Discriminating between standard Romanian and Moldavian tweets using filtered character ngrams

Andrea Ceolin and Hong Zhang

Seventh workshop on NLP for similar languages, varieties and dialects (VarDial) - December 13, 2020

Introduction

- Language identification can be challenging for NLP techniques when languages are hardly distinguishable.
- The VarDial 2020 RDI shared task (Găman et al. 2020) asked participants to distinguish standard Romanian and Moldavian tweets.
- The distinction between standard Romanian and Moldavian is very subtle, and only motivated by the presence of a political boundary in a linguistically uniform region.
- An additional challenge of the task is the problem of cross-domain generalization.
- We implemented both shallow and deep models for this task. Results suggest that the deep models suffer from greater degradation when the training and testing data are from different domains.

Models and Results

We implemented the following models:

- **Multinomial Naïve Bayes - Words**
- **Multinomial Naïve Bayes - Character Ngrams**
- **Linear SVM - Words**
- **Linear SVM - Character Ngrams**
- **Character CNN**
- **Character TDNN**

The models were trained on a News training dataset. Here is a summary of the results we obtained using a News evaluation dataset and a Tweets evaluation dataset:

Model	News	Tweets
MNB - Word unigrams	0.891	0.637
MNB - Word unigrams- TFIDF	0.892	0.665
MNB - Char. ngrams [5-8]	0.883	**0.674**
MNB - Char. ngrams [5-8] - TFIDF	0.351	0.322
Linear SVM - Word unigrams	0.693	0.504
Linear SVM - Word unigrams - TFIDF	**0.942**	0.502
Linear SVM - Char. ngrams [6-8]	0.795	0.599
Linear SVM - Char. ngrams [6-8] - TFIDF	0.351	0.345
CNN	0.931	0.485
TDNN	0.771	0.515

Cross-domain degradation

The model that yields the best cross-domain generalization is the **Multinomial Naïve Bayes - Character Ngrams** model. Shallow MNB models suffer less from cross-domain degradation, as shown in Figure 1.

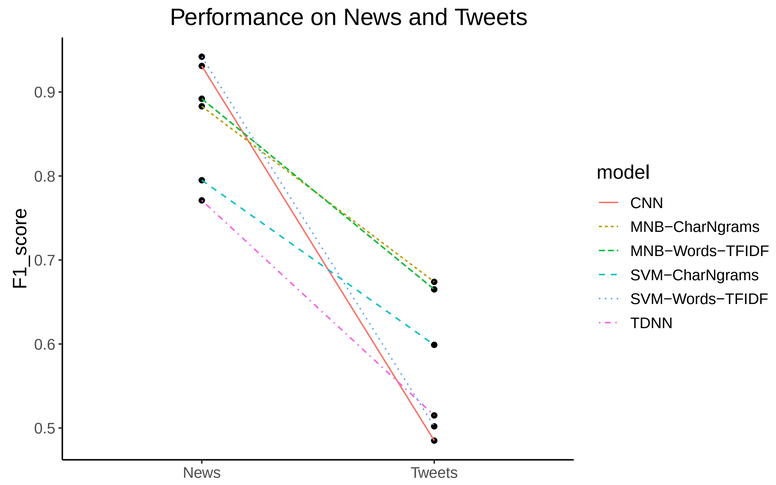

Figure 1: Performance comparison between news and tweets. All models were trained on the news data.

Our deep models, albeit being able to achieve comparable performance when tested on the same domain, did not appear to generalize to the different domain, suggesting the possibility of overfitting the training data.

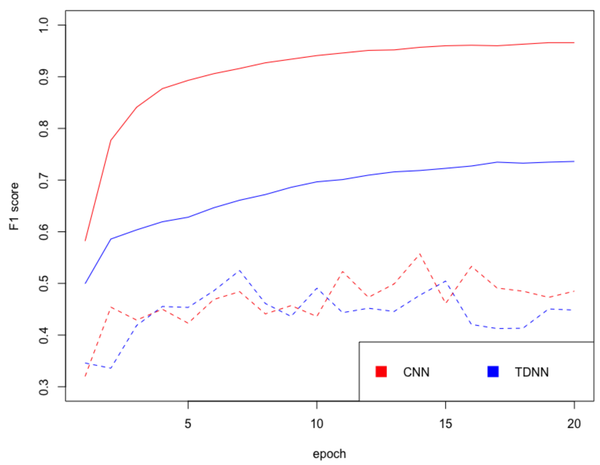

Figure 2: Testing performance as a function of training epochs. Solid line: testing and training data from same domain. Dashed line: testing data from a different domain.

Feature selection

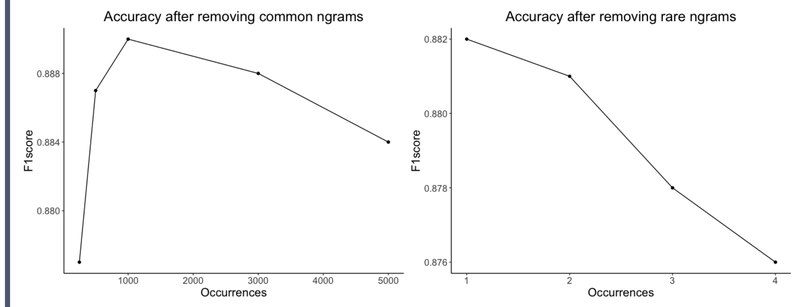

Figure 3: Left: Testing performance as the function of number of high frequency ngrams removed. Right: Testing performance as the function of number of low frequency ngrams removed. Testing was based on the news domain.

Removing the most common ngrams led to improved testing performance of the **MNB - Character Ngrams** model (up to F1=**0.890**). Best performance achieved when removing all the ngrams appeared more than 1000 times.

Discussion and Conclusion

We submitted three runs to the VarDial 2020 RDI shared task:

1. **MNB - Char. ngrams, [5-8], filter <1000, alpha=0.0001**. This was the best ngram model on the news articles.
2. **MNB - Char. ngrams, [6-8], filter <250, alpha=0.001**. This was one of the two best models on the tweets evaluation data set.
3. **MNB - Char. ngrams, [5-7], filter <200, alpha=0.001**. This was one of the two best models on the tweets evaluation dataset.

Our shallow model #1 achieved third place in the held-out testing data (F1=**0.666**). The score goes up to F1=**0.692** after preprocessing.
Some observations:

- The best performing model across all teams is a shallow SVM based on word and character ngrams, which splits the training sentences in smaller sentences to improve cross-domain adaptation, with dramatic performance improvement (F1=**0.788**). Good job, Team Tübingen!
- Deep models from other teams appear to suffer from the same problem as our own implementation of the two neural architectures.
- Thanks to the organizers of the task!