Detecting Gender by Full Name: Experiments with the Russian Language

Alexander Panchenko[1,2] and Andrey Teterin[1]

[1] Digital Society Laboratory LLC, Moscow, Russia
[2] Université catholique de Louvain, Louvain-la-Neuve, Belgium
alexander.panchenko@uclouvain.be

Abstract. This paper describes a method that detects gender of a person by his/her full name. While some approaches were proposed for English language, little has been done so far for Russian. We fill this gap and present a large-scale experiment on a dataset of 100,000 Russian full names from Facebook. Our method is based on three types of features (word endings, character n-grams and dictionary of names) combined within a linear supervised model. Experiments show that the proposed simple and computationally efficient approach yields excellent results achieving accuracy up to 96%.

Keywords: gender detection, short text classification.

1 Introduction

The Web is full of user-generated content: a plethora of platforms and technologies let a user create comments, posts and other types of textual messages. Most of the time, some information about author of a given text is available. In its simplest form, an author is represented with a *name string*, being either a real name or an alias. While some platforms, such as Facebook and Google+, let the user indicate gender, age and other information, others, such as Twitter or Web forums do not. Furthermore, in most of the platforms, only a name string is required, while other fields, such as gender, can be left unspecified. This is why often only two things can be used to describe a user: her name string and her texts.

However, in many cases it is desirable to know more about an author of a given piece of user-generated content. For instance, in the Internet marketing information about gender and age helps to improve targeting of advertisements [1]. In cyber security, user profiling can help track down Internet predators and assist in investigations of crimes [2]. In information retrieval, sociodemographic attributes can help customize user search experience and provide more relevant results [3, 4].

All these factors motivate the need for systems that infer gender, age and other latent sociodemographic attributes of a user. In this paper, we investigate one particular task in this research direction – *gender detection.* In particular, we focus on gender recognition of Russian full names. The goal of our method is to guess gender of a person by her name. Such a technology can be of use when a true user name is known, but gender was not specified, e.g. for analysis of Twitter users.

There have been some attempts to propose a method for automatic gender recognition (see Section 2). Major limitations of the prior researches are following: (i) most of the studies focus on gender recognition using text written by a person, neglecting full name of a user; (ii) most of the prior works deal with English language, neglecting particularities of other languages, such as Russian. Indeed, most researchers have focused on gender detection by text [2, 5–10] with some exceptions, such as [11]. Accuracy of the state-of-the-art approaches in this field is about 80-90%. However, in practical tasks higher accuracy is desirable. We show that one can recognize gender of a person with accuracy higher than 95% if a full name is available (it is normally the case in social networks and blogs).

Our work fills the gaps mentioned above, as we study gender recognition methods for Russian language based on full name of a person. The main contributions of this paper are as follows. First, we show that for Russian language, the problem is not very difficult. Even a simplest statistical model yields accuracy of 85%. Second, we propose a more sophisticated, yet simple and efficient, method that is able to recognize gender of a Russian name with accuracy and precision up to 96%.

It is possible that some Russian Internet companies, such as Yandex [3] and Mail.ru [4] already developed technologies similar to ours [12]. However, to the best of our knowledge, we are the first to openly describe details of such technology for the Russian language. A live demo of our method is freely available online [5].

2 Related Work

Gender of a text author is often known. This makes it easy to build a training corpus of articles, blogs or posts labeled with gender tags. Provided that the gender detection technology has immediate applications ranging from marketing to cyber-security, no wonder many researchers tried to build supervised models predicting gender by text.

Koppel et al. [5] describe an approach that identifies gender of an author of a written document. In this experiment a genre-labeled subset of the BNC corpus [6] was used. The proposed method relies on combination of lexical and syntactic features and yields accuracy of roughly 80%. The best performance in this experiment was achieved by a linear model based on function words and parts-of-speech n-grams.

Goswami et al. [6] describe a gender detection experiment with 9,660 gender-labeled blog posts from the `blogger.com` platform. Features used in this ex-

[3] `http://www.yandex.com/`

[4] `http://www.mail.ru/`

[5] `http://research.digsolab.com/gender`

[6] `http://www.natcorp.ox.ac.uk/`

periment include slang words statistics, sentence length and other stylometric parameters. The proposed model yields accuracy of 89.3%.

Mukherjee and Liu [13] propose two novel approaches to gender classification. The first is based on variable length POS sequences, while the second relies on automatic feature selection. The authors report increase in accuracy due to these features from 79.6% to 88.6% on a collection of 3,100 gender-labeled blogs from the `blogger.com`.

Peersman et al. [2] describe a method for short text classification by gender. The authors deal with a gender-balanced corpus of messages coming from the Dutch social network *Netlog*. The proposed technique relies on an SVM classifier [14] with features based on word/character unigrams, bigrams and trigrams. The best accuracy score of 88,8% in this experiment was achieved with a model based on 50,000 most informative word unigrams selected with χ^2 test.

Daniel and Zelenkov [12] performed a statistical analysis of the spoken subcorpus of the Russian National Corpus [7]. It appeared that in public communication there is a statistically significant difference between the speech of men and women (men talk more), while the same difference is absent in private communication. The article also mentions a gender recognition system for written texts with accuracy "about 90%" trained on the same corpus.

It is worth mentioning that age and gender prediction have much in common. First, often researchers tackle two these problems in the same study [7, 2, 6]. Second, the state-of-the-art techniques for age and gender prediction are fairly similar. In their simplest form these methods are based on supervised linear models trained on character and/or lexical unigrams. Nguyen et al. [8] also points out an interaction between age and gender variables. Furthermore, the authors study impact of gender on quality of age prediction. For instance, it was found that age prediction works better for females.

Ciot et al. [9] were among the first to present an experiment on non-English data. The authors tackled the gender recognition problem of French, Japanese, Indonesian and Turkish texts. This study revealed that (i) the methods working well in English yield good results as well for French, Turkish and Indonesian; (ii) baseline methods provide poor results for Japanese; (iii) language-specific features can boost accuracy of the baseline approach.

The work of Burger et al. [11] is arguably the one most similar to our research. The authors proposed a model based on features extracted both from texts written by a person and his full name. The study is based on a dataset of 184 thousand Twitter users speaking more than 13 languages. However, Russian-speaking users were not studied in this experiment. Similarly to other researchers, Burger et al. used supervised models trained on character and word n-grams. Their model based on full names achieved an accuracy of 89%, while the model based on all text fields of Twitter profile provided an accuracy of 92%.

Thus, recently gender detection technology received a significant attention in the literature. Some further related experiments include Rao et al. [10], Rangel and Rosso [7], Al Zamal et al.[15] and Lui et al. [16].

[7] `http://www.ruscorpora.ru/en/`

3 Dataset

We performed our experiments on a dataset of 100,000 names of Facebook users with publicly available profiles. Each such *name string* contains first and last name of a user or whatever information the user inserted into this field instead. Each name string has a gender label: *male* or *female*. We did not consider users with unknown gender. The dataset was collected with the Facebook API [8] from publicly available profiles of Russian-speaking users. The dataset contains both names written in Cyrillic and Latin alphabets, e.g. "Alexander Ivanov" and its Cyrillic equivalent "Александр Иванов".

Fig. 1 lists the most frequent names and surnames in the dataset. Here and below we present transliterated versions of Cyrillic characters. In our experiment, we considered the first token of a name string as a given name and the second one as a surname. It is clear from the table that in Russian language: (i) information about gender is encoded in the endings; (ii) there is a gender agreement. For instance, "Alexandr Ivanov" is a man's name, "Alexandr**a** Ivanov**a**" is a woman's name, and "Alexandr Ivanov**a**" is an ungrammatical name.

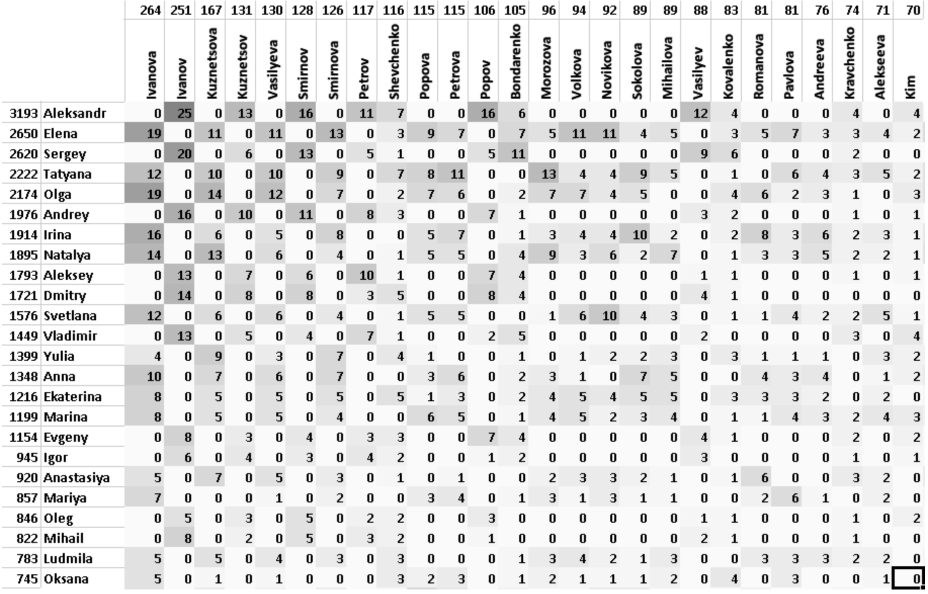
While our dataset represents a significant number of common Russian names, some rare names are under-represented. For instance, the female name "Oksana Kim", is a relatively rare, but a perfectly valid female name. Yet, it is not present in our dataset (see Fig. 1). However, there are 745 people with the first name "Oksana" and 70 persons with the last name "Kim".

In order to assess difficulty of the gender recognition task, we analyzed endings of first and last names in a sample of 10,000 objects. In this context, an *ending* is a substring composed of last two characters of a first/last name. According to this experiment, 72% of first names and 68% of surnames have typical male/female ending. Here a *typical male/female ending* is an ending that splits males from females with an error less than 5% (see Table 1). It appears that gender of more than 50% given names from our sample can be robustly detected with 8 endings. Furthermore, gender of more than 50% second names can be recognized with only 5 endings (see Table 1). These observations suggest that a simple symbolic ending-based method cannot robustly classify about 30% of names. This motivates the need for a more sophisticated statistical approach.

4 Gender Detection Method

The gender detection method takes as input a string representing a name of a person and outputs a gender (male or female). A name string is usually extracted from a user profile. Thus, we tackle the problem as a binary classification task. A label with unknown gender can be obtained by means of the reject option [17, p.42]. Below we describe features and the model used in our gender recognition approach. This section is concluded with a description of a simple rule-based baseline.

[8] `https://developers.facebook.com/tools/explorer`



		264	251	167	131	130	128	126	117	116	115	115	106	105	96	94	92	89	89	88	83	81	81	76	74	71	70
		Ivanova	Ivanov	Kuznetsova	Kuznetsov	Vasilyeva	Smirnov	Smirnova	Petrov	Shevchenko	Popova	Petrova	Popov	Bondarenko	Morozova	Volkova	Novikova	Sokolova	Mihailova	Vasilyev	Kovalenko	Romanova	Pavlova	Andreeva	Kravchenko	Alekseeva	Kim
3193	Aleksandr	0	25	0	13	0	16	0	11	7	0	0	16	6	0	0	0	0	0	12	4	0	0	0	4	0	4
2650	Elena	19	0	11	0	11	0	13	0	3	9	7	0	7	5	11	11	4	5	0	3	5	7	3	3	4	2
2620	Sergey	0	20	0	6	0	13	0	5	1	0	0	5	11	0	0	0	0	0	9	6	0	0	0	2	0	0
2222	Tatyana	12	0	10	0	10	0	9	0	7	8	11	0	0	13	4	4	9	5	0	1	0	6	4	3	5	2
2174	Olga	19	0	14	0	12	0	7	0	2	7	6	0	2	7	7	4	5	0	0	4	6	2	3	1	0	3
1976	Andrey	0	16	0	10	0	11	0	8	3	0	0	7	1	0	0	0	0	0	3	2	0	0	0	1	0	1
1914	Irina	16	0	6	0	5	0	8	0	0	5	7	0	1	3	4	4	10	2	0	2	8	3	6	2	3	1
1895	Natalya	14	0	13	0	6	0	4	0	1	5	5	0	4	9	3	6	2	7	0	1	3	3	5	2	2	1
1793	Aleksey	0	13	0	7	0	6	0	10	1	0	0	7	4	0	0	0	0	0	1	1	0	0	0	1	0	1
1721	Dmitry	0	14	0	8	0	8	0	3	5	0	0	8	4	0	0	0	0	0	4	1	0	0	0	0	0	0
1576	Svetlana	12	0	6	0	6	0	4	0	1	5	5	0	0	1	6	10	4	3	0	1	1	4	2	2	5	1
1449	Vladimir	0	13	0	5	0	4	0	7	1	0	0	2	5	0	0	0	0	0	2	0	0	0	0	3	0	4
1399	Yulia	4	0	9	0	3	0	7	0	4	1	0	0	1	0	1	2	2	3	0	3	1	1	1	0	3	2
1348	Anna	10	0	7	0	6	0	7	0	0	3	6	0	2	3	1	0	7	5	0	0	4	3	4	0	1	2
1216	Ekaterina	8	0	5	0	5	0	5	0	5	1	3	0	2	4	5	4	5	5	0	3	3	3	2	0	2	0
1199	Marina	8	0	5	0	5	0	4	0	0	6	5	0	1	4	5	2	3	4	0	1	1	4	3	2	4	3
1154	Evgeny	0	8	0	3	0	4	0	3	3	0	0	7	4	0	0	0	0	0	4	1	0	0	0	2	0	2
945	Igor	0	6	0	4	0	3	0	4	2	0	0	1	2	0	0	0	0	0	3	0	0	0	0	1	0	1
920	Anastasiya	5	0	7	0	5	0	3	0	1	0	1	0	0	2	3	3	2	1	0	1	6	0	0	3	2	0
857	Mariya	7	0	0	0	1	0	2	0	0	3	4	0	1	3	1	3	1	1	0	0	2	6	1	0	2	0
846	Oleg	0	5	0	3	0	5	0	2	2	0	0	3	0	0	0	0	0	0	1	1	0	0	0	1	0	2
822	Mihail	0	8	0	2	0	5	0	3	2	0	0	1	0	0	0	0	0	0	2	1	0	0	0	1	0	0
783	Ludmila	5	0	5	0	4	0	3	0	3	0	0	0	1	3	4	2	1	3	0	0	3	3	3	2	2	0
745	Oksana	5	0	1	0	1	0	0	0	3	2	3	0	1	2	1	1	1	2	0	4	0	3	0	0	1	0

Fig. 1. Name-surname co-occurrences: rows and columns are sorted by frequency.

4.1 Features

In our experiments, we used three types of features based respectively on endings, character n-grams and a dictionary of male/female names and surnames.

Word endings As we already mentioned above, Russian language, unlike English, has a gender agreement. Thus, the same name or surname often has different endings for a male and a female:

- males: Alexander Yaroskavski, Oleg Arbuzov
- females: Alexandr**a** Yaroskavsk**aya**, Nayal**iya** Arbuzov**a**

Thus, some Russian surnames are transliterated differently for males and females (see above). However, other surnames are spelled in the same way for both genders, e.g. “Sidorenko”, “Moroz” or “Bondar”.

The two most common one-character endings of female names/surnames are “a” and “ya” (“я” in Cyrillic). We use four features that indicate on female gender in Russian language: (1) first name ends with “a”, (2) first name ends with “я”, (3) last name ends with “a”, (4) last name ends with “я”.

Character n-grams These features rely on character unigrams, bigrams or trigrams extracted from the name strings. We represent a name with k its most frequent n-grams. The extraction is done with help of the NLTK module [18].

Type	Ending	Gender	Error, %	Example
first name	na (на)	female	0.27	Ekateri**na**
first name	iya (ия)	female	0.32	Anastas**iya**
first name	ei (ей)	male	0.16	Serg**ei**
first name	dr (др)	male	0.00	Alexan**dr**
first name	ga (га)	male	4.94	Sere**ga**
first name	an (ан)	male	4.99	Iv**an**
first name	la (ла)	female	4.23	Luidmi**la**
first name	ii (ий)	male	0.34	Yur**ii**
second name	va (ва)	female	0.28	Morozo**va**
second name	ov (ов)	male	0.21	Objedk**ov**
second name	na (на)	female	2.22	Matyushi**na**
second name	ev (ев)	male	0.44	Serge**ev**
second name	in (ин)	male	1.94	Teter**in**

Table 1. Most discriminative and frequent two character endings of Russian names.

Most frequent trigrams are listed below (here “_” denotes a beginning or an end of a name string): a__, va_, v__, na , ova, __A, ov_, ina, kov, nov, _Al, n__, a__, o__, __V, ndr, Ale, iya , ei , lek, eks, ko_, nko, rin, _An, enk, __C, na_, “ii ”, __E, eva, __N, __M, san, ksa, __I , ev_ , in_, and, len, __O, va_, rov, _Na. Note that the most frequent trigrams are not necessarily the most discriminative, e.g. “ko_” is a common surname ending for both males and females.

Dictionaries of first and last names This type of features is based on dictionaries of first and last names. Each entry of a dictionary contains a first/last name and a probability that it belongs to the male gender:

$$P(c = male|w) = \frac{n^w_{male}}{\sum_{c \in \{male, female\}} n^w_c},$$

where n^w_{male} is a number of male profiles with the first/last name w in the dictionary. We use two dictionary-based features: (1) probability that first name is of male gender $P(c = male|firstname)$, (2) probability that the last name is of male gender $P(c = male|lastname)$.

We used 90,000 name strings to build the two dictionaries. The training set used in our experiments does not contain any of these 90,000 samples. In order to remove noisy entries, we deleted all names and surnames that occurred only once. The full dictionary of first names contained 3,427 entries, while the dictionary of last names contained 11,411 entries. We used several versions of these full dictionaries in our study. Each version included top $\gamma\%$ most frequent given names and surnames.

4.2 Model

In all our experiments we used L2-regularized *Logistic Regression* model [19] as it generally yields reasonable results for the NLP-related problems [20]. Let $y_i \in \{-1, +1\}$ be a gender label and $\mathbf{x} = (x_1, \ldots, x_n)$ be a set of features representing a name string. The logistic regression combines features in a linear combination with weights $\mathbf{w}$. This weight vector is obtained by minimizing the following unconstrained optimization problem [21]:

$$\min_{\mathbf{w}} \sum_i log(1 + e^{-y_i \mathbf{w}^T \mathbf{x}_i}) + \frac{1}{C} ||\mathbf{w}||_2$$

We use the `scikit-learn` module in this experiment [9]. This implementation relies on the Dual Coordinate Descent Method for training of the model [22]. Default meta-parameters of the model were used. The model uses nearly no regularization – the inverse of the regularization strength C was set to 100,000. Optimization of the meta-parameters, such as C, or using more sophisticated models, such as SVM [14] can lead to significant improvements in results. However, in this paper we focus on a comparison of different features used withing the framework of one model.

The model $\mathbf{w}$ can be applied to perform classification of a name string represented with a feature vector $\mathbf{x}$ as follows:

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}}}.$$

4.3 Rule-based Baseline

There exist several available spelling dictionaries of Russian male and female first names, such as:

- a reference dictionary of personal names [10];
- a dictionary of personal names of Russian language by F. L. Ageenko [11];
- category “Names” of Russian Wiktionary [12];
- Russian spelling dictionary of Wikisource [13].

These dictionaries can be used to classify full names by gender. We compiled a dictionary of 1,428 Russian first names labeled with gender from Wiktionary and Wikisource [14]. This dictionary has no names assigned to both male and female categories.The dictionary was used to implement a rule-based baseline

[9] `http://scikit-learn.org/`

[10] `http://imena-list.ru/`

[11] `http://www.gramota.ru/slovari/info/ag/`

[12] `http://ru.wiktionary.org/wiki/`Категория:Имена

[13] `http://ru.wikisource.org/wiki/`Орфографический_словарь_русского_языка

[14] Available at `http://panchenko.me/gender/wiki-gender-dict.csv`

that works as follows. First, an input name string is transformed into a set of tokens t. Let d_f be a set of female first names and d_m be a set of male first names. Second, gender c is assigned to a full name with the following rule:

$$c = \begin{cases} male, & \text{if } (t \cap d_m \neq \emptyset) \text{ and } (t \cap d_f = \emptyset), \\ female, & \text{if } (t \cap d_m = \emptyset) \text{ and } (t \cap d_f \neq \emptyset), \\ unknown, & \text{else.} \end{cases}$$

Thus, a person with a female first name will be considered as a female, while a person with an unknown first name will have no gender label.

5 Results and Discussion

In this section, we present results of the experiments with the gender classification approaches described above. We start with the results of the rule-based baseline. Next, we proceed to statistical models based on one type of features: endings, character trigrams and dictionary. Finally, we present results of the statistical models that rely on several kinds of features at the same time.

Table 2 presents main results of our experiments in terms of the standard performance metrics calculated on a sample of 10,000 name strings. Among these 10,000 names we found 127 (1.27%) duplicate names. Fig. 2 illustrates how size of the training set affects accuracy.

Model	**Accuracy**	**Precision**	**Recall**	**F-measure**
rule-based baseline	0,638	**0,995**	0,633	0,774
endings	0,850 ± 0,002	0,921 ± 0,003	0,784 ± 0,004	0,847 ± 0,002
3-grams	0,944 ± 0,003	0,948 ± 0,003	0,946 ± 0,003	0,947 ± 0,003
dicts	0,956 ± 0,002	**0,992 ± 0,001**	0,925 ± 0,003	0,957 ± 0,002
endings+3-grams	0,946 ± 0,003	0,950 ± 0,002	0,947 ± 0,004	0,949 ± 0,003
3-grams+dicts	0,956 ± 0,003	0,960 ± 0,003	0,957 ± 0,004	0,959 ± 0,003
endings+3-grams+dicts	**0,957 ± 0,003**	0,961 ± 0,003	**0,959 ± 0,004**	**0,960 ± 0,002**

Table 2. Results of the experiments on the training set of 10,000 names (10-fold cross-validation). Here *endings* – 4 Russian female endings, *trigrams* – 1000 most frequent 3-grams, *dictionary* – name/surname dictionary with $\gamma = 80\%$ top entries. This table presents precision, recall and F-measure of the female class.

5.1 Rule-based Baseline

As one can see, the rule-based classifier is very precise. It achieves precision of 0.995. However recall of this method is only 0.663. This is due to a large number of unclassified examples: the dictionary used by this classifier does not list many first names common on Facebook.

5.2 Word endings

As to the statistical classifiers, even the simplest model that relies on two female endings “a” and “ya” (“я”) yields reasonable results, achieving accuracy up to 0.850. However, while precision of such a naive model is relatively high (0.921), its recall is only 0.784. Therefore, a significant fraction of female names do not follow simple ending-based rules. Naturally, performance of the ending-based model improves a little as training set grows (see Fig. 2). A hundred of examples is sufficient for training.

5.3 Character *n*-grams

In these experiments we focus on trigrams, as according to our results they worked better than bigrams and unigrams. Table 2 and Fig. 2 present performance of the models based on 1,000 most frequent trigrams. Character trigrams significantly outperform word endings if a training set is larger than 30 samples. Furthermore, unlike the ending-based model, the trigram-based model improves as the training set grows reaching accuracy of 0.944 on a dataset of 10,000 samples.

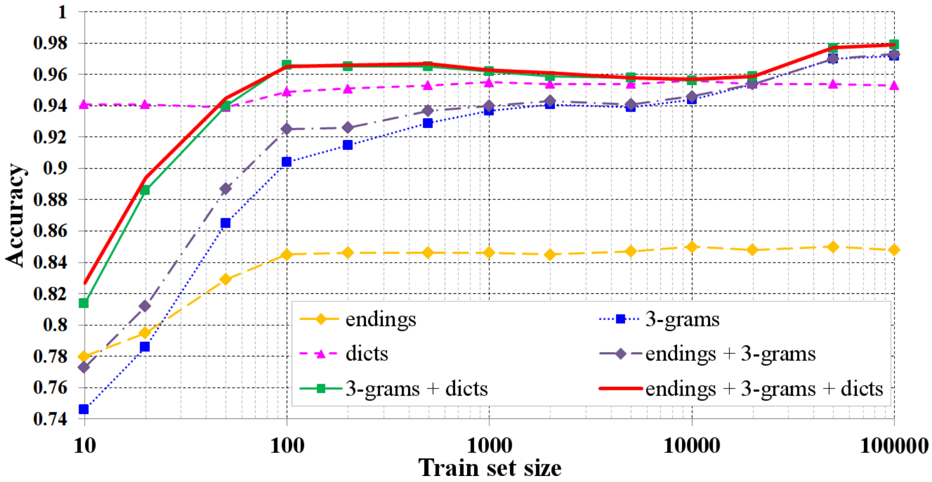


Fig. 2. Learning curves of single and combined models. Accuracy was estimated on separate sample of 10,000 names.

Fig. 3 plots accuracy of the *3-grams* model function of the number of most frequent trigrams used. In our further experiments (the combined models e.g.*3-grams+dicts*) we used a model based on top 1000 trigrams as a good trade-off between computational complexity and accuracy.

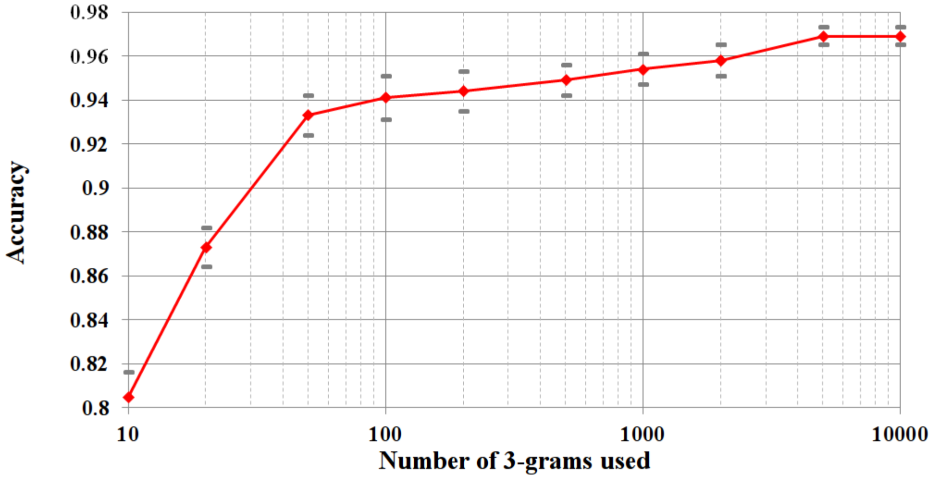


Fig. 3. Accuracy of the model *3-grams* function of the number of features used k.

5.4 Dictionary of first and last names

The model *dicts* that relies on a dictionaries of given names and surnames yields very competitive results (accuracy up to 0.956). Furthermore, this model provides the best precision among all statistical models (0.992). However, its recall is significantly lower than that of trigrams and combined models. As one may expect, only several dozens of training examples are enough to train this model (see Fig. 2). Further increase of the training set naturally does not improve accuracy as (i) it has only two features; (ii) the dictionaries are extracted from an independent part of data, not from the training set.

Fig. 4 plots accuracy of the *dicts* model function of the dictionary size γ. The maximal accuracy is achieved if the full dictionary is used ($\gamma = 100\%$). However, in our experiments we used $\gamma = 80\%$ of the dictionaries (2,741 first names and 9,128 last names) as the difference between the two respective dictionary-based models is minimal. In fact, even if one would use only $\gamma = 60\%$ of the dictionaries (2,056 first names and 6,846 last names) one will get nearly the same results.

5.5 Combined Models

We experimented with three combined models: *endings+3-grams*, *3-grams+dicts* and *endings+3-grams+dicts*. Combination of trigrams with word endings yields nearly the same performance as the model based on trigrams only. Indeed, the two female endings "a" and "ya" ("я") are present among frequent trigrams, such as "a__", "va_", "na_", "ya__", "aya_", or "iya_".

On the other hand, the model that relies on both dictionary and trigrams outperforms both trigram- and dictionary-based models, reaching accuracy of 0.956 due to increase in recall. However, precision of this combined model is significantly lower than that of the single dictionary-based model (0.960 versus

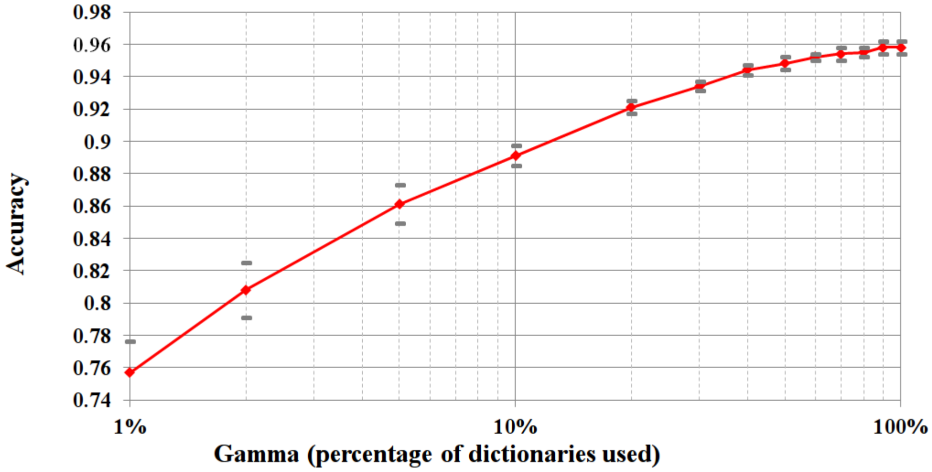


Fig. 4. Accuracy of the *dicts* model function of the fraction of dictionaries used γ.

0.992). This is so as an n-gram model can capture noisy sequences or estimate poorly weights of some n-grams due to sparsity of the training data.

Finally, the model *endings+3-grams+dicts* that makes use of all three types of features shows slightly higher performance than *3-grams+dicts*. However, this difference is not statistically significant. As we already discussed above, trigrams model well word endings. Table 3 lists some errors produced by our best model *endings+3-grams+dicts* on test and train sets. Train set error is very small (about 0.3%), while test set error is bigger (about 4%). There are several types of errors:

- Inconsistent annotation, such as "Anna Kryukova (male)" or "Boris Krolchansky (female)".
- Name string is neither male nor female, but rather a name of a group, e.g. "Wikom Tools", "Kazakh University of Humanities" or "Privat Bank".
- Name string represents a foreign name, e.g. "Abdulloh Ibn Abdulloh", "Brooke Alisson", "Ulpetay Niyetbay" or "Yola Dolson". Our model was not trained to deal with such names.
- Meaningless or partially anonymized names names, e.g. "Crazzy Ma", "Un Petit Diable", "Vv Tt", "Vio La Tor" or "Muu Muu". Additional information is required to derive gender of such users.
- People with rare names or surnames, e.g. "Guldjan Reyzova", "Yagun Zumpelich" or "Akob Saakan". These are people with common names and surnames in other countries, such as Georgia, Kazakhstan, Azerbaijan or Tajikistan. In our dataset, people from these countries are under-represented.

– Full names that can denote both males and females, e.g. “Jenya Chekulenko”, “Jenya Sergienko”, “Sasha Sidorenko” or “Sasha Radchenko”. Additional information is required to infer gender of such people.

– Misclassifications of common names, e.g. “Ilya Nasorshin”, “Oleg Dubovik” or “Elena Antropova”.

	Train Set Errors		**Test Set Errors**	
	name	**true class**	**name**	**true class**
1	Lea Shraiber	female	Ilya Nadorshin	male
2	Profanum Vulgus	female	Rustem Saledinov	male
3	Anna Kryukova	male	Erkin Bahlamet	male
4	Gin Amaya	male	Gocha Lapachi	male
5	Gertrud Gallet	female	Muttaqiyyah Abdulvahhab	female
6	Dolores Laughter	female	Yola Dolson	female
7	Di Nolik	male	Heiran Gasanova	female
8	Jlija Hotieca	female	Hadji Murad	male
9	Gic Globmedic	female	Jenya Chekulenko	female
10	Ulpetay Niyetbay	female	Tury.Ru Domodedovskaya Metro Office	male
11	Olga Shoff	male	Elmira Nabizade	female
12	Phil Golosoun	male	Niko Liparteliani	male
13	Tsitsino Shurgaya	female	Oleg Grin’	male
14	Anna Grobov	female	Santi Zarovneva	female
15	Linguini Incident	female	Misha Badali	male
16	Toma Oganesyan	female	Che Serega	male
17	Swon Swetik	female	Petr Kiyashko	male
18	Adel Simon	female	Sandugash Botabaeva	female
19	Ant Kam-	male	Jenya Sergienko	female
20	Xristi Xitrozver	female	Abdulloh Ibn Abdulloh	female
21	Anii Reznookova	female	Naikaita Laitvainenko	male
22	Aurelia Grishko	male	Fil Kalnitskiy	male
23	Alex Bu	female	Helen Hovel’	female
24	Karen Karine	female	Valery Kotelnikov	male
25	Russian Spain	female	Max Od	male
26	Lucy Walter	male	Jean Kvartshelia	male
27	Aysah Ahmed	female	Adjedo Trupachuli	female
28	Kiti Iz	female	Ainur Serikova	female
29	Cutejilian Juka	female	Privat Bank	female
30	Azer Dunja	male	No Limit	female

Table 3. Examples of train and test set errors of the model *endings+3-grams+dicts*. Here Cyrillic characters were transliterated into English in the standard way.

6 Conclusion

We presented several simple and computationally efficient models for gender detection by a full name of a person. The methods yields excellent results on a dataset Russian-speaking Facebook users, achieving accuracy up to 96%. In our further research, we plan to complement the developed approach by a gender detection method based on texts written by a person, as a full name is not always available or can give no clue about gender. For instance, the name "Sasha Sidorenko" can refer to both a male and a female.

Acknowledgements

This research was supported by Digital Society Laboratory LLC. We thank Kirill Shileev, Segei Objedkov and three anonymous reviewers for their helpful comments that significantly improved quality of this paper.

References

1. Underwood, A.: Gender targeting for promoted products now available. (October 2012)
2. Peersman, C., Daelemans, W., Van Vaerenbergh, L.: Predicting age and gender in online social networks. In: Proceedings of the 3rd international workshop on Search and mining user-generated contents, ACM (2011) 37–44
3. Kharitonov, E., Serdyukov, P.: Gender-aware re-ranking. In: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, ACM (2012) 1081–1082
4. Bi, B., Shokouhi, M., Kosinski, M., Graepel, T.: Inferring the demographics of search users: social data meets search queries. In: Proceedings of the 22nd international conference on World Wide Web, International World Wide Web Conferences Steering Committee (2013) 131–140
5. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing **17**(4) (2002) 401–412
6. Goswami, S., Sarkar, S., Rustagi, M.: Stylometric analysis of bloggers' age and gender. In: Third International AAAI Conference on Weblogs and Social Media. (2009)
7. Rangel, F., Rosso, P.: Use of language and author profiling: Identification of gender and age. Natural Language Processing and Cognitive Science (2013) 177
8. Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T.: "how old do you think i am?": A study of language and age in twitter. In: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media. (2013)
9. Ciot, M., Sonderegger, M., Ruths, D.: Gender inference of twitter users in non-english contexts. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Wash. (2013) 18–21
10. Rao, D., Yarowsky, D., Shreevats, A., Gupta, M.: Classifying latent user attributes in twitter. In: Proceedings of the 2nd international workshop on Search and mining user-generated contents, ACM (2010) 37–44

11. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics (2011) 1301–1309
12. Daniel, M. A. Zelenkov, Y.: Russian national corpus as a playground for sociolinguistic research. episode iv. gender and length of the utterance. In: Proceedings of Dialog-2012. (2012) 51–62
13. Mukherjee, A., Liu, B.: Improving gender classification of blog authors. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics (2010) 207–217
14. Vapnik, V.: The nature of statistical learning theory. Data Mining and Knowledge Discovery (6) 1–47
15. Al Zamal, F., Liu, W., Ruths, D.: Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In: ICWSM. (2012)
16. Liu, W., Zamal, F.A., Ruths, D.: Using social media to infer gender composition of commuter populations. Proceedings of the When the City Meets the Citizen Worksop (2012)
17. Bishop, C.M., Nasrabadi, N.M.: Pattern recognition and machine learning. Volume 1. springer New York (2006)
18. Bird, S.: Nltk: the natural language toolkit. In: Proceedings of the COLING/ACL on Interactive presentation sessions, Association for Computational Linguistics (2006) 69–72
19. Agresti, A.: Categorical data analysis. Volume 359. John Wiley & Sons (2002)
20. Panchenko, A., Beaufort, R., Naets, H., Fairon, C.: Towards detection of child sexual abuse media: categorization of the associated filenames. In: Advances in Information Retrieval. Springer (2013) 776–779
21. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. The Journal of Machine Learning Research **9** (2008) 1871–1874
22. Yu, H.F., Huang, F.L., Lin, C.J.: Dual coordinate descent methods for logistic regression and maximum entropy models. Machine Learning **85**(1-2) (2011) 41–75