EuroCC@Turkey

https://eurocc.truba.gov.tr/

This document is prepared by EuroCC@Turkey for EuroCC under GA NO 951732

CASE STUDY REPORT

Text Processing of Social Media Messages

NCC Partner	*METU*
Company* ***Contact***	*Somera - somera.com.tr* *Can Eroğul - can.erogul@somera.com.tr*
NCC Expert	*Pınar Karagöz - karagoz@ceng.metu.edu.tr*
Start & End Date	*01.01.2021 - 01.09.2021*
Approved by	*NCC Project Management Team - Date*

***Company** accepts that the Case Study Report is shared with the EuroCC Project and the community through the EuroCC@Turkey awareness creation activities and platforms.

Date	Author	Comments	Version
30.12.2020	Pınar Karagöz	First version of the document is created	1
22.02.2021	Pınar Karagöz	Document is migrated to the new template with minor revisions	2
27.02.2021	Pınar Karagöz	Milestones are added to the document	3
06.09.2021	Pınar Karagöz	Milestone dates and final status are updated	4
20.09.2021	Pınar Karagöz	Obtained results are updated	4

1. Problem Identification

Somera is an SME who provides in-house developed systems for processing social media posts and web pages for its clients and business partners. Their solutions involve big data analysis and live streaming of the analysis results through dashboards. For big data processing, Somera uses its own servers as well as cloud resources such as AWS. Within the scope of this project, we will investigate the use of TRUBA's HPC infrastructure as an alternative.

2. First Suggestion

We plan to follow the following steps as the solution:

1. Going over the current tasks of Somera that involve cloud service.
2. Learning about TRUBA's HPC infrastructure
3. Determining the task(s) to be migrated
4. Task migration to TRUBA's infrastructure
5. Performance Analysis

Milestones:

- 31.03.2021 - The first module ported on TRUBA's HPC infrastructure
- 15.05.2021 - Results of pretrained model obtained on TRUBA's HPC infrastructure
- 01.09.2021 - Results of the trained model obtained on TRUBA's HPC infrastructure, and the results integrated in the company's solution.

3. Solution Stage – I

In the analysis conducted in step 1, Somera proposed to port a new module on text analysis on TRUBA's HPC infrastructure, rather than migrating existing modules. The company has a vast collection of social media messages in Turkish and involves a rich set of text processing functionalities such as search for a given pattern, message grouping, named entity recognition, sentiment analysis etc. Somera aims to deploy deep learning based language models on HPC infrastructure and train them for supervised learning problems specific for the company.

As the language model, among several alternatives, BERT language model proposed by Google (https://github.com/google-research/bert) is considered. The reason for this choice is due to recent attempts for developing language model through BERT specific for Turkish. As the first milestone, it is aimed to install BERT models on TRUBA's HPC infrastructure. As the second milestone, we plan to adapt the pretrained models for functionalities needed by Somera. As the last milestone, we aim to obtain a model trained with Somera's data. Additionally it is also aimed to analyze how useful the results are for Somera's commercial services.

4. Solution Stage – II

In the second stage, the BERT module is deployed on TRUBA HPC systems and trained with public data. As the specific task to be modelled, spam/fraud detection is selected. More specifically, this is a classification task on social media response sent on a specific post such that the response is classified as fraud or not.

Loan sharks in Turkey are sending their contact information and stories of their customers as advertisements. They are sending these advertisements as responses to posts in Facebook accounts of legal banks. Banking Regulation and Supervision Agency of Turkey enforces Turkish banks to delete such loan shark advertisements from their social media accounts, since users might think that the advertisements are sent from the bank or bank is approving these texts. Therefore, such postings are considered as fraud in our task and it is aimed to detect them in an automated way.

In order to work on another similar task on Turkish with a balanced data set, clickbait detection is also explored with BERT infrastructure. Clickbait detection is about identifying advertisement-like content that aims to entice the users to follow a link mostly leading to deceptive or misleading content.

5. Results and Achievements

As the final stage of the case study, the BERT model trained for Turkish texts is fine-tuned for the fraud detection task described in Stage II with a data set of about 52 K instances in the raw form. The data set includes 6 K fraud instances and 46 K non-fraud instances. The obtained accuracy is 98.70 % on average under 10-fold cross validation. In comparison to the inhouse computation resources (Nvidia 1080 8 GB) of Somera, with TRUBA system 38% computation time improvement has been observed. For the clickbait detection, the data set includes 24 K click-bait and 24 K regular instances, 48 K samples on the total. For this task, 30% computation time improvement has been obtained. As seen in these figures, the basic advantage brought by using TRUBA system was the use of GPU servers available among the set of resources.