UDC 340.14:004.8:519.6

LITHUANIAN HATE SPEECH CLASSIFICATION USING DEEP LEARNING METHODS

Eglė Kankevičiūtė[1], Milita Songailaitė[2], Bohdan Zhyhun[3], Justina Mandravickaitė[4]
[1,2,3,4]Vytautas Magnus University, Kaunas, Lithuania
ORCID: https://orcid.org/0000-0002-9352-9709[1], https://orcid.org/0000-0003-4315-2461[2], https://orcid.org/0000-0002-1159-2975[3], https://orcid.org/0000-0001-9426-6165[4]
E-mail: egle.kankeviciute@vdu.lt[1], milita.songailaite@vdu.lt[2], bohdan.zhyhun@vdu.lt[3], justina.mandravickaite@vdu.lt[4]

DOI: 10.15673/atbp.v15i3.2621

Abstarct. *The ever-increasing amount of online content and the opportunity for everyone to express their opinions online leads to frequent encounters with social problems: bullying, insults, and hate speech. Some online portals are taking steps to stop this, such as no longer allowing user-generated comments to be made anonymously, removing the possibility to comment under the articles, and some portals employ moderators who identify and eliminate hate speech. However, given the large number of comments, an appropriately large number of people are required to do this work. The rapid development of artificial intelligence in the language technology area may be the solution to this problem. Automated hate speech detection would allow to manage the ever-increasing amount of online content, therefore we report hate speech classification for Lithuanian language by application of deep learning.*

Анотація. *Постійно зростаюча кількість онлайн-контенту та можливості для кожного висловити свою думку в Інтернеті призводять до частих зустрічей із соціальними проблемами: залякуванням, образами та ворожнечею. Деякі онлайн-портали вживають заходів, щоб зупинити це, наприклад, більше не дозволяють анонімно створювати коментарі користувачів, усувають можливість коментувати під статтями, а деякі портали наймають модераторів, які виявляють і усувають мову ненависті. Однак, враховуючи велику кількість коментарів, для виконання цієї роботи потрібна відповідна кількість людей. Вирішенням цієї проблеми може стати стрімкий розвиток штучного інтелекту в області мовних технологій. Автоматизоване виявлення мови ворожнечі дозволить керувати постійно зростаючою кількістю онлайн-контенту, тому ми повідомляємо про класифікацію мови ворожнечі для литовської мови за допомогою глибокого навчання.*

Keywords: Deep Learning, Transformers, Hate Speech, Text Classification
Ключові слова: глибоке навчання, трансформери, мова ворожнечі, класифікація тексту.

1. Introduction

Online hate speech (HS) has become a threat because of violence based on racial, ethnic, sexual, social, and religious distinctions which occur due to the ever-growing volume of online content and the capacity of every user to voice their opinions through comments or other means of expression. Controlling the spread of HS, especially on social networks, is notoriously challenging. It is particularly troublesome since there is debate over how to monitor the content that contains HS, including whether it should be flagged, removed, or otherwise managed [1]. However, it has been demonstrated that HS offenses can result in hate crimes [2], therefore limiting HS should be a priority. Automatic hate speech recognition tools could facilitate this task. There are many distinct hate targets, therefore different languages offer conditions for comparison, revealing variations in linguistic and communicative acts against these targets, allowing HS detection systems to be used in a number of settings [3].

In this paper we present hate speech detection experiments for Lithuanian, where we fine-tune and compare *Multilingual* BERT, *LitLat* BERT and *Electra* models. Section 2 briefly describes relevant literature, definitions and existing challenges. Section 3 presents data, research methods and experimental setup. Section 4 shows the results of fine-tuning and comparison of selected pre-trained models. Section 5 ends this paper with conclusions.

2. Literature review on HS detection

As HS is a complex and non-trivial phenomenon, it is difficult to detect. Researchers contribute to the detection of this phenomenon by designing frameworks, annotating corpora, extracting meaningful features, and testing automatic classifiers. Also, a few evaluation tasks for HS detection in different languages resulted in released benchmark corpora to encourage further developments in HS detection as it may aid in confronting escalation of online violence and hatred or the spread of fake news [4]. Online HS is suspected to be an important factor in political and ethnic violence such as the Rohingya crisis in Myanmar [5], [6]. Therefore, media platforms are pressured to timely detection and elimination of HS as well as related phenomena, such as cyber-bullying and offensive content [7]. Researchers contributed to a considerable amount of work on HS detection, e.g. as in [8]–[13]. However, most of it is based on hand-crafted features, user information and/or a variety of

metadata which is usually platform-specific [5] and that limits HS detection generalization in terms of new data sets as well as new data sources.

To contribute to HS detection, a number of shared tasks have been organized [14]–[17]. Each of them concentrated on different aspects of HS. For example, [9] introduced a typology on the abusive nature of HS, distinguishing it into generalized, explicit, and implicit abuse. Meanwhile, [17] studied hateful and aggressive messages which targeted women and immigrants. Furthermore, [14] explored the identification of targeted and untargeted insults, therefore proposing the classification of HS into hateful, offensive and profane. Moreover, [15] examined aggression and misogynistic content identification in terms of trolling and cyberbullying. Finally, [18] concluded that most of the shared tasks cover individual-directed abuse, identity-directed abuse, and concept-directed abuse.

As for language coverage in HS detection, English is by far the most investigated as initial work in this area started with English [19]. However, with the advances in natural language processing (NLP) and deep learning, non-English HS detection solutions steadily increase. Therefore, there exist HS detection models for Arabic (e.g., [20]–[24]), Turkish (e.g., [24], [25]), Greek (e.g., [24], [26], [27], Danish (e.g., [24], [28], Hindi (e.g., [29]–[31]), German (e.g., [15], [32]), Malayalam (e.g., [33], [34]), Tamil (e.g., [34], [35]), Chinese (e.g., [36]–[38]), Italian (e.g., [39]), Urdu (e.g., [40]–[42]), Bengali (e.g., [43]–[45]), Korean (e.g., [46]), French (e.g., [47]–[49]), Indonesian and Portugese (e.g., [50]), Spanish (e.g., [51]), Polish (e.g., [52]) and some others as well, not mentioned due to limits of scope of this article.

Regarding challenges in HS detection, one of them is dataset accessibility and availability. A common problem in this is that datasets, available publicly, become unavailable after some time due to a variety of reasons. To this problem data degradation issue comes hand in hand, i.e., when a dataset, published in some encrypted format, needs to be regenerated on-demand, and after such generation, this dataset does not produce the same volume of data as reported on the publication [53]. This can happen when, e.g., Twitter data is released in the form of tweetIDs and the user deletes the original account or tweets.

Another challenge is class imbalance issue as the hate class in most cases makes less than 12% for the multi-class datasets and for the binary datasets – much less than the preferable half of the total dataset [53]. This dataset structure is one of the causes of lesser accuracy in HS detection. Also, a variety of HS definitions may raise a challenge. HS belongs to a set of related concepts, such as abusiveness, aggressiveness, racism, etc. See the detailed representation of these related concepts in Fig. 1.

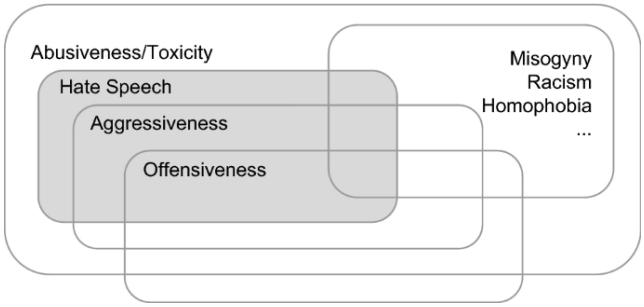

Fig. 1 – Relations between HS and related concepts [4]

However, even in the variety of HS definitions, there were consistencies among them. For example, [54] have analysed available definitions of HS and identified the following similarities:

- HS has a target;
- HS induces violence or hate;
- HS attacks or demeans;
- HS can have some types of humour, such as sarcasm.

A variety of definitions assumes that there are differences in perception of HS. Therefore, existing datasets are affected by these varied definitions because they are annotated based on them, and similar instances can be assigned to different annotation categories based on these differences. For example, [55] investigated the effects of definition on annotation reliability. They concluded that HS requires a stronger, more uniform definition. Similarly, [54] found that most of the publicly available datasets are incompatible due to different definitions attributed to similar concepts.

Also, HS datasets occasionally have very similar labels and some studies merge some of them together into one class (usually the purpose is to reduce class imbalance) [53]. However, this practice could have a negative impact on research as the distinction between classes is necessary. For example, this happened for [12] dataset with offensive and hate classes and [56] – with racist and sexist classes. Classes in the former dataset were merged in [7] and [57] where they merged hate and offensive classes into one class. Meanwhile, [58] merged the offensive and neither class into a non-hate class. Similarly, [59] and [60] merged classes in the dataset introduced in [56]. In HS studies abusive language or toxic comments can surround several paradigms [53], therefore it was proposed to use the terms strictly following the available definitions. Similarly, it was suggested that offensive language is not the same as HS and should not be merged [12].

Finally, methods for detecting HS and related abusive behaviour have become popular and getting better in terms of

performance and generalization [61], [62]. However, current state-of-the-art solutions still have their limitations in accuracy and therefore their practical real-time applications are restricted [63]. HS detection is still an extremely difficult task, especially when the expression of hate is implicit [64].

3. Data, methods and experimental setup

3.1. Annotated Lithuanian Hate Speech Corpus

About 60,000 user-generated comments from various news portals (15min.lt, alkas.lt, delfi.lt) were collected to create a solution for recognizing HS in Lithuanian. They were also supplemented with 226,776 user-generated comments from the news portal lrytas.lt and thousands of manually collected HS comments from various social media pages and news portals. The latter comments have been collected by a focused search. A total of 25,219 comments were annotated by four annotators. The annotation scheme consisted of 3 classes (tags): neutral language, offensive language and hate speech.

The data in the collected dataset was not evenly distributed. Most user-generated comments were *non-hate* (neutral), which accounts for 60.7 percent of the total dataset. The lesser part is the category *offensive* (offensive language), which is 31 percent, and the smallest part is the category *hate* (hate speech), which accounts for only 8.26 percent of all annotated data (see Fig. 2).

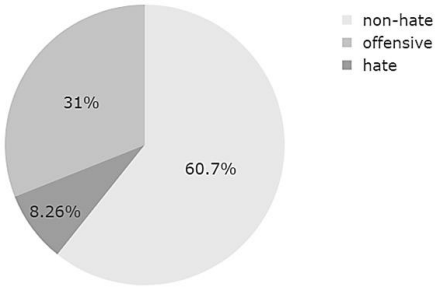

Fig. 2 – Distribution of classes in annotated text

There were exceptional cases when marking comments with manifestations of HS, for example, the content of the comment was racist, insulting and inciting hatred, but there were no clearly expressed words, there was a lack of context, for example, the comment: "on fire them!", taken from a topic about women of another nationality, but without context, it can simply be assigned to a neutral class. Another exceptional case was where the comment has a racist, hateful meaning, but is expressed in a figurative sense. To determine the class of such a user-generated comment requires a background of social knowledge. All four annotators decided together what category to assign this type of comment to.

There were also comments which did not correspond to any of the three annotated categories, for example, when the content of the comment consisted only of an internet link, name or nickname, various symbols and emoticons. Such comments were skipped by assigning them the *skip* category. This category also included user-generated comments which the annotators could not agree on their decisions.

3.2. HS detection methodology

For HS detection we used three popular deep learning models: *Multilingual* BERT, *LitLat* BERT and *Electra*. These artificial intelligence models were trained to work with Lithuanian language data. All three models were further trained to classify Lithuanian user-generated comments, i.e., detect those comments that may contain HS. The selected models have been briefly introduced in the following subsections.

3.3.1. BERT architecture models

Transformer BERT (Bidirectional Encoder Representations from Transformers) is a deep learning model based on the attention mechanism [65], which is usually applied to solve various language technology problems [66]. This model works on the principles of transfer learning [67]. A neural network is trained to generate word embeddings, which are then used as input functions for models that solve mainstream language technology tasks. One of the most significant advantages of the BERT architecture models over other neural network models is understanding the context between words in the text. The model learns the context using the attention mechanism characteristic of transformer models, which consists of encoding and decoding mechanisms [65]. In our case, two BERT models were used in developing the hate speech detection solution: *Multilingual* BERT and *LitLat* BERT.

***Multilingual* BERT** model uses the architecture of the BERT model and is trained by the Google team using 104 different languages, including Lithuanian [68]. The model was trained with various texts from Wikipedia. These texts were not annotated, or otherwise processed [69] for the training process.

***LitLat* BERT** is a trilingual model which was built using the XLM-RoBERTa-base (Robustly optimized) multilingual model architecture [70]) and trained on Lithuanian, Latvian, and English data. According to the scientific literature, *LitLat* performs better than *Multilingual* BERT [71], since in this case, BERT focuses on only three languages.

3.3.2. *Electra* transformer

Electra is a transformer model that uses a pre-training method that trains two neural network models: a generator and a discriminator. The purpose of the generator is to replace lexemes in a sequence, so it is trained as a masked language

model. Meanwhile, the discriminator tries to determine which lexemes have been replaced by the generator [72] (see Fig. 3). The generator can be any language model that produces an output that is a distribution of lexemes. However, the most common choice is a Masked Learning model trained along with a discriminator. After initial training, the generator is discarded, and only the discriminator (*Electra* model) is refined in subsequent tasks [73]. This proposed training method is significantly more efficient than the masked training method used in BERT models. This is why the *Electra* model requires less data and computer resources for training.

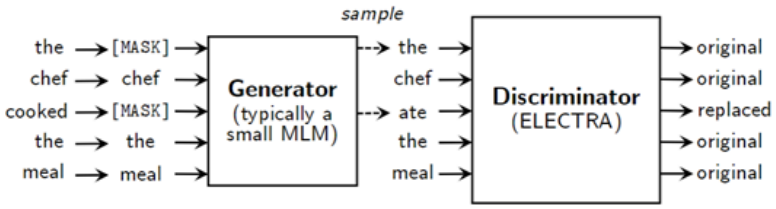

Fig. 3 – An example of detecting a modified lexeme when using *Electra* model [73]

The main disadvantage of this model, compared to the previously reviewed models *Multilingual* and *LitLat* BERT, is that there has been no pre-trained *Electra* model for the Lithuanian language. This means that we needed to use as many collected Lithuanian texts as possible in order to pre-train the model ourselves. Nevertheless, for training this model from scratch we needed fewer computer resources than we would have needed for training BERT models.

3.3.3. Fine-tuning the models for the classification task

All three embedding models were additionally trained (fine-tuned) on an annotated dataset of Lithuanian user-generated comments. This dataset consisted of 25 219 comments, annotated into the three classes mentioned earlier:

1. *Hate speech* (2 082 comments);
2. *Neutral language* (15 316 comments);
3. *Offensive language* (7 821 comments.

The datasets were divided into training, validation, and testing sets by the ratio 0.6:0.2:0.2. Comments for the hateful and abusive language classes were replicated (duplicated) in each of the subsets to compare the number of user-generated comments in each class. Since the generated embeddings were vectors of length 512 or 128 symbols, any comments longer than 512 characters were discarded. The BERT models were trained for ten epochs and the *Electra* model – for 30 epochs. The learning rate of the models was 1e-3, and the *Adam* optimizer [74] was used to optimize the network weights.

3.4. Implementation of HS detection methodology

The HS detection methodology was implemented by creating a prototype that detects HS in user-generated comments in Lithuanian news and social media. The model we selected for developing a prototype was the *LitLat* transformer, which classifies the comments into three classes: *hate*, *neutral* or *offensive* language. Upon detecting user-generated comments marked with hateful or offensive language, a model built into a particular system will be able to alert the administrators and let them know that the corresponding user-generated comment is inappropriate. Then the system administrators will have the opportunity to remove the comment or to deal with it differently in order to monitor that such user-generated comments would not exist in the system.

The workflow of the prototype is presented in Fig. 4. In the first stage of the prototype, data extraction takes place, during which user-generated comments are collected from various Lithuanian news portals and social networking sites. Then, the collected comment texts are processed in the prototype's second stage. At this stage, the text is processed in the exact same way as when preparing user-generated comments for model training. The third stage is the classification of processed comments. Here, the model integrated into the prototype classifies user-generated comments into three classes described earlier. Finally, in the fourth step, the results generated by the prototype are obtained. The result is a list of classified comments, where one can see user-generated comments that may contain HS.

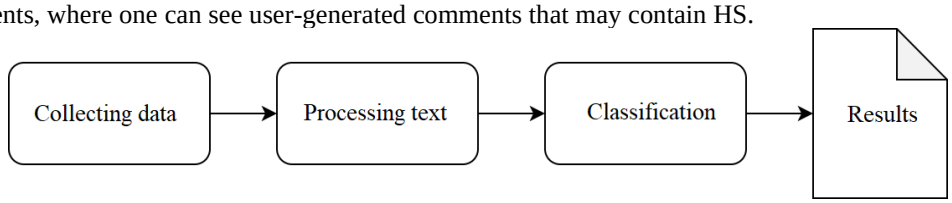

Fig. 4 – Prototype workflow for detecting HS in Lithuanian in user-generated comments published in news portals and social media

Since the prototype is adapted for integration into news portals and social networking sites, no graphical user interface was developed. The integrated prototype is expected to work at the code level on the servers. The results of the prototype are meant to be reviewed by the admins of these sites and further dealt with according to their needs.

4. Results

During the experiments, the models were trained to classify annotated texts. The performance of these models was evaluated using accuracy, precision, recall, and F1-score metrics. Each classification model was trained three times with

different random number generation parameters (Seed). Once the models have been trained, the aim was to select the model with the highest accuracy estimate to be used for the next stage of testing. This method of selecting trained models avoids randomly obtaining sub-optimal initial network weights, which can severely compromise the classification results during training [75].

First, the results of the three *LitLat* BERT test runs are reviewed. For training this model, 10 epochs were used, which was sufficient to train the models for the classification task. In the graph below (see Fig. 5) we can see that the best performance of the *LitLat* BERT model was achieved in the third trial, where F1-score at epochs 2 and 3 reached almost 71%. However, the completeness result shows the opposite. In comparison to the other curves, the completeness curve is the lowest in the third trial, which means a worse estimate, but still an estimate higher than 0.7. Since F1-score combines precision and recall, the third test run of *LitLat* base_986 was chosen for further testing based on the result of this metric.

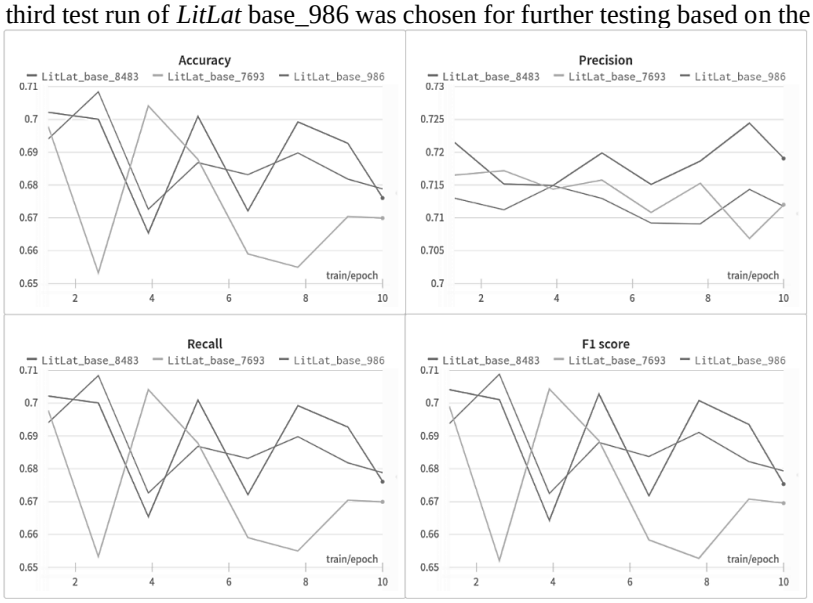

Fig. 5 – Classification accuracy (top left), precision (top right), recall (bottom left) and F1 score (bottom right) values for all *LitLat* BERT model tests

Meanwhile, when training the *Multilingual* BERT model, the best F1-score is seen in the first test run, which reached almost 0.66 at epochs 5-6 (see Fig. 6). The first test run resulted in a higher score in either accuracy or recall. As far as the precision is concerned, the best result was achieved in the second test run, which reached an estimate of 0.69. However, considering the overall results, the first test run, *Multilingual*_base_7458, was chosen for further testing.

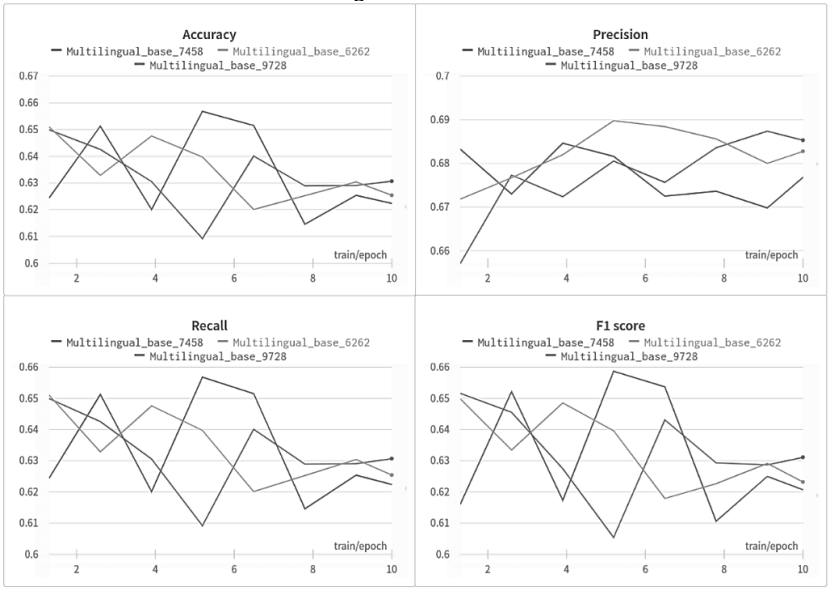

Fig. 6 – Classification accuracy (top left), precision (top right), recall (bottom left) and F1 score (bottom right) values for all *Multilingual* BERT model tests

The *Electra* model showed a slightly different trend in the curves (see Fig. 7). BERT models obtain high metric estimates from the very beginning, i.e., at epoch 1, in some cases even the best model estimate is reached at epoch 1 or 2 (for example, in the second test run *LitLat* BERT reached a peak of precision at epoch 1 and then the result only got

worse). In contrast, the *Electra* model started from the lowest estimate and with each epoch the result improved rapidly up to a certain threshold. For this reason, the *Electra* model needed more epochs for training (30 epochs were chosen) to reach a similar level as the BERT models. However, it is important to mention that the training time of the *Electra* model was shorter, so the increased number of epochs did not affect the time.

From the line graphs in Fig. 7 it is observed that the *Electra* model performed best in the first test. The first test achieved the highest scores in both accuracy, precision, recall and F1-score. However, we can see that neither accuracy nor F1-score curves for the first test reached an estimate of 0.54, which means that this model performed worse as compared to the BERT models.

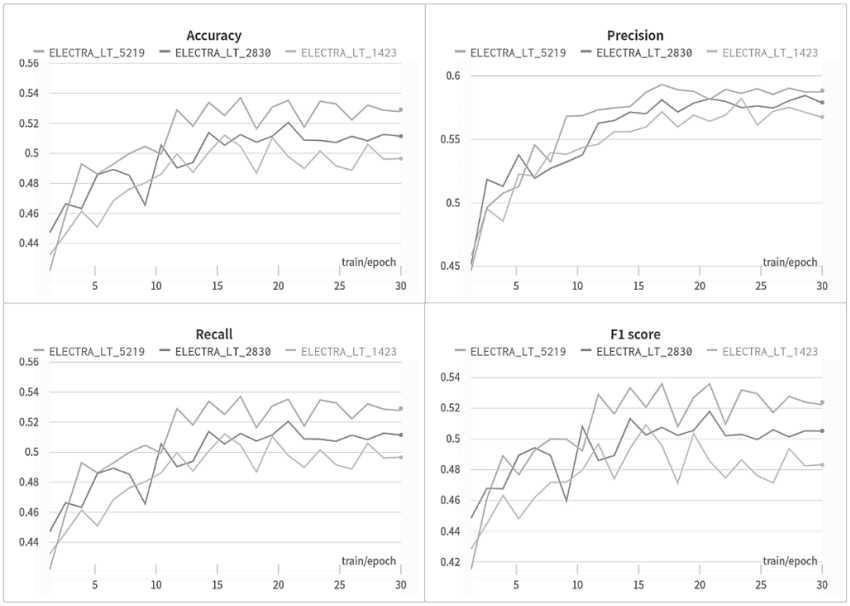

Fig. 7 – Classification accuracy (top left), precision (top right), recall (bottom left) and F1 score (bottom right) values for all *Electra* model tests.

Summary of all the model tests is presented in see Fig. 8. After plotting the first 10 epochs, we can see that the *Electra* model is still well behind the two BERT models in all test estimates. The peak of the *Electra* model estimates occured between epochs 10 and 20.

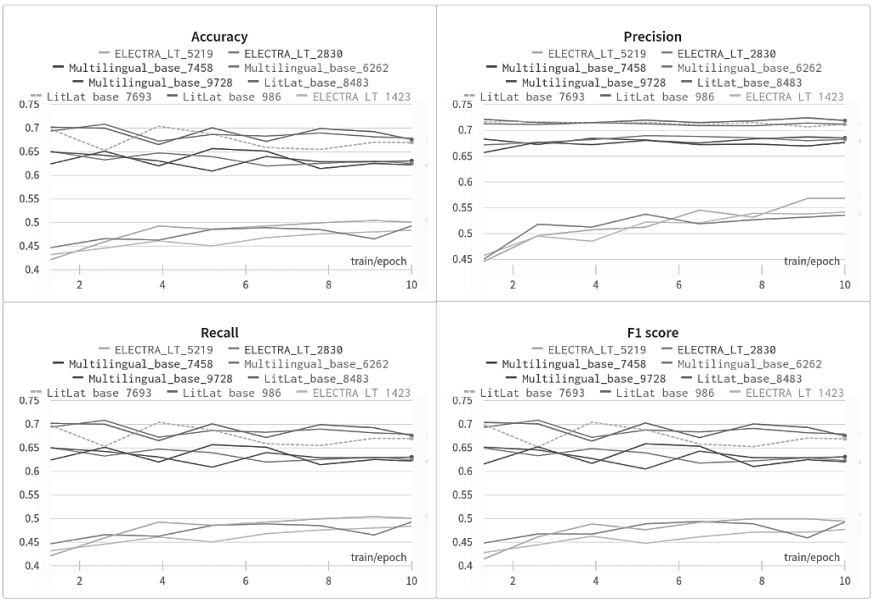

Fig. 8 – Classification accuracy (top left), precision (top right), recall (bottom left) and F1 score (bottom right) values for all models across all tests

Accuracy for all three models is shown in Fig. 9. We can see that the best performing model was the *LitLat* BERT, and the worst one was the *Electra* transformer. The reason why *Electra* model's accuracy is so low is that it was trained on only 70 million Lithuanian words. In comparison, *LitLat* BERT model was trained on 1.21 billion Lithuanian words. So, even though the structure of the *Electra* transformer allows for the model to be trained with smaller amounts of data,

the amount of data available was still insufficient to train the model accurately.

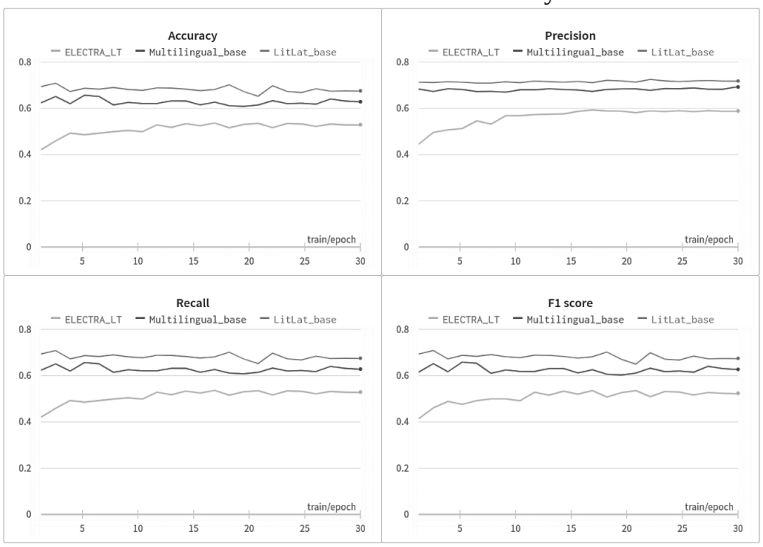

Fig. 9 – Classification evaluation curves for accuracy (top left), precision (top right), recall (bottom left) and F-score (bottom right) metrics.

Since accuracy testing of the models showed that the best HS detection model is *LitLat* BERT, we chose to use this model for the prototype of HS detection in user-generated comments published in Lithuanian news portals and social media.

5. Conclusions

1. An annotated corpus consisting of 25 219 user-generated comments has been developed for training and testing the hate speech recognition model. In this dataset, 8.26% of the comments contain hate speech, 31% of the comments contain offensive speech and 60.7% of the comments contain neutral speech.
2. Based on precision, recall, accuracy and F1-score, we found that the *LitLat* BERT model performed the best, with an accuracy of 72%, which is why it was chosen for prototype implementation.

6. References

[1] B. Kalsnes and K. A. Ihlebæk, “Hiding hate speech: political moderation on Facebook,” *Media, Culture and Society*, vol. 43, no. 2, pp. 326–342, 2021, doi: 10.1177/0163443720957562.

[2] J. C. Pereira-Kohatsu, L. Quijano-Sánchez, F. Liberatore, and M. Camacho-Collados, “Detecting and monitoring hate speech in twitter,” *Sensors (Switzerland)*, vol. 19, no. 21, pp. 1–37, 2019, doi: 10.3390/s19214654.

[3] C. A. Martinez, J.-W. van Prooijen, and P. A. M. Van Lange, “Hate: Toward understanding its distinctive features across interpersonal and intergroup targets.,” *Emotion*, vol. 22, no. 1, p. 46, 2022.

[4] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti, “Resources and benchmark corpora for hate speech detection: a systematic review,” *Language Resources and Evaluation*, vol. 55, no. 2, pp. 477–523, 2021, doi: 10.1007/s10579-020-09502-8.

[5] L. Yuan, T. Wang, G. Ferraro, H. Suominen, and M.-A. Rizoiu, “Transfer Learning for Hate Speech Detection in Social Media,” 2019.

[6] S. Stecklow, “Why Facebook is losing the war on hate speech in Myanmar,” 2018. https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/ (accessed Jan. 24, 2023).

[7] Z. Zhang, D. Robinson, and J. Tepper, “Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network,” in *The Semantic Web*, A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, and M. Alam, Eds., Cham: Springer International Publishing, 2018, pp. 745–760.

[8] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, “Antisocial behavior in online discussion communities,” *Proceedings of the 9th International Conference on Web and Social Media, ICWSM 2015*, pp. 61–70, 2015, doi: 10.1609/icwsm.v9i1.14583.

[9] Z. Waseem, “Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter,” in *Proceedings of the First Workshop on {NLP} and Computational Social Science*, Austin, Texas: Association for Computational Linguistics, 2016, pp. 138–142. doi: 10.18653/v1/W16-5618.

[10] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, “Mean birds: Detecting aggression and bullying on Twitter,” *WebSci 2017 - Proceedings of the 2017 ACM Web Science Conference*, pp. 13–22, 2017, doi: 10.1145/3091478.3091487.

[11] E. Wulczyn, N. Thain, and L. Dixon, “Ex machina: Personal attacks seen at scale,” *26th International World Wide Web Conference, WWW 2017*, pp. 1391–1399, 2017, doi: 10.1145/3038912.3052591.

[12] T. Davidson, D. Warmsley, M. Macy, and I. Weber, “Davidson,” *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, no. Icwsm, pp. 512–515, 2017.

[13] E. F. Unsvåg and B. Gambäck, "The Effects of User Features on Twitter Hate Speech Detection," *2nd Workshop on Abusive Language Online - Proceedings of the Workshop, co-located with EMNLP 2018*, no. 2012, pp. 75–85, 2018, doi: 10.18653/v1/w18-5110.
[14] S. Modha, T. Mandl, P. Majumder, and D. Patel, "Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in indo-european languages," *CEUR Workshop Proceedings*, vol. 2517, pp. 167–190, 2019.
[15] R. Kumar, B. Lahiri, A. K. Ojha, and A. Bansal, "ComMA@FIRE 2020: Exploring multilingual joint training across different classification tasks," *CEUR Workshop Proc*, vol. 2826, pp. 823–828, 2020.
[16] J. M. Struß, M. Siegel, J. Ruppenhofer, M. Wiegand, and M. Klenner, "Overview of GermEval task 2, 2019 shared task on the identification of offensive language," *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019*, pp. 354–365, 2020.
[17] V. Basile *et al.*, "SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter," *NAACL HLT 2019 - International Workshop on Semantic Evaluation, SemEval 2019, Proceedings of the 13th Workshop*, pp. 54–63, 2019, doi: 10.18653/v1/s19-2007.
[18] B. Vidgen, A. Harris, D. Nguyen, R. Tromble, S. Hale, and H. Margetts, "Challenges and frontiers in abusive content detection," no. Section 2, pp. 80–93, 2019, doi: 10.18653/v1/w19-3509.
[19] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying," *ACM Trans. Interact. Intell. Syst.*, vol. 2, no. 3, 2012, doi: 10.1145/2362394.2362400.
[20] H. Mubarak, K. Darwish, and W. Magdy, "Abusive Language Detection on {A}rabic Social Media," in *Proceedings of the First Workshop on Abusive Language Online*, Vancouver, BC, Canada: Association for Computational Linguistics, 2017, pp. 52–56. doi: 10.18653/v1/W17-3008.
[21] S. Hassan, Y. Samih, H. Mubarak, and A. Abdelali, "{ALT} at {S}em{E}val-2020 Task 12: {A}rabic and {E}nglish Offensive Language Identification in Social Media," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Barcelona (online): International Committee for Computational Linguistics, 2020, pp. 1891–1897. doi: 10.18653/v1/2020.semeval-1.249.
[22] H. Alami, S. O. El Alaoui, A. Benlahbib, and N. En-Nahnahi, "LISAC FSDM-USMBA Team at SemEval-2020 Task 12: Overcoming AraBERT's pretrain-finetune discrepancy for Arabic offensive language identification," *14th International Workshops on Semantic Evaluation, SemEval 2020 - co-located 28th International Conference on Computational Linguistics, COLING 2020, Proceedings*, pp. 2080–2085, 2020, doi: 10.18653/v1/2020.semeval-1.275.
[23] S. Sikora, B. Hurley, and A. G. Tharakan, "Automation with intelligence," *Deloitte*, p. 28, 2019.
[24] S. Wang, J. Liu, X. Ouyang, and Y. Sun, "Galileo at SemEval-2020 Task 12: Multi-lingual Learning for Offensive Language Identification using Pre-trained Language Models," Online, 2020.
[25] A. Ozdemir and R. Yeniterzi, "{SU}-{NLP} at {S}em{E}val-2020 Task 12: Offensive Language {I}dentifi{C}ation in {T}urkish Tweets," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Barcelona (online): International Committee for Computational Linguistics, 2020, pp. 2171–2176. doi: 10.18653/v1/2020.semeval-1.288.
[26] H. Ahn, J. Sun, C. Y. Park, and J. Seo, "NLPDove at SemEval-2020 Task 12: Improving Offensive Language Detection with Cross-lingual Transfer," *14th International Workshops on Semantic Evaluation, SemEval 2020 - co-located 28th International Conference on Computational Linguistics, COLING 2020, Proceedings*, pp. 1576–1586, 2020, doi: 10.18653/v1/2020.semeval-1.206.
[27] K. Socha, "{KS}@{LTH} at {S}em{E}val-2020 Task 12: Fine-tuning Multi- and Monolingual Transformer Models for Offensive Language Detection," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Barcelona (online): International Committee for Computational Linguistics, 2020, pp. 2045–2053. doi: 10.18653/v1/2020.semeval-1.270.
[28] M. Pàmies, E. Öhman, K. Kajava, and J. Tiedemann, "LT@Helsinki at SemEval-2020 Task 12: Multilingual or language-specific BERT?," *14th International Workshops on Semantic Evaluation, SemEval 2020 - co-located 28th International Conference on Computational Linguistics, COLING 2020, Proceedings*, pp. 1569–1575, 2020, doi: 10.18653/v1/2020.semeval-1.205.
[29] R. Raj, S. Srivastava, and S. Saumya, "NSIT & IIITDWD @ HASOC 2020: Deep learning model for hate-speech identification in Indo-European languages," *CEUR Workshop Proc*, vol. 2826, pp. 161–167, 2020.
[30] M. A. Bashar and R. Nayak, "QutNocturnal@HASOC'19: CNN for hate speech and offensive content identification in Hindi language," *CEUR Workshop Proceedings*, vol. 2517, no. December, pp. 237–245, 2019.
[31] P. Mishra, H. Yannakoudakis, and E. Shutova, "Tackling Online Abuse: A Survey of Automated Abuse Detection Methods," no. 2013, 2019.
[32] Q. Que, R. Sun, and S. Xie, "Simon@HASOC 2020: Detecting hate speech and offensive content in German language with BERT and ensembles," *CEUR Workshop Proceedings*, vol. 2826, pp. 283–289, 2020.
[33] S. Sai and Y. Sharma, "Siva@HASOC-Dravidian-CodeMix-FIRE-2020: Multilingual offensive speech detection in code-mixed and romanized text," *CEUR Workshop Proceedings*, vol. 2826, pp. 336–343, 2020.

[34] V. Pathak, M. Joshi, P. Joshi, M. Mundada, and T. Joshi, "KBCNMUJAL@HASOC-Dravidian-CodeMixFIRE2020: Using machine learning for detection of hate speech and offensive code-mixed social media text," *CEUR Workshop Proceedings*, vol. 2826, pp. 351–361, 2020.
[35] G. Arora, "Gauravarora@HASOC-Dravidian-CodeMixFIRE2020: Pre-training ULMFiT on synthetically generated code-mixed data for hate speech detection," *CEUR Workshop Proceedings*, vol. 2826, pp. 362–369, 2020.
[36] H.-P. Su, Z.-J. Huang, H.-T. Chang, and C.-J. Lin, "Rephrasing Profanity in {C}hinese Text," in *Proceedings of the First Workshop on Abusive Language Online*, Vancouver, BC, Canada: Association for Computational Linguistics, 2017, pp. 18–24. doi: 10.18653/v1/W17-3003.
[37] X. Tang, X. Shen, Y. Wang, and Y. Yang, "Categorizing Offensive Language in Social Networks: A Chinese Corpus, Systems and an Explanation Tool," in *Chinese Computational Linguistics*, M. Sun, S. Li, Y. Zhang, Y. Liu, S. He, and G. Rao, Eds., Cham: Springer International Publishing, 2020, pp. 300–315.
[38] H. Yang and C.-J. Lin, "{TOCP}: A Dataset for {C}hinese Profanity Processing," in *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, Marseille, France: European Language Resources Association (ELRA), 2020, pp. 6–12.
[39] M. Polignano, V. Basile, P. Basile, M. de Gemmis, and G. Semeraro, "AlBERTo: Modeling Italian Social Media Language with BERT," *Italian Journal of Computational Linguistics*, vol. 5, no. 2, pp. 11–31, 2019, doi: 10.4000/ijcol.472.
[40] H. Rizwan, M. H. Shakeel, and A. Karim, "Hate-Speech and Offensive Language Detection in {R}oman {U}rdu," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, 2020, pp. 2512–2522. doi: 10.18653/v1/2020.emnlp-main.197.
[41] R. U. Mustafa and P. M. Saqib Nawaz, Fournier-viger, "Early Detection of Controversial Urdu Speeches from Social Media," *Data Science and Pattern Recognition*, vol. 1, no. 2, pp. 26–42, 2017.
[42] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed, and M. T. Sadiq, "Automatic Detection of Offensive Language for Urdu and Roman Urdu," *IEEE Access*, vol. 8, pp. 91213–91226, 2020, doi: 10.1109/ACCESS.2020.2994950.
[43] A. M. Ishmam and S. Sharmin, "Hateful Speech Detection in Public Facebook Pages for the Bengali Language," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 2019, pp. 555–560. doi: 10.1109/ICMLA.2019.00104.
[44] Md. R. Karim, B. Raja Chakravarthi, J. P. McCrae, and M. Cochez, "Classification Benchmarks for Under-resourced Bengali Language based on Multichannel Convolutional-LSTM Network," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, 2020, pp. 390–399. doi: 10.1109/DSAA49011.2020.00053.
[45] N. Romim, M. Ahmed, H. Talukder, and Md. Saiful Islam, "Hate Speech Detection in the Bengali Language: A Dataset and Its Baseline Evaluation," in *Proceedings of International Joint Conference on Advances in Computational Intelligence*, M. S. Uddin and J. C. Bansal, Eds., Singapore: Springer Singapore, 2021, pp. 457–468.
[46] J. Moon, W. I. Cho, and J. Lee, "BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection," pp. 25–31, 2020, doi: 10.18653/v1/2020.socialnlp-1.4.
[47] S. Saketh Aluru, B. Mathew, P. Saha, and A. Mukherjee, "Deep Learning Models for Multilingual Hate Speech Detection *," 2020.
[48] N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, and D. Y. Yeung, "Multilingual and multi-aspect hate speech analysis," *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 4675–4684, 2019, doi: 10.18653/v1/d19-1474.
[49] N. Ghanghor, R. Ponnusamy, P. K. Kumaresan, R. Priyadharshini, S. Thavareesan, and B. R. Chakravarthi, "{IIITK}@{LT}-{EDI}-{EACL}2021: Hope Speech Detection for Equality, Diversity, and Inclusion in {T}amil , {M}alayalam and {E}nglish," in *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, Kyiv: Association for Computational Linguistics, 2021, pp. 197–203.
[50] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, "Hate speech detection in the Indonesian language: A dataset and preliminary study," in *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2017, pp. 233–238. doi: 10.1109/ICACSIS.2017.8355039.
[51] J. Á. González, L.-F. Hurtado, and F. Pla, "TWilBert: Pre-trained deep bidirectional transformers for Spanish Twitter," *Neurocomputing*, vol. 426, pp. 58–69, 2021, doi: https://doi.org/10.1016/j.neucom.2020.09.078.
[52] M. Ptaszynski, A. Pieciukiewicz, and P. Dybała, "Results of the PolEval 2019 Shared Task 6 : first dataset and Open Shared Task for automatic cyberbullying detection in Polish Twitter," in *Proceedings of the PolEval 2019 Workshop*, M. Ogrodniczuk and Ł. Kobyliński, Eds., Warszawa: Institute of Computer Sciences. Polish Academy of Sciences, 2019, pp. 89–110.
[53] K. Madukwe, X. Gao, and B. Xue, "In Data We Trust: A Critical Analysis of Hate Speech Detection Datasets," in *Proceedings of the Fourth Workshop on Online Abuse and Harms*, Online: Association for Computational Linguistics, 2020, pp. 150–161. doi: 10.18653/v1/2020.alw-1.18.
[54] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Comput Surv*, vol. 51, no.

4, 2018, doi: 10.1145/3232676.
[55] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki, “Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis,” 2017, doi: 10.17185/duepublico/42132.
[56] Z. Waseem and D. Hovy, “Hateful symbols or hateful people? predictive features for hate speech detection on twitter,” *HLT-NAACL 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Student Research Workshop*, pp. 88–93, 2016, doi: 10.18653/v1/n16-2013.
[57] Z. Zhang and L. Luo, “Hate speech detection: A solved problem? The challenging case of long tail on Twitter,” *Semantic Web*, vol. 10, pp. 925–945, 2019, doi: 10.3233/SW-180338.
[58] K. Miok, B. Škrlj, D. Zaharie, and M. Robnik-Šikonja, “To BAN or Not to BAN: Bayesian Attention Networks for Reliable Hate Speech Detection,” *Cognitive Computation*, vol. 14, no. 1, pp. 353–371, 2022, doi: 10.1007/s12559-021-09826-9.
[59] H. Watanabe, M. Bouazizi, and T. Ohtsuki, “Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection,” *IEEE Access*, vol. 6, pp. 13825–13835, 2018, doi: 10.1109/ACCESS.2018.2806394.
[60] M. Wiegand, J. Ruppenhofer, and T. Kleinbauer, “{D}etection of {A}busive {L}anguage: the {P}roblem of {B}iased {D}atasets,” in *Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 602–608. doi: 10.18653/v1/N19-1060.
[61] S. Mishra and S. Mishra, “3Idiots at HASOC 2019: Fine-tuning transformer neural networks for hate speech identification in indo-european languages,” *CEUR Workshop Proceedings*, vol. 2517, pp. 208–213, 2019.
[62] A. Schmidt and M. Wiegand, “A Survey on Hate Speech Detection using Natural Language Processing,” in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, Valencia, Spain: Association for Computational Linguistics, 2017, pp. 1–10. doi: 10.18653/v1/W17-1101.
[63] E. Mosca, M. Wich, and G. Groh, “Understanding and Interpreting the Impact of User Context in Hate Speech Detection,” in *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, Online: Association for Computational Linguistics, 2021, pp. 91–102. doi: 10.18653/v1/2021.socialnlp-1.8.
[64] Z. Waseem, T. Davidson, D. Warmsley, and I. Weber, “Understanding Abuse: A Typology of Abusive Language Detection Subtasks,” pp. 78–84, 2017, doi: 10.18653/v1/w17-3012.
[65] S. Lei, W. Yi, C. Ying, and W. Ruibin, “Review of attention mechanism in natural language processing,” *Data Analysis and Knowledge Discovery*, vol. 4, no. 5, pp. 1–14, 2020.
[66] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” [Online]. Available: https://github.com/tensorflow/tensor2tensor
[67] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf, “Transfer Learning in Natural Language Processing,” in *Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Tutorials*, Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 15–18. doi: 10.18653/v1/N19-5004.
[68] T. Pires, E. Schlinger, and D. Garrette, “How multilingual is multilingual BERT?,” *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 4996–5001, 2020, doi: 10.18653/v1/p19-1493.
[69] Hugging Face, “bert-base-multilingual-cased.” https://huggingface.co/bert-base-multilingual-cased (accessed Dec. 20, 2022).
[70] Y. Zhao and X. Tao, “ZYJ123@DravidianLangTech-EACL2021: Offensive Language Identification based on XLM-RoBERTa with DPCNN,” *Proceedings of the 1st Workshop on Speech and Language Technologies for Dravidian Languages, DravidianLangTech 2021 at 16th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2021*, pp. 216–221, 2021.
[71] EMBEDDIA, “litlat-bert.” https://huggingface.co/EMBEDDIA/litlat-bert (accessed Dec. 20, 2022).
[72] Hugging Face, “ELECTRA.” https://huggingface.co/docs/transformers/model_doc/electra (accessed May 27, 2022).
[73] K. Clark, M.-T. Luong, G. Brain, Q. V Le Google Brain, and C. D. Manning, “ELECTRA: PRE-TRAINING TEXT ENCODERS AS DISCRIMINATORS RATHER THAN GENERATORS.” [Online]. Available: https://github.com/google-research/
[74] M. Mosbach, M. Andriushchenko, and D. Klakow, “On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines,” 2020.
[75] Y. Hao, L. Dong, F. Wei, and K. Xu, “Investigating Learning Dynamics of {BERT} Fine-Tuning,” *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 87–92, 2020.

Отримана в редакції 08.08.2023. Прийнята до друку 09.09.2023. Received 08 August 2023. Approved 09 September 2023. Available in Internet 12 September 2023.