was to arrive at estimate the number of police-recorded cyber and digitized crime and to describe characteristics of suspects by using a multilabel machine learning (ML) classifier on bag-of-words, metatextual features and NLP-feature data. Moreover, we wanted to see whether deep learning could yield an advantage over classical machine learning. We focused on cybercrime (hacking, ransomware and DDoS attacks) and digitized crime (online threats, stalking, online libel, online identity fraud and online buying and selling fraud).

3 Method

3.1 Sample

A random sample of 7.439 from 156.829 query-selected police incident texts was acquired from the total police data base of 2016 (*N*=3,945,964), to mitigate imbalance in the cyber- and digitized crime labels for training a text classification model. These data were manually labelled by human annotators. Of these, 400 registrations were scored by two annotators to ascertain the interrater reliability for each label separately. The registrations were split into training and testing data using a 60/40 ratio stratified splitting.

3.2 Feature construction

The data was pre-processed to obtain a term-feature matrix with lemma unigrams and bigrams and NLP features. On the basis of constructing the NLP features was the integrated suite of language modules frog for the Dutch language [2], that includes tokenization, lemmatization, part-of-speech tagging, syntactic parsing and named entity recognition. The NLP features consisted of the following derived feature classes:

- Lexicographic features (p = 27,292), consisting of lemma uni- and bigrams;
- Meta-textual features (p = 27), like text/sentence length, number of sentences [3][4];
- Syntactic NLP features (p = 3,996), including part-of-speech uni- and bigrams and lemma bigrams based on syntactic proximity;
- Semantic NLP features (p = 4,374), like synonym set unigrams where lemma's are replaced by their set of all synonyms, combinations of synonymset – lemma bigrams based on syntactic proximity.

Too frequent and too infrequent terms were filtered out to keep the size of the document-feature matrix manageable. No feature selection was performed, as we only used embedded feature selection methods in the modelling phase.

3.3 Machine Learning Models

The following classical algorithms were used for obtaining a text classification model:

- Multivariate random forests [5]
- Classifier chains [6][7] of the following algorithms:
  — L_1-penalized logistic regression [8], with automatic penalty selection
  — Random forests [9]
  — Stochastic gradient boosting [10], with separate tuning per label

For deep learning, a fully connected feed forward deep neural network (DNN) with an output node per label was used. The model was tuned by varying the number of layers, nodes in each layer, the activation functions, amount of dropout regularization and number of epochs using a validation split by iterative stratified splitting in an 80/20 ratio.

4 Key results

The labels that had the lowest interrater reliability were hardest to predict for every model. These were the digitized crimes like online threats. For the goal of identifying suspects of cyber and digitized crime, the model should perform well on both the precision and recall. However, only three of the eight labels had a sufficiently high precision and recall: hacking, ransomware and online purchase/selling fraud.

Of the classical machine learning models, the best performance was attained by a classifier chain of random forests. The DNN model attained a similar predictive performance on every separate label. However, the DNN model did substantially better on multilabel fit criteria, like the multilabel accuracy and the hamming loss. This result suggests that the DNN is superior in predicting label patterns to classical multilabel machine learning.

Tf-idf weighting of the lemma counts and NLP-feature counts had a negative effect on classification performance for all models, compared to raw counts. The use of meta-textual features based on NLP provided only a slight improvement of classification performance for most of the models.

References

[1] McGuire, M., & Dowling, S. (2013). Cyber crime: A review of the evidence. Summary of key findings and implications. Home Office Research report, 75, 1-35.

[2] Van den Bosch, A., Busser, G.J., Daelemans, W., and Canisius, S. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch, In F. van Eynde, P. Dirix, I. Schuurman, and V. Vandeghinste (Eds.), *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, Leuven, Belgium, pp. 99-114.

[3] Zhang, C., Wu, W., Niu, Z., & Ding, W. (2014). Autorship identification from unstructured texts. *Knowledge-Based Systems, 66,* 99-111.

[4] van der Lee, C., & van den Bosch, A. (2017). Exploring lexical and syntactic features for language variety identification. In *Proceedings of the fourth workshop on NLP for similar languages, varieties and dialects (VarDial)* (pp. 190-199).

[5] Cheng, W., Hüllermeier, E., & Dembczynski, K. J. (2010). Bayes optimal multilabel classification via probabilistic classifier chains. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (pp. 279-286), Haifa, Israel.

[6] Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2011). Classifier chains for multi-label classification. *Machine learning*, *85*(3), 333-359.

[7] Segal, M., & Xiao, Y. (2011). Multivariate random forests. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1*(1), 80-87.

[8] Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society Series B-Methodological, 58*(1), 267-288.

[9] Breiman, L. (1999). *Random forests*. Berkeley, CA: University of California,Statistics Department.

[10] Friedman J.H. (1999). Stochastic gradient boosting. *Computational statistics and data analysis*, 38(4), 367-378.