Deep Learning Based Topic Classification for Sensitivity Assignment to Personal Data

Apdullah Yayık\textsuperscript{a,*}, Hasan Apik\textsuperscript{a}, Ayse Tosun\textsuperscript{b}, Enver Ozdemir\textsuperscript{c}

\textsuperscript{a}Mobildev Research and Development, Istanbul
\textsuperscript{b}Faculty of Computer and Informatics Engineering, Istanbul Technical University, Istanbul
\textsuperscript{c}National Center for High Performance Computing, Istanbul Technical University, Istanbul

Abstract

Knowing the topic of textual content before performing a natural language processing task enables the design of topic-specific pipelines. Since the topic is represented by all the sentences and words of the document, it can be accepted as a reference point that can describe the document alone. In this partnership, Mobildev successfully completed the construction and deployment of a topic classification model in order to assign a sensitivity level to the extracted personal data within the topic context. Since justice, health, and religion topics are considered as highly sensitive data by Personal Data Protection Rule, it is essential for the model to identify documents in these stated topics. Therefore, a publicly available dataset was chosen and, new document instances from these three important categories were added. Two state-of-the-art machine learning models for natural language processing tasks were assessed on the extended dataset: fasttext and bidirectional encoder representations transformers (BERT). The performance of the models and the computational costs for training at the server side are reported. After testing at the client side the most suitable model for a lightweight client operation is determined and deployed into the Mobildev’s existing platform. The model training and assessment have been successfully completed in collaboration with the National Center for High Performance Computing at Istanbul Technical University.

1. Introduction

Mobildev, founded in 2002, offers efficient solutions in short message services and e-mail marketing, as well as providing AI-driven solutions for Personal Data Protection Rule (PDPR) \cite{1} compliance through discovering personal data and obtaining legal consents. Since 2016, Mobildev has focused on providing solutions for PDPR compliance to its customers. In 2017, an R&D Center was established within its structure and two main products were developed and commercialised for a wide range of customers. The first product, namely IVT\textsuperscript{1}, was developed for obtaining consent for the personal data, whereas the second product, namely Datamin\textsuperscript{2}, provides PDPR compliance through discovering and grouping personal data entities in both structured and unstructured business documents. Datamin has a client-server architecture shown in Figure 1. Starting from the original implementation, the red blocks in this figure are newly developed and reported in this paper. On the client-side of Datamin, entity type classification and person-based grouping operations are performed. The entity type classification is employed using a named entity recognition model that relies on a bidirectional long short-term memory artificial neural network powered by rule sets and look-up tables. The rule sets consist of optimised regular expressions together with numerical verification (e.g. id number, tax number), and a window method (e.g. student number, blood group) that relies on the context in a predefined window or dictionary control (e.g. bank name, university name) to avoid false positives. The person-based grouping operation aims to associate classified personal data entities through a rule-based algorithm. Datamin’s machine learning model runs at the

\textsuperscript{*}Corresponding author, e-mail: apdullah.yayik@mobildev.com

\textsuperscript{1}https://www.mobildev.com/ivt

\textsuperscript{2}https://www.mobildev.com/datamin, supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under Grant No. 3190083
Fig. 1: The general architecture of the system. Modules that are shown in color were developed within the scope of this research collaboration. The others already exist in the Datamin product. Personal data entities from a document are classified with the help of a hybrid approach that is a combination of a named entity recognition model based on bidirectional long short-term memory and rule sets. The high-sensitivity level is assigned to a personal data entity only if the topic of the document obtained by the machine learning model is religion, justice, or health. In the last stage, extracted personal data entities are grouped on a personal basis.

client-side. It is crucial to keep the model execution at the client-side since we would like to minimise the circulation of the documents between client and server side and to comply with PDPR. Each client later sends only the extracted and grouped personal data entities for each document to the server through a Rest-API.

On the server-side of Datamin, grouped personal data entities of the documents obtained from several clients
are verified and merged. The verification operation relies on ignoring 10 or more grouped profiles sharing the same personal data that directly points to a real person’s identity. The grouped personal data entities for documents from several clients are merged only if they share at least one personal data entity that directly identifies a real person. Currently, the system can achieve operations for more than 77 types of personal data entities in 34 different types of file formats.

1.1. Problem Definition

Although the initial Datamin project succeeded in identifying categorised personal data from the documents, it would be more useful when the system could also assign a sensitivity level to these personal data with respect to the context in which the personal information is revealed. According to the Turkish PDPR, religious affiliations, diagnosed health problems, and criminal records have been identified as special categories of personal data. In the case of processing, storing, or sharing such categories of personal data, it is necessary to manage the processes by taking into account their high level of privacy. According to the law, it is expected that other non-special categories of personal data, extracted from the documents having a religion, health, or justice topics, should also be considered as highly sensitive. Therefore, we aim to extend the functionality of Datamin by revealing the topic information of the documents in which personal data entities have been extracted, and decided on the level of sensitivity of personal data in the identified context.

1.2. Proposed Solution

Topic classification, i.e. revealing the topic of a textual content that it intends to convey in a document by labeling pre-defined classes, is a challenging task in natural language processing (NLP) [2]. The classification process is possible with the detection of nonlinear relationships between words, but it is error-prone when the classical methods such as sparse term frequency-inverse document frequency (TF-IDF) vector models are used. For this reason, we employed machine learning models using input representations based on dense word embedding vectors and n-gram models. More specifically, a fasttext model that employs logistic regression using word embeddings has been trained and a pre-trained bidirectional encoder representations from transformers (BERT) model has been fine-tuned. Since the annotated data for the Turkish language is not sufficient in this field, an existing Turkish text dataset [3] is extended by adding new documents from the religion, health, and justice categories. A dataset including 31 thousand instances for eight classes has been constructed. The model that performs the best is determined by comparing macro-averaged F1-measure and additional statistical difference analysis with Nemenyi [4] and Mann-Whitney [5] post-hoc tests. This research has been conducted in collaboration with the National Center for High Performance Computing (UHeM) at Istanbul Technical University. The computational platform where the models are trained and assessed, and model validation and deployment of the selected new model into the existing Datamin platform are reported in this paper.

2. Computational Platform

The training phase of the project uses fasttext and BERT-based models. These models are run on dual-socket CPU systems in the National Center for High Performance Computing, Istanbul Technical University. Each node in the system includes four NVIDIA Tesla V100 General-Purpose Graphics Processing Units (GPGPU) that are connected via NVLink. Each GPU has 32 GB of memory and each node has 384 GB total system memory. The interconnect network between nodes is provided via Enhanced Data Rate InfiniBand. The hardware platform is mainly used for the training stage of the fasttext and BERT-based models.

3. Study Design

3.1. Selected Models for Topic Classification

Since words do not only express the meaning they carry, finding the most relevant structure to represent them by considering their textual context is a crucial step in NLP. During the last decade, researchers focused on creating word representation models using neural networks. In 2013, Mikolov et.al. [6], introduced the Word2Vec model having two different architectures, the first one relying on predicting words from its neighbors (continuous bag-of-word - CBOW) while the second one being a skip-gram that is the exact opposite of the first one. Fasttext [7] is a machine learning model, proposed by Facebook AI Research, based on the logistic regression method. Since the model has a shallow architecture and implements a simple but efficient operation that is specially designed for text classification tasks, it has attracted researchers since 2016 [8, 9]. The instance representation of the model relies on the average n-gram word embeddings, which is similar to the CBOW architecture in [6], and the loss function
in logistic regression is minimised using a stochastic gradient descent algorithm. In the last layer of the model, hierarchical softmax layers are deployed.

BERT [10], proposed by Google AI Brain, is an unsupervised, pre-trained, general purpose machine learning model that can be fine-tuned to solve almost all downstream supervised NLP tasks, such as named entity recognition (NER) [11], machine translation [12], and text classification [13]. Fine-tuning is a transfer learning method for NLP tasks, like ImageNet [14] and VGG-NET [15] models in computer vision. The BERT model uses the WordPiece model [16] that considers the probability of the sub-words to obtain token embeddings for the words. The input representation of the BERT model (that is shown in Figure 2 at [10]) is the sum of the token embeddings, and the segmentation and position embeddings that indicate the sentence and sub-word orders, respectively. Pre-training of the BERT model is performed by optimising two objective functions together. The first one is formed by using the masked language model that uses randomly assigned 15% of the word representations in given instances as unknown and predicts the model parameters to achieve the actual word representations. The second one is formed by predicting consecutive sentences in the given instances. In this study, a BERT-based model that consists of almost 110 million parameters in 12 transformers blocks [17], 12 attention heads, and 768 hidden sizes have been fine-tuned end-to-end. Additionally, a fasttext model having a relatively much shallower architecture is trained.

3.2. Dataset Construction

During our project, we used a publicly available dataset published by Yıldırım [3] to construct our dataset with Turkish documents. This dataset consists of instances from different contexts including but not limited to politics, economy, culture, and sport. We extended this dataset so that our proposed model for detecting topics from textual data could learn from as many classes as possible, and so its generalisation ability would increase. Since justice, health, and religion topics are considered as highly sensitive data by PDPR [1], it is essential for our model to identify documents on those topics. Therefore we chose the stated publicly available dataset and added new document instances from these three important categories for our study. Instances of the health topic were collected from academic theses in medicine published between 2019 and 2020 in Turkey. Instances of the religion class were collected from the verses of the Bible [18] and Quran [19]. Instances of the justice class were collected from the constitution of the Republic of Turkey [20] and the criminal procedure law [21]. The distribution of the dataset over different topics/classes is shown in Table 1. The dataset consists of 31 thousand instances and 6 million words, 30 thousand of which are unique. This is a relatively clean dataset compared to the social media posts used in other prior studies; as our objective is to classify the topics of official business documents, we would not need to work on a noisy data to accomplish this task. Typical NLP preprocessing operations were implemented on the instances such as case-folding to lower case, discarding punctuation marks, numbers, and white spaces, and removing stop-words [22].

Table 1: Instance distribution of the classes in the dataset.

| Class      | Number of instances |
|------------|---------------------|
| Politics   | 2549                |
| Economy    | 3962                |
| Culture    | 1852                |
| Health     | 3078                |
| Sport      | 10679               |
| Technology | 1471                |
| Religion   | 4914                |
| Justice    | 2574                |

3.3. Model Construction

In this study, 90% of the dataset is used in the training set and the remaining is reserved as the testing set. The fasttext model was fed with 100-dimensional embeddings each of which represents averaged bi-gram word embeddings in an instance in the training set. 15% of the training set is used as an evaluation set. The training operation is employed 50 epochs with a learning rate of 0.7 in the stochastic gradient algorithm. The BERT-based model is fine-tuned using an Adam optimiser [23] with $\beta_1$ of 0.9 and $\beta_2$ of 0.999, epsilon of $10^{-8}$, and a learning rate of $3 \cdot 10^{-5}$. 10% of the training set is used as an evaluation set to check the model performance after each epoch. The model was set to fine-tune 20 epochs, but the operation is terminated in the case of 3 consecutive losses on the evaluation set.
4. Model Evaluation

In Table 2, performance of the two machine learning models are reported. The first one is trained with fasttext and the second one is fine-tuned with BERT-based using 90-10 % train-test split. In Table 2, it is seen that macro-averaged F1-measure of the fasttext model is almost the same as that of the BERT-based model. Considering the individual performance on the topics, the fasttext model seems to have a slightly higher F1-measure in all except justice and religion topics. This might be related to the structure of justice and religion documents as they have a context-specific vocabulary and sentence forms compared to documents related to technology or culture news.

### Table 2: Performance of the fasttext and BERT-based models. ste stands for standard error.

| Class    | Model     | Precision (%) | Recall (%) | F-1 Measure (%) |
|----------|-----------|---------------|------------|-----------------|
| Technology          | fasttext  | 91.88         | 99.32      | 95.45           |
|                      | BERT-based| 92.16         | 95.27      | 93.69           |
| Culture            | fasttext  | 94.87         | 100.00     | 97.36           |
|                      | BERT-based| 94.36         | 99.46      | 96.84           |
| Justice            | fasttext  | 87.85         | 87.85      | 87.85           |
|                      | BERT-based| 96.58         | 91.50      | 93.97           |
| Economy            | fasttext  | 94.69         | 98.74      | 96.67           |
|                      | BERT-based| 95.06         | 92.19      | 93.61           |
| Health             | fasttext  | 87.14         | 99.03      | 92.71           |
|                      | BERT-based| 83.01         | 96.75      | 89.36           |
| Religion           | fasttext  | 96.45         | 83.10      | 89.28           |
|                      | BERT-based| 97.99         | 89.41      | 93.50           |
| Sport              | fasttext  | 99.43         | 97.28      | 98.34           |
|                      | BERT-based| 99.81         | 98.69      | 99.25           |
| Politics           | fasttext  | 95.47         | 99.22      | 97.31           |
|                      | BERT-based| 91.82         | 96.86      | 94.27           |
| Macro-avg ± ste    | BERT-based| 93.47 ± 1.50  | 95.57 ± 2.26 | 94.37 ± 1.40 |

In Table 3 the results of Nemenyi and Mannwhitney post-hoc tests on the predictions of the models are shown. From Table 3, it is concluded that the decisions of the fasttext model are not statistically different from that of the BERT-based model \( (p > 0.05) \). This situation confirms that two models perform significantly the same for classifying the topics in the selected documents.

### Table 3: The results of the post-hoc tests.

| Post-Hoc Test | p value |
|---------------|---------|
| Nemenyi       | 0.900   |
| Mannwhitney   | 0.606   |

Besides, the dataset is divided into 10 folds where at each fold 90 % of the dataset is used in training and the rest is used for testing the model. In Figure 2, cross-validated performance metrics of the fasttext models are shown in a box plot. It is seen in Figure 2 that in 10-fold cross-validated fasttext models, the mean of the performance metrics are around 95 %, and the outliers are in the range of 70 % and 85 %. These ranges show that the fasttext model makes accurate and robust predictions across the whole dataset.

The duration and memory usage that is required for the training and validation of each model are given in Table 4. In both cases, the same hardware is used. The storage size of each trained model is also given in the table. Executions of the models are tested on the same textual content having 370 characters and 52 words (after the preprocessing, these numbers are changed to 292 and 38 respectively). From Table 4 it is concluded that both the creation and execution of the fasttext model is much less costly than that of the BERT-based model. The fasttext model creation is 294x faster and its memory usage is 14x less than that of the BERT-based model. Contrary to the expected, the low-cost of creating the fasttext model does not lead to high-cost in execution. The fasttext model executes 355x faster and uses 82x less memory than the BERT-based model. Besides the storage size of the fasttext model is 68x smaller than that of the BERT-based model. During training, the BERT-based model requires a significant amount of memory, whereas fasttext takes only minutes to train its parameters.
Fig. 2: Boxplots of the performance metrics of a 10-fold cross-validated fasttext model. Mean values of the metrics are close to 95% and the outliers are in the range of 70% and 85%.

Table 4: Comparison of the fasttext and BERT-based models in several parameters in the cases of model training and test execution.

| Model Training | fasttext | BERT-based |
|----------------|----------|------------|
| duration       | 1 minute 35 sec | 7 hour 45 minutes 56 sec |
| memory usage   | 1 GB     | 14 GB      |
| hardware       | GPU Nvidia Tesla V4, 256 GB RAM | GPU Nvidia Tesla V4, 256 GB RAM |
| duration of an instance | 1.6 ms | 569 ms |

| Model Test Execution | hardware       | CPU 8 Core, 16 GB RAM |
|----------------------|----------------|-----------------------|
| model size           | 6.4 MB         | 438 MB                |
| memory usage         | 12 MB          | 987 MB                |
| CPU level            | 1              | 46                    |

5. Model Deployment

The modules shown in red in Figure 2 depict the operations developed and integrated into the Datamin platform during this study. The newly integrated modules at the client side assign a sensitivity level to each personal data entity. It is assumed that each document has only one topic. For efficiency, the topic of a document is assigned by considering only the first N components (i.e. pages, slides, lines) with the fasttext model, where N is empirically determined as 5. The minimum number of input characters sent to the fasttext model was limited to 20 to prevent erroneous decisions. Besides considering the output of the softmax layer, the decisions with probabilities higher than 75% are accepted only. The topic with a probability higher than 75% is assigned to the document and combined with the personal data classification, the information is highlighted as highly sensitive when the personal data is referred in one of the special categories. This way the business organisations would comply with PDPR by automatically processing their documents considering their contexts. The integration of the new modules with the existing Datamin components has been completed and served as a complete product to the customers of MobilDev.

6. Conclusion

This study aims to assign a high-sensitivity level to the personal data by considering the topic of the textual content they are extracted from. For this aim, we investigated classifying textual contents with respect to their topics using the fasttext and BERT-based machine learning models. A dataset consisting of news documents on different topics has been extended by adding new documents from religious, health, and justice topics. Then, using this dataset, the models were trained and assessed, and the most suitable one in terms of accuracy and computational and memory constraints has been selected. Our collaboration with the National Center for High Performance Computing located in Istanbul Technical University encouraged us to evaluate multiple models on larger hardware and memory computing platforms employed in the center. Since its rapid execution, having a tiny model size for deploying on the client-side, and performing on par with the BERT-based model, we agreed to integrate the fasttext
model into production. In this sense, this study shows that more useful and practical results can be accomplished with a simple but task-specific model rather than highly-complicated, general-purpose designed models. It is noteworthy that if the same models had been compared on the dataset including relatively noisy instances such as social media posts, the model choice would be different.

In the literature, it is seen that when these models were trained with similar clean datasets (Dbpedia [24]) by different researchers, and similar accuracy rates were reported. Armand et.al. [25] achieved an accuracy of 98.40% with a fasttext model, while Qizhe et.al. [26] achieved an accuracy of 98.67% with a BERT-based model. Our results confirmed the previously achieved performance of both models with the macro-averaged F1-measures of 94.37% and 94.31%.

Generally, for sequence classification tasks i.e. language, news, or topic detection, the machine learning model is fed with all the words and sentences of the instance at a time to obtain a decision that represents the instance itself. However, for sequence labeling tasks such as named entity recognition and machine translation, feeding the machine learning model with short or long-termed windowed words and computing the gradients from left to right and/or right to left is required to memorise and transfer the temporal information [2]. Such a model cannot observe all the words and sentences at a time, in theory. Both transformers and recurrence neural network models aim to solve sequence classification and labeling tasks together, so their architecture is unnecessarily more complex for sequence classification tasks. In our opinion, for sequence classification tasks classical approaches like linear regression and multi-layer perceptron fed with enriched word representations perform well. The fasttext model internally uses logistic regression with a modern deep learning approach, for this reason, it is empirically observed that the fasttext model can perform on par with the BERT-based model.

Since the technology and machine learning field for topic classification is continually evolving, a static model cannot adapt to new unseen patterns. For this reason, metrics such as CPU and memory usage together with the latency and the accuracy of the deployed model will be logged. These metrics will be monitored to reveal the necessity for updating the model with a more representative dataset. Contributions to the deployed model will continue through collaborations between Mobildev and the National Center for High Performance Computing, Istanbul Technical University.

Acknowledgements

This work is supported by PRACE project funded by the European Union Horizon 2020 Research and Innovation Program under grant agreement 823767.

References

1. Grand National Assembly of Turkey. Personal data protection rule no. 6698, 2016. https://www.mevzuat.gov.tr/MevzuatMetin/1.5.6698.pdf.
2. Daniel Jurafsky and James H Martin. Speech and language processing 3rd Edition (draft). 2019.
3. MF Amasyalı and T Yıldırım. Otomatik haber metinleri sınıflandırma. SIU 2004, pages 224–226, 2004.
4. Peter Nemenyi. Distribution-free multiple comparisons. In Biometrics, volume 18, page 263. International Biometric Soc 1441 I ST, NW, SUITE 700, WASHINGTON, DC 20005-2210, 1962.
5. Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. The annals of mathematical statistics, pages 50–60, 1947.
6. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.
7. Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759, 2016.
8. Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. Give your text representation models some love: the case for basque. arXiv preprint arXiv:2004.00033, 2020.
9. Fahim Mohammad. Is preprocessing of text really worth your time for toxic comment classification? In Proceedings on the International Conference on Artificial Intelligence (ICAI), pages 447–453. The Steering Committee of The World Congress in Computer Science, Computer . . . , 2018.
10. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
11. Yu Wang, Yining Sun, Zuchang Ma, Lisheng Gao, Yang Xu, and Ting Sun. Application of pre-training models in named entity recognition. In 2020 12th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), volume 1, pages 23–26. IEEE, 2020.
12. Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. Incorporating bert into neural machine translation. *arXiv preprint arXiv:2002.06823*, 2020.

13. Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. Docbert: Bert for document classification. *arXiv preprint arXiv:1904.08398*, 2019.

14. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

15. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image reada: A method for stochastic optimization cognition. *arXiv preprint arXiv:1409.1556*, 2014.

16. Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

17. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

18. Incil. https://incil.info/. Accessed: 2021-01-21.

19. Kur’an. https://webdosya.diyanet.gov.tr/kuran/kuranikerim/dosyalar/document/kuran.pdf. Accessed: 2021-01-21.

20. Türkiye cumhuriyeti anyası. https://www.mevzuat.gov.tr/mevzuat?MevzuatNo=2709&MevzuatTur=1&MevzuatTertip=5. Accessed: 2021-01-21.

21. Ceza muhakemeleri kanunu. https://www.mevzuat.gov.tr/MevzuatMetin/1.5.5271.pdf. Accessed: 2021-01-21.

22. Zemberek-nlp turkish stop words. https://github.com/ahmetaa/zemberek-nlp/blob/master/experiment/src/main/resources/stop-words.tr.txt. Accessed: 2021-01-21.

23. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

24. Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195, 2015.

25. Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hervé Jégou, and Tomas Mikolov. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.

26. Qihe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019.