Automatic Arabic Dialect Classification

- Dictum Ltd, Nizhny Novgorod, Russia
- Lobachevsky State University of Nizhny Novgorod (UNN)

Oleg Durandin, Dmitriy Strebkov, Nadezhda Hilal

Background: What is Dialectal Arabic

Dialectal Arabic (**DA**) represents different variations of Arabic language widely used in everyday life, at live communication or chatting.

DA differs from standard Arabic in:

- more complex word formation system;
- lack of spelling standards;
- additional functional words and particles.

Most popular: Levantine, Egyptian, Saudi, Algerian and Gulf-Arabic.

Dialect identification task

Task: identification (classification) of Arabic dialect

Treat that task as a statistical language identification problem.

Practical usage:

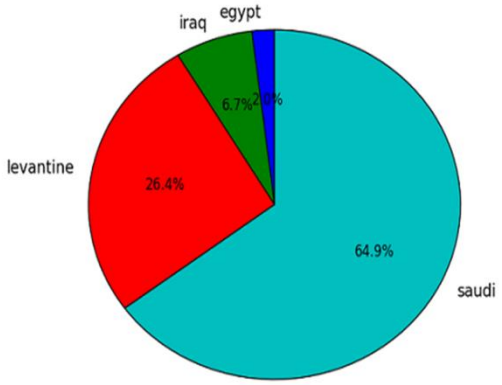
- identifying nationality of an author of social media text;
- revealing common ideas/preferences;
- finding actual topics that are actively discussed in concrete Arabic country or city.

Geolocation based classification is **not** relevant.

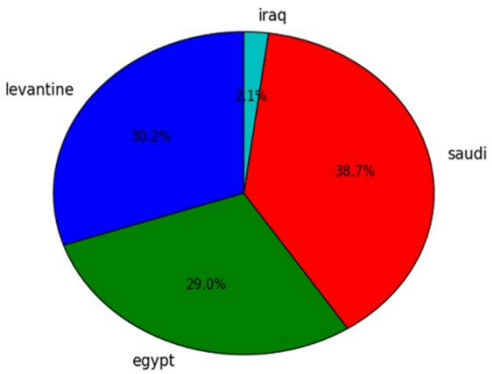
Mining Dialectal Arabic

Mined DA from Twitter using special words – **word-marks** to get 2 datasets:

- Big set of **automatically** labeled tweets:



- Small set of tweets labeled **manually**:



Experiment

Solving classification problem.

Train set: instances labeled **automatically** and **manually**.

Test set: instances labeled **manually**.

Features: bigrams, trigrams, 4-grams, word-marks vocabulary.

Varying size of train data

How varying of the size of train data affects precision and recall for our model?

- Varying size of **manually** annotated subset
- Varying size of **automatically** annotated subset

Finally…

- Results
- Confusion matrix
- Benefits
- Plans