THE EFFECT OF DATA SIZE ON THE RESULTS OF MACHINE LEARNING ALGORITHMS FOR BINARY TEXT CLASSIFICATION

A PREPRINT

Fatih Beyhan
EMW Project
Koç University
beyhanf@outlook.com

September 16, 2020

ABSTRACT

Data sharing is one of the main points of the projects in Machine Learning and Deep Learning society. Koç University EMW Project's team shared a data set, which is consist of URLs for articles and their labels for different tasks, on CLEF-2019. These URLs may be expired in the future and this may affect the shared results. We tried different ratios of data and compared their results to see how much the data size is affecting the results. Those results will be shown on this paper.

***K*eywords** data size · machine learning

1 Introduction

Sharing the data is one of the non-technical problems of Machine Learning and Deep Learning projects. On EMW Project, we are sharing only the URLs of the articles and the script to extract the texts from URLs. However, some of the URLs can be expired or the content can be moved to another URL. In this paper, we are focusing on only one of the tasks in CLEF 2019. The task we are focusing is a classification problem. Participants are asked to build a classifier that will classify the news whether they are related to a protest event or not. The dataset for this task consists of 3500 news, 2500 non-protest and 900 protest news. The question is "Do we need to share all this data?" or "Can we change the ratio of the news?". Due to this reason, we did some experiments on our data set with classical and advanced machine learning algorithms to see how the missing URLs can affect the results of the models. The results of these experiments can help us to simplify the data set and resolve the copyright issues.

On our previous work, we were asked to build protest-classifier. Different algorithms were tried and due to properties of the shared dataset, which is not the same with CLEF 2019 dataset, classical machine learning algorithms such as support vector machine, did better than advanced algorithms such as multilayer neural network, BERT.

Hence, we will do our experiments with five different algorithms which are Naïve Bayes, Support Vector Machine, Multi-Layer NN, Bi-LSTM and BERT. The complexity of algorithms is increasing, respectively.

To sum up, we are working on a text classification problem. TfIdfVectorizer() method of scikit-learn library is used to prepare the dataset for the algorithms, except Bi-LSTM and BERT.

2 Data

Dataset, for the protest classification task, is shared with 3 sections: train, validation and test sets. There are article URLs and their labels, which is binary. Corpus is coming from 4 different sources. The data is binary, and it is labelled by annotators from EMW Project. They distributed the data set in a way that does not violate copyright of the news sources. This involves only sharing information that is needed to reproduce the corpus from the source for task 1 and task 2 and only relevant snippets for task 3 [1].

	Protest	Non-protest	Total
Train	856	2723	**3579**
Dev.	102	355	**457**
Test	154	533	**684**

Table 1: *Number of instances in each set can is shown here.*

The dataset is imbalanced. The ratio of protest news is 20% for each set. However, our previous work showed that this ratio is not generalizable. Since, we dismiss this ratio while building our models.

	Train	Dev.	Test	Total
newindianexpress	528	68	105	**701**
indiatimes	1804	233	379	**2416**
thehindu	792	110	143	**1045**
indianexpress	455	46	60	**561**

Table 2: *Number of instances for each source in each set can is shown here.*

The dataset is coming from 4 different sources which are news agencies from India. In the original task of CLEF 2019, there was a second test set which was from China. This set was testing generalizability of the model. The gold standard corpus of EMW Project consists of English news articles from various local and international sources from India, China, and South Africa. Variety of the sources has allowed studying cross context robustness and generalizability, therefore addressing style and content change across sources, which are critical requirements of the ML models [2]. However for this specific task, the dataset consists of only Indian news.

Having a wider dataset will be the topic of another work of ours.

3 Methodology

The main purpose of this paper is to show how data size will affect the machine learning approaches on our specific task. Hence, different machine learning models are used for experiments. As it is mentioned above, four different algorithms are used; `GaussianNB()` and `SVM()` from `scikit-learn`, Multi-Layer NN from `PyTorch`. There are couple of common approaches that should be applied before any machine learning on text data by starting any model training or further steps. Decided stop-words are dismissed from articles. Each word in corpus is lemmatized and any non-alphabetic character is thrown away. These steps are applied to each of the 3 sets. After these steps, the data is ready to be fed into models. There are two different experiment for each machine learning model. The first experiment is sampling from whole training data. The training data has 3579 training sample. Only the 856 of the training set is protest news and the training data is shuffled. It is expected that any random sample of the training set will have the same ratio of protest news and non-protest news. First experiment has 5 steps:

1. Create different sample sizes for training set by starting from small ratio and adding the same amount till the whole data set, e.g. starting from 500 and increment data size by 500: 500, 1000 ... 3500.
2. For each sample size, sample the exact amount from training set.
3. Create Pipeline objects which are consist of tfidf vectorizers and a machine learning model and start a grid search. Repeat the 3rd step multiple times to have consistency of results.
4. Create the models with best parameters of the grid search and test on the validation and test sets.
5. Save the precision, recall and F1-score of each step and each iteration.

The second experiment is almost the same with the first experiment with a minor change. On this task, we keep the whole set of a label and sample different amounts from other label and apply this for both two labels, respectively. Support Vector Machine and Gaussian Naive Bayes models are used in the pipeline objects. Since grid search is being applied on the sampled data set, the best result on that data is accomplished with given algorithm. Moreover, sampling same amount of data multiple times gives opportunity to have more reliable results. On the other hand, slightly different approach applied for Multi-Layer NN. A grid-search algorithm was built from scratch. Similar to the classical algorithms, it has `TfIdfVectorizer()` and Multi-Layer NN from `PyTorch`.

Lastly, validation and test sets are not being changed. The models are being trained on different sample sizes and tested on same validation and test sets.

4 Results

After two experiments, which are mentioned above, precision, recall, and F1-scores are obtained for different algorithms on test set. The sample size steps varies from algorithm to algorithm and experiment to experiment for the sake of reliability.

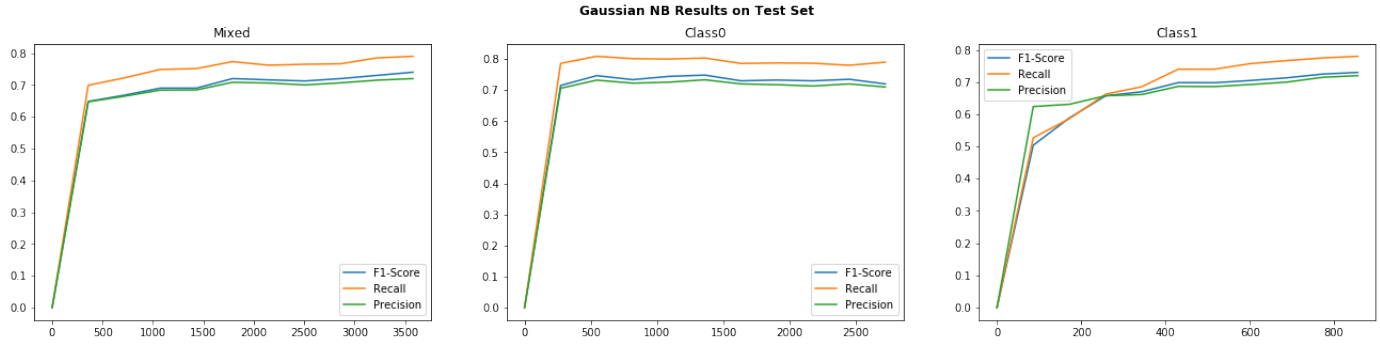

Figure 1: *Scores of Naive Bayes Classifier for test set. Standard deviation of scores ranges between (0-0.03) and averages of each sample size is on the figure. The graph on the* ***left*** *shows the scores for sampling of mixed articles. The graph on the* ***middle*** *shows the scores for sampling of non-protest articles. The graph on the* ***right*** *shows the scores for sampling of protest articles.*

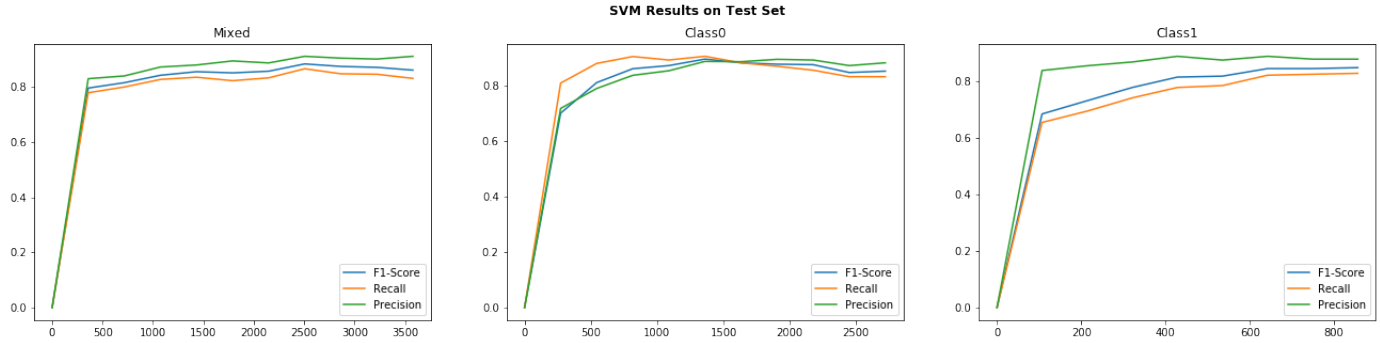

Figure 3: *Scores of SVM Classifier for test set. Standard deviation of scores ranges between (0-0.03) and averages of each sample size is on the figure. The graph on the* ***left*** *shows the scores for sampling of mixed articles. The graph on the* ***middle*** *shows the scores for sampling of non-protest articles. The graph on the* ***right*** *shows the scores for sampling of protest articles.*

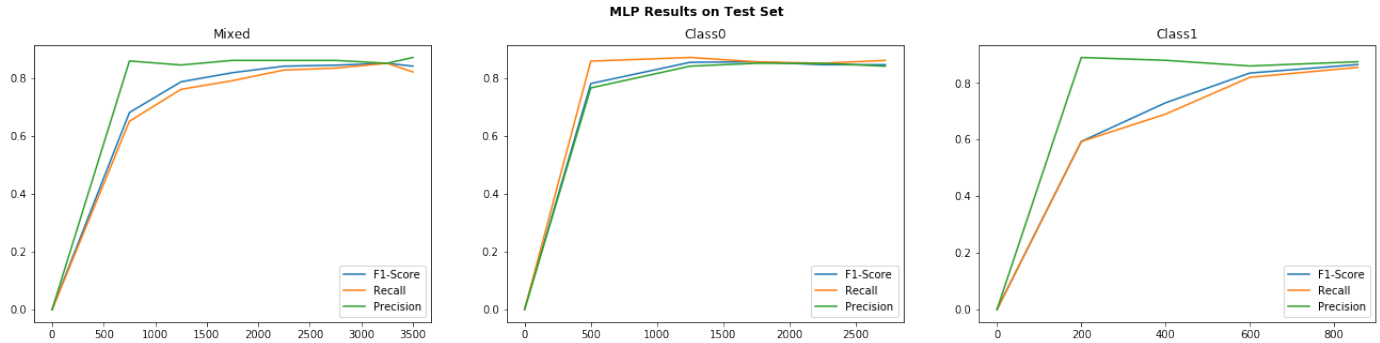

Figure 3: *Scores of Multi-Layer NN Classifier for test set. Standard deviation of scores ranges between (0-0.04) and averages of each sample size is on the figure. The graph on the* ***left*** *shows the scores for sampling of mixed articles. The graph on the* ***middle*** *shows the scores for sampling of non-protest articles. The graph on the* ***right*** *shows the scores for sampling of protest articles.*

The results of each algorithm can be seen above. Despite that all algorithms have different ways of training, they showed a similar pattern with the increase of the sample size. The common thing among these algorithms is that none of them is working with word-vectors. Semantic relations between words are not caught.

Despite the common belief in the machine learning community, unbalanced dataset did not end up with worse results. The middle graph on each figure shows the effect of non-protest news. The original dataset is imbalanced. After having the same amount of news with protest news, it keeps increasing and the scores are not going down. Until one point, it even improves the scores. The thing that can be understood is that sharing less non-protest news while keeping the same amount of protest news, would not affect the results for participants.

On the other hand, it is not expected to see half of the data suddenly disappear in a short period. Results for sampling of mixed classes are on the left graph of each figure. The disappearance of 10% or even more, which is not likely, would not have a big effect on the obtained results.

5 Future Work

All experiments mentioned above are using a statistical method to represent the articles. TfIdf vectorizer is turning each article into a vector with arbitrary dimensions. Hence, semantic relations are not being caught. Same experiments are going to be applied on models which are able to catch some semantic relations, such as Bi-LSTM, BERT, ELMO, etc. Currently, pre-trained models are being fine-tuned on the same dataset with the same methodological structure. Results will be the topic of the next paper.

Beside changing representations and models, another scenario of URL disappearance will be simulated. All URLs of a specific source can be unreachable due to policy change on URLs. For example, some sources can move their content from one URL to an archive URL. With this scenario, all URLs will disappear from a source. The effect of this on results will be simulated.

Last but not the least, all experiments above applied with one condition: while the training set was being changed with experiments, the test set was not being changed. In order to make this whole experiment more realistic, the same ratio will be shifted from the test set.

References

[1] Hürriyetoğlu, A., Yörük, E., Yüret, D., Mutlu, O., Yoltar, Ç., Duruşan, F., Gürel, B., (2020). Cross-context News Corpus for Protest Events Related Knowledge Base Construction. Automated Knowledge Base Construction 2020.

[2] Hürriyetoğlu, A., Yörük, E., Yüret, Yoltar, Ç., D., Gürel, B., Mutlu, O., Akdemir, A.,(2019). Overview of CLEF 2019 Lab ProtestNews: Extracting Protests from News in a Cross-context Setting. CEUR.