Discriminating between standard Romanian and Moldavian tweets using filtered character ngrams

Andrea Ceolin
University of Pennsylvania
ceolin@sas.upenn.edu

Hong Zhang
University of Pennsylvania
zhangho@sas.upenn.edu

Abstract

We applied word unigram models, character ngram models, and CNNs to the task of distinguishing tweets of two related dialects of Romanian (standard Romanian and Moldavian) for the VarDial 2020 RDI shared task (Găman et al., 2020). The main challenge of the task was to perform cross-genre text classification: specifically, the models must be trained using text from news articles, and be used to predict tweets. Our best model was a Naïve Bayes model trained on character ngrams, with the most common ngrams filtered out. We also applied SVMs and CNNs, but while they yielded the best performance on an evaluation dataset of news article, their accuracy significantly dropped when they were used to predict tweets. Our best model reached an F1 score of 0.715 on the evaluation dataset of tweets, and 0.667 on the held-out test dataset. The model ended up in the third place in the shared task.

1 Introduction

Language identification can be challenging for NLP techniques when languages are hardly distinguishable. One example of this challenge is the identification of Moldavian, a dialect of Romanian which exhibits almost no difference with standard Romanian. The distinction between Romanian and Moldavian is only motivated by the presence of a political boundary, which corresponds to no real isogloss. In spelling, the two languages are almost identical, with a minor exception involving the distribution of the letters 'â' and 'î', although other grammatical distinctions can be found in number, gender and case morphology, and in lexical choices. In particular, the lexical divergence was the result of Moldavian being under the influence of Russian, in the years in which it was part of the Soviet Union.

An additional challenge for language identification is that existing resources might belong to domains that are different from the domain on which one needs to perform a classification task. For instance, in certain cases one can find data from textbooks, encyclopedias and newspaper articles, but not from social media, even though language identification is often used to classify online messages (Tromp and Pechenizkiy, 2011; Bergsma et al., 2012; Barman et al., 2014; Lui and Baldwin, 2014). This raises the question of how to use out-of-domain data when performing language identification in a restricted domain.

The VarDial 2020 RDI shared task (Găman et al., 2020) invited participants to perform cross-genre language identification by training a classifier on newspaper articles, and using it to distinguish standard Romanian from Moldavian tweets. In this paper, we present the contribution of the team Phlyers to the task.

2 Methods

Previous methods used for language identification typically involve bag-of-words models (Huang and Lee, 2008), Naïve Bayes models applied to word and character ngrams (Jauhiainen et al., 2016) and Support Vector Machines (Zampieri et al., 2019). Deep learning methods based on CNNs and LSTMs

 The code for the models employed in this work is found at https://github.com/AndreaCeolin/VarDial2020. We thank Monica-Alexandrina Irimia for comments about the project.

Proceedings of the 7th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects, pages 265–272
Barcelona, Spain (Online), December 13, 2020

have also been successfully applied to similar tasks (Jaech et al., 2016; Butnaru and Ionescu, 2019; Hu et al., 2019; Tudoreanu, 2019).

Last year's VarDial edition (Zampieri et al., 2019) proposed the first shared task based on distinguishing standard Romanian from Moldavian. The best model achieved an F1 score of 0.895 on the test set, using an ensemble method based on CNNs and Support Vector Machines (Tudoreanu, 2019). The task consisted in training a classifier on news article in Romanian and Moldavian from the MOROCO corpus (Butnaru and Ionescu, 2019), and using it to classify other news articles yet to be added to the corpus.

This year's task asked participants to train a classifier on the news articles of the MOROCO corpus in order to distinguish standard Romanian from Moldavian in a test dataset of tweets (Găman and Ionescu, 2020). The task was particularly challenging because the organizers provided a large evaluation dataset based on news articles (5923) and a small evaluation dataset based on tweets (215) (cf. Table 1). This made the evaluation stage particularly delicate, because on the one hand a good model tested on the news evaluation dataset could fail to generalize to a different domain, while on the other hand the size of the tweets evaluation dataset was so small that the risk of overfitting was considerable.

VarDial (2020) RDI Shared task	Sentences
Training (News)	33564
Evaluation (News)	5923
Evaluation (Tweets)	215
Test (Tweets)	5022

Table 1: Summary statistics of the VarDial 2020 shared task.

We decided to train a variety of models, and to study their generizability to genres different from those in the training data. The models that we trained for the task are the following:

- **Multinomial Naïve Bayes - Words**. This is a standard Naïve Bayes model applied to word unigrams. The best performance on the news evaluation dataset was reached by using a TFIDF matrix instead of word counts. The optimal alpha was 0.0001 for both the unigram- and the TFIDF- based model.

- **Multinomial Naïve Bayes - Character Ngrams**. This is a standard Naïve Bayes model applied to character ngrams. The best performance on the news evaluation set was reached by a model which calculates ngrams in the window [5-8], with alpha=0.0001. Padding symbols (n-1) are added both before and after each word in order to retrieve ngrams for each value of n.

- **Linear SVM - Words**. This is a standard Support Vector Machine model with a linear kernel. The best performance on the news evaluation set was reached by using a TFIDF matrix instead of word counts. The optimal regularization parameter C was 2.

- **Linear SVM - Character Ngrams**. Our best Support Vector Machine model with a linear kernel uses character ngrams in the window [6-8], with C=1. Padding symbols (n-1) are added both before and after each word in order to retrieve ngrams for each value of n.

- **Character CNN**. We used the character-based CNN proposed in Zhang et al. (2015), and modified it according to the baseline model in Butnaru and Ionescu (2019). We created an alphabet of 76 symbols representing all the characters that appear at least 50 times in the training data, plus a 'NA' symbol, and then we used one-hot encoding vectors as input to the CNN. The three hyperparameters we fine-tuned were the batch size (10), the learning rate (0.0001), and the size of the fully-connected layers (1000), while the other parameters were taken from Butnaru and Ionescu (2019). Training was performed for 20 epochs. See Figure 1 for a summary of the model.

- **Character TDNN**. Inspired by the research in speech recognition community, we implemented a Time Delay Neural Network (TDNN) (Peddinti et al., 2015), in order to better capture the morphological features of the two languages. Our model contains 2 stacked convolution blocks. Each block

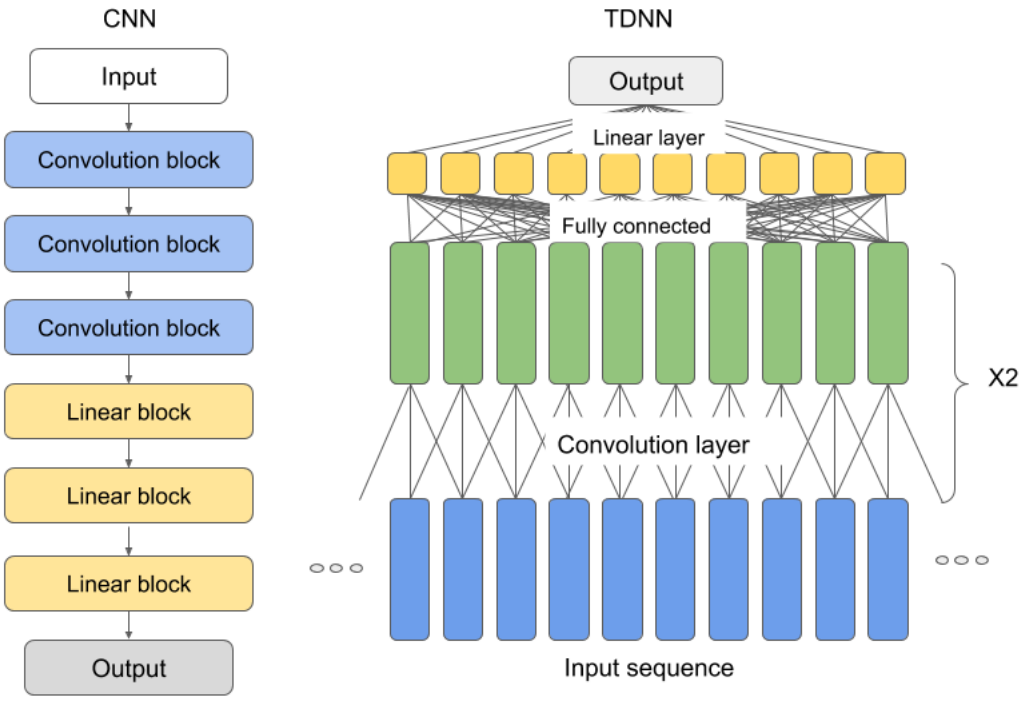

Figure 1: The architectures of our neural network models (CNN on the left, TDNN on the right).

learns 100 k-by-3 filter banks, where k is the dimension of the input vectors and 3 is the window length, that maps a trigram character window to a scalar. Input sequences were padded to equal length. The output 1-by-n vector, where n is the input sequence length, is then passed through a fully connected layer. Subsampling was not performed in training. Training was performed for 50 epochs.

The architectures of the CNN and the TDNN are shown in Figure 1.

3 Results

3.1 Evaluation

The results of the models are summarized in Table 2. All the models exhibit a drop in performance when used to classify tweets instead of news articles (Figure 2). The best model on the news evaluation dataset is the Linear SVM model trained on TFIDF transformed word unigrams, with an F1 score of 0.942. By inspecting the weight matrix, we were able to identify the words which have the highest contribution in determining the class of the news articles. Among the best words that are used to identify the Moldavian class, 'sînt' ('are') and 'cînd' ('when') have the largest weights. These are frequent words which are spelled differently in Romanian ('sunt' and 'când'). As for Romanian, 'news' and 'foto', which are loanwords, as well as the frequent word 'sâmbătă' ('Saturday'), which has a different spelling in Moldavian, ('sîmbătă'), carry the largest weights.

The CNN model reaches a similar accuracy (F1 score: 0.931), which is almost identical to the accuracy obtained by Butnaru and Ionescu (2019) on the same dataset. An interesting observation for the neural network-based models is that although the accuracy of the models on the news dataset increases by iterating through the training set, the improvement does not generalize to the testing domain (see Figure 3). The lack of cross-domain generalizability might indicate that naive implementations of deep neural network architectures are not well-suited for cross-genre classifications.

The models that best generalize to Twitter data are the Multinomial Naïve Bayes (MNB) models trained on TFIDF transformed word unigrams (F1 score=0.892) and character ngrams (F1 score=0.883). In particular, the highest accuracy is reached by the character ngram model, which yields an F1 score of

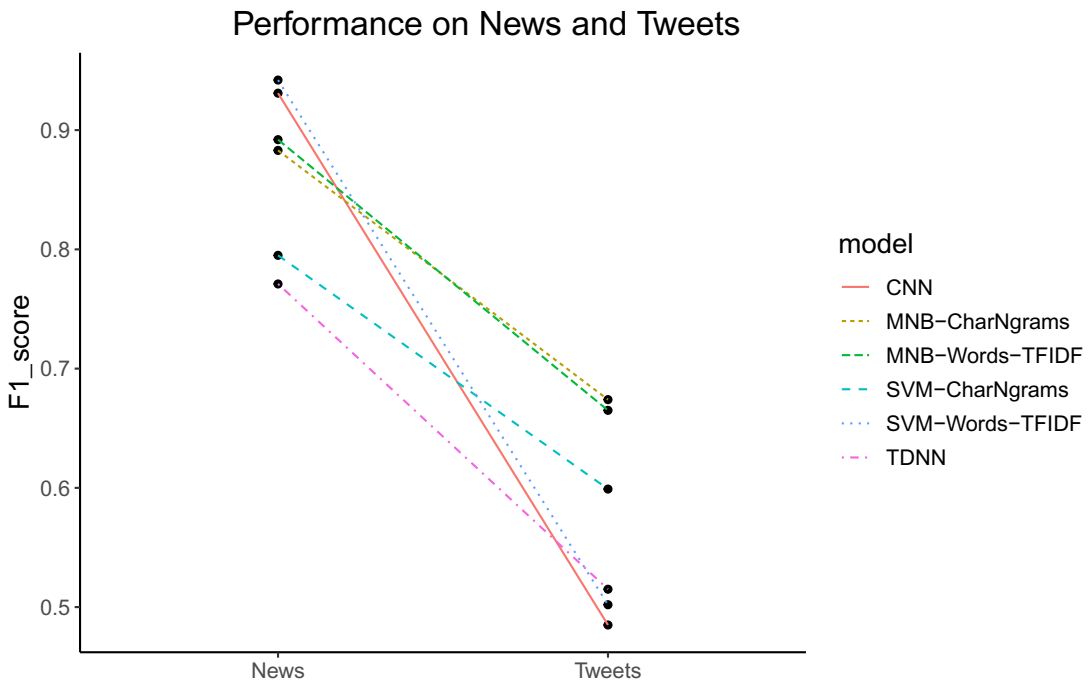

Figure 2: Drop in performance of the best performing models from the News development set to the Tweets development set.

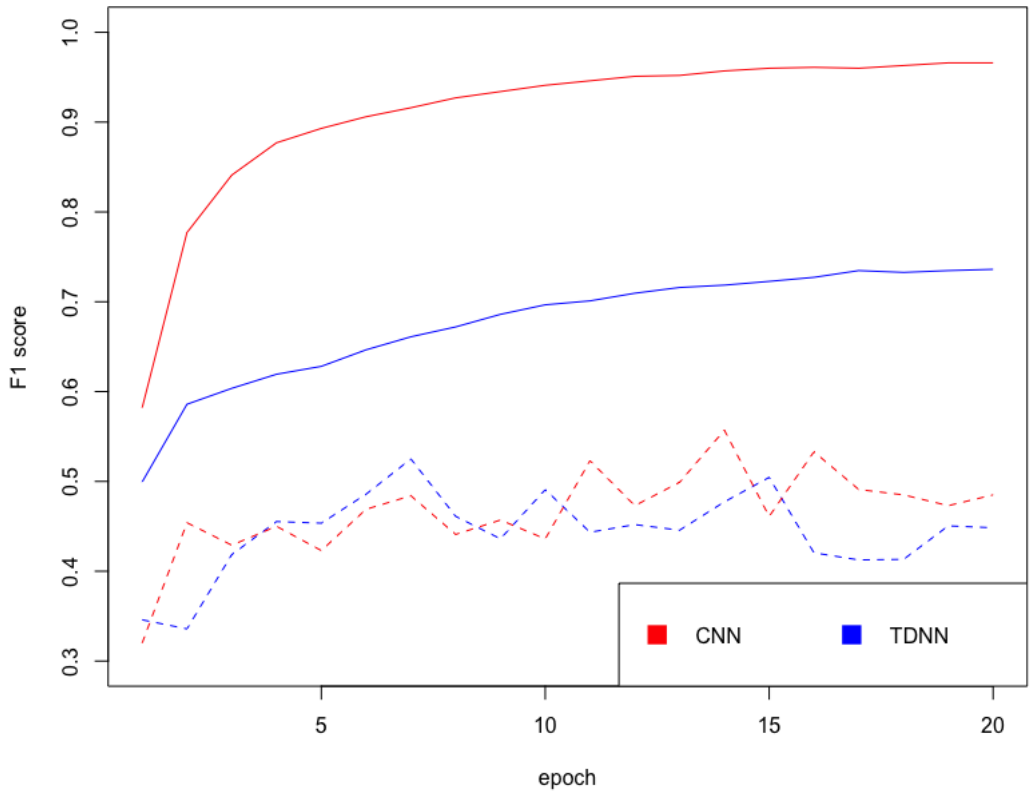

Figure 3: Comparing the training and validation performance measured by F1 score after 20 epochs of training. Solid line: validation score on news articles; dotted line: validation score on tweets.

	News articles (2019)	Tweets (2020)
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

Table 2: Performance of the models tested (F1 scores).

0.674 on the tweets dataset.

For these reasons, we decided to use the Multinomian Naïve Bayes model based on character ngrams for the task, and we proceeded to the fine-tuning stage.

3.2 Fine-tuning on the News set

We fine-tuned the MNB character ngram model on the tweets dataset (Table 3). We noticed that by removing the most common ngrams, performance improved. In particular, removing all the ngrams which appeared more than 1000 times overall improved the performance up to an F1 score of 0.890 on the news dataset (see Figure 4). On the contrary, removing less frequent ngrams did not improve performance. This result is interesting, because usually performance is increased by removing the tail of the frequency distribution, not the head: in this case, since the TFIDF transformation was not sufficient to normalize the behavior of high-frequency ngrams, removing them turned out to be a better strategy to increase the performance of the classifier.

Filter applied based on total occurences	Ch-ngram filtered	News articles (19)	Tweets (20)
MNB - Char. ngrams [5-8], filter < 5000	1638 (0.2%)	0.884	0.674
MNB - Char. ngrams [5-8], filter < 3000	3580 (0.4%)	0.888	0.674
MNB - Char. ngrams [5-8], filter < 1000	13766 (1.5%)	**0.890**	0.683
MNB - Char. ngrams [5-8], filter < 500	26384 (2.8%)	0.887	**0.702**
MNB - Char. ngrams [5-8], filter < 250	45102 (4.8%)	0.877	0.692
MNB - Char. ngrams [5-8], filter > 4	564021 (60.5%)	0.876	0.692
MNB - Char. ngrams [5-8], filter > 3	523377 (56.7%)	0.878	0.683
MNB - Char. ngrams [5-8], filter > 2	461784 (49.6%)	0.881	0.684
MNB - Char. ngrams [5-8], filter > 1	340764 (36.7%)	0.882	0.674

Table 3: Fine-tuning of the models tested. Total char. ngrams: 931786.

3.3 Fine-tuning on the Tweets set

After further fine-tuning of the ngram window and the threshold of ngrams to filter, we obtained two best models on the tweets dataset (Table 4). Both models reached an F1 score of 0.715. The performance did not improve after including in the training set the data coming from the evaluation set containing news articles. The two best models had the following settings:

- MNB - Char. ngrams, [6-8], filter <250, alpha=0.001
- MNB - Char. ngrams, [5-7], filter <200, alpha=0.001

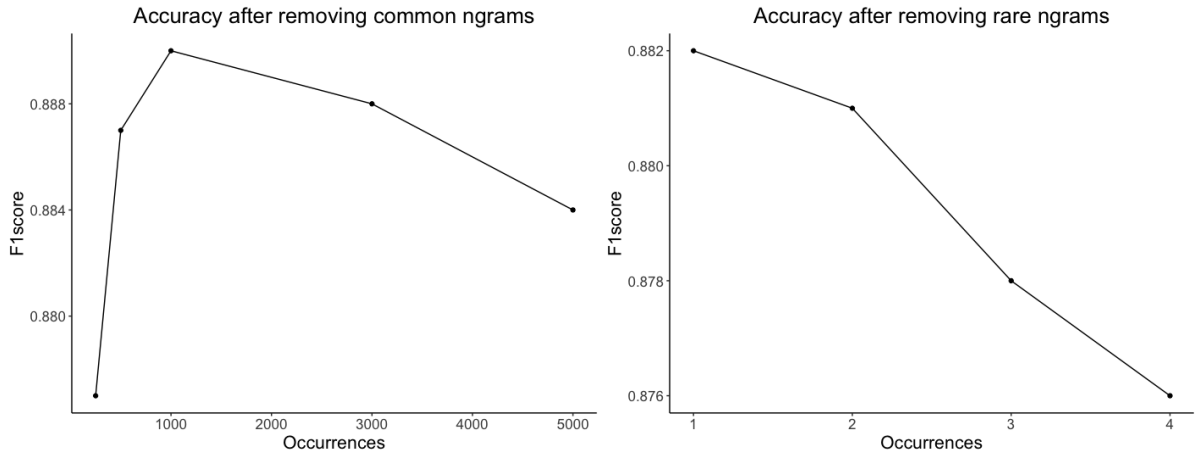

Figure 4: Accuracy change on the news evaluation dataset after ngram filtering. Left: common ngrams removed; right: common ngrams preserved.

Tweets (2020)		
	Train	Train+Dev1
MNB - Char. ngrams [6-8], filter <250	**0.715**	**0.715**
MNB - Char. ngrams [5-7], filter <200	**0.715**	**0.715**

Table 4: Final performance on the tweets evaluation dataset after fine-tuning the parameters of the MNB - ngrams model.

3.4 VarDial 2020 - RDI Shared Task

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.

The results for our models on the test dataset of the task are summarized in Table 5. The best model was the one which was fine-tuned on the news dataset. This result suggests that our fine-tuning strategy on the tweets dataset led to overfitting on the news articles, and thus poorer performance on tweets. Additionally, after our submissions, we realized that the tweets in the test data had not being preprocessed to remove punctuation, numbers and other symbols. Preprocessing increases the submission F1 score up to 0.692.

Both the Multinomial Naïve Bayes model based on TFIDF word unigrams, which was the second best model on the tweets dataset, and the one based on unfiltered character ngrams, performed worse than the best model submitted. The same was true for the Linear SVM model based on character ngrams.

4 Conclusion

We applied word unigram models, character ngram models, and two neural network models to classify tweets of two related dialects (standard Romanian and Moldavian) for the VarDial 2020 RDI shared task (Găman et al., 2020), with training data from a different domain. Two of the models we proposed, a Linear SVM model based on TFIDF word unigrams and a CNN model, reached a high accuracy on the news evaluation dataset, but failed to generalize to the tweets evaluation dataset. On the contrary,

Model	Tweets (2020) test data results	
	no preprocessing, submitted	preprocessing
1. MNB - Char. ngrams [5-8], filter <1000	**0.666**	**0.692**
2. MNB - Char. ngrams [6-8], filter <250	0.651	0.678
3. MNB - Char. ngrams [5-7], filter <200	0.645	0.675
MNB - Word unigrams - TFIDF	0.630	0.677
MNB - Char. ngrams [5-8]	0.651	0.676
Linear SVM - *n*grams [6-8]	0.593	0.590

Table 5: Results on the test dataset.

Multinomial Naïve Bayes models turned out to be the best performing models on the task. The more complex neural network models suffered from the problem of poor generalizability. In addition, we showed that removing high frequency ngrams can be a valid alternative when working on datasets for which a TFIDF transformation does not improve classification accuracy.

References

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the first workshop on computational approaches to code switching*, pages 13–23.

Shane Bergsma, Paul McNamee, Mossaab Bagdouri, Clayton Fink, and Theresa Wilson. 2012. Language identification for creating language-specific twitter collections. In *Proceedings of the second workshop on language in social media*, pages 65–74.

Andrei M Butnaru and Radu Tudor Ionescu. 2019. Moroco: The Moldavian and Romanian dialectal corpus. *arXiv preprint arXiv:1901.06543*.

Mihaela Găman and Radu Tudor Ionescu. 2020. The Unreasonable Effectiveness of Machine Learning in Moldavian versus Romanian Dialect Identification. *arXiv preprint arXiv:2007.15700*.

Mihaela Găman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, and Marcos Zampieri. 2020. A Report on the VarDial Evaluation Campaign 2020. In *Proceedings of the Seventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Hai Hu, Wen Li, He Zhou, Zuoyu Tian, Yiwen Zhang, and Liang Zou. 2019. Ensemble Methods to Distinguish Mainland and Taiwan Chinese. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 165–171.

Chu-Ren Huang and Lung-Hao Lee. 2008. Contrastive approach towards text source classification based on top-bag-of-word similarity. In *Proceedings of the 22nd Pacific Asia conference on language, information and computation*, pages 404–410.

Aaron Jaech, George Mulcaire, Mari Ostendorf, and Noah A Smith. 2016. A neural model for language identification in code-switched tweets. In *Proceedings of The Second Workshop on Computational Approaches to Code Switching*, pages 60–64.

Tommi Sakari Jauhiainen, Bo Krister Johan Linden, Heidi Annika Jauhiainen, et al. 2016. HeLI, a word-based backoff method for language identification. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2016).*

Marco Lui and Timothy Baldwin. 2014. Accurate language identification of twitter messages. In *Proceedings of the 5th workshop on language analysis for social media (LASM)*, pages 17–25.

Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth Annual Conference of the International Speech Communication Association.*

Erik Tromp and Mykola Pechenizkiy. 2011. Graph-based n-gram language identification on short texts. In *Proc. 20th Machine Learning conference of Belgium and The Netherlands*, pages 27–34.

Diana Tudoreanu. 2019. DTeam@ VarDial 2019: Ensemble based on skip-gram and triplet loss neural networks for Moldavian vs. Romanian cross-dialect topic identification. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 202–208.

Marcos Zampieri, Shervin Malmasi, Yves Scherrer, Tanja Samardžic, Francis Tyers, Miikka Pietari Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, et al. 2019. A report on the third VarDial evaluation campaign. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2019)*. The Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.