AF and VT Detection From Doctors Notes

Noam Ben-Moshe and **Hagay Michaeli**
b.noam@campus.technion.ac.il
hagaymi@campus.technion.ac.il
https://github.com/noambenmoshe/NLP_Project.git

Abstract

Using Rambam hospital unlabeled database of ECG-Holters ids and corresponding doctor descriptions we propose an algorithm that classifies from the description if the doctor thinks that in a specific holter Atrial Fibrillation (AF) is recorded or not or whether the recording has Ventricular tachycardia (VT) or not. In order to succeed we had to over come some challenges. The first is that the doctor descriptions are in Hebrew. To over come this challenge we checked two different methods. One is using AlephBert which is a model that was trained to work with Hebrew Text. The other approach was to use translation from Hebrew to English and then classify. The second challenge is that we are dealing with medical data which means that the data contains new vocabulary the models were not trained on. To deal with this challenge we experimented with BlueBERT which was trained on medical data. As a baseline we used previous work done in AIMLab which used Google Translate to translate descriptions from Hebrew to English and then defined rules in order to classify. Our novelty is using deep learning methods to tackle this problem and our algorithm reached better F_1-score on the test set.

1 Introduction

Medical data sets such as physical examination results are often organized such as each sample is documented with a doctor's description. In some cases such data sets can be used for training AI models for medical condition classification. This kind of training is often done in Supervised Learning setting which requires an explicit label for each sample, which does not exist in the original data set but can be extracted from the descriptions. In The following work we introduce a method of such label extraction from Doctor's free text description using Natural Language Processing (NLP) tools. The data used consists of descriptions of Holter-ECG examination and the following doctor's descriptions. Holter-ECG examination is the output of a portable device worn for 24 hours or more that is continuously recording using typically two or three ECG leads. The objective was to classify weather AF or VT examples are recorded in the ECG examination. Where AF and VT are heart diseases.

2 Related Work

2.1 Bert

Bert is a Bidirectional Transformer used for langauge modeling (LM), first introduced in (Devlin et al., 2018). The LM model was trained on the tasks of masked word and next sentence prediction on Wikipedia and Toronto Book Corpus data sets. The original paper introduced the model's ability to be used as a backbone of models for many NLP tasks such as question answering and text classification, similar to the objective of the following work. Although for many tasks the usage of Bert requires only short fine tuning, this is not the case for tasks that are based on different vocabulary.

2.2 AlephBert

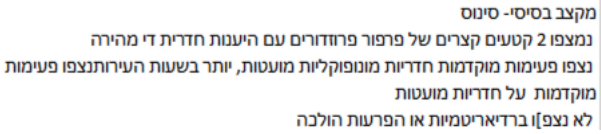
AlephBert (Seker et al., 2021) is a State-of-the-art language model for Hebrew. It is based on Bert-base architecture and was trained on Hebrew text from Wikipedia and Twitter.

2.3 BlueBert

BlueBert (Peng et al., 2019) is a BERT model pretrained on PubMed abstracts. The benefits of this model is that it was trained on medical data and so might be more familiar with medical vocabulary used in the doctor's descriptions.

3 Data

Rambam hospital unlabeled database of ECG-Holters ids and corresponding doctor descriptions in Hebrew was used. The database consists of 1982 valid examples, each example consists of a Holter examination and the description of a cardiologist describing the significant findings seen in the examination and if a certain disease is observed during the examination. An example of a Holter examination doctor's note is seen in figure 1.



מקצב בסיסי- סינוס
נמצפו 2 קטעים קצרים של פרפור פרוזדורים עם היענות חדרית די מהירה
נצפו פעימות מוקדמות חדריות מונופוקליות מועטות, יותר בשעות הערותנצפו פעימות
מוקדמות על חדריות מועטות
לא נצפ]ו ברדיאריטמיות או הפרעות הולכה

Figure 1: Input example - Hebrew description of a Holter examination

3.1 Pre-Processing

3.2 Annotations

To be able to evaluate the performances a labeled test set and validation set was needed. A keywords dictionary was created by a cardiologist from Rambam hospital, the dictionary keys is a list of synonyms for every abnormality that can appear in the Holter's open text. Some of the synonyms are in Hebrew and some are in English. Two annotators labeled 1000 examples independently and based their label on the key word dictionary. Then the labels of the two annotators were compared and examples that the annotators disagreed on were checked again until all labels were agreed upon. For our purpose 2 labels were given for each description - AF or None-AF and VT or None-VT. AF meaning there is an example of AF in the recording and None-AF meaning there is no example of AF in the recording. The same goes for VT and None-VT. For example the labels of the example in 1 is AF and None-VT.

3.3 Train, Validation and Test Sets

500 labeled examples were used as test set, 100 labeled examples were used as a validation set. The remaining 400 labeled examples were used for training. The percentage of examples from each disease is listed in table 1.

Set	AF %	VT %
Labeled Training Set	10.7%	5.7%
Validation Set	17%	7%
Test Set	12%	5.8%

Table 1: Distribution on labeled training set (n=400), validation set (n = 100) and test set (n=500).

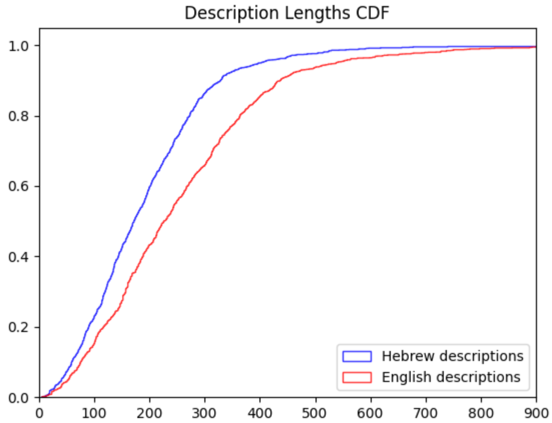


Figure 2: Training and Validation tokenized description lengths

4 Performance Measures

The following statistics were computed to assess the models performance: sensitivity (Se), specificity (Sp), positive predictive value (PPV), the area under the receiver operating characteristic (AUROC) and the harmonic mean of the precision and recall (F_1-Score).

5 Methods

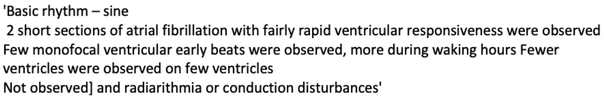
The main idea consist of three steps. The first was to load a pre-trained model. The second was to do a domain adaptation to the model through language modeling with all the training set data. The last step was to train a classifier with the labeled training data. Training was done in two steps:

1. Language Modeling - masked different parts of the sentence and made the model learn what is the missing word. This was done to tune the parameters to work with the current data, domain adaptation. Another reason for this was to take advantage of the unlabeled data.

2. Classification - Use a classification head and the labeled training data to classify if an examination has an example of the disease searched.

We experimented with two different pre-trained models. One is called AlephBert and the second is called BlueBert. The input of models based on AlephBert are the original Hebrew doctor descriptions. The input of the BlueBert based models need to be in English. To deal with this we used Google Translate library to translate the doctor's notes from

Hebrew to English. An example of the translated data is seen in figure 3 which is the translation of the example in figure 1.



'Basic rhythm – sine
2 short sections of atrial fibrillation with fairly rapid ventricular responsiveness were observed
Few monofocal ventricular early beats were observed, more during waking hours Fewer ventricles were observed on few ventricles
Not observed] and radiarithmia or conduction disturbances'

Figure 3: Input example - Translated description of a Holter examination

5.1 Training

Training was done using Hugging Face and torch packages. We used for all tasks effective batch size of 12 and trained for 10 Epochs. Because of the imbalanced classes we used a weighted binary cross entropy loss for the classification task, such that the positive samples weight was higher. We chose the maximal input length that was used in pre-trained Bert - 512. We can see in figure 2 that this length satisfies 94% of the English inputs and 97% of the Hebrew inputs.

6 Results

The results are presented in tables 2 and 3. Presented in table 2 are the results of the language modeling (LM) task. We compare the performance of the two models by using the accuracy measure and calculating how many of the masked words were correctly filled by the model. The results show that BlueBert got a higher accuracy than AlephBert. In table 3 the F_1-score of 2 classification taskts: classification for AF and classification on VT are compared between 5 different models. The 5 models are: a Baseline model which is a model that is not based on deep learning, 2 models that use AlephBert as the pre-trained model and 2 models that use BlueBert as the pre-trained model. We wanted to check that the LM is helping the model to learn so on each of the pre-trained models the Non-LM model means that the pre-trained model was used with a classifying head and in the LM model the model was first trained for the LM task and then for the classification task. The best model is the AlephBert -LM model which has the best F_1-score both in classifying the AF disease and classifying the VT disease. The results also show that the LM task was necessary and helps convert the pre-trained model to the specific-domain needed. This can be seen by that the LM models got higher resultes than the NO-LM models apart from the BlueBert classifying AF.

Model	LM-Accuracy
AlephBert	76.3%
BlueBert	**77.3%**

Table 2: Evaluation of LM task

Model	AF-F_1-score		VT-F_1-score	
Baseline	0.91		0.85	
AlephBert	NO-LM	0.823	NO-LM	0.951
	LM	**0.931**	LM	**0.983**
BlueBert	NO-LM	0.86	NO-LM	0.852
	LM	0.852	LM	0.967

Table 3: Evaluation on test set (n=500)

7 Discussions and Conclusions

7.1 Feature Extraction

Since most of the data was unlabeled, one of the suggested solutions was using NLP model for extracting features from the free text which will be used for an unsupervised clustering algorithm. In order to examine the possibility of this solution we used BERT output for each description as its feature vector and visualised it. For that matter we used t-SNE, which is a dimensionalty reduction algorithm that preserves the local structure of the data and is considered to have good performance at visualizing high dimensional data with manifold structure. We used t-SNE to visualise the feature extracted from pre-trained Bert, Bert after LM training and Bert after classification training. We can see in figure 4 the results of AlephBert trained to classify AF and it's results on the test set. It can be seen that the features are separated w.r.t the classes only after classification training, and that the separation is more significant when using LM training before the classification training (bottom line). However, there is no evidence for a separable local structure in the features of pre-trained or LM trained BERTs (upper line).

7.2 Sensitivity

The models were evaluated and selected according to F_1-score which consider FP and FN errors equally. However, the risk of FN errors in our model is arguably higher than FP errors for two reasons: (i) There are naturally much more negative records in the data set (ii) The Holter records which

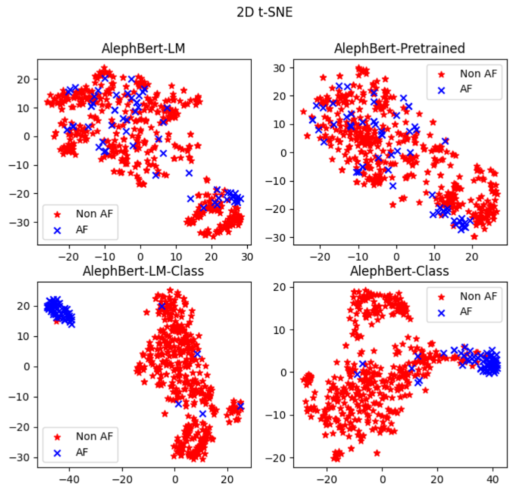


Figure 4: 2D t-SNE of BERT output in different pahses. A separable local structure is observed only after classification training.

Classification AF	Se	Sp
Baseline	0.86	0.995
AlephBert-LM	0.9	0.995

Table 4: Performance measures on AF classification

description are found positive are verified again by a Cardiologist. We address this issue in our model by using weighted loss in the classification training phase, where the weight of positive samples is higher. Note that the Baseline model consider this issue as well, and is designed to achieve high sensitivity in the cost of lower F_1-score. We can see in tables 4, 5 that the sensitivity of AlephBert-LM is higher or equal to the baseline while the overall F_1-score improved.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, pages 58–65.

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Greenfeld, and Reut Tsarfaty. 2021. Alephbert:a hebrew large pre-trained language model to start-off your hebrew nlp application with.

Classification VT	Se	Sp
Baseline	1	0.98
AlephBert-LM	1	0.998

Table 5: Performance measures on VT classification