

A Framework for Sexism Detection on Social Media via ByT5 and TabNet

Arjumand Younus¹, Muhammad Atif Qureshi¹

¹Technological University Dublin, Ireland

Abstract

Hateful and offensive content on social media platforms particularly content directed towards a specific gender is a great impediment towards equality, diversity and inclusion. Social media platforms are facing increasing pressure to work towards regulation of such content; and this has directed researchers in text mining to work towards hate speech identification algorithms. One such attempt is sexism detection for which mostly transformer-based text methods have been proposed. We propose a combination of byte-level model ByT5 with tabular modeling via TabNet that has at its core an ability to take into account platform and language aspects of the challenging task of sexism detection. Despite not performing well in the sexism detection task for IberLEF our approach shows promise for future research in the area.

Keywords

token-free, tabular, ByT5, TabNet

1. Introduction

Microblogs have emerged as a popular mechanism for expressing various opinions and as a valuable forum for freedom of expression. Sometimes, however these freedoms assume a toxic nature to the point of targeting a particular community, race and/or gender [1]. It is for these reasons that there is a mounting pressure on social media service providers to regulate content and improve hate speech detection algorithms within their platforms [2].

Among various hateful content, sexism is a type that is extremely challenging to deal with [3, 4]. The main aim of sexist is to spread prejudice against women, and act offensively towards them. The following characteristics of sexist content on social media is what makes it extremely difficult to detect:

- Its expressed in a subtle form with various microaggressions embedded in the language.
- There's a lot of context involved often in the form of an ongoing popular event or conversation.
- The platform or language used also plays a role in cultural positioning of a stance as sexist or non-sexist.

The challenges listed above have implied successful deployment of deep learning systems within, and most previous works have successfully built on these approaches [5]. Most deep


IberLEF'22, September 2022, A Coruña, Spain.

✉ arjumand.younus@tudublin.ie (A. Younus); muhammadatif.qureshi@tudublin.ie (M. A. Qureshi)

🌐 <https://arjumandyounus.github.io/> (A. Younus); <https://matifq.github.io/> (M. A. Qureshi)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

learning systems however ignore the various idiosyncrasies of the platform or language itself when proposing a classification system. Moreover when it comes to multilingual data expressions on a platform such as Twitter the chance of data being noisy is immensely high and it is for this reason that traditional language models need extra tuning.

Our framework proposes a combination of ByT5 which is a token-free language model with an attention-based neural network over tabular data. Our underlying intuition lies in the notion of taking into account language and platform dependencies in a unified framework which may yield rich representations for sexism detection datasets that have more detail. Our results for the EXIST task [6] despite being poor show promise towards future directions in the area.

The remainder of this paper is organized as follows. In Section 2, we present an overview of related work. In Section 3, we present details of our framework. In Section 4, we present a discussion of our results with some concluding remarks.

2. Related Work

Machine learning in all its glory has enabled tremendous progress in text classification tasks; and similarly detecting sexism in social media texts also falls in this area. Within this area researchers have directed efforts in two directions i.e. collection of datasets for sexism detection [7] [8] and identifying the presence of a sexist agenda within a piece of text[9]. Within the second class of approaches, the most successful ones have been those that rely on deep learning methods such as fastText, RNN and CNN [10] or some ensemble of these [11]. Finally transformed-based models on top of BERT have also exhibited a good performance [12]. Our work is uniquely positioned in terms of taking a non-traditional setup towards how sexist data is utilized whereby richer row-level information can be incorporated.

3. Methodology

Before delving into our proposed framework we present an overview of the individual components of our framework i.e. ByT5 and TabNet. Each of these serves a completely different purpose, and we combine them in order to work towards richer forms of data when it comes to text classification in social media.

3.1. ByT5: A Pre-Trained Byte-to-Byte Model

ByT5 is essentially a byte-level model and relies on a small vocabulary of size around 256 (with few special IDs). The fundamental innovation within ByT5 is the tokenization process whereby it is reduced to simple encoding from string to bytes [13]. Other aspects of ByT5 architecture are similar to mT5 [14], but the split between encoder and decoder layers are not half; and in fact the number of encoder layers is three times more than the decoders. This is done to ensure deeper encoder stacks to make up for the decreased embedding capacity for the vocabulary. In short ByT5 has the capability to move text classification tasks to a more language-independent approach which has been our intuition in this work.

3.2. TabNet: Attentive Interpretable Tabular Learning

TabNet is essentially designed for tabular data; and its underlying principle is encoding raw data representations into meaningful representations via neural network architectures. Despite the data for the sexism detection task within EXIST comprising text there are some aspects of the data that can be modeled in a tabular format as we explain in the next subsection.

TabNet [15] uses a kind of soft feature selection and by means of that performs instance-wise feature selection i.e. at the level of a row. This feature selection step is accomplished through a sequential multi-step decision mechanism. That is, the input information is processed top-down in several steps. The sequential attention is performed via blocks of transformers while the final soft feature selection uses sparsemax function.

3.3. Combination of ByT5 and TabNet

The various techniques for token-free models such as ByT5 demonstrate the potential of standardizing model configurations across modalities, tasks, and languages. It is these configurations we aim to combine by means of making use of different data-driven aspects

Figure 1 presents an overview of our mechanism. In the first step the textual content of the tweet/gab itself is passed on to ByT5 model for training whereby a label on whether or not it is sexist is obtained. This is then fed into the tabular structure which contains other information about a tweet/gab such as the language of the tweet (Spanish or English), source of the tweet (i.e. the platform from which we obtain it and for this particular dataset the two platforms are Gab and Twitter), and readability score of the tweet obtained from textstat library; it can be accessed at <https://github.com/shivam5992/textstat>. All this information is passed to TabNet from which final prediction outcomes are obtained.

It is worth mentioning that while our method to combine the two forms of data is simple there can be more ways to combine these two using ensembles of decision trees or some other combination of neural networks. Furthermore, since tabular data has the capability to deal with various forms of numeric and categorical data representations using our method can be enriched with more information towards a better performance for sexist content detection.

4. Results and Further Experiments

The second shared task on sEXism Identification in Social neTworks (EXIST) at IberLEF 2022 is an international competition in the field of Natural Language Processing (NLP) with the aim to automatically identify sexism in social media content by applying machine learning methods. Two tasks were specifically defined with the first being binary classification and the second being fine-grained classification that distinguishes multiple types of sexist content (e.g., dominance, stereotyping, and objectification). Our team xaiTUD, explainable AI group based in TU Dublin participated in binary classification task, and obtained an f1 score of 46%.

Due to its byte-level nature, the ByT5 model is slower to compute on account of fine-grained tokenization producing more tokens for the same text and finally requiring more time for the model to digest. It is for this reason that before the submission on the test data we were able to complete our ByT5 runs with two epochs which led to an overall poor performance. Table 1

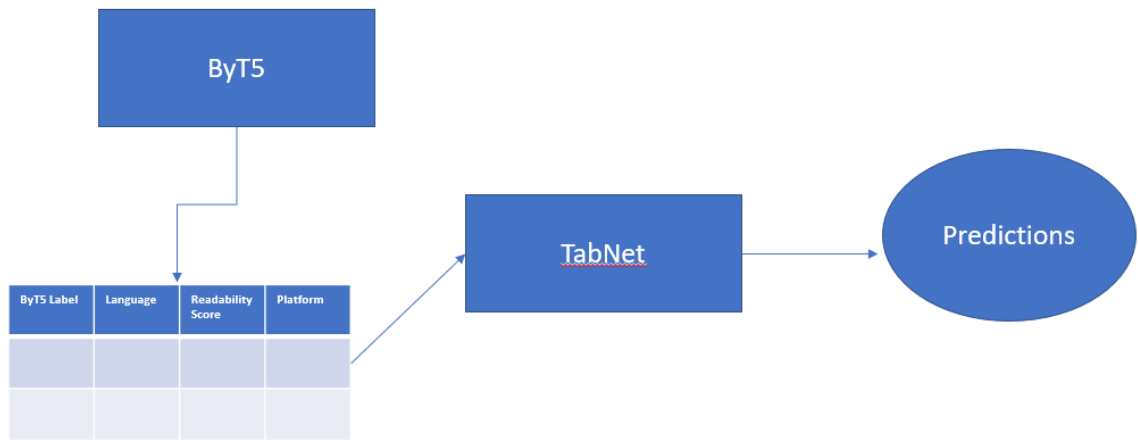


Figure 1: Flow Diagram for Combination of ByT5 and TabNet

Epochs for ByT5	CV Accuracy
2	49%
10	52%
25	54%
50	61%

Table 1

Cross-Validation Accuracy with Epochs for ByT5

shows the cross-validation accuracy for the various epoch settings for ByT5; note that the final cross-validation accuracy is shown for TabNet output.

5. Discussion and Conclusion

As can be seen from Table 1 the epochs play a huge role in improving the performance of our model which is essentially on account of ByT5 learning a cleaner and finer representation as the model continues to run. Moreover, our model shows promise on account of its ability to include rich categorical and numeric data into the prediction process via TabNet. In future work our aim is to explore the use of these approach for different multilingual classification tasks.

References

- [1] Z. Zhang, L. Luo, Hate speech detection: A solved problem? the challenging case of long tail on twitter, *Semantic Web* 10 (2019) 925–945.
- [2] E. Barendt, What is the harm of hate speech?, *Ethical Theory and Moral Practice* 22 (2019) 539–553.
- [3] W. Yin, A. Zubiaga, Towards generalisable hate speech detection: a review on obstacles and solutions, *PeerJ Computer Science* 7 (2021) e598.
- [4] P. Chiril, V. Moriceau, F. Benamara, A. Mari, G. Origgi, M. Coulomb-Gully, An annotated corpus for sexism detection in french tweets, in: *Proceedings of the 12th language resources and evaluation conference*, 2020, pp. 1397–1403.
- [5] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, Automatic classification of sexism in social networks: An empirical study on twitter data, *IEEE Access* 8 (2020) 219563–219576.
- [6] F. Rodriguez-Sanchez, J. Carrillo-de Albornoz, A. Plaza, Laura Mendieta-Aragon, G. Marco-Remon, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2022: sexism identification in social networks., in: *Procesamiento del Lenguaje Natural*, vol 69, 2022.
- [7] N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, D.-Y. Yeung, Multilingual and multi-aspect hate speech analysis, *arXiv preprint arXiv:1908.11049* (2019).
- [8] O. El Ansari, Z. Jihad, M. Hajar, A dataset to support sexist content detection in arabic text, in: *International Conference on Image and Signal Processing*, Springer, 2020, pp. 130–137.
- [9] P. Parikh, H. Abburi, P. Badjatiya, R. Krishnan, N. Chhaya, M. Gupta, V. Varma, Multi-label categorization of accounts of sexism using a neural framework, *arXiv preprint arXiv:1910.04602* (2019).
- [10] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep learning for hate speech detection in tweets, in: *Proceedings of the 26th international conference on World Wide Web companion*, 2017, pp. 759–760.
- [11] S. Zimmerman, U. Kruschwitz, C. Fox, Improving hate speech detection with deep learning ensembles, in: *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018.
- [12] S. Butt, N. Ashraf, G. Sidorov, A. Gelbukh, Sexism identification using bert and data augmentation-exist2021, in: *International Conference of the Spanish Society for Natural Language Processing SEPLN 2021, IberLEF 2021*, 2021.
- [13] L. Xue, A. Barua, N. Constant, R. Al-Rfou, S. Narang, M. Kale, A. Roberts, C. Raffel, Byt5: Towards a token-free future with pre-trained byte-to-byte models, *Transactions of the Association for Computational Linguistics* 10 (2022) 291–306.
- [14] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A massively multilingual pre-trained text-to-text transformer, *arXiv preprint arXiv:2010.11934* (2020).
- [15] S. O. Arik, T. Pfister, Tabnet: Attentive interpretable tabular learning, in: *AAAI*, volume 35, 2021, pp. 6679–6687.