CLASSIFYING THE STANCE & ARGUMENT TOWARD COVID-RELATED FIELDS USING NLP METHODS

RUARG22 COMPETITION

"Dialogue" is the largest international conference on computational linguistics and intellectual technologies.
The conference has been held annually since 1995 and is the successor of the seminar "Models of Communication", held since the 1970s.

AUTHOR

Daniil Karzanov,
HSE University

CONTACTS

dvkarzanov@gmail.com

INTRODUCTION

The topic of covid, quarantine and masks is very relevant today despite the decrease in the epidemics. Being able to classify people's positions in an automatic manner can still benefit different healthcare institutions in developing new strategies for promoting measures against COVID.

As part of the RuArg22 competition, we are given data consisting of Russian texts posted in different media and addressing three different topics related to COVID-19 pandemics. We are offered to evaluate a speaker's position on each topic and its premise.

RELEVANT WORKS

Insight from NLP Analysis: COVID-19 Vaccines Sentiments on Social Media 2021

Fake News Detection in Social Media using Graph Neural Networks and NLP Techniques: A COVID-19 Use-case 2020

COVIDLies: Detecting COVID-19 Misinformation on Social Media 2020

METHODOLOGY

Baseline. Fully-connected neural network (256-768-4) with Adam optimizer and "DeepPavlov/rubert-base-cased-sentence" embeddings.

Logistic Regression. Applies a sigmoid function to the linear transformation of the initial features.

SVM. Draws a separating hyperplane in the initial or augmented feature space using a sigmoid kernel.

FastText. Hierarchical classifier, hashing and binary trees underlying the model reduce significantly computation time. Contains an internal word embedder.

Neural Networks. 3-layer (768-40-3) fully-connected neural network with ReLu activation function, Adam optimizer, and sparse categorical cross-entropy loss.

Random Oversampling. Randomly selects examples from the minority class, with replacement, and adds them to the training dataset.

Vectorization Techniques:

- count vectorization + tf-idf transformation

Pre-trained HuggingFace word-embeddings:

- DeepPavlov/distilrubert-tiny-cased-conversational
- cointegrated/rubert-tiny-bilingual-nli
- cointegrated/rut5-small
- cointegrated/rubert-tiny

DATA ANALYSIS & PREPARATION

The texts are very versatile. The messages contain official announcements and informal comments-like messages some of which contain offensive and aggressive lexicon.
The classes are highly unbalanced and "irrelevant" appears approximately as often as in the other three classes.
We lower the text and remove special substrings (e.g. "[USER]") and punctuation symbols.

Original Text	Translation
О несоблюдении карантинных мер контактными лицами можно сообщить на на горячую линию…	Non-compliance with quarantine measures by contact persons can be reported to the hotline…
[USER], подождите недели две после карантина, не долго осталось!	[USER], wait two weeks after quarantine, not long left!
…вот из-за таких идиотов, которые ходят без масок и не сидят на карантине страдают все!	…that's because of such idiots who do not wear masks and do not sit in quarantine, everyone suffers!

Table 1: Examples of texts in the training set.

class	quarantine stance	quarantine argument	vaccines stance	vaccines argument	masks stance	masks argument
-1 "irrelevant"	4627	4627	5059	5059	3587	3587
1 "other"	1341	1756	866	1238	1832	2451
2 "for"	587	217	374	149	704	339
0 "against"	172	127	418	271	594	340

Table 2: Value counts of the target columns.

STRATEGIES

For each topic, we fit the first model (M0) to distinguish a new bool variable "topic" which checks if the stance is -1 or not. Thereafter, we fit two other models (M1 and M2) separately on the balanced slice of the dataframe where the topic is not 0 i.e., where the stance is 1, 2, or 3. In such a manner we obtain 9 different models overall and the procedure is shown in Figure 1. The strategy brings us to many variations due to options in the choice of models M0, M1, M2, and the vectorization techniques.
As for the prediction procedure (see Figure 2), we forecast the value for the topic variable first for each entry in the test set. If it's 0, we understand that this topic was not mentioned in the sentence, so stance and argument are both saved as irrelevant. Otherwise, we apply two separate neural networks to predict balanced stance and argument.

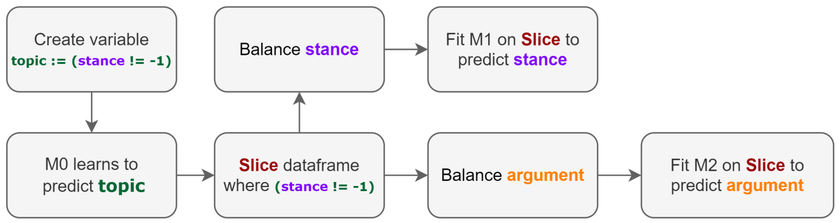

Figure 1: Scheme of the fitting procedure.

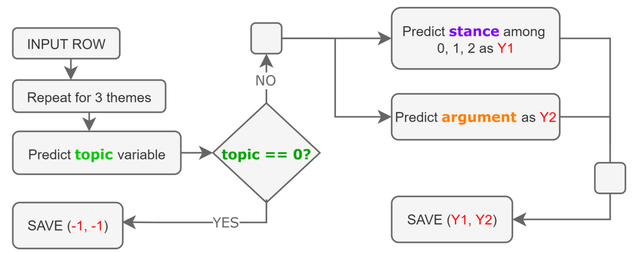

Figure 2: Scheme of the forecasting procedure for each entry in the test set.

RESULTS

№	M0	M0-Emb	M1	M1-Emb	M2	M2-Emb	F1- Stance	F1- Premise
0	Baseline						0.392	0.451
1	FastText						0.463	0.462
2	LOGIT	tf-idf	LOGIT	tf-idf	LOGIT	tf-idf	0.430	0.361
3	LOGIT	tf-idf	NN	DeepPavlov	NN	DeepPavlov	0.499	0.525
4	SVM	tf-idf	NN	DeepPavlov	NN	DeepPavlov	0.496	0.494
5	SVM	DeepPavlov	NN	DeepPavlov	NN	DeepPavlov	0.509	0.529
6	NN	DeepPavlov	NN	DeepPavlov	NN	DeepPavlov	0.530	0.559
7	NN	bilingual-nli	NN	bilingual-nli	NN	bilingual-nli	0.478	0.521
8	NN	rut5-small	NN	rut5-small	NN	rut5-small	0.470	0.483
9	NN	rubert-tiny	NN	rubert-tiny	NN	rubert-tiny	0.495	0.495

Table 3: Performance comparison.

Labeling premise is the most problematic for all models. While all the approaches successfully pass the stance baseline notably, just a few of them outran the premise baseline significantly. Since the powerful FastText model, as well as the baseline, were outperformed by most pipeline combinations, we consider the idea of extracting the topic variable and consequently reducing the number of classes in the argument variable quite an efficient approach. Having attempted each combination without class balancing, we observed that the scores drop dramatically by around 0.08-0.01.

The most powerful in terms of predictability appears to be three neural networks with the DeepPavlov/distilrubert-tiny-cased-conversational word vectorization.

cointegrated/rubert-tiny-bilingual-nli demonstrates one of the highest F1-Premise, none of the "cointegrated/.." embeddings improve the scores.