Estimating Confidence of Predictions of Individual Classifiers and Their Ensembles for Genre Classification

Mikhail Lepekhin, Serge Sharoff

MIPT, University of Leeds

1. Contributions

- Confidence of prediction for genres
- Robustness of various classifiers
- Assessment of social media samples for genres

2. Training data

Code	Genre label	Prototypes	LJ train	LJ test	FTD train	FTD test
A1	Argument	Argumentative blogs or opinion pieces	1858	599	207	77
A4	Fiction	Novels, myths, songs, film plots	698	232	62	23
A7	Instruction	Tutorials, FAQs, manuals	1617	478	59	17
A8	News	Reporting newswires	2255	787	379	103
A9	Legal	Laws, contracts, terms&conditions	17	12	69	13
A11	Personal	Diary entries, travel blogs	2291	709	126	49
A12	Promotion	Adverts, promotional postings	195	61	222	85
A14	Academic	Academic research papers	34	10	144	49
A16	Information	Encyclopedic articles, definitions, specifications	695	221	72	33
A17	Review	Reviews of products or experiences	681	219	107	34
	Total		10341	3288	1447	483

3. Confidence of prediction

A classifier can only be confident on a text example if it predicts the correct labels for texts with *similar* embeddings. Application of dropout to the inference stage n times generates the corresponding probability distributions $p_1, ..., p_n$. Then we pool them into single distribution $\widehat{p}$. The maximum value of probability in $\widehat{p}$ is the confidence value.

Mann-Whitney test for the confidence of correct vs incorrect predictions

Genre	XLM-R stat	XLM-R delta	RuBERT stat	RuBERT delta	LogReg stat	LogReg delta	Ensemble 2 stat	Ensemble 2 delta	Ensemble 3 stat	Ensemble 3 delta
Argument	**0.697**	**0.123**	0.624	0.08	0.596	0.052	0.67	0.101	0.638	0.075
Fiction	0.705	0.130	**0.749**	0.148	0.635	0.071	0.732	**0.158**	0.716	0.126
Instruction	0.838	0.224	0.800	0.179	**0.878**	**0.254**	0.834	0.229	0.845	0.236
News	0.914	0.282	0.919	0.225	0.909	0.325	0.898	0.254	**0.929**	**0.335**
Legal	0.740	0.145	0.728	0.119	0.513	0.016	0.743	0.149	**0.763**	**0.153**
Personal	0.830	0.193	**0.897**	0.233	0.848	0.213	0.858	**0.234**	0.863	0.228
Promotion	0.878	0.309	0.878	0.282	0.870	0.254	0.899	0.336	**0.915**	**0.349**
Academic	0.854	**0.278**	0.786	0.125	**0.885**	0.249	0.778	0.17	0.828	0.243
Information	0.683	0.109	0.685	0.124	0.586	0.044	**0.721**	**0.141**	0.648	0.079
Review	**0.690**	**0.118**	0.557	0.038	0.576	0.047	0.641	0.094	0.640	0.080
Total	**0.816**	0.211	0.808	0.183	0.792	0.199	0.815	0.214	0.815	**0.217**

4. Accuracy of classifiers and ensembles

Train	Test	Genre	XLM-R	RuBERT	LR	Ensemble2	Ensemble3
LJ+FTD	FTD	Argument	**0.708**$^{\pm 0.030}$	0.693$^{\pm 0.022}$	0.550	**0.732**$^{\pm 0.012}$	**0.716**$^{\pm 0.022}$
LJ+FTD	FTD	Fiction	**0.780**$^{\pm 0.047}$	**0.784**$^{\pm 0.053}$	0.739	**0.809**$^{\pm 0.034}$	**0.807**$^{\pm 0.017}$
LJ+FTD	FTD	Instruction	**0.761**$^{\pm 0.062}$	0.665$^{\pm 0.037}$	0.371	**0.742**$^{\pm 0.034}$	**0.714**$^{\pm 0.039}$
LJ+FTD	FTD	News	**0.945**$^{\pm 0.008}$	**0.929**$^{\pm 0.012}$	0.770	**0.945**$^{\pm 0.011}$	0.913$^{\pm 0.020}$
LJ+FTD	FTD	Legal	**0.795**$^{\pm 0.081}$	**0.773**$^{\pm 0.037}$	0.000	**0.788**$^{\pm 0.056}$	**0.750**$^{\pm 0.133}$
LJ+FTD	FTD	Personal	**0.714**$^{\pm 0.024}$	**0.661**$^{\pm 0.041}$	0.635	**0.711**$^{\pm 0.027}$	**0.713**$^{\pm 0.018}$
LJ+FTD	FTD	Promotion	**0.946**$^{\pm 0.011}$	0.908$^{\pm 0.018}$	0.455	**0.946**$^{\pm 0.015}$	**0.937**$^{\pm 0.010}$
LJ+FTD	FTD	Academic	**0.880**$^{\pm 0.015}$	**0.852**$^{\pm 0.056}$	0.000	**0.892**$^{\pm 0.017}$	**0.852**$^{\pm 0.041}$
LJ+FTD	FTD	Information	**0.650**$^{\pm 0.053}$	**0.627**$^{\pm 0.038}$	0.321	**0.670**$^{\pm 0.043}$	**0.624**$^{\pm 0.036}$
LJ+FTD	FTD	Review	**0.685**$^{\pm 0.041}$	0.590$^{\pm 0.048}$	0.418	**0.687**$^{\pm 0.048}$	**0.653**$^{\pm 0.028}$
LJ+FTD	LJ	Argument	**0.590**$^{\pm 0.020}$	**0.563**$^{\pm 0.035}$	0.500	**0.603**$^{\pm 0.023}$	**0.609**$^{\pm 0.012}$
LJ+FTD	LJ	Fiction	**0.734**$^{\pm 0.028}$	**0.716**$^{\pm 0.026}$	0.637	**0.756**$^{\pm 0.028}$	**0.762**$^{\pm 0.023}$
LJ+FTD	LJ	Instruction	**0.795**$^{\pm 0.021}$	0.787$^{\pm 0.011}$	0.768	**0.810**$^{\pm 0.014}$	**0.818**$^{\pm 0.011}$
LJ+FTD	LJ	News	**0.907**$^{\pm 0.006}$	**0.904**$^{\pm 0.007}$	0.864	**0.912**$^{\pm 0.005}$	**0.914**$^{\pm 0.004}$
LJ+FTD	LJ	Legal	**0.331**$^{\pm 0.216}$	**0.394**$^{\pm 0.156}$	0.154	**0.355**$^{\pm 0.143}$	**0.332**$^{\pm 0.126}$
LJ+FTD	LJ	Personal	**0.807**$^{\pm 0.011}$	0.783$^{\pm 0.018}$	0.742	**0.816**$^{\pm 0.010}$	**0.817**$^{\pm 0.004}$
LJ+FTD	LJ	Promotion	**0.465**$^{\pm 0.049}$	**0.479**$^{\pm 0.047}$	0.289	**0.495**$^{\pm 0.038}$	**0.488**$^{\pm 0.051}$
LJ+FTD	LJ	Academic	**0.324**$^{\pm 0.104}$	**0.332**$^{\pm 0.107}$	0.000	**0.409**$^{\pm 0.084}$	**0.350**$^{\pm 0.158}$
LJ+FTD	LJ	Information	0.642$^{\pm 0.014}$	**0.641**$^{\pm 0.030}$	0.539	**0.658**$^{\pm 0.018}$	**0.674**$^{\pm 0.012}$
LJ+FTD	LJ	Review	**0.640**$^{\pm 0.017}$	0.604$^{\pm 0.026}$	0.520	**0.653**$^{\pm 0.015}$	**0.658**$^{\pm 0.016}$
LJ+FTD	FTD	Accuracy	**0.822**$^{\pm 0.005}$	0.789$^{\pm 0.009}$	0.532	**0.828**$^{\pm 0.007}$	**0.828**$^{\pm 0.007}$
LJ+FTD	LJ	Accuracy	**0.757**$^{\pm 0.011}$	**0.741**$^{\pm 0.010}$	0.694	**0.769**$^{\pm 0.010}$	**0.774**$^{\pm 0.006}$

https://github.com/MikeLepekhin/GenreClassifierEnsembles

5. Prediction on social media samples

Prediction distribution on Livejournal (in thousands of texts)

Genre label	XLM-R	RuBERT	LogReg	Ensemble3	Percente
Argument	11192	21568	11773	15454	20.8
Fiction	5182	6801	5471	5836	7.86
Instruction	5583	3635	3465	3531	4.76
News	6925	7548	11079	8674	11.68
Legal	33	133	23	83	0.11
Personal	34127	24427	37315	32099	43.22
Promotion	2722	2548	1489	2059	2.77
Academic	67	126	30	74	0.1
Information	1118	1337	1214	1156	1.56
Review	5666	6009	2408	4021	5.41
Non-text	384	134	0.04	10	0.01
Total	74267	74267	74267	74267	100

6. Conclusions and future research

1. The transformer-based classifiers (XLM-RoBERTa or RuBert) are generally accurate in non-topical classification tasks provided that enough training data is available for each label.
2. For most genres, the ensembles of several classifiers obtain a higher f1-score than any of the separated classifiers.
3. Adding even a weaker classifier, in our case, Logistic Regression, to the ensemble does benefit the classification accuracy.
4. Mann-Whitney statistics shows that the ensemble with Logistic Regression is more reliable for most genres than the ensemble of the two models and each individual classifier.
5. Applying of the classifiers to large social media sample reveals their distribution of genres, such as the greater rate of Argumentative texts in LJ in comparison to the greater rate of Personal reporting in VK.

LREC, 20-24 June 2022, Marseille, France.