Gender Identification in Russian Texts

Rupal Bhargava, Gunjan Goel, Anjali Shah, Yashvardhan Sharma
WiSoc Lab, Department of Computer Science and Information Systems
Birla Institute of Technology and Science, Pilani, India 333031
Email: {rupal.bhargava, h2016068, h2016066, yash}@pilani.bits-pilani.ac.in

***Abstract*—Gender Identification is a task where we have to identify the gender of the author for written texts. An hybrid approach has been designed by combining deep neural network and a rule-based classifier for russian texts. LSTM and Bi-LSTM have been used as a part of Neural Network due to their capability to learn long-term dependencies.**

***Index Terms*—Author Profillation, Deep Learning, NLP, Gender Identification, Rule-based Classification**

I. Introduction

The last few years have seen a massive research related to automatic retrieving of information from the text, mainly the information about its author (authorship profiling) like gender, age etc. The automatic extraction of the information from text related to gender is essential to forensics, security, and marketing. For example, companies may be interested to learn about the gender of the people who likes or dislikes their products which can then be analyzed to know which section of the market is disliking their products. It helps in improving the sales of a company.

Most of the studies that have been done for classifying the gender of the authors have been conducted using English texts, like blogs, twitter posts etc. but there have been very few studies dealing with other languages, especially for Slavic languages like Russian. The Subtask [1] in the shared task of the PAN in FIRE-2017 addresses the Cross-genre Gender Identification in Russian text (RusProfiling shared task). The objective of this paper is to explore the possibility of automatically classifying Russian written texts according to their authors' gender using the parameters that are rather context-independent.

Let us give a brief description of the text corpora that was given by PAN for evaluation. The corpus was divided into five subgroups : First group contains Offline texts (picture descriptions, letter to a friend etc.) from RusPersonality Corpus. The second one has posts from Facebook then third contains tweets from various users. The fourth group has data related to products and service online reviews. While the last group contains gender imitation corpus where women are imitating men and the other way round. The test corpus is widely distributed on the different type of datasets to make the classifier be context independent.

II. Related Work

The gender of the authors is one of the characteristics that may affect the style of writing texts. There are a lot of papers on automatic detection of personality traits using texts. Research in identifying author's gender started with extensions of this work on categorization and classification of text [7]. With regard to the shared task on Author Profiling at PAN [2, 3], most participants used combinations of style-based features such as frequency of punctuation marks, capital letters, quotations, together with POS tags and content-based features such as Latent Semantic Analysis, bag-of-words, TF-IDF, dictionary-based words, topic-based words. Using the various methods and features, researchers have automated prediction of an author's gender with accuracies ranging from 80% to 90%. For instance, the winners of PAN 2015 obtained models to classify texts according to the gender of their authors with the accuracy as high as 0.97 for Danish and Spanish and 0.86 for English [4].

Most of the studies have been done in English language for this problem and very few details regarding slavic language have been studied. [5] have talked about the property of the Russian Language. The gender of the speaker can be known in Russian texts if a verb in a sentence is in the past form and the subject is a singular first person pronoun "I". This property of the Russian language can be used in identifying the gender and it has been used for classifying the gender in this paper. Many authors have considered deep neural network for sentiment analysis but very few[6] have addressed gender classification problem using neural network. It explores the potential of deep learning network for the PAN task of Gender Identification in Russian texts.

III. Data Analysis

The training data provided [1], consisted of 600 XML files containing the Russian text. Each file contains the tweets in the Russian language with the author's id as the file name. A separate file containing the labels for each sample data was provided. As the problem is binary classification, the data was divided into 2 labels as "male" and "female". Class Type "male" is assigned to the tweets written by male author and Class Type "female" was assigned to the tweets written by the female author. The dataset contains 300 samples for each class.

Further, analysis of the data showed that the Russian tweets given in the training dataset, also contains many English words. So using only Russian embedding for neural network won't be sufficient so English glove embeddings were also added to enhance our model. Later, translation of data was done into English language and after analysis it was concluded that the Russian language can distinguish the gender of the

speaker using the verb if the given statement is in past form. This analysis has improved our results greatly.

IV. Proposed Technique

A hybrid approach has been proposed which is used to identify the gender from Russian texts. The proposed terminology involves the following techniques which are combined together to generate the final classifier :

- Preprocessing
- Rule-Based Classifier
- Neural Network
- Classification

A. Preprocessing

The given training data was preprocessed for efficient application and precise classification of the classifiers.The data was also preprocessed for separating the class labels. The preprocessing of the dataset includes:

1) *Retrieving the text from XML file :* Each XML file contains XML tags. These tags aren't required for the training and hence text needs were extracted from the XML tags.
2) *Removal of emoticons, numbers, punctuation marks :* After extracting the tweet text, emoticons and numbers were filtered using TweetTokenizer API of NLTK. This module also includes the filtering of the unnecessary punctuation marks.
3) *Filtering stopwords :* Some common terms were dropped before creating the word embedding in order to train the model precisely. An available list of Russian stop words and customized it according to our need. For e.g, for applying rule-based classifier, the verb shouldn't be considered as a stop word.
4) *Case Conversion :* All the data entries are converted into lowercase. This technique involves the replacement of the upper case letters to their lowercase counterparts.

B. Rule-Based Classifier

The author's gender is known to be explicitly expressed in Russian texts if a verb in a sentence is in the past form and the subject is a singular first-person pronoun я.Compare: "Прошлой зимойя ездила в Альпы" (Last winter I went in the Alps - a female speaker); "Прошлой зимойя ездил в Альпы" (Last winter I travelled to the Alps - a male speaker). If the subject is not the pronoun "я" or if the verb is not in the past form, the gender of the speaker is not explicit. Compare:"Я поеду в Альпы" (I'll go to the Alps); the gender of the speaker is not explicit.

It is worth emphasizing that the existence of grammatical forms which reflect the speaker's gender does not automatically make gender identification in Russian texts a trivial task. Any non-first person narrative does not indicate the gender of its author. Besides, it is easy for the author to imitate the speech of an individual of the other gender using the above forms. NLTK toolkit has been used to generate Parts of Speech(POS) tagging. If a term has a tag "VBD" or "VBN" then it determines past tense verb and then it checks whether the verb ends with "ла" or "л". If any of the prefixes is found then the corresponding result is saved otherwise it moves ahead to check for other terms.

The above rule has been applied on the available training dataset and classified the texts based on that.. If the text contains "Я" and verb is in past sense then it is classified to male or female. The results are stored in a file, 1 represents female and 0 represents the male. But, if a rule is not satisfied by a text then it is passed to our pre-trained neural network for classification.

C. Neural Network

The main advantage of using deep learning in classification procedure is that it doesn't need to identify the features on our own. Neural network trains the model based on the attributes provided. LSTMs(Long Short Term Memory) is a special kind of RNN, capable of learning long-term dependencies. These are explicitly designed to avoid the long-term dependency problem that exists in RNN. LSTMs can remember information for long periods of time which makes it more useful for texts. We have utilized the behaviour of LSTMs in our model so that it can learn more effectively.

The entire training data is split into training and validation data with VALIDATION_SPLIT of 0.2. The input to the neural network is the text provided in the training dataset. Each text acts as one input to the neural network. We padded the text sequences to a maximum length which allows to having a fixed dimension in the final matrix. An embedding is created for the words present in the text that uses the available pre-trained word embeddings of Russian and English.

This embedding matrix is fed to the very first layer of the network i.e Embedding Layer. The output is then fed to the LSTM layer with filter width of 5. The value was decided after experimentation. Then, a DENSE layer is used with sigmoid as an activation function. The optimizer used to learn the neural network is RMSProp with a learning rate of 0.1. The final predictions are verified with the ground truth and loss is calculated using binary cross entropy and the validation data is used to calculate this loss and the model is trained accordingly.The final model is converted to JSON and stored. The weights corresponding to the model are stored in an h5 file. Table 1 contains all the required hyperparameters for the neural network. These parameters were decided after rigorous training and experimenting with various values.

Table I
Hyperparameters for neural network

HyperParameter	Value
Learning Rate	0.1
BatchSize	60
Epoch	15
Filter Width	5

D. Classification

The final step of any classification problem is finding out the class of the input data. Here. in author profillation we

need to find out the gender of the author which is a binary classification problem. The value associated with the female is 1 and for the male, it's 0. A hybrid approach has been used where the rule-based classifier and neural network are combined together to solve the problem.

The input test data is preprocessed using the technique mentioned in 4A. After preprocessing the text is given to rule-based classifier and it checks whether the text can be classified based on it. If yes, then store the result in a file. If not, then fed the data to the pre-trained neural network and store the prediction in a separate file. Finally, we merge the two files to store the results of entire test data.

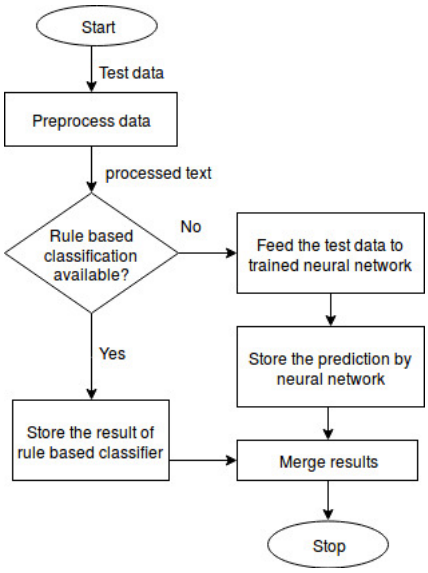

Figure 1. Classifier for gender identification

The final output file contains file name of the test data and the corresponding prediction is provided as 1 or 0 where 1 represents female and 0 represents the male.

V. Experiments

The FIRE task involves the classification of the author into male and female. The training dataset contained 600 tweets sample in the Russian language. Among those, 480 samples were used for training the model and remaining 120 samples were used for testing. The test dataset included tweets, Facebook posts, online products reviews, texts describing images, or letters to a friend and gender imitation corpus. The main objective of the task was to identify the gender of the author of this text. Total five teams have participated and each team used three different approaches for classification and generated results as shown in Figure 2. Our proposed hybrid approach which contains rule-based classifier along with Deep Neural Network has achieved an accuracy of 87.28%. We submitted a total of 5 runs based on 5 different classifiers. (Rule-based classifier, Neural Network using LSTM, Bi-LSTM and combining both with rule-based classifier). The hybrid approach has performed well in all the test datasets. The results are discussed further in the subsequent section.

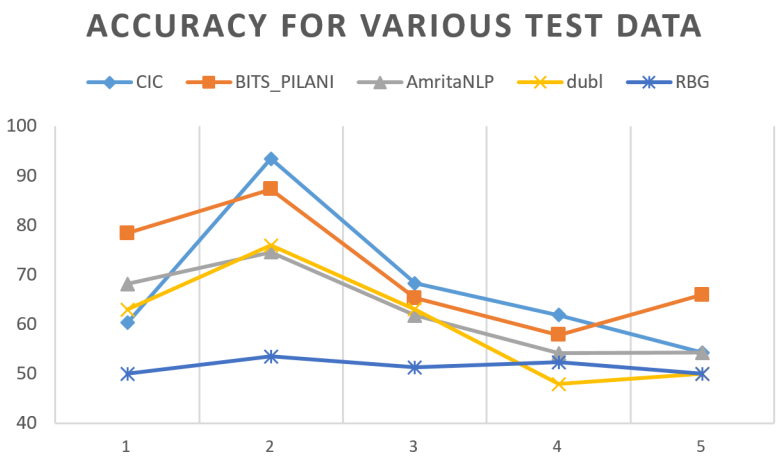

Figure 2. Comparison of highest accuracy for participating teams

The task was evaluated on the basis of accuracy and our team BITS_PILANI has achieved an accuracy of 87.28% for facebook posts. Combining the rule-based classifier along with the Neural Network resulted in the highest accuracy among all the test runs that were submitted as a part of the task. It is because of the neural network, that the model performs good for "Gender Imitation" dataset as well. Figure 5.1 shows the accuracy of all the participating teams for the various test datasets. The test dataset are as follows:

1) Offline Texts (picture descriptions, letter to a friend) from RUSPROFILLING corpus
2) Facebook Posts
3) Twitter tweets
4) Product and Service online reviews
5) Gender Imitation Corpus

The proposed approach provides good accuracy as the fact that in Russian languages the gender of the author can be distinguished using verbs.

VI. Conclusion

In this paper, the hybrid approach of rule-based classifier and Neural Network has been proposed for task of PAN, FIRE 2017. Presented approach uses LSTM and Bi-LSTM as the Neural Network trained with the given dataset. The trained network can still perform better if trained with large dataset as neural network performs better when trained with huge dataset. The results can still be further improved if morphological, stylistic and content based features can also be added to the rule-based classifier. It would be interesting to find certain language-based features and hyperparameters for Neural Network that may further improve the accuracy.

References

[1] Tatiana Litvinova, Francisco Rangel, Paolo Rosso, Pavel Seredin, Olga Litvinova. Overview of the RUSProfiling PAN at FIRE Track on Cross-genre Gender Identification in Russian. In: Notebook Papers of FIRE 2017, FIRE-2017, Bangalore, India, December 8-10, CEUR Workshop Proceedings. CEUR-WS.org

[2] Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G. “Overview of the Author Profiling Task at PAN 2013 Notebook for PAN at CLEF 2013”, Forner et al.
[3] Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W. “Overview of the 2nd author profiling task at pan 2014”, CLEF 2014 Labs and Workshops, Notebook Papers. CEUR-WS.org
[4] Rangel, F., Fabio, C., Rosso, P, Potthast, M., Stein, B., Daelemans W. “Overview of the 3rd Author Profiling Task at PAN 2015”, CEUR Workshop Proceedings. Toulouse, France
[5] T. Litvinova, P.Seredin, O. Litvinova, O. Zagorovskaya, A. Sboev, D. Gudovskih, I. Moloshnikov, R. Rybka “Predicting The Gender of an Author of a Russian Text Using Regression and Classification Techniques”, in Proc of CDUD 2016
[6] Sboev A., Litvinova T., Voronina I., Gudovskikh D., Rybka R. “Deep Learning Network Models to Categorize Texts According to Author’s Gender and to Identify Text Sentiment”, Proceedings of 2016 International Conference on Computational Science and Computational Intelligence
[7] Koppel, M., Argamon, S., Shimoni, A.R. “Automatically categorizing written texts by author gender”, Literary Linguist. Comput.