Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6247–6254
Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Detecting Troll Tweets in a Bilingual Corpus

Lin Miao[1], Mark Last[1], Marina Litvak[2]
Ben-Gurion University of the Negev[1], Shamoon College of Engineering[2]
P.O.B. 653 Beer-Sheva 8410501 Israel[1], 56 Bialik St. Be'er Sheva 8410802 Israel[2]
miaol@post.bgu.ac.il, mlast@bgu.ac.il, litvak.marina@gmail.com

Abstract

During the past several years, a large amount of troll accounts has emerged with efforts to manipulate public opinion on social network sites. They are often involved in spreading misinformation, fake news, and propaganda with the intent of distracting and sowing discord. This paper aims to detect troll tweets in both English and Russian assuming that the tweets are generated by some "troll farm." We reduce this task to the authorship verification problem of determining whether a single tweet is authored by a "troll farm" account or not. We evaluate a supervised classification approach with monolingual, cross-lingual, and bilingual training scenarios, using several machine learning algorithms, including deep learning. The best results are attained by the bilingual learning, showing the area under the ROC curve (AUC) of 0.875 and 0.828, for tweet classification in English and Russian test sets, respectively. It is noteworthy that these results are obtained using only raw text features, which do not require manual feature engineering efforts. In this paper, we introduce a resource of English and Russian troll tweets containing original tweets and translation from English to Russian, Russian to English. It is available for academic purposes.

Keywords: Bilingual Training, Authorship Verification, Social Network Analysis, Natural Language Processing

1. Introduction

With the rise of social media, people are exposed to large amounts of information on social media platforms, which creates the opportunity for some organizations to distribute rumors, misinformation, and speculation, in an attempt to manipulate the opinion of the public. A Kremlin-linked propaganda organization, known as the Internet Research Agency (IRA), has been assessed by the U.S. Intelligence Community to be part of a Russian state-run effort to influence the outcome of the 2016 U.S. presidential race[1]. As a result, Twitter suspended the IRA-connected accounts and deleted 200,000 Russian troll tweets. Twitter has been flooded with false news, and propaganda-spreading trolls. This dedicates the importance of detecting troll tweets to protect the public from the inappropriate content in the social network.

Considering that there are many languages in Twitter data, the detection of troll tweets needs to be able to handle multiple languages, to cover a larger portion of available troll tweets. The dataset[2] we use is constructed by the tweets from IRA-related troll accounts. Because English and Russian represent the largest percentages of this dataset, we will handle English and Russian tweets to cover a larger portion of available troll tweets.

The aim of this work is to determine whether a given tweet is a troll tweet, based solely on its content. The task of computational detection of troll tweets in a multilingual corpus is a linguistic and machine learning problem. It can be considered as an authorship verification task, based on the assumption that the troll tweets are generated by some sort of "troll farm." Some reports about a "troll farm" suggest that work at the "troll farm" is strictly regulated by a set of guidelines. The employees of "troll farm" are expected to manage 10 accounts and tweet 50 times a day[3], and using specific keywords in the posts is mandatory[4]. Considering all this evidence, we concluded that the tweets from the same troll farm may have similar writing style, as well as features that can be identified. In this case, we used authorship verification to determine whether each tweet is authored by a "troll farm" account.

In this work, we compare monolingual, cross-lingual, and bilingual learning, using different algorithms on different feature sets, with the goal of detecting troll tweets more efficiently. We evaluated two types of predictive features: stylometric features and n-grams. The main contributions of this work are:

- Provide a bilingual dataset with troll tweets. The dataset we built contains tweets from troll and legitimate accounts in two languages (including translation from one language to another) and can be used for training models for automatic detection of troll tweets/accounts in either English, Russian, or both languages.
- Troll detection in bilingual domain.
- Focusing only on textual features to build classification models for datasets with limited data.
- Using machine translation in bilingual setting for enriching training knowledge.
- Experiments with different bilingual settings for exploring how classification models can benefit from an available bilingual domain.

[1] https://www.nbcnews.com/tech/social-media/now-available-more-200-000-deleted-russian-troll-tweets-n844731

[2] https://github.com/Lin1202/TrollDetection.git.

[3] https://www.buzzfeednews.com/article/maxseddon/documents-show-how-russias-troll-army-hit-america

[4] https://www.bbc.com/news/world-europe-31962644

The experimental results show that bilingual learning with deep learning methods is the most effective approach to detect troll tweets in a bilingual corpus.

2. Related Work

2.1. Troll Activity Detection

A lot research works have been conducted on IRA-related troll data. Most of the past work focuses on the troll accounts. They use account profiles, tweet text (Galán-García et al., 2016) or behavior of accounts (Zannettou et al., 2018) to detect troll accounts. Llewellyn et al. (2018) investigated the IRA troll accounts, and found Brexit-related content from 419 of these accounts. Zannettou et al. (2018) compared the behavior of troll accounts with a random set of Twitter users to analyze the influence of troll accounts on social media. However, new accounts can be opened at any time, and troll accounts can be suspended or deleted at any time. It will be problematic to monitor dynamic accounts. Besides, the lifetime of a troll account can be very short. Therefore, detecting troll tweets only using text should be considered. Ghanem et al. (2019) conducted IRA-troll accounts detection from textual perspective using textual and profiling features. But they only worked on English tweets instead of handling the multilingual tweets. In our work, we use IRA-troll data to detect troll tweets in English and Russian.

2.2. Text Representation Models

Our work is to verify if a tweet is produced by a "troll farm", which is similar to the Authorship Verification (AV) if we reduce it to the task of recognizing writing style of the analysed text. Authorship verification refers to the situation when an investigator is given examples of the writing of a single author and is asked to determine if certain other texts were written by this author (Koppel and Schler, 2004; Koppel et al., 2004). Most authorship studies used stylometry and relied on shallow classification models.

Stylometry aims at reflecting personal writing style (Brocardo et al., 2015; Juola and Mikros, 2016). Stylometry analysis captures the unique writing style of the author and uses textual content to help determine the true author (Brocardo et al., 2013). Stylometry features are able to capture the distinctive aspects of someone's writing style, and are consistent even when the author is writing in different languages. Bogdanova and Lazaridou (2014) proposed cross-language stylometric features for cross-lingual Authorship Attribution (AA). They also tried to build a monolingual AA system in one language and then use machine translation to translate any other testing data to that language. However, most of these works focus on long documents. It is more challenging to do authorship verification for short and noisy texts from social media, especially when more than one language is considered.

Recently, deep learning methods have also started to be applied to the authorship analysis task. González (2017) and (Mohsen et al., 2016) has demonstrated that deep learning can be successfully applied in author identification and performing better than other techniques. It is well known that deep learning methods can take *raw text* as input directly. The hidden features are automatically discovered and composed together to produce the final text representation(LeCun et al., 2015). To approach multilingual authorship task, Peng et al. (2003) used character level n-grams and conducted on three different languages. Character and word *n-grams* can capture important morphological properties and discover useful inter-word and inter-phrase features. Therefore, they have been used as the core of many authorship analysis systems (Jankowska et al., 2014; Schwartz et al., 2013; Layton et al., 2010). However, character n-grams sometimes are too short to capture entire words, although some types can capture partial words and other word-relevant tokens. So we also combine character and word n-grams to help determine the author of a text by capturing the syntax and style of that author.

Past works have shown that the most useful features for classifying tweets are word and character n-grams, emoticons, part-of-speech (POS) features, punctuation, and the raw tweet text. González-Gallardo et al. (2015) used stylistic features represented by character n-grams and POS n-grams to classify tweets in four different languages.

2.3. Learning from Multilingual Corpus

To leverage the language resources, we thought to use multilingual learning for training model. Multilingual learning is to joint different languages to learn a single multilingual model rather than handling one language at a time. Multilingual learning has been applied on many researches because it can capture regularities across languages for the mutual benefit of each language (Snyder and Barzilay, 2008). Zapotoczny et al. (2017) showed that a recently proposed neural dependency parser could be improved by joint training on multiple languages. Using multilingual learning can significantly improve the performance of the parser. Ghoshal et al. (2013) used multilingual learning for hybrid Deep Neural Network-hidden Markov model systems, showing that training the hidden layer by using data from multiple languages leads to improved recognition accuracy. Duong et al. (2017) used a multilingual joint training model to build high quality cross-lingual word embeddings for many languages in a unified vector space. Because the tweets in our corpus are mainly in English and Russian, so we will conduct bilingual learning in this work.

Some multilingual classification works rely on automatic machine translation to translate documents from the source language to the target language or vice versa, and then apply classification methods (Shanahan et al., 2004; Banea et al., 2008; Prettenhofer and Stein, 2010). There are some features of the original text that survive translation with enough intact identification (Stuart et al., 2013). Translated texts preserve some of the stylometric features of the original texts (Caliskan and Greenstadt, 2012). Bel et al. (2003) tested the hypothesis of cross-lingual training. That explained a classifier for language A plus a translation from language A to B can enable the classifier to classify texts written in B. Machine translation will be used in our work to enrich the language resources for classification.

3. Methodology

When given a tweet written in a particular language, we need to detect whether it is produced by a "troll farm". We

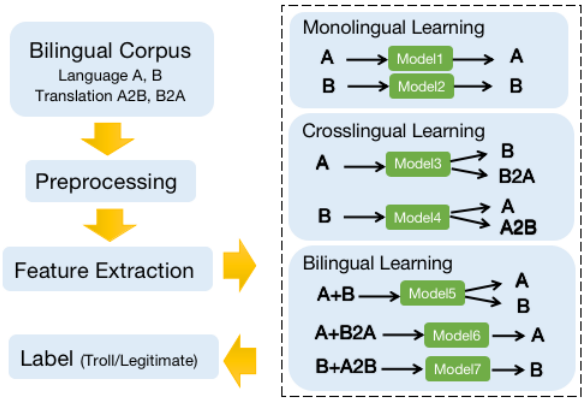


Figure 1: Three evaluated learning scenarios

Feature	Troll	Random
number of emoji	0.054	0.127
number of hashtag	0.784	1.367
contain url or not	0.449	0.482
number of mentions	0.227	1.471
length based on word	5.028	7.072
length based on character	24.158	33.618
number of sentences	0.463	0.773
number of letters	5.028	7.072
number of digits	3.889	5.675
number of nouns	2.308	2.958
number of verbs	1.235	1.631
number of pronouns	0.651	1.058
number of prepositions	0.939	1.142
number of adverbs	0.554	0.788
number of conjunctions	0.408	0.650
number of adjectives	1.189	1.300
number of interjections	0.027	0.069
number of lower case	4.023	5.961
number of upper case	0.095	0.198
number of punctuations	1.394	2.246

Table 1: Standard deviation of stylometric features

evaluated three types of learning to detect troll tweets.

1 **Monolingual Learning.** A classifier is trained on labeled tweets in each language and applied on the tweets in the same language.

2 **Cross-lingual Learning.** A classifier is trained on labeled tweets in one available language but applied on the tweets in another language.

3 **Bilingual Learning.** Classifier is trained on two different languages, and also applied on the tweets in those two languages. It can be reduced to monolingual learning if translation is applied.

Figure 1 shows three evaluated learning scenarios.

3.1. Machine Learning Models

In this work, we considered Naive Bayes (NB), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) classifier to detect troll tweets based on stylometric features. We used Logistic Regression classifier for n-gram features. Also, the following deep learning methods were applied—Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and their combination—on texts of tweets.

3.2. Text Features

We used machine translation in our work, and extracted features that must be preserved in translated text. The following features were extracted from tweets and used in our work:
(1) The stylometric features (Juola and Mikros, 2016) are described in Table 1. We can see that all standard deviations of random tweets are greater than troll tweets, which implies that troll tweets have a relatively consistent writing style both in English and Russian. In contrast, random tweets show higher style diversity. The methods based on stylometric features usually require advanced text preprocessing and explicit features extraction. Such features usually present in every language.
(2)In this work, we used word n-grams (n∈[1,5]), character n-grams (n∈[1,4]), and their combination, to extract the features, and then to compare the different results.
(3) Raw text was submitted as input to deep neural networks.

3.3. Text Preprocessing

Our preprocessing of tweets contains such basic steps as tokenization and POS tagging. Considering the influence of stop words, we found that the stylometric features we extracted would not be unduly influenced by the stop words. That is because only two features: length_word, and length_character, can be different when removing the stop words. To some extent, the use of the stop words is also part of the writing style of the author. So we extracted the stylometric features without removing the stop words. However, for word n-grams, we removed the stop words both for English and Russian.

4. Experiments

4.1. Dataset

We built a new dataset (Miao et al., 2019) based on the data provided by (Zannettou et al., 2018). In the original dataset, 61% are in English, 27% are in Russian. We removed tweets in other languages because the original dataset contains too few tweets in those languages, and this amount of data is insufficient for building an accurate classification model. Retweets and duplicates were filtered out to avoid over-fitting. After content filtering, we obtained 9,257 English and 4,307 Russian original tweets. In order to build a balanced dataset for binary classification, we randomly collected the same amount of normal tweets. The list of random tweets used the same hashtags as those of the troll tweets.
After addition of tweets from legitimate accounts, our dataset contains 18,514 English and 8,614 Russian original tweets.
In addition, we used Google Translate to translate the Russian corpus into English and English corpus into Russian. After these changes, our final dataset contains another 8,614 English tweets (translated from Russian tweets), and 18,514 Russian tweets (translated from English tweets).

4.2. Evaluation Metrics

We used 80% of the data for training, 10% for validation, and the remaining 10% for testing.

Given the unbalanced distribution of tweets in the real world, we computed Receiver Operating Characteristic (ROC) curves and Area Under the ROC Curve (AUC) to quantify the performance of all the methods.

4.3. Experimental Tools

We used the following tools for building an experimental framework: (1) Google Translate[5] to translate English tweets into Russian, and Russian tweets into English; (2) Tokenizer[6] for tokenization, (3) nltk.pos_tag[7] for POS tagging of English and Russian; and (4) Keras[8] for implementing deep learning methods.

To implemented the deep learning methods, first, we uses Keras to pad digital sequences with 0s to a maximum length. We set this length to 60. An embedding dimension of 300 was used to randomly initialized each word into a 300-dimension dense vector. We relied on the sigmoid activation function[9], and learn the weights using the adam optimizer[10]. We used a typical batch size of 256.

CNN We experimented using a multi-channel convolutional neural network with three region sizes (2,3,4) and two filters for each region size applying these six filters on the embedding matrix. The output of each convolution was a column vector, which was subsampled using maxpooling. The outputs of maxpooling were concatenated.

RNN We first used an embedding layer. Then, a spatial dropout was applied to help avoid focusing on specific words in an attempt to generalize well. We set the number of units of GRU to 200. On top of every batch, we applied a global average pooling, and a global max pooling, then concatenated the outputs of the two previous steps.

RNN+CNN We combined the RNN and CNN together by adding a convolutional layer after the layer of bidirectional GRU.

4.4. Results

Our experiments had multiple aims, as follows:

- To compare between different n-grams and decide which ones should be used for tweets classification.
- To examine performance of evaluated machine learning methods applied on tweets represented by stylometric features.
- To compare between different neural networks applied to the troll dataset.
- To examine all evaluated models in monolingual, cross-lingual, and multilingual learning.

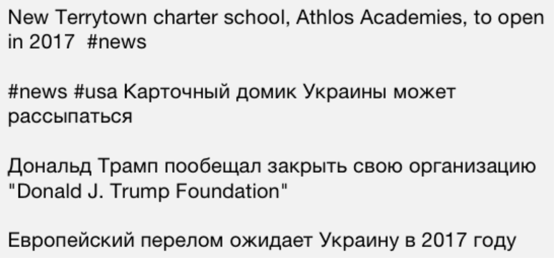
4.4.1. N-grams Analysis

In order to analyse the predictive capability of different n-grams, we used them for the classification task with logistic regression. Table 2 contains their classification scores. The results approve that most of the time, word+character n-grams produce better results. Because they can capture lexical, syntactic, and structural characteristics of an author's writing, increasing the ability to distinguish author features.

Training	Testing	Word	Char	W+C
English		0.803	0.844	**0.861**
Russian		0.764	0.808	**0.813**
English	Russian	0.524	0.619	**0.636**
	Ru2En	0.557	**0.663**	0.640
Russian	English	**0.683**	0.500	0.650
	En2Ru	0.640	0.654	**0.665**
En+Ru	English	0.802	**0.845**	0.835
	Russian	0.755	**0.795**	0.782
En+Ru2En	English	0.809	0.848	**0.861**
Ru+En2Ru	Russian	0.773	0.797	**0.823**

Table 2: Comparison of different levels of n-grams. Word: word-based n-grams; Char: character-based n-grams; Word+Char: word n-grams+character n-grams.

Specifically, using English word n-grams we find that troll tweets are more related to the words, such as "news", "trump", "politics", "breaking", "hilary", "police", etc. Among determinative word n-grams extracted from Russian tweets, both Russian words, such as "ученые", "россии", "тюмень", "сирии", "петербурге", etc., and some English words, such as "breaking", "news", "business", etc. are found. Since we did not filter out the stop words when extract character n-grams, we actually captured more stop words in normal tweets using character n-grams models for English and Russian, such as, "the", "and", "you", "that", "your", "my", "это", "я", "так", "не", "что", "и", etc. Besides, some English letters and digits are extracted from Russian character n-grams. Therefore, combining the word and character n-grams features can enable the models to capture the writing styles more comprehensively, which leads to better results. More importantly, n-grams features are language-independent, so, when training corpus mixed another language, it can also be extracted. For this reason, we are able to use n-grams for cross-lingual and multilingual learning, considering each single tweet can contain more than one language. To be specific, when applying the English model on Russian tweets, English n-grams are used to classify English text (such as hashtags), rather than Russian text in the Russian tweets, and vice versa. Some troll tweets examples are shown in Figure 2, seeing that troll tweets in English and Russian have some words or digits in common.



New Terrytown charter school, Athlos Academies, to open in 2017 #news

#news #usa Карточный домик Украины может рассыпаться

Дональд Трамп пообещал закрыть свою организацию "Donald J. Trump Foundation"

Европейский перелом ожидает Украину в 2017 году

Figure 2: Troll Tweets Examples

[5] https://cloud.google.com/translate/

[6] https://keras.io/preprocessing/text/

[7] https://www.nltk.org/api/nltk.tag.html

[8] https://keras.io/

[9] https://keras.io/activations/

[10] https://keras.io/optimizers/

Training	Testing	SVM	KNN	NB
English		**0.836**	0.794	0.747
Russian		**0.757**	0.704	0.713
English	Russian	**0.706**	0.654	0.689
	Ru2En	0.679	0.624	**0.689**
Russian	English	**0.719**	0.706	0.718
	En2Ru	0.698	0.657	**0.703**
En+Ru	English	**0.820**	0.780	0.735
	Russian	**0.752**	0.691	0.690
En+Ru2En	English	**0.821**	0.776	0.746
Ru+En2Ru	Russian	**0.751**	0.701	0.689

Table 3: Results of classifiers on stylometric features. Ru2En: Russian tweets translated into English; En2Ru: English tweets translated into Russian.

4.4.2. Classification with Stylometric Features

Table 3 shows the comparative performance of three machine learning methods applied on tweets represented by stylometric features. We can see that SVM outperforms other methods. The best results for both English and Russian are obtained when training and testing are conducted on the same language. However, we do not observed any improvements when we translated the tweets into the other language, or mixed the tweets of different languages together for training. Surprisingly, in the cross-lingual learning scenario, the best results were achieved by using the stylometric features, as it can be seen in Section 4.4.5. We found out that the following features have the highest impact for both English and Russian models: ‘number of digits’, ‘number of words’, ‘length based on words’, ‘number of sentences’, ‘number of conjunctions’, ‘number of pronouns’. After analysis of troll and normal tweets in terms of these high importance features, we found out that both English and Russian troll tweets are more likely to use larger number of digits than normal tweets. As an example, ‘2016’ is found to have high effects to classifier troll tweets and normal tweets in character n-grams model. In addition, English and Russian troll tweets are usually shorter (both in terms of number of words and number of sentences) than normal tweets. On the other hand, the conjunctions and pronouns are less frequent in troll tweets compared with normal tweets for both English and Russian. It can be seen that these stylometric features behave highly consistently in both English and Russian. Therefore, they can be successfully applied in cross-lingual classification. We also observed that normal tweets are more likely to contain stop words, which can be explained by the high number of pronouns and conjunctions. As such, we can conclude that these selected stylometric features can be successfully transferred from one language to another. However, most of the stylometric features are language-dependent and will also rely on external natural language processing techniques.

4.4.3. Classification with Deep Neural Networks

Based on the results of our experiments with n-gram and stylometic features, it can be seen that troll tweets have distinctive writing styles compared with normal tweets. Recently, multiple works have shown that deep learning methods can learn the writing styles from raw texts automatically. The comparative results of three deep neural networks (NNs) are shown in Table 4. In general, RNN has better results than CNN, because RNN has the advantage of dealing with sequential data. In our task, the order of words is very important to capture the meaning and the writing style of the author. It also can be seen that better results are usually obtained from RNN+CNN. This is because stacking CNN and RNN together takes advantage of both architectures. Surprisingly, the deep learning methods achieved lower scores than other ML methods. One possible reason may be that the limited corpus size is not enough to learn accurate and representative model with deep neural networks. However, deep learning methods enable us simply pass the raw text directly to the networks. This can totally eliminate the challenges of feature engineering. The latter can explain why the deep learning algorithms have shown themselves more adaptable and efficient in bilingual learning.

Training	Testing	CNN	RNN	R+C
English		0.823	**0.846**	0.824
Russian		0.750	**0.795**	0.781
English	Russian	0.556	0.534	**0.594**
	Ru2En	**0.598**	0.543	0.567
Russian	English	0.580	0.576	**0.592**
	En2Ru	0.657	0.656	**0.666**
En+Ru	English	0.798	0.831	**0.839**
	Russian	0.745	0.763	**0.787**
En+Ru2En	English	0.819	0.845	**0.875**
Ru+En2Ru	Russian	0.776	0.793	**0.828**

Table 4: Comparison of deep learning methods.

We output several tweets to make analysis using SHAP (SHapley Additive exPlanations)(Lundberg and Lee, 2017), which can explain the output of the machine learning model, showing in Figure 3. Figure 3 (a) shows a troll tweet that was correctly classified. This example demonstrates that “politics”, “americans”, etc. have strong effects to classify this tweet as troll tweet. Two examples of normal tweets, misclassified as troll tweets, are shown in Figure 3 (b) and Figure 3 (c). It can be seen that tweets contain such high impact words as, "trump", "killed", etc., and therefore were misclassified. Figure 3 (d) shows a troll tweet wrongly classified as normal, with vocabulary which is more typical for normal tweets.

4.4.4. Monolingual Learning

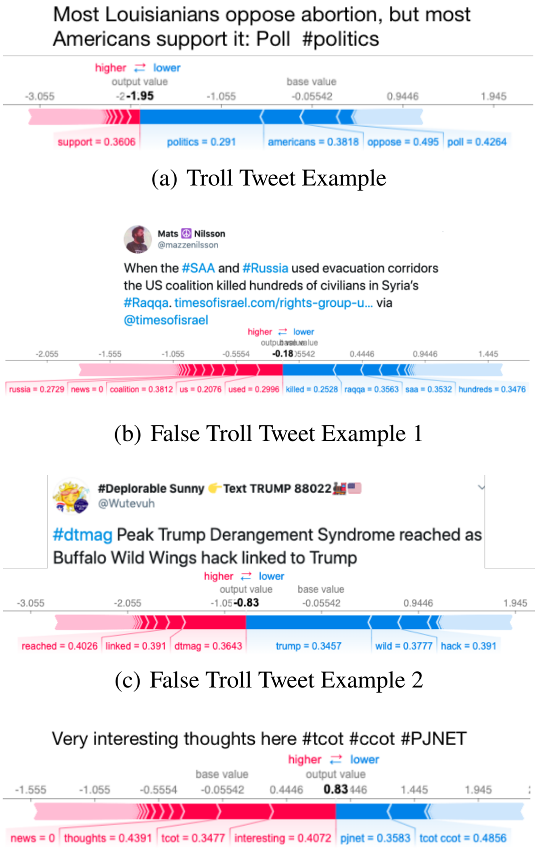
It can be seen from Table 5 that the best results for monolingual scenario is from word+char n-grams using logistic regression. These results suggest that n-grams features are quite good for the classical authorship verification task. Because the corpus is not large, especially for monolingual learning, the data is quite small. In this scenario, using word and character n-grams can be more efficient for authorship verification. Besides, n-grams features are very easy to compute and the logistic regression runs faster than deep learning methods. But in a bilingual corpus, these monolingual methods needed to separate different languages, then, apply on monolingual learning.

4.4.5. Cross-lingual Learning

The cross-lingual learning did not perform very well, as shown in Table 5. Shown in the previous analysis, there are some common features from different languages, but

Best Algorithms		SVM	Logistic Regression	RNN+CNN
Training	Testing	Stylometric	N-grams	Raw Text
Monolingual Training				
English		0.836	**0.861**	0.824
Russian		0.757	**0.813**	0.781
Cross-lingual Training				
English	Russian	**0.706**	0.636	0.594
	Ru2En	**0.679**	0.640	0.567
Russian	English	**0.719**	0.650	0.592
	En2Ru	**0.698**	0.665	0.666
Bilingual Training				
English+Russian	English	0.820	0.835	**0.839**
	Russian	0.752	0.782	**0.787**
English + Ru2En	English	0.821	0.861	**0.875**
Russian + En2Ru	Russian	0.751	0.823	**0.828**

Table 5: Comparison of the best results for each feature sets in different training scenarios.


Most Louisianians oppose abortion, but most Americans support it: Poll #politics

(a) Troll Tweet Example

Mats ☮ Nilsson
@mazzenilsson

When the #SAA and #Russia used evacuation corridors the US coalition killed hundreds of civilians in Syria's #Raqqa. timesofisrael.com/rights-group-u... via @timesofisrael

(b) False Troll Tweet Example 1

#Deplorable Sunny 👉Text TRUMP 88022🚂🇺🇸
@Wutevuh

#dtmag Peak Trump Derangement Syndrome reached as Buffalo Wild Wings hack linked to Trump

(c) False Troll Tweet Example 2

Very interesting thoughts here #tcot #ccot #PJNET

(d) False Normal Tweet Example 1

Figure 3: Error Analysis

not all the features learned from one language can be easily transfered to another languages. Even using machine translation does not affect the results significantly. Using stylometric features provided us the best results in cross-lingual scenario. However, machine translation methods may be less appropriate for cross-lingual learning using stylometric features. This is because the translation could mask the features from the source language, which may lead to poor performance when attempting to distinguish the authorship of tweets.

4.4.6. Bilingual Learning

Combining the target language tweets and translated tweets together as the training set, we obtained the best result of 0.875 for English and 0.828 for Russian. One of the possible explanations may be that when target language tweets are combined with translated tweets for training, the training set provides more knowledge, despite the bad quality of machine translation. Despite this, deep learning methods use raw text, and when used jointly with the translated tweets, to learn comprehensive features of the dataset. This strengthens the distinguishing ability of the classifier without explicit feature extraction. Especially, for each individual tweet, it may contain more than one language. Therefore, bilingual learning can strengthen tweets classification, since it can capture features in both languages comprehensively. Analyzing the English and Russian troll tweets, we found that some of them share the same English or Russian hashtags. This gave a good explanation, showing that when we use bilingual learning, the classifier can learn from all hashtags. In addition, some Russian tweets contain English words. This is especially true when the training set is extended with translated tweets, thereby improving the results.

5. Conclusion and Future Work

In this work, we sought to detect troll tweets in a bilingual corpus. We applied monolingual learning, cross-lingual learning and bilingual learning, using several classification algorithms to approach this task. Bilingual learning achieves the best results when using deep learning algorithms and translated tweets, without feature engineering. Additionally, we provide a bilingual troll tweets resource of English and Russian including the translation of the other language. We believe that future research should extend to multilingual tweets detection, since the multilingual scenario represents the real Twitter environment. Moreover, additional language resources will be made in the future.

Bibliographical References

Banea, C., Mihalcea, R., Wiebe, J., and Hassan, S. (2008). Multilingual subjectivity analysis using machine translation. In Proceedings of the Conference on Empirical

Methods in Natural Language Processing, pages 127–135. Association for Computational Linguistics.

Bel, N., Koster, C. H., and Villegas, M. (2003). Cross-lingual text categorization. In International Conference on Theory and Practice of Digital Libraries, pages 126–139. Springer.

Bogdanova, D. and Lazaridou, A. (2014). Cross-language authorship attribution. In LREC, pages 2015–2020. Citeseer.

Brocardo, M. L., Traore, I., Saad, S., and Woungang, I. (2013). Authorship verification for short messages using stylometry. In Computer, Information and Telecommunication Systems (CITS), 2013 International Conference on, pages 1–6. IEEE.

Brocardo, M. L., Traore, I., and Woungang, I. (2015). Authorship verification of e-mail and tweet messages applied for continuous authentication. *Journal of Computer and System Sciences*, 81(8):1429–1440.

Caliskan, A. and Greenstadt, R. (2012). Translate once, translate twice, translate thrice and attribute: Identifying authors and machine translation tools in translated text. In Semantic Computing (ICSC), 2012 IEEE Sixth International Conference on, pages 121–125. IEEE.

Duong, L., Kanayama, H., Ma, T., Bird, S., and Cohn, T. (2017). Multilingual training of crosslingual word embeddings. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, volume 1, pages 894–904.

Galán-García, P., Puerta, J. G. d. l., Gómez, C. L., Santos, I., and Bringas, P. G. (2016). Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. *Logic Journal of the IGPL*, 24(1):42–53.

Ghanem, B., Buscaldi, D., and Rosso, P. (2019). Textrolls: Identifying russian trolls on twitter from a textual perspective. *arXiv preprint arXiv:1910.01340*.

Ghoshal, A., Swietojanski, P., and Renals, S. (2013). Multilingual training of deep neural networks. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pages 7319–7323. IEEE.

González-Gallardo, C. E., Montes, A., Sierra, G., Núnez-Juárez, J. A., Salinas-López, A. J., and Ek, J. (2015). Tweets classification using corpus dependent tags, character and pos n-grams. In CLEF (Working Notes).

González, P. S. . N. S. . F. A. (2017). Convolutional neural networks for authorship attribution of short texts. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2.

Jankowska, M., Milios, E., and Keselj, V. (2014). Author verification using common n-gram profiles of text documents. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 387–397.

Juola, P. and Mikros, G. K. (2016). Cross-linguistic stylometric features: A preliminary investigation. In International Conference on Statistical Analysis of Textual Data, Nice, France.

Koppel, M. and Schler, J. (2004). Authorship verification as a one-class classification problem. In Proceedings of the twenty-first international conference on Machine learning, page 62. ACM.

Koppel, M., Schler, J., and Mughaz, D. (2004). Text categorization for authorship verification. In Eighth International Symposium on Artificial Intelligence and Mathematics. Fort Lauderdale, Florida, http://rutcor. rutgers. edu/˜ amai/aimath04/SpecialSessions/Koppel-aimath04. pdf.

Layton, R., Watters, P., and Dazeley, R. (2010). Authorship attribution for twitter in 140 characters or less. In Cybercrime and Trustworthy Computing Workshop (CTC), 2010 Second, pages 1–8. IEEE.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Llewellyn, C., Cram, L., Favero, A., and Hill, R. L. (2018). For whom the bell trolls: Troll behaviour in the twitter brexit debate. *arXiv preprint arXiv:1801.08754*.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., pp. 4765–4774.

Mohsen, A. M., El-Makky, N. M., and Ghanem, N. (2016). Author identification using deep learning. In Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on, pages 898–903. IEEE.

Peng, F., Schuurmans, D., Wang, S., and Keselj, V. (2003). Language independent authorship attribution using character level language models. In Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1, pages 267–274. Association for Computational Linguistics.

Prettenhofer, P. and Stein, B. (2010). Cross-language text classification using structural correspondence learning. In Proceedings of the 48th annual meeting of the association for computational linguistics, pages 1118–1127.

Schwartz, R., Tsur, O., Rappoport, A., and Koppel, M. (2013). Authorship attribution of micro-messages. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1880–1891.

Shanahan, J. G., Grefenstette, G., Qu, Y., and Evans, D. A. (2004). Mining multilingual opinions through classification and translation. In Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text.

Snyder, B. and Barzilay, R. (2008). Unsupervised multilingual learning for morphological segmentation. In Proceedings of ACL-08: HLT, pages 737–745.

Stuart, L. M., Tazhibayeva, S., Wagoner, A. R., and Taylor, J. M. (2013). Style features for authors in two languages. In Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 01, pages 459–464. IEEE Computer Society.

Zannettou, S., Caulfield, T., De Cristofaro, E., Sirivianos, M., Stringhini, G., and Blackburn, J. (2018). Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web. *arXiv preprint arXiv:1801.09288*.

Zapotoczny, M., Rychlikowski, P., and Chorowski, J. (2017). On multilingual training of neural dependency parsers. In International Conference on Text, Speech, and Dialogue, pages 326–334. Springer.

Language Resource References

Lin Miao and Mark Last and Marina Litvak. (2019). Bilingual Troll Tweets. https://github.com/Lin1202/TrollDetection.git.