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

Mikhail Lepekhin, Serge Sharoff
MIPT, University of Leeds
Moscow, UK
lepehin.mn@phystech.edu, S.Sharoff@leeds.ac.uk

Abstract

Genre identification is a kind of non-topic text classification. The main difference between this task and topic classification is that genres, unlike topics, usually cannot be expressed just by simple keywords, and thus they need to be defined in terms of their functions in communication. Neural models based on pre-trained transformers, such as BERT or XLM-RoBERTa, demonstrate SOTA results in many NLP tasks, including non-topical classification. However, in many cases, their downstream application to very large corpora, such as those extracted from social media, can lead to unreliable results because of dataset shifts, when some raw texts do not match the profile of the training set. To mitigate this problem, we experiment with individual models as well as with their ensembles. To evaluate the robustness of all models we use a prediction confidence metric, which estimates the reliability of a prediction in the absence of a gold standard label. We can evaluate robustness via the confidence gap between the correctly classified texts and the misclassified ones on a labeled test corpus, higher gaps make it easier to improve our confidence that our classifier made the right decision. Our results show that for all of the classifiers tested in this study, there is a confidence gap, but for the ensembles, the gap is wider, meaning that ensembles are more robust than their individual models.

Keywords: Document Classification, Evaluation Methodologies, Language Modeling, pre-trained transformers

1. Introduction

Non-topical text classification includes a wide range of tasks aimed at predicting a text property that is not connected directly to a text topic. For example, predicting a text style, politeness, difficulty level, the age or the first language of its author, etc. Automatic genre identification (Santini et al., 2010) is one of the standard problems of non-topical text classification. It is applied in many areas such as information retrieval, language teaching or linguistic research.

Compared to topical text classification, genre classification has additional difficulties. First of all, the concept of genre is more complex than that of a topic. Every topic has a set of keywords and, therefore, one can define whether a text belongs to a topic or not based on the occurrence of the keywords in the text. Genre is a functional dimension of texts and in most cases, it cannot be defined just by a set of its keywords. Besides, genre typologies are often very large, while the training datasets offer few examples per more specific genres, which makes the genre harder to be classified by neural models.

Text genre identification has a long history of research. (Sharoff et al., 2010) is one of the first works on text genre classification containing a comparison of various datasets, models, and data features. It shows how traditional ML models such as SVM with various hyperparameters and features performs for the genre classification of texts. Since then, traditional ML models have been superseded by neural models. (Devlin et al., 2019) introduced BERT – (Bidirectional Encoder Representations from Transformers), an efficient language representation model based on the Transformer architecture (Vaswani et al., 2017). It achieves state-of-the-art results for various NLP tasks, including text classification. XLM-RoBERTa (Conneau et al., 2019) is an improved variant of BERT. It has a similar architecture but uses a bigger and more genre-diverse corpus based on Common Crawl (instead of Wikipedia for the multilingual BERT). Therefore, we choose XLM-RoBERTa as the classifier for the experiments in our research.

One of the most significant problems in genre classification concerns topical shifts, according to (Petrenz and Webber, 2010). If a topic is more frequent in the training corpus for a given specific genre, then a classifier tends to predict the genre by the keywords of the topic, for example, *hurricane* can be associated with the genre of FAQs (Sharoff et al., 2010). This causes numerous unreasonable mistakes in genre classification. Conversely, a topical shift within the same genre can effect the predictions. For this reason, Petrenz and Webber (2010) check the accuracy of the genre classifiers via testing on the datasets from different domains, and so do we in our work. However, we introduce neural methods for confidence estimation of the genre classifiers.

Ensembles has not been experimented with in the Genre identification domain. Our paper is the first one that contains such sort of research.

We take the baseline classifiers from our work (Lepekhin and Sharoff, 2021). Then the following steps are conducted in this paper to investigate the properties of ensembles for the problem of genre classification and to get a genre classifier with a higher accuracy score:

1. finding the optimal hyperparameters for training the genre classifiers (subsection 4.2),

2. analysing the distribution of genres in Social Media corpora (subsection 4.3),
3. measuring of the confidence of prediction for each classifier on each pair (train dataset, test dataset) in section 5.

The tools to replicate the experiment are available.[1]

2. Confidence of prediction

To estimate robustness of our predictions at the inference stage, we use a *confidence* metric, which equals $1 - uncertainty$, as introduced in (Tsymbalov et al., 2020). The key idea is that a classifier can only be confident on a text example if it predicts the correct labels for texts with *similar* embeddings. If the genre predicted by a classifier is not the same for many texts in the neighbourhood of an original text, we cannot consider the classifier confident. To simulate variation in embeddings, we apply dropout with probability 0.1 to all of the model layers, including the embedding layer and the final dense classification layer (Sun et al., 2019). A softmax classifier returns a probability distribution of the possible text labels. Normally, the label with the maximal probability is considered the answer of a classifier. Application of dropout to the classifier disturbs the probability distribution. We perform dropout n times, and thereby we generate 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 called *the confidence of prediction*. The intuition of this metric is that if a classifier is confident in the predicted text label, it is unlikely to distribute the likelihood on other class labels significantly often. In our study, we use $n = 10$ - the same as used by the authors of (Tsymbalov et al., 2020), since this provides a good balance between assessing the confidence value and the speed of computation.

While the method is applicable to unlabeled datasets, it is beneficial to estimate the reliability of how it deals with the predicted labels. Therefore, we apply this procedure to the test dataset to compare the difference in the confidence of correct and incorrect predictions made by different classifiers. We denote this difference as *confidence delta* and in our study we test the hypothesis that the value of delta helps to understand whether the text is classified correctly or not. For a given classifier model, the higher the delta, the more reliable threshold we can choose to cut off misclassified texts. It means that even in the absence of a gold-standard label, we can use the prediction confidence to say whether a classifier is likely to predict the text genre correctly or not.

3. Training data

In our experiments, we use two Russian datasets annotated with the same set of genre labels – FTD genre dataset (Sharoff, 2018) and texts from LiveJournal which are labeled manually for this paper. The first one has nearly 2000 labeled texts. The LiveJournal corpus is much bigger, it consists of more than 10500 texts see Table 1.

LiveJournal is a Russian blogging platform. It consists mostly of posts and comments, sharing news, describing personal experiences or opinions. FTD is a collection of documents from multiple sources, such as Wikipedia, Reddit, and online newspapers. These two corpora are quite different in terms of genre distribution. FTD is more or less balanced when the LJ corpus is not. The legal and academic texts are especially underrepresented in LJ. It means that we have to treat the results of the classification of the Legal texts in the LJ corpus as unreliable. The largest category in both LJ and FTD concern news reporting. We collect the text from LiveJournal by random sampling of the authors and their posts. Each text in the LiveJournal dataset is labeled by two assessors. Those texts for which the labels did not coincide were additionally discussed.

Our intention is to test our genre classifiers on a domain that is different from that of the training data similarly to (Petrenz and Webber, 2010). We conduct experiments, training models on LiveJournal and FTD separately, to show how the selection of the training corpus affects metrics on the testing data, as well as on their concatenation. We split the corpora in the same ratio. For each corpus, 75% is used for training, 25% - for testing.

4. Baseline Classifiers

4.1. Train XLM-RoBERTa, RuBERT and Logistic Regression

We fine-tune base RuBERT pretrained by DeepPavlov (Kuratov and Arkhipov, 2019), base XLM-RoBERTa (Conneau et al., 2019) and a Logistic Regression classifier with character- and word-based features. In some experiments, we also trained SVM and Random Forest classifiers on the same features to compare their performance with Logistic Regression.

We use XLM-RoBERTa because it achieves the best accuracy on the Russian section of the XNLI corpus. We also used the RuBERT model because it attained the highest accuracy on the RuSentiment classification dataset among all the monolingual models for the Russian language. In each experiment, we use the Adam optimiser and $learning\_rate = 5 \cdot 10^{-5}$ as advised in (Sun et al., 2019). But our computational resources are too limited to use a $batch\_size$ value proposed in (Sun et al., 2019), we set $batch\_size = 16$. Another important thing in any task of text classification is tackling the long texts. We have to take into account the fact that both RuBERT and XLM-RoBERTa cannot process texts that are longer than 512 tokens. In our study, we just take the first 512 tokens, also known as *head-only* truncation method from (Sun et al., 2019).

[1] `https://github.com/MikeLepekhin/GenreClassifierEnsembles`

Genre short	Genre label	Prototypes	LJ		FTD	
			train	test	train	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

Table 1: Training and testing corpora

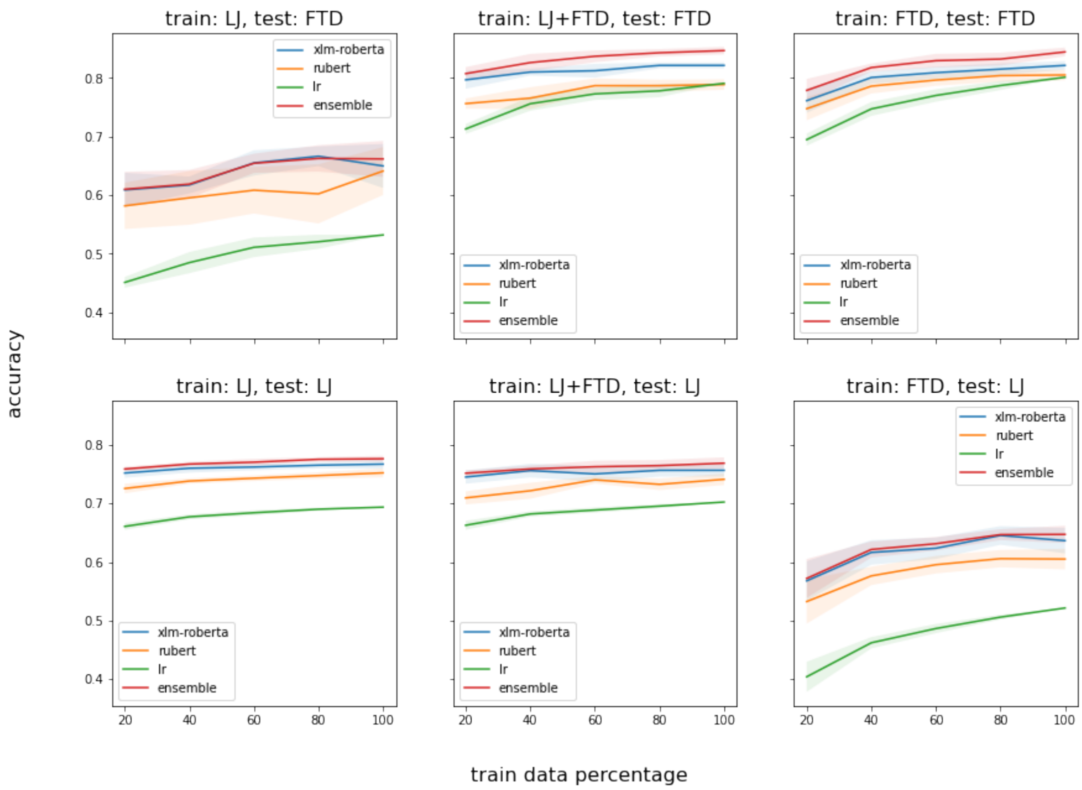

Figure 1: Dependence of the model accuracy on the train data size

As a traditional classifier we focus Logistic Regression because it is a traditional ML algorithm which differs drastically from that of both XLM-RoBERTa and RuBERT. Moreover, it is quite a lightweight classifier that has a decent speed of training and inference. We use 5000 word-based features and 10000 char-based features, all of which are selected by the highest tf-idf value. The char-based and the word-based features are represented by 2-, 3- and 4-gramms of chars and words correspondingly.

We train two types of ensembles – one of them consists of the XLM-RoBERTa and the RuBERT genre classifiers, the second one additionally includes a traditional ML classifier, mostly Logistic Regression, while in some experiments we also ran SVM and RF for comparison with LR. We name them correspondingly *ensemble2* and *ensemble3*. The ensemble weights are selected independently for each pair (train dataset, test dataset) on the basis of the performance on a validation set. A relatively high value of the LR weights for Ensemble 3 (even though this is not the strongest classifier) comes from the difference in the probability levels, as the median probability for the predicted class is around 0.98 for the transformer classifiers, while it is 0.68 for LR. Among the transformer classifiers, the XLM-RoBERTa weight is typically higher. This matches our findings in Table 3, where XLM-RoBERTa obtains the highest f1-score among the separate classifiers.

As we can see in Table 3, the f1 metric of each model strongly depends on the genre. Each classifier is worse in predicting informational texts than most other genres. Such a suboptimal result on the informational texts is likely to be caused by their heterogeneity as they include texts of very different classes, such as encyclopedias, dictionaries, CVs, biographies, product specifications, etc.

Figure 1 shows how much the models' accuracy depends on the training data size. The models continue to improve with the increase in training data. We can also see than each ensemble almost always works better than every individual model, regardless of the size of the training data.

The domain mismatch between the FTD and the LiveJournal corpora is shown in the drop of accuracy when training on FTD and testing on LiveJournal or vice versa Table 3. Each classifier trained on FTD performs much worse on LiveJournal than on FTD. The same situation appears for any classifier trained on LiveJournal. This can be seen clearly for the genre Promotion. The classifiers show relatively high accuracy on the FTD testing promotion texts and a suboptimal one on the promotion texts from LiveJournal. One reason is the topical shift as FTD does not include promotion texts that advertise websites or internet services. Since most of such Promotion texts have numerous keywords on the topic *internet*, a classifier that was not trained on the Promotion texts on this topic is unlikely to identify the genre correctly. We can see, that even the classifiers trained on the LJ training subset do not perform well on the LJ testing data, though the result is not that bad as it is for the FTD-trained models. It is an example of domain shifts that distort the prediction of genres by models.

In order to balance the mismatch between LiveJournal and FTD, we train the classification architectures on the concatenation of LiveJournal and FTD. As we can see in Table 5, these classifiers perform much better on LiveJournal than those trained only on FTD. Similarly, the classifiers are more accurate on FTD than those trained solely on LiveJournal. It shows that including texts from multiple domains into the training set helps to cope with topical shifts in the test.

To train each neural model in the experiments, we use single NVIDIA TITAN RTX. Table 4 shows that the training and inference time for XLM-RoBERTa and RuBERT are at the same level, whilst the Logistic Regression is much faster to train and test per text.

Table 3 also show how much the transformer classifiers depend on the random seed during model training, so we report the confidence intervals over ten random seeds and we highlight in bold the best F-scores if there is no other confidence interval that lies entirely to the right of it without intersection. Often the intersection of the confidence intervals for the best F-scores are quite frequent.

In most cases, the ensemble attains the highest f1-measure. Among the individual classifiers, XLM-RoBERTa seems to be the most reliable one. Table 3 shows that the number of genres for which XLM-RoBERTa is the best is lower than that for the ensemble but higher than the values for the RuBERT and the Logistic Regression genre classifiers.

We also tried to replace Logistic Regression with other weak classifiers, including SVM (with standard hyperparameters from sklearn) with linear and RBF kernels and Random Forest (with 100 and 1000 trees). The ensemble with linear SVM attains total accuracy 0.828 ± 0.004 on FTD and 0.771 ± 0.009 on LJ. The ensemble with Random Forest attains total accuracy 828 ± 0.004 on FTD and 0.770 ± 0.011 on LJ. The Random Forest classifier of 1000 trees does not increase accuracy compared to Random Forest of 100 trees but it requires 10 times more time for training and testing. Ensembles with them show higher accuracy than that of individual XLM-RoBERTa and RuBERT but lower than the ensemble with Logistic Regression. Moreover, the training and inference time of SVM and Random Forest turn out to be much higher than that for Logistic Regression. These results support choosing Logistic Regression as a classifier for further experiments.

The most frequent kinds of mistakes are shown in Table 6. We show the percentage of the texts in the test data that are misclassified in a certain way. In the table, we use the shortcuts for the genres as per Table 1. Surprisingly, confusion pairs do not vary too much with changing of genre classifier. Besides, the nature of the mistakes is quite predictable and it is related to the similarities between the genres, often leading to genre hybrids. Functions of some Argumentation and Informational texts might be similar. The same can be told about genre pairs (Review, Personal), (Research, Argumentation), (Info, Research), etc. The most important difference between the confusion matrices of different genre classifiers is the relative number of mistakes. For example, the Ensemble3 and the XLM-RoBERTa show lower confusion numbers in Table 6.

4.2. Optimal hyperparameters

XLM-RoBERTa and RuBERT start to over-fit after 3 epochs, see Figure 2 for the LJ corpus, the behaviour for FTD is similar. Since the third epoch, RuBERT continues improving on the training subset, while its accuracy on the test does not change. It looks like RuBERT is slightly more vulnerable to over-fitting beyond 3-4 epochs on the genre classification task. The

Train	Val/ Test	Ensemble2				Ensemble3			
		XLM-R	RuBERT	LR	Acc	XLM-R	RuBERT	LR	Acc
FTD	FTD	0.692	0.308	0	0.840	0.158	0.222	0.620	0.882
FTD	LJ	0.821	0.179	0	0.671	0.684	0.017	0.299	0.674
LJ	FTD	0.513	0.487	0	0.701	0.474	0.139	0.388	0.708
LJ	LJ	0.692	0.308	0	0.775	0.421	0.274	0.305	0.785
LJ+FTD	FTD	0.769	0.231	0	0.826	0.211	0.125	0.665	0.875
LJ+FTD	LJ	0.718	0.282	0	0.770	0.474	0.166	0.360	0.776

Table 2: Ensemble weights

Train	Test	Genre	XLM-R	RuBERT	LR	Ensemble2	Ensemble3
FTD	FTD	Argument	**0.729**$^{\pm0.021}$	**0.695**$^{\pm0.027}$	**0.721**	**0.744**$^{\pm0.016}$	**0.775**$^{\pm0.016}$
FTD	FTD	Fiction	**0.690**$^{\pm0.051}$	**0.728**$^{\pm0.043}$	**0.789**	**0.757**$^{\pm0.058}$	**0.795**$^{\pm0.030}$
FTD	FTD	Instruction	**0.760**$^{\pm0.104}$	**0.790**$^{\pm0.049}$	**0.741**	**0.811**$^{\pm0.060}$	**0.810**$^{\pm0.039}$
FTD	FTD	News	**0.944**$^{\pm0.011}$	**0.920**$^{\pm0.014}$	0.904	**0.937**$^{\pm0.016}$	**0.936**$^{\pm0.009}$
FTD	FTD	Legal	0.757$^{\pm0.039}$	0.793$^{\pm0.043}$	**0.923**	0.789$^{\pm0.042}$	0.826$^{\pm0.041}$
FTD	FTD	Personal	**0.725**$^{\pm0.028}$	**0.698**$^{\pm0.030}$	**0.708**	**0.732**$^{\pm0.034}$	**0.750**$^{\pm0.024}$
FTD	FTD	Promotion	**0.937**$^{\pm0.012}$	0.918$^{\pm0.008}$	**0.924**	**0.939**$^{\pm0.010}$	**0.937**$^{\pm0.006}$
FTD	FTD	Academic	**0.883**$^{\pm0.023}$	**0.876**$^{\pm0.022}$	0.820	**0.892**$^{\pm0.018}$	**0.886**$^{\pm0.014}$
FTD	FTD	Information	**0.657**$^{\pm0.047}$	**0.587**$^{\pm0.075}$	0.293	**0.670**$^{\pm0.054}$	**0.654**$^{\pm0.047}$
FTD	FTD	Review	0.711$^{\pm0.030}$	0.698$^{\pm0.037}$	**0.763**	0.743$^{\pm0.019}$	**0.782**$^{\pm0.022}$
FTD	LJ	Argument	**0.475**$^{\pm0.019}$	**0.425**$^{\pm0.037}$	**0.429**	**0.477**$^{\pm0.021}$	**0.480**$^{\pm0.020}$
FTD	LJ	Fiction	**0.675**$^{\pm0.033}$	0.605$^{\pm0.050}$	0.346	**0.697**$^{\pm0.033}$	**0.699**$^{\pm0.035}$
FTD	LJ	Instruction	**0.539**$^{\pm0.084}$	**0.534**$^{\pm0.036}$	0.381	**0.548**$^{\pm0.064}$	**0.544**$^{\pm0.061}$
FTD	LJ	News	**0.877**$^{\pm0.010}$	0.853$^{\pm0.008}$	0.745	**0.876**$^{\pm0.010}$	**0.876**$^{\pm0.010}$
FTD	LJ	Legal	**0.495**$^{\pm0.079}$	0.415$^{\pm0.077}$	0.563	**0.495**$^{\pm0.072}$	0.489$^{\pm0.059}$
FTD	LJ	Personal	**0.716**$^{\pm0.027}$	0.685$^{\pm0.021}$	0.616	**0.727**$^{\pm0.015}$	**0.730**$^{\pm0.010}$
FTD	LJ	Promotion	**0.255**$^{\pm0.046}$	**0.252**$^{\pm0.024}$	0.209	**0.265**$^{\pm0.036}$	**0.265**$^{\pm0.027}$
FTD	LJ	Academic	**0.111**$^{\pm0.063}$	**0.122**$^{\pm0.048}$	**0.105**	**0.151**$^{\pm0.042}$	**0.127**$^{\pm0.037}$
FTD	LJ	Information	**0.541**$^{\pm0.027}$	0.364$^{\pm0.129}$	0.093	**0.534**$^{\pm0.043}$	**0.520**$^{\pm0.050}$
FTD	LJ	Review	**0.486**$^{\pm0.044}$	0.443$^{\pm0.031}$	0.341	**0.495**$^{\pm0.026}$	**0.501**$^{\pm0.023}$
LJ	FTD	Argument	**0.565**$^{\pm0.027}$	**0.589**$^{\pm0.045}$	**0.550**	**0.594**$^{\pm0.029}$	**0.594**$^{\pm0.029}$
LJ	FTD	Fiction	**0.710**$^{\pm0.043}$	**0.774**$^{\pm0.038}$	**0.739**	**0.770**$^{\pm0.034}$	**0.770**$^{\pm0.034}$
LJ	FTD	Instruction	**0.493**$^{\pm0.067}$	**0.472**$^{\pm0.105}$	0.371	**0.505**$^{\pm0.068}$	**0.505**$^{\pm0.068}$
LJ	FTD	News	**0.866**$^{\pm0.017}$	**0.857**$^{\pm0.034}$	0.770	**0.857**$^{\pm0.011}$	**0.857**$^{\pm0.011}$
LJ	FTD	Legal	**0.402**$^{\pm0.305}$	**0.518**$^{\pm0.242}$	0.000	**0.529**$^{\pm0.298}$	**0.529**$^{\pm0.298}$
LJ	FTD	Personal	**0.632**$^{\pm0.037}$	**0.610**$^{\pm0.039}$	**0.635**	**0.646**$^{\pm0.019}$	**0.646**$^{\pm0.019}$
LJ	FTD	Promotion	**0.782**$^{\pm0.090}$	**0.728**$^{\pm0.104}$	0.455	**0.790**$^{\pm0.076}$	**0.790**$^{\pm0.076}$
LJ	FTD	Academic	**0.249**$^{\pm0.218}$	**0.300**$^{\pm0.182}$	0.000	**0.270**$^{\pm0.138}$	**0.270**$^{\pm0.138}$
LJ	FTD	Information	**0.461**$^{\pm0.037}$	**0.444**$^{\pm0.034}$	0.321	**0.462**$^{\pm0.033}$	**0.462**$^{\pm0.033}$
LJ	FTD	Review	**0.571**$^{\pm0.054}$	**0.492**$^{\pm0.075}$	**0.418**	**0.569**$^{\pm0.042}$	**0.569**$^{\pm0.042}$
LJ	LJ	Argument	**0.591**$^{\pm0.021}$	**0.580**$^{\pm0.026}$	0.500	**0.605**$^{\pm0.021}$	**0.605**$^{\pm0.021}$
LJ	LJ	Fiction	**0.769**$^{\pm0.018}$	**0.740**$^{\pm0.020}$	0.637	**0.776**$^{\pm0.012}$	**0.776**$^{\pm0.012}$
LJ	LJ	Instruction	**0.804**$^{\pm0.014}$	**0.794**$^{\pm0.013}$	0.768	**0.815**$^{\pm0.008}$	**0.815**$^{\pm0.008}$
LJ	LJ	News	**0.912**$^{\pm0.006}$	0.899$^{\pm0.004}$	0.864	**0.912**$^{\pm0.005}$	**0.912**$^{\pm0.005}$
LJ	LJ	Legal	**0.172**$^{\pm0.165}$	**0.236**$^{\pm0.221}$	0.154	**0.153**$^{\pm0.140}$	**0.153**$^{\pm0.140}$
LJ	LJ	Personal	**0.813**$^{\pm0.012}$	**0.795**$^{\pm0.011}$	0.742	**0.820**$^{\pm0.008}$	**0.820**$^{\pm0.008}$
LJ	LJ	Promotion	**0.480**$^{\pm0.059}$	**0.455**$^{\pm0.048}$	0.289	**0.491**$^{\pm0.051}$	**0.491**$^{\pm0.051}$
LJ	LJ	Academic	**0.378**$^{\pm0.187}$	**0.376**$^{\pm0.142}$	0.000	**0.361**$^{\pm0.138}$	**0.361**$^{\pm0.138}$
LJ	LJ	Information	**0.667**$^{\pm0.013}$	**0.664**$^{\pm0.019}$	0.539	**0.681**$^{\pm0.010}$	**0.681**$^{\pm0.010}$
LJ	LJ	Review	**0.641**$^{\pm0.017}$	**0.622**$^{\pm0.032}$	0.520	**0.665**$^{\pm0.019}$	**0.665**$^{\pm0.019}$

Table 3: F1-measures for classifiers trained and tested on different datasets

authors of (Sun et al., 2019) claim that for their data and BERT-like models, the optimal number of epochs equals 4 on their text classification tasks. Since (Sun et al., 2019) do not use RuBERT or XLM-RoBERTa, the novelty of our study is also in the fact that we show optimality of this number of epochs for other BERT-like

Dataset	XLM-R		RuBERT		LR	
	train	test	train	test	train	test
FTD	101.3	34.7	96.7	34.7	18.8	0.1
LJ	115	51.1	101.5	51	8.7	0.1

Table 4: Training and inference time, milliseconds averaged per text

Train	Test	Genre	XLM-R	RuBERT	LR	Ensemble2	Ensemble3
LJ+FTD	FTD	Argument	$\mathbf{0.708}^{\pm0.030}$	$0.693^{\pm0.022}$	0.550	$\mathbf{0.732}^{\pm0.012}$	$\mathbf{0.716}^{\pm0.022}$
LJ+FTD	FTD	Fiction	$\mathbf{0.780}^{\pm0.047}$	$\mathbf{0.784}^{\pm0.053}$	0.739	$\mathbf{0.809}^{\pm0.034}$	$\mathbf{0.807}^{\pm0.017}$
LJ+FTD	FTD	Instruction	$\mathbf{0.761}^{\pm0.062}$	$0.665^{\pm0.037}$	0.371	$\mathbf{0.742}^{\pm0.034}$	$\mathbf{0.714}^{\pm0.039}$
LJ+FTD	FTD	News	$\mathbf{0.945}^{\pm0.008}$	$\mathbf{0.929}^{\pm0.012}$	0.770	$\mathbf{0.945}^{\pm0.011}$	$0.913^{\pm0.020}$
LJ+FTD	FTD	Legal	$\mathbf{0.795}^{\pm0.081}$	$\mathbf{0.773}^{\pm0.037}$	0.000	$\mathbf{0.788}^{\pm0.056}$	$\mathbf{0.750}^{\pm0.133}$
LJ+FTD	FTD	Personal	$\mathbf{0.714}^{\pm0.024}$	$\mathbf{0.661}^{\pm0.041}$	0.635	$\mathbf{0.711}^{\pm0.027}$	$\mathbf{0.713}^{\pm0.018}$
LJ+FTD	FTD	Promotion	$\mathbf{0.946}^{\pm0.011}$	$0.908^{\pm0.018}$	0.455	$\mathbf{0.946}^{\pm0.015}$	$\mathbf{0.937}^{\pm0.010}$
LJ+FTD	FTD	Academic	$\mathbf{0.880}^{\pm0.015}$	$\mathbf{0.852}^{\pm0.056}$	0.000	$\mathbf{0.892}^{\pm0.017}$	$\mathbf{0.852}^{\pm0.041}$
LJ+FTD	FTD	Information	$\mathbf{0.650}^{\pm0.053}$	$\mathbf{0.627}^{\pm0.038}$	0.321	$\mathbf{0.670}^{\pm0.043}$	$\mathbf{0.624}^{\pm0.036}$
LJ+FTD	FTD	Review	$\mathbf{0.685}^{\pm0.041}$	$0.590^{\pm0.048}$	0.418	$\mathbf{0.687}^{\pm0.048}$	$\mathbf{0.653}^{\pm0.028}$
LJ+FTD	LJ	Argument	$\mathbf{0.590}^{\pm0.020}$	$\mathbf{0.563}^{\pm0.035}$	0.500	$\mathbf{0.603}^{\pm0.023}$	$\mathbf{0.609}^{\pm0.012}$
LJ+FTD	LJ	Fiction	$\mathbf{0.734}^{\pm0.028}$	$\mathbf{0.716}^{\pm0.026}$	0.637	$\mathbf{0.756}^{\pm0.028}$	$\mathbf{0.762}^{\pm0.023}$
LJ+FTD	LJ	Instruction	$\mathbf{0.795}^{\pm0.021}$	$0.787^{\pm0.011}$	0.768	$\mathbf{0.810}^{\pm0.014}$	$\mathbf{0.818}^{\pm0.011}$
LJ+FTD	LJ	News	$\mathbf{0.907}^{\pm0.006}$	$\mathbf{0.904}^{\pm0.007}$	0.864	$\mathbf{0.912}^{\pm0.005}$	$\mathbf{0.914}^{\pm0.004}$
LJ+FTD	LJ	Legal	$\mathbf{0.331}^{\pm0.216}$	$\mathbf{0.394}^{\pm0.156}$	0.154	$\mathbf{0.355}^{\pm0.143}$	$\mathbf{0.332}^{\pm0.126}$
LJ+FTD	LJ	Personal	$\mathbf{0.807}^{\pm0.011}$	$0.783^{\pm0.018}$	0.742	$\mathbf{0.816}^{\pm0.010}$	$\mathbf{0.817}^{\pm0.004}$
LJ+FTD	LJ	Promotion	$\mathbf{0.465}^{\pm0.049}$	$\mathbf{0.479}^{\pm0.047}$	0.289	$\mathbf{0.495}^{\pm0.038}$	$\mathbf{0.488}^{\pm0.051}$
LJ+FTD	LJ	Academic	$\mathbf{0.324}^{\pm0.104}$	$\mathbf{0.332}^{\pm0.107}$	0.000	$\mathbf{0.409}^{\pm0.084}$	$\mathbf{0.350}^{\pm0.158}$
LJ+FTD	LJ	Information	$0.642^{\pm0.014}$	$\mathbf{0.641}^{\pm0.030}$	0.539	$\mathbf{0.658}^{\pm0.018}$	$\mathbf{0.674}^{\pm0.012}$
LJ+FTD	LJ	Review	$\mathbf{0.640}^{\pm0.017}$	$0.604^{\pm0.026}$	0.520	$\mathbf{0.653}^{\pm0.015}$	$\mathbf{0.658}^{\pm0.016}$
LJ+FTD	FTD	Total accuracy	$\mathbf{0.822}^{\pm0.005}$	$0.789^{\pm0.009}$	0.532	$\mathbf{0.828}^{\pm0.007}$	$\mathbf{0.828}^{\pm0.007}$
LJ+FTD	LJ	Total accuracy	$\mathbf{0.757}^{\pm0.011}$	$\mathbf{0.741}^{\pm0.010}$	0.694	$\mathbf{0.769}^{\pm0.010}$	$\mathbf{0.774}^{\pm0.006}$

Table 5: F1-measures for classifiers trained on the concatenation of FTD and LJ

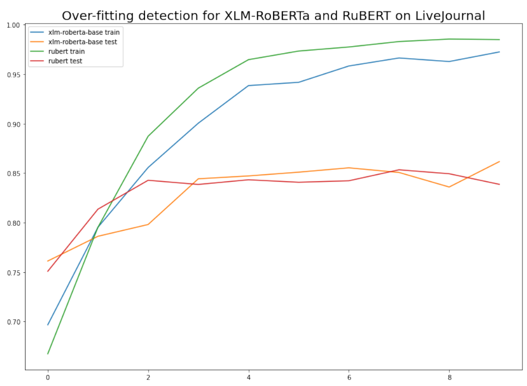

Figure 2: Learning curve of xlm-roberta-base and rubert-base-cased on LiveJournal

models.

4.3. Experiments on the big datasets

We apply the classifiers to large social media samples from VKontakte and LJ datasets from the General Internet Corpus of Russian (Piperski et al., 2013) to get an estimate of the genre distributions in social media. Table 7 and Table 8 list the corresponding distributions of the predicted classes. The most frequent class in both cases is Personal reporting with the second most common genre is Argumentation, while there are far fewer Legal and Academic texts. This result correlates well with the expected distribution of texts in social media and gives some validation for the accuracy of our predictions in the absence of a very large test set. We use chi-squared test to find the most significant differences in the shares of genres in these two samples. The LiveJournal sample has a bigger share of Legal and Argumentative texts, as it is often used as a blogging platform rather than the more traditional social network of VKontakte. In turn, VKontakte has a significantly higher share of Personal reporting as well as Promotional texts in comparison to LJ.

5. Experiment on the confidence of predictions

The evaluation of confidence depends significantly on text genre and hence requires a lot of texts for each genre for getting a stable estimation. For this reason, we apply only those classifiers which are trained in the concatenation of the LiveJournal and FTD training corpora, while also testing their concatenation.

Table 9 shows the confidence delta in the cases when the classifiers make a correct prediction, and when they

Train	Test	XLM-R	RuBERT	LogReg	Ensemble3
FTD	FTD	(A4, A1, 0.217) (A16, A1, 0.212) (A11, A1, 0.122) (A16, A9, 0.121)	(A4, A1, 0.130) (A16, A14, 0.121) (A16, A17, 0.121) (A11, A17, 0.102)	(A16, A1, 0.303) (A4, A1, 0.217) (A16, A8, 0.182) (A1, A17, 0.163) (A16, A14, 0.152)	(A16, A1, 0.242) (A4, A1, 0.217) (A4, A11, 0.130) (A17, A11, 0.118)
FTD	LJ	(A14, A1, 0.5) (A4, A11, 0.308) (A1, A11, 0.305) (A7, A1, 0.288) (A4, A1, 0.269)	(A14, A1, 0.5) (A1, A11, 0.274) (A4, A1, 0.225) (A4, A12, 0.225) (A16, A12, 0.211)	(A4, A11, 0.615) (A14, A8, 0.5) (A16, A8, 0.421) (A7, A12, 0.338) (A17, A11, 0.297)	(A14, A1, 0.5) (A4, A11, 0.346) (A1, A11, 0.263) (A7, A1, 0.25) (A16, A12, 0.237)
LJ	FTD	(A16, A1, 0.242) (A17, A11, 0.206) (A11, A1, 0.163) (A4, A11, 0.130) (A7, A17, 0.118)	(A17, A11, 0.206) (A11, A1, 0.184) (A9, A7, 0.154) (A16, A1, 0,152) (A17, A1, 0.147)	(A16, A1, 0.242) (A4, A11, 0.217) (A17, A11, 0.206) (A11, A1, 0.184) (A17, A1, 0.147)	(A16, A1, 0.212) (A17, A11, 0.206) (A11, A1, 0.163) (A4, A11, 0.130) (A7, A17, 0.118)
LJ	LJ	(A14, A1, 0.5) (A12, A7, 0.286) (A12, A8, 0.286) (A1, A11, 0.190) (A12, A17, 0.143)	(A14, A1, 0.5) (A12, A17, 0.286) (A1, A11, 0.211) (A17, A11, 0.162) (A4, A1, 0.154)	(A9, A7, 1.000) (A12, A7, 0.571) (A14, A1, 0.500) (A17, A11, 0.324) (A12, A16, 0.286)	(A14, A1, 0.500) (A12, A7, 0.286) (A12, A8, 0.286) (A17, A11, 0.270) (A4, A1, 0.192)

Table 6: The most common items in the confusion matrices

Genre label	XLM-R	RuBERT	LogReg	Ensemble3	Percentage
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

Table 7: Prediction distribution on the LJ sample, divided by 10^3

Genre label	XLM-R	RuBERT	LogReg	Ensemble3	Percentage
Argument	19941	46514	20344	30325	13.63
Fiction	5182	33320	22780	26917	12.01
Instruction	5583	12706	7960	9579	4.31
News	6925	8841	28150	12487	5.61
Legal	33	168	45	86	0.04
Personal	123338	89444	124520	116983	52.59
Promotion	25816	17604	13073	17260	7.76
Academic	284	249	131	186	0.08
Information	2341	2106	931	1397	0.62
Review	9859	11351	4526	7214	3.24
Non-text	481	156	0.015	24	0.01
Total	222459	222459	222459	222459	100

Table 8: Prediction distribution on the VKontakte sample, divided by 10^3

make a mistake. The delta value is positive, which indicates that the higher confidence does correspond to a correct prediction.

We also conducted an unpaired two-sample Mann-

Genre	XLM-RoBERTa		RuBERT		LogReg		Ensemble 2		Ensemble 3	
	stat	delta	stat	delta	stat	delta	stat	delta	stat	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**

Table 9: Mann-Whitney test for the confidence of the genre classifiers

Whitney test for each classifier and genre for the confidence levels for the correct and incorrect predictions. For 3 genres of 10, Ensemble3 achieves the highest value of Mann-Whitney statistics and confidence delta, while it is nearly always at least the second best. Moreover, it has the highest delta on the entire testing subset. The genre classifier based on Logistic Regression performs the worst.

6. Related Work

Genre classification is not a new task. Up to date, a lot of attempts have appeared to build a precise classifier of genres based on various architectures from linear discrimination (Karlgren and Cutting, 1994) to SVM (Sharoff et al., 2010) and recurrent neural networks (Kunilovskaya and Sharoff, 2019). Our study differs in the way that instead of using classical ML techniques, we apply advanced transformer-based neural architectures – XLM-RoBERTa and RuBERT. (Rönnqvist et al., 2021) produced a closely-related experiment which involved fine-tuning an XLM-RoBERTa model to predict genres for a range of languages. Among other things they showed the multilingual potential of XLM-RoBERTa, a genre classifier can be trained on one language and tested on another one. However, they have not tested robustness of predictions.

(Sharoff et al., 2010) contains a study on the performance of SVM for the problem of genre classification. (Sharoff et al., 2010) and (Kunilovskaya and Sharoff, 2019) use the Russian FTD corpus for training genre classifiers. In our study, we use the same FTD corpus. The validation accuracy in both (Sharoff et al., 2010) and (Kunilovskaya and Sharoff, 2019) is significantly lower than in our paper. Our work is different in that we use the LiveJournal corpus to validate our models by taking into account the topical shifts.

(Tsymbalov et al., 2020) describes the method to measure prediction confidence that we use in our study. The key idea of this metric presumes that a classifier can only be confident on some text examples if it predicts the correct labels to the texts with similar embeddings. We do not bring any change to the definition and the algorithm of calculation of confidence of prediction. Our major contribution is the application of this metric to the task of genre segmentation. Moreover, our study reaffirms the result of (Tsymbalov et al., 2020) that ensembles are more confident in their predictions, since the confidence delta for the ensembles is higher than that for each individual classifier.

In the original paper on BERT (Devlin et al., 2019), there are some tips for training and fine-tuning the BERT architecture for text classification given by the authors. The paper recommends using 2-4 epochs for training or fine-tuning and claims that using more epochs leads to over-fitting. It coincides completely with our experiment on the optimal number of epochs. But, since our training and testing data are in Russian, we use RuBERT and XLM-RoBERTa instead of the base English BERT. Thus, we manage to approve the pieces of advice from (Devlin et al., 2019) and show that they are valid for Russian and multilingual BERT-based models.

(Sun et al., 2019) is also an important work on how to apply the BERT architecture to the task of text classification. It researches how multiple hyperparameters for training BERT affect its performance on various Natural Language Understanding tasks. The authors propose 3 different text truncation strategies - *head-only*, *tail-only*, *head-only*, and *head+tail*. Their experiment shows that the *head+tail* method attains the best result. But we decide to use *head-only* instead of it since it is not that obvious how the *head+tail* method affects the topical shifts in data. Another hyperparameter we change is $batch\_size$ due to our limited computational resources. All the other recommendations from the paper are followed.

7. Conclusions and future research

We show that:

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. Moreover, the training and inference time for logistic regression is insignificant compared to the corresponding costs for XLM-RoBERTa and RuBERT. It means that adding a weak classifier is worth it.
4. The ensembles have more reliable predictions in terms of their confidence, as they provide the biggest confidence gap between the correctly and incorrectly classified texts.
5. 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.
6. Applying of the classifiers to large social media sample reveals their distribution of genres. Using chi-squared test, we reveal how the genre distributions vary in texts from different social media sources, such as the greater rate of Argumentative texts in LJ in comparison to the greater rate of Personal reporting in VK.

Thus, using ensembles is helpful to boost the accuracy and robustness of predictions while their only disadvantage concerns an increase in the computing costs. Though ensembles are more reliable for text genre classification, it is still unclear how efficient they are on related tasks. In future, we consider training models for seq2seq genre segmentation when we need to detect genre boundaries within a single text, for example, when a personal story includes a quote from fiction or when an expression of opinions cites informative news reports. We hypothesise that the ensembles will confirm their advantage over individual models. The confidence measure can be used to make predictions for open-set genre classification tasks (Pritsos and Stamatatos, 2018), when the classifier should refrain from making any prediction for a text, if its confidence on this text is lower than a threshold.

8. Bibliographical References

Conneau, A., Khandelwal, K., Goyal, N., Vishrav, C., Wenzek, G., Francisco Guzman´, E. G., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv*, arXiv: 1911.02116.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Karlgren, J. and Cutting, D. (1994). Recognizing text genres with simple metrics using discriminant analysis. In *COLING '94: Proc. of the 15th. International Conference on Computational Linguistics*, pages 1071 – 1075, Kyoto, Japan.

Kunilovskaya, M. and Sharoff, S. (2019). Building functionally similar corpus resources for translation studies. In *Proc RANLP*, Varna, September.

Kuratov, Y. and Arkhipov, M. (2019). Adaptation of deep bidirectional multilingual transformers for russian language. In *Proceedings of Dialogue: Computational Linguistics and Intellectual Technologies*, pages 333–339.

Lepekhin, M. and Sharoff, S. (2021). Experiments with adversarial attacks on text genres. *arXiv preprint arXiv:2107.02246*.

Petrenz, P. and Webber, B. (2010). Stable classification of text genres. *Computational Linguistics*, 34(4):285–293.

Piperski, A., Belikov, V., Kopylov, N., Selegey, V., and Sharoff, S. (2013). Big and diverse is beautiful: A large corpus of Russian to study linguistic variation. In *Proc* 8^{th} *Web as Corpus Workshop (WAC-8)*.

Pritsos, D. and Stamatatos, E. (2018). Open set evaluation of web genre identification. *Language Resources and Evaluation*, 52(4):949–968.

Rönnqvist, S., Skantsi, V., Oinonen, M., and Laippala, V. (2021). Multilingual and zero-shot is closing in on monolingual web register classification. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 157–165, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Santini, M., Mehler, A., and Sharoff, S. (2010). Riding the rough waves of genre on the web. In Alexander Mehler, et al., editors, *Genres on the Web: Computational Models and Empirical Studies*. Springer, Berlin/New York.

Sharoff, S., Wu, Z., and Markert, K. (2010). The Web library of Babel: evaluating genre collections. In *Proc Seventh Language Resources and Evaluation Conference, LREC*, Malta.

Sharoff, S. (2018). Functional text dimensions for the annotation of Web corpora. *Corpora*, 13(1):65–95.

Sun, C., Qiu, X., Xu, Y., and Huang, X. (2019). How to fine-tune BERT for text classification? *arXiv preprint arXiv:1905.05583*.

Tsymbalov, E., Fedyanin, K., and Panov, M. (2020). Dropout strikes back: Improved uncertainty estimation via diversity sampling. *arXiv preprint arXiv:2003.03274s*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.