Author Gender Prediction in Russian Social Media Texts

Dmitry Gudovskikh, Tatiana Litvinova, Aleksandr Shovel, Pavel Sterdin, Olga Litvinova, Dina Pisarevskaya, Paolo Renso

Manual labeled data:
- Twitter corpus. 1000 words from each user. Retweets were removed.
- Facebook corpus. 1000 words per user.

Twitter corpus is available at http://en.ruaprofilinglab.ru/ruaprofiling-at-pan/korpus/

DATA

Preprocessing:
1. Non-Russian texts were removed;
2. Citations were removed;
3. Accounts of public people were not used as they might have someone else writing for them;
4. http references were removed.
5. All hashtags (marked with #) were replaced with the `<hashtag>` tag;
6. Named entities (marked with @) are replaced with the NER tag.

Morpho-syntactic parser

Features selection:
1. morphological features
2. syntactic features
3. psycholinguistic features
4. Top 250 lemmas
5. Top 15 POS trigrams

TEXT PROCESSING

- **Top 250 lemmas.** We used 250 most frequent lemmas in the corpus. Note that we chose to employ lemmas as Russian is a morphologically rich language where gender is expressed explicitly in a range of grammatical structures. However, grammatical gender markings are easily falsified. As we have in mind as a general goal to build a system for gender prediction efficient, even in the case of gender imitation (valuable for security reasons), we made a decision to avoid using the token-based approach.
- A set of morphological (the frequencies of POS), syntactical features (frequencies of different types of syntactic relationships between heads and dependents) and psycholinguistic markers (derivative coefficients which reflect different ratios of POS), 56 in total. For example: The Coefficient of Trager (CTR). This is the ratio of the number of verbs to the number of adjectives in the text document; Coefficient of readiness to action (CRA). This is the ratio of the number of verbs to the number of nouns; Coefficient of Aggressiveness (CA). This is the ratio of the number of verbs to the number of adjectives. The ratio of words; The average size of sentences in words; The ratio of verbs to the number of nouns; The number of exclamation marks; The presence of emotional etc.
- **Top 15 frequent POS trigrams** which averaged values are different in males and female texts. At the first stage we chose POS trigrams which occurred in 75 % of the documents of the class and then calculated the difference between the average values of the frequencies of POS trigrams in texts by males and females and as a result, 15 POS trigrams with the largest difference in the average values were selected.

FEATURES

<table>
<thead>
<tr>
<th>Twitter test</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>Average Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM (linear)</td>
<td>0.466</td>
<td>0.741</td>
<td>0.604</td>
<td>0.665</td>
</tr>
<tr>
<td>SVM (RBF)</td>
<td>0.442</td>
<td>0.768</td>
<td>0.67</td>
<td>0.693</td>
</tr>
<tr>
<td>SVM (poly)</td>
<td>0.405</td>
<td>0.807</td>
<td>0.698</td>
<td>0.693</td>
</tr>
<tr>
<td>ExtraTreesClassifier (max_estimators = 100)</td>
<td>0.762</td>
<td>0.666</td>
<td>0.746</td>
<td>0.725</td>
</tr>
<tr>
<td>RandomForestClassifier (max_estimators = 100)</td>
<td>0.779</td>
<td>0.690</td>
<td>0.743</td>
<td>0.721</td>
</tr>
<tr>
<td>KNN</td>
<td>0.799</td>
<td>0.699</td>
<td>0.738</td>
<td>0.723</td>
</tr>
<tr>
<td>Adaboosting</td>
<td>0.804</td>
<td>0.790</td>
<td>0.799</td>
<td>0.777</td>
</tr>
<tr>
<td>DecisionTreeClassifier</td>
<td>0.804</td>
<td>0.790</td>
<td>0.799</td>
<td>0.777</td>
</tr>
</tbody>
</table>

**TRAIN - TEST**

90% samples of Twitter corpus

StratifiedKFolds strategy with 30 folds

10% for test all corpus

RESULTS