AmritaNLP@PAN-RusProfiling : Author Profiling using Machine Learning Techniques

Vivek Vinayan, Naveen J R, Harikrishnan NB, Anand Kumar M and Soman KP
Center for Computational Engineering and Networking
Amrita University, Coimbatore, India
vivekvinayan82@gmail.com,naveenaksharam@gmail.com,harikrishnannb07@gmail.com
m_anandkumar@cb.amrita.edu,kp_soman@amrita.edu

ABSTRACT

This paper illustrates work done on "Gender Identification in Russian texts (RusProfiling)" shared task, hosted by PAN in conjunction with FIRE 2017. The task is to predict the author's gender, based on the Twitter data corpus which is in Russian.

We will give a brief introduction to the task at hand, elaborate on the data-set provided by the competition organizers, discuss various feature selection methods, provided experimental analysis that we followed for feature representation and show comparative outcomes of different classifiers that we used for validation.
We submitted a total of 3 models and their respective prediction for each test data-set with slightly different pre-processing technique based upon the test corpus content. As each of the test corpus were sourced from various platforms, this made it challenging to stick to one representation alone.
As per the global ranking published for the shared task[6] our team secured 2nd position overall (Concatenating all Data-set) and our 3rd submission model performed the best among the 3 submission models from the overall test data corpus.
Further under extended work we discuss in brief how hyper parameter tuning of certain attributes extend our validation accuracy by 6% from baseline.

KEYWORDS

Author Profiling, Russian Language,
Text Classification, Semi-supervised Classifiers

ACM Reference Format:
Vivek Vinayan, Naveen J R, Harikrishnan NB, Anand Kumar M and Soman KP. 2017. AmritaNLP@PAN-RusProfiling : Author Profiling using Machine Learning Techniques. In .

1 INTRODUCTION

The Internet, it is a vast platform where anyone can have access to myriads of information, from online news media articles to various social media platforms, from personal blogs to personalized websites, all this literally at the end of our fingertips, and in this present age, life is becoming unimaginable without it. With the availability to all this resources, people are writing and share information more avidly over the internet than ever before, and it also provides a certain degree of anonymity while doing so. Access to such multitudinous information brings in certain set of problems like theft of identity/content and plagiarism to name a few and this we are trying to address with tasks such as "Author Profiling".

In "Author profiling" [9] sharedtask, we examine the style of an individual author and thus distinguish between classes of authors by studying their sociolect aspects. In an even broader manner it helps in predicting an author's demographic, personality, education and socio-networks through classification of texts into classes, based on the stylistic choices of the author.
With this paper on RusProfiling sharedtask, we focus on cross-genre gender identification in Russian texts[6], which is becoming a part of, one of the most upcoming trending task in NLP domain, under "Author Profiling"[7, 8].
In this task we have Twitter as training data corpus and as test data corpus we have dataset from multiple social media domain platforms like Twitter, Facebook, online reviews (where texts are describing images, or letters to a friend), product and services. The focus with this task is on gender profiling in social media and the main interest is in the everyday language and on how the basic social and personality skills reflects on their writing [3, 5].

The main challenge in this task is the language itself, as it is not a native, thus we used certain pre-processing methods and built our baseline representation on which we implemented classical machine learning algorithms for this text classification task.

2 CORPUS

The Corpus for the training data was mainly sourced from social media website Twitter and the labels were annotated for each of the document with the author gender "male" or "female". The training corpus is a collection of 600 data file in XML format which consists of exactly half female and half male genre documents, the file name are annotated by their associated gender label in a separate file called "truth" which is in text format.

Table 1: Training Dataset Statistics

Training Dataset	
Total number of documents	600
Total number of male documents	300
Total number of female documents	300

A cursory analysis of the training corpus reviled that each training data file has a combination of different tags and hyperlinks, further the documents varied in count of content words i.e one document went from no text to others over 3000 plus words in a single document. Few of the files had mixed data of Russian and English, where as few other where completely in English language.

The test corpus is presented in 5 folders varied by the category of different sources. Each set contains different amount of files, the

Table 2: Testing Dataset Statistics

Testing Dataset	
DS1- Offline Texts (picture description etc)	370
DS2 - Facebook	228
DS3 - Twitter	400
DS4 - Online Reviews	776
DS5 - Gender Imitation Corpus	94
Total number of documents	1868

Table 3: Vocabulary Size based on min_df and n-gram range

(min, n-gram)	VOCABULARY SIZE					
	Training Dataset	DS1	DS2	DS3	DS4	DS5
(1,1)	183119	96544	52098	16077	9184	6797
(1,2)	852380	390683	234284	80991	39732	25139
(1,3)	1698948	746445	451658	164853	75790	46435
(1,4)	2583433	1114714	672974	252602	111942	68167
(2,1)	45223	22482	17278	5989	3438	2151
(2,2)	91646	38869	29643	13529	6528	3562
(2,3)	106192	43324	31965	16190	7291	3898
(2,4)	109234	44206	32232	17074	7565	3987
(3,1)	28365	13188	9973	3801	2190	1233
(3,2)	106192	19702	15086	7202	3381	1767
(3,3)	52940	20923	15708	8098	3526	1864
(3,4)	53646	21065	15745	8381	3545	1886
(4,1)	20677	9228	6910	2805	1594	821
(4,2)	32403	12953	9931	4903	2305	1101
(4,3)	34876	13551	10226	5380	2376	1139
(4,4)	35184	13604	10237	5512	2383	1145

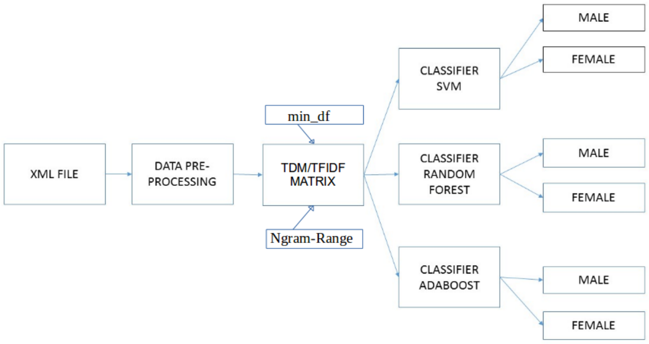
count of documents for each category varies from 96 to 776 files each. On further inspection the text format provided in each folders apart from the 3rd folder is different when comparing with the training corpus, namely offline texts, Facebook, Twitter, product and online reviews and gender imitation corpus in order of their folder number respectively as shown in Table 1-2.

We have also taken the statistical data of the complete vocabulary size that we gained from grid search of attributes namely n-gram_range and min_df count. In each combination their respective corpus size is found, and the statistics are tabulated in Table 3.

3 METHODOLOGY

The Figure 1 gives a rudimentary picture of the architecture that we have implemented for our 3 submissions, in all of these models we mainly focused on data pre-processing methods to incorporate various features and build upon each one of them to improve the feature representation. We started from a simple count based model, the same methods are discussed next.

Figure 1: Architecture of our model for the Sharedtask



3.1 Feature Selection

The feature selection was a process in which we started by building a baseline model [1] and improved on the accuracy of the model with step by step empirical procedure of combining and modifying the existing feature representation [10].

- **Count Based Matrix** :
  The 1st approach from the dataset was to form a simple count based Term Document(TD) and Term Frequency Inverse Document Frequency (TFIDF) matrix which became the baseline for our accuracy and further went with adding general features to previous representation.

- **Feature Extraction** :
  With the knowledge of the social media network "Twitter", we essentially narrowed down our focus on search for features to tags, like '@' which is mainly used to address people/gathering and hash tag '#' which is based on the context or the image of the adjoining content. Moving on, we found that URLs were being used widely across most of the dataset which linked to various internet sources, so we then incorporated these as a feature to the earlier feature representation which proved to show slight improvement on all of the classification algorithms, captured below in Figure 3-4.

- **Data Normalization** :
  On further analysis we found that, individual URL's in itself as a feature seemed fruitless, thus only considering the hyperlink itself, we focused on normalizing these across the dataset and went with the count of the URL and those of the tag's as feature to represent a document. It proved to increase the accuracy little more, This further led to normalizing of various emoticons represented by a keyword and various other punctuation like the exclamation mark '!', period '.' and hyphen '-' which occurred multiple time or in continuous repetition were converted to a single instance of each.
- **Word Average:** :
  As we were not familiar with the language, we considered the average word length as the total number of character per document to the total number of feature instance in that document and appended that list as an average word length

Figure 2: Pre-Processing

Training Data:

@BorisVasilevski3 главных вопроса для постановки целей https://t.co/mDjHguJBaK@timarina2 привет!!! @RinatDecorator1 Ринат, как ты?Есть такие слезы, которые надо выплакать обязательно...
В любое время дня и ночи...
Чтобы в нутри перегорело...
https://t.co/VD9sHFX0nE@BorisVasilevski, а точку...)))
@timarina2@tunukbek3@tunukbek3@fadin_ivan
@timarina2@70e8afbc3f2349418 уже есть | Красота | Новости | VOGUE http://t.co/cfoSlqQHvRCлов нет.... http://t.co/E2Cy5AJcd1@ksorbs Nor when I was not at this game and did not see the game, now

Processing Data:

@ borisvasilevski3 главных вопроса для постановки целей https привет. @ rinatdecorator1 Ринат, как ты?Есть такие слезы, которые надо выплакать обязательно.
В любое время дня и ночи.
Чтобы в нутри перегорело. https а точку.))) @ timarina2 @ tunukbek3 @ tunukbek3 @ fadin_ivan @ timarina2 @ 70e8afbc3f2349418 уже есть | Красота | Новости | vogue https нет. https nor when i was not at this game and did not see the game, now

per document making it an independent feature. This is to accommodate for the fact that the average vocabulary word length that gender used can also be taken as a discriminative feature between the 2 classes.

4 EXPERIMENT AND DISCUSSIONS

As a part of experimental analysis we manually ran over few random training documents based on the individual sizes of the file to gather a glimpse of the overall change in data, then ran snippets on these training set data to see the scale of, improvement of, accuracy with the above considered parameters. Thus to distinguish for better feature representation for the classification.

After going through various transitions, the selected certain features were extracted and then used as a part of pre-processing of the entire training corpus these features were individually added one set at a time to show the increase or decrease in their accuracy corresponding to various classifiers by cross validation with different classical ML classifiers, namely Logistic regression (LR), Support Vector Machine (SVM) using linear kernel, Decision tree (DT), Adaboost and Random forest(RF)[2] the result are as displayed in Table 4.

Table 4: Cross-validation Result with Different Classifier

SN	Count Matrix	LR	SVM linear	DT	Adaboost	RF
1	TD	63.33	79.66	74.00	83.00	82.66
1	TFIDF	70.33	72.50	70.00	83.16	81.83
2	TD	66.70	81.83	75.00	85.16	84.16
2	TFIDF	72.16	75.83	75.33	80.83	81.66
3	TD	61.70	81.83	74.83	83.99	84.99
3	TFIDF	72.80	78.00	68.16	82.49	80.83
4	TD	66.70	81.00	74.10	85.66	82.99
4	TFIDF	74.00	78.00	68.00	81.66	82.83
5	TD	70.00	79.83	74.49	85.33	83.66
5	TFIDF	74.00	77.50	67.16	82.83	81.66

The following are the feature we considered one step at a time and consecutively added the next feature to the previous set as mentioned below:

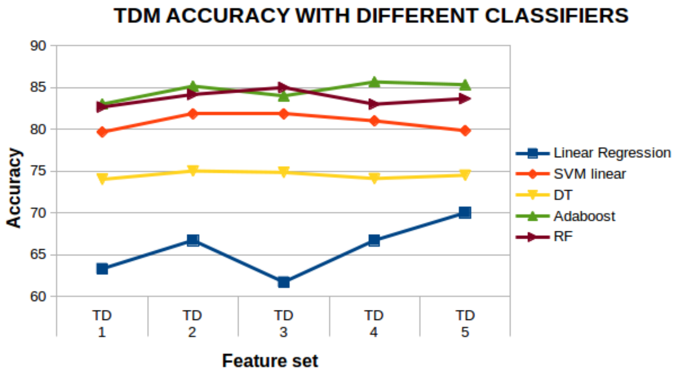


Figure 3: TD Classifier Accuracy

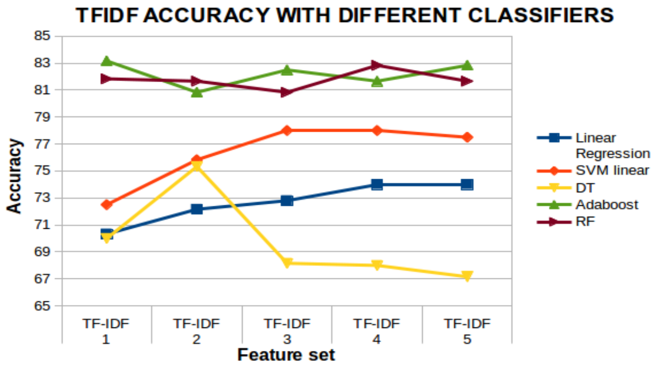


Figure 4: TFIDF Classifier Accuracy

(1) A simple count based matrix is taken to achieve baseline accuracy, for this we have considered both TD and TFIDF matrix representation from which we set a base-line accuracy of 80.5% (We randomly initialized few attributes like n-gram_range = 2 , min_df = 3 and used a linear SVM classifier to get the baseline).
(2) Count of 'http' and 'https' are taken and converted to a single key word 'https' as this will help in adding feature to see the usage of URLs between the 2 class distinguishing which gender base might have used more number of hyperlinks within their tweets.
(3) Count of '#' tags was further attributed to the previous representation.
(4) Replaced emoticons with keyword.
(5) Took the average word length in a document i.e count of character to number of feature instances as the language, this we chose as a preferred method.

5 FEATURE REPRESENTATIONS MODELS

We submitted a total of 3 models/run's and for each individual run the following pre-processing method have been followed:

- **Submission 1** : We have considered feature representation 2, 3, 4, 5 and also the normalization of '@' followed by content tags to simple key word(Splitting tags from their context otherwise to preserve the word content in particular did not show much difference in validation accuracy), and used SVM classifier for classification. Based on learned model from training corpus the prediction for the test corpus's were taken.
- **Submission 2** : The same feature representation as the 1st run was considered, but we used a different classifier, we took Adaboost based training model and the prediction for the test corpus's were taken
- **Submission 3** : In this run, we considered mostly with regard to the other test datasets 1,2,4 and 5 where the content are in longer and in paragraph form rather than the shorter version and there was less to no use of tags and or hyperlinks. Thus to normalize this we disregarded the above used tags and reduced any extended repeat of punctuation's to a single count(e.g:'...' is shortened to a single '.')
  A sample of this is shown in Figure 2.

6 RESULTS

As per the global ranking published for the shared task by the organizers[6] our team secured 2nd position overall (Concatenating all dataset). From the rankings our 3rd submission performed the best compared to our team's previous 2 submissions by a margin of 1%, 2% respectively whereas from the leading team we trailed by margin of 6%, this was w.r.t to the facts that we mentioned in our submission 3 and also we got better validation accuracy for the submission Model 3 for datasets 1,2,4 and 5.
Individually, submission 3 gaining 2nd best accuracy in "off line texts (picture descriptions, letter to a friend etc.)from RusProfiling corpus" whereas the submission 1 gained our team 2nd place for "gender imitation corpus" and 3rd in "product and service online reviews corpus".

7 EXTENDED WORK

In our earlier experiments we randomly initialized our attributes like max feature length, n_gram and min_df with 10000, 2 and 3 respectively. As a motive to increase validation accuracy we performed a grid search for the hyper-parameter namely word count, n_gram and min_df based values with the TFIDF model, where we considered the following range of data values for each:

Word count : 10000 - 50000
ngram-range : 2 - 6
Min-df : 1 - 4

After applying grid search we pushed the baseline accuracy to 82.5% when initializing max_feature with 10,000, n_gram with 2 and min_df with 1 and applying a linear SVM classifier. We further pushed our validation to 86.49% by applying Adaboost classifier. Over all we found that the trend of accuracy of TD feature representation model decreased with increase in all the attributes, and the accuracy of TFIDF feature representation model increased but saturated after n_gram value exceeds 6 and the min_df value exceeds 4, the same is show in Figure 5 where the best of the attributes, feature combination were taken for each TD and TFIDF representation.

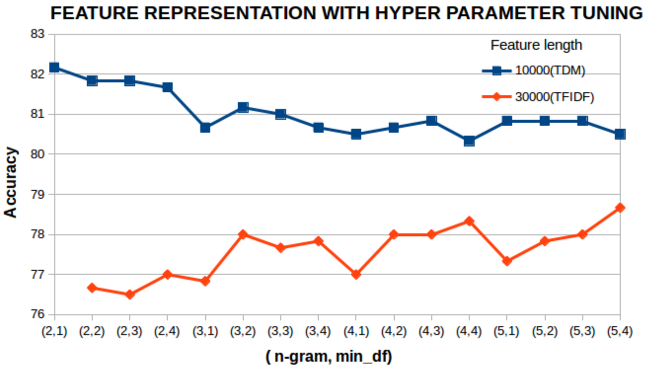


Figure 5: Hyper-parameter tuning for feature representation

8 CONCLUSION & FUTURE WORK

The challenge in this shared task we faced was the fact that we were working on a language corpus that is non native to us, thus we mainly focused on pre-processing and normalizing the data corpus to get improved feature representation. We built from a basic count representation and incorporated simple modification on iterating feature representation and observed the various accuracy changes involved with those features. Based on the experimental analysis and further discussion on optimizing of the various attributes in the extended work, we could make an inference that the baseline can further be increased which could better improve the prediction, fetching us better gender identification model.

As a future study we considered making various embedded representation for the Russian corpus and use deep learning techniques for categorizing author gender [11]. As these methods require more number of training instances we are considering including certain additional corpus provided by PAN [4] for this task and also consider certain portions of labelled test dataset based on the variety of the source that they are taken from.

REFERENCES

[1] H.B. Barathi Ganesh, M. Anand Kumar, and K.P. Soman. 2016. Statistical semantics in context space: Amrita CEN@author profiling. *CEUR Workshop Proceedings* 1609 (2016), 881–889.

[2] H.B. Barathi Ganesh, U. Reshma, and M. Anand Kumar. 2015. Author identification based on word distribution in word space. *2015 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2015* (2015), 1519–1523. https://doi.org/10.1109/ICACCI.2015.7275828

[3] Fabio Celli, Bruno Lepri, Joan-Isaac Biel, Daniel Gatica-Perez, Giuseppe Riccardi, and Fabio Pianesi. 2014. The Workshop on Computational Personality Recognition. (2014).

[4] Tatiana Litvinova, Olga Litvinlova, Olga Zagorovskaya, Pavel Seredin, Aleksandr Sboev, and Olga Romanchenko. 2016. " Ruspersonality": A Russian corpus for authorship profiling and deception detection. In *Intelligence, Social Media and Web (ISMW FRUCT), 2016 International FRUCT Conference on*. IEEE, 1–7.

[5] Tatiana Litvinova and Olga Litvinova. 2016. Authorship Profiling in Russian-Language Texts. In *Proceedings of 13th International Conference on Statistical Analysis of Textual Data (JADT 2016), University Nice Sophia Antipolis, Nice.* 793–798.

[6] Tatiana Litvinova, Francisco Rangel, Paolo Rosso, Pavel Seredin, and Olga Litvinova. 2017. 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

[7] Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. 2013. Overview of the author profiling task at PAN 2013. *Notebook Papers of CLEF* (2013), 23–26.

[8] Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. 2016. Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations. *Working Notes Papers of the CLEF* (2016).

[9] Francisco Manuel Rangel Pardo, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. 2015. Overview of the 3rd Author Profiling Task at PAN 2015. In *CLEF 2015 Evaluation Labs and Workshop Working Notes Papers*. 1–8.

[10] Aleksandr Sboev, Tatiana Litvinova, Dmitry Gudovskikh, Roman Rybka, and Ivan Moloshnikov. 2016. Machine Learning Models of Text Categorization by Author Gender Using Topic-independent Features. *Procedia Computer Science* 101 (2016), 135–142.

[11] Aleksandr Sboev, Tatiana Litvinova, Irina Voronina, Dmitry Gudovskikh, and Roman Rybka. 2016. Deep Learning Network Models to Categorize Texts According to Author's Gender and to Identify Text Sentiment. In *Computational Science and Computational Intelligence (CSCI), 2016 International Conference on*. IEEE, 1101–1106.