Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2537–2542
Marseille, 11–16 May 2020

Data Augmentation using Machine Translation for Fake News Detection in the Urdu Language

Maaz Amjad[1], Grigori Sidorov[1], Alisa Zhila[2]
[1]Center for Computing Research (CIC), Instituto Politécnico Nacional, Mexico, [2] Independent researcher
maazamjad@phystech.edu

Abstract

The task of fake news detection is to distinguish legitimate news articles that describe real facts from those which convey deceiving and fictitious information. As the fake news phenomenon is omnipresent across all languages, it is crucial to be able to efficiently solve this problem for languages other than English. A common approach to this task is supervised classification using features of various complexity. Yet supervised machine learning requires substantial amount of annotated data. For English and a small number of other languages, annotated data availability is much higher, whereas for the vast majority of languages, it is almost scarce. We investigate whether machine translation at its present state could be successfully used as an automated technique for annotated corpora creation and augmentation for fake news detection focusing on the English-Urdu language pair. We train a fake news classifier for Urdu on (1) the manually annotated dataset originally in Urdu and (2) the machine-translated version of an existing annotated fake news dataset originally in English. We show that at the present state of machine translation quality for the English-Urdu language pair, the fully automated data augmentation through machine translation did not provide improvement for fake news detection in Urdu.

Keywords: fake news detection, Urdu language, language resources, data augmentation, benchmark dataset, classification

1. Introduction

Urdu belongs to the Indo-Aryan language group, and it is the eights most commonly spoken language in the world with more than 100 million speakers. It is the national language of Pakistan and is widely spoken in the Indian subcontinent[1]. Nevertheless, Urdu is considered a severely low-resource language, i.e., only a tiny amount of machine-readable data is publicly accessible. Compiling a dataset originally in Urdu is an expensive task. However, there is a small number of languages including English with more resources available for various natural language processing (NLP) tasks. This provides an opportunity for text data augmentation in a low-resource language using machine translation (MT) from a high-resource language. So, additional annotated text data in one language is produced via machine translation of annotated texts from the other. This paper examines the impact of the MT-based approach to text data augmentation for the fake news detection in Urdu.

So far, fake news detection has been developed to a larger extent for the English language where a variety of different features have been explored. Pérez-Rosas et al. (2018) examined readability, knowledge bases, punctuation. Numerous linguistic techniques have been used to predict different language aspects such as author's writing style, grammatical structure of sentences, and content domain in order to determine truthfulness of a news article (Ireland and Pennebaker, 2010). Based on the available research for English, recent evidence suggests that linguistic features such as word choice and parts-of-speech play a significant role in discriminating satire from real news (Ott et al., 2011; Rubin et al., 2016) and are extremely helpful to differentiate deceptive texts from truthful.

Due to the limited availability of language processing tools (e.g., parsers) for the Urdu language, the research on automated fake news detection in Urdu has been mainly relying on the character and word *n*-gram features (Amjad et al., 2020). Previously, text representation using *n*-grams had been shown successful for fake news classification in Spanish (Posadas-Durán et al., 2018). Yet more complex features, for example, syntactic features based on context-free grammar trees derived using Stanford Parser (Klein and Manning, 2003) that proved efficient for English (Pérez-Rosas et al., 2018), are not readily available for Urdu.

Text data augmentation techniques have been widely investigated for word sense disambiguation (Banea and Mihalcea, 2011), low-resource morphological inflection (Bergmanis et al., 2017), speech recognition (Cui et al., 2015), and audio augmentation for speech recognition (Ko et al., 2015) among others. In these works, automated data augmentation resulted in significant training improvement and more robust models, especially when only smaller datasets were initially available. Text data augmentation techniques include synonym replacement using predictive language models (Kobayashi, 2018) or based on word similarity calculation (Wang and Yang, 2015), data noising as smoothing (Ziang et al., 2017), etc. However, the synonym-based augmentation might not necessarily be helpful for annotated fake news dataset as different synonyms might drastically distort the style.

Text data augmentation through MT remains one of the promising techniques. Previous studies (Yu et al., 2018) used this data augmentation technique to generate new data by translating sentences into French and back into English for a reading comprehension task. Noticeably, the English-French language pair is resource rich including the available parallel corpora, and the MT quality between these two languages is one of the highest.

In this paper, we analyze the opportunities of using English-Urdu MT for annotated text data augmentation in application to fake news detection in Urdu. To the best of our knowledge, this is the first study that undertakes a detailed analysis of data augmentation technique for automated fake

[1] https://en.wikipedia.org/wiki/Urdu

news detection. We generate new annotated data in Urdu by automatically translating the fake news dataset introduced in (Pérez-Rosas et al., 2018) from English. We use the freely available MT from Google, Google Translate[2], which is a *de facto* standard for non-commercial use. We do not perform any manual post-editing after the MT step as it is about as costly as manual annotation. Then, we compare performance of the best *n*-gram based classifier from (Amjad et al., 2020) by training and testing on various combinations of the original Urdu and MT datasets.

We found that the classifier trained on the original Urdu dataset shows better results than the purely MT-translated and the augmented (the combination of the two) datasets despite the 20% size increase in the augmented dataset. We also show that the transferability of the fake news classification model trained on the MT data and applied to inference on the original Urdu text and vice versa is low. Description of the datasets, the detailed augmentation procedure, experimental setup, and the results follows.

2. Datasets

Two existing datasets of news articles annotated with veracity, one originally in Urdu (Amjad et al., 2019) and the other originally in English (Pérez-Rosas et al., 2018), were used as a basis for this study. Both datasets contain news articles in the following domains: sports, entertainment, business, health, and technology.

We applied English-to-Urdu machine translation to the English fake news dataset. Then, we combined the two Urdu datasets, the original one and the machine-translated one, in several ways for more extensive evaluation. The description follows. All resulting datasets in Urdu are available for download [3].

Original Urdu Dataset (Amjad et al., 2019) contains 500 real and 400 fake news originally written in Urdu.

Machine-Translated (MT) Dataset is a machine-translated version of the English Fake News dataset created by Pérez-Rosas et al. (2018), which contains 200 legitimate and 200 fake news. The free version of Google Translate [2] was used for translation from English to Urdu. We intentionally did not perform any manual post-editing because our goal was to investigate the potential opportunities of using machine translation for data augmentation as a less costly and faster alternative to manual dataset creation.

Augmented Dataset is created by combining the Original Urdu and the MT datasets. It contains 700 real news and 600 fake news. Of those, 900 articles are originally in Urdu, and 400 articles are translated.

Augmented Downsized Dataset is created by randomly removing articles from the Augmented Dataset to bring its size to the original Urdu dataset for a more fair comparison. It has 636 articles in Urdu and 272 translated articles.

More detailed statistics for each dataset is provided in Table 1. When training the classifiers, the train-test split was 70/30 for all datasets.

Additionally, to verify the transferability of the fake news detection models trained on one dataset and applied to another, we introduce two **"cross"** datasets.

Dataset	**# news**	**total tok.**	**tok./ news**	**F**	**MT**
Original	900	153K	686	0.44	0.00
MT	400	31K	312	0.50	1.00
Augm.	1300	184K	570	0.46	0.31
Augm.$_{ds}$	908	128K	565	0.46	0.29

Table 1: Dataset statistics includes **F**ake news article fraction as well as **M**achine **T**ranslated article fraction.

Original to MT trains on the Original Urdu training set (638 articles) and tests on the 120 MT articles.

MT to Original trains on the 240 machine-translated articles and tests on the Original Urdu test set of 262 articles.

3. Experimental Settings

Recently, Amjad et al. (2020) conducted research on application of n-gram feature based classifiers to fake news detection in an original Urdu news corpus. In this paper, we use five best performing classifiers from the previous work and compare their performance on the described datasets.

Features. As explained in 1., we use only on n-gram features including character n-grams, word n-grams, and n-grams of function words.

Character n-grams are sequences of n characters, e.g., unigrams, bigrams, trigrams, etc. They had been previously shown to capture morphological and syntactical information well in text. Recent work showed that character n-grams achieved significant improvements in detecting fake news (Potthast et al., 2018; Posadas-Durán et al., 2018).

Word n-grams highlight whether certain phrases are characteristic for fake news (Pérez-Rosas et al., 2018).

N-grams of function word include articles, prepositions, determiners, conjunctions, and auxiliary verbs. It has been argued that the use of function words as features provided notable results to differentiate factually incorrect text from factually correct (Posadas-Durán et al., 2018; Potthast et al., 2018). Likewise, a number of authors (Sanchez-Perez et al., 2017; Gómez-Adorno et al., 2018a; Gómez-Adorno et al., 2018b) suggested that n-grams of function words achieved significant improvements in detecting author's writing style. Posadas-Durán et al. (2018) suggested that word and character n-grams without stop words provided worse results in the fake news detection task.

In this work, we use several combinations of different types of n-grams with n varying from 0 to 2, which previously showed best results for fake news detection using the Original Urdu dataset (Amjad et al., 2020). The combinations are encoded as (Xc-Yw-Zf) standing for character X-gram, word Y-gram, and function word Z-gram.

One possible implication of machine translation is that the standard word order can be changed in the target language in an unnatural way. Additionally, words may be misused. All this may distort the original meaning. We attempt to verify whether different n-gram feature types, namely, character, word, or function word n-grams, can differentiate fake news in the machine-translated dataset in Urdu.

Feature Normalization. As n-grams are count-based features, their values need to be *normalized* so that they are in-

[2] https://translate.google.com

[3] https://github.com/MaazAmjad/Urdu-News-Augmented-Dataset.git

dependent from the text length. It was shown (Amjad et al., 2020) that global weighting schemes such as *TF-IDF* and *log-entropy* decrease the classification performance. Global weighting schemes failed for fake news detection task because the frequent word n-grams (as well as character and function word, respectively) provide relevant information to the classifier, while their values tend to be suppressed by the global normalization schemes. We use $binary$ weighting scheme for most of the features, and $L2-norm$ weighting scheme only for character bi-grams.

The **binary weighting scheme** of a feature constrains feature's values to only two possible values, 1 if the n-gram is present in text or 0 otherwise. More precisely, we followed the scheme per (Pang et al., 2002), where w_i = 1, if $tf_i > 0$ and w_i = 0, if $tf_i = 0$, where tf_i is defined as the number of times that term i appears in document D.

The $L2$ **norm** (Horn and Johnson, 1990) relies on the euclidean distance for normalization:

$$||x|| = \sqrt{x^T x} = \sqrt{\sum_{i=1}^{d} x_i^2}$$

The unit $L2$ norm means that for each vector representation x, the length of vector $||x||$ = 1. Hence, vector x can be normalized as $\frac{x}{||x||}$. The $L2$-normalized frequency weighting scheme scales each datapoint's feature vector representation to have length of 1, or in other words, a unit norm.

Classifier Algorithms. The primary focus of this work is investigating the effect that dataset augmentation can have on the fake news classifier efficiency. Therefore, we conducted our experiments on the fixed selection of the best performing machine learning algorithms from the previous work on fake news discrimination for both Urdu (Amjad et al., 2020) and Spanish (Posadas-Durán et al., 2018), which are *Support Vector Machine (SVM)* and *AdaBoost*. These algorithms are well known for their high performance for a range of NLP tasks in various languages obtaining state-of-the-art performance in opinion mining (Sidorov et al., 2012), authorship attribution (Stamatatos, 2009), author profiling (Markov et al., 2017), and sentiment analysis (Pang and Lee, 2008; Loukachevitch et al., 2015).

Metrics. We use $F1$-measure and ROC-AUC as the main metrics for evaluation. Following the paper by Pérez-Rosas et al. (2018), we evaluate our classification models for detection on both directions: (i) out of all news, detect the legitimate ones, and (ii) out of all news, detect the fake ones. We use F1_{legit} and F1_{fake} correspondingly. The ROC-AUC metrics stays the same for both directions of detection as it evaluates the overall quality of a model.

4. Results and Analysis

From Tables 3. and 3. we observe that the highest results were consistently achieved for the fake news dataset originally in Urdu across all experiments. The MT and Augmented datasets failed to provide better results. Interestingly, some experiments on the smaller $\text{Augmented}_{\text{Downsized}}$ dataset performed slightly better ($< 1\%$) than when trained on the full-sized Augmented dataset. However, we do not claim statistical significance for this difference. It might be due to the varying fraction of the original Urdu texts, 71% for $\text{Augmented}_{\text{Downsized}}$ *vs.* 69% for Augmented.

The worst performance was observed in the experiments using the cross datasets where training was performed on one type of texts, either original Urdu or purely MT, and testing was done on the opposite test set. Here the performance was close to random as is revealed by the almost straight corresponding lines on Figure 3).

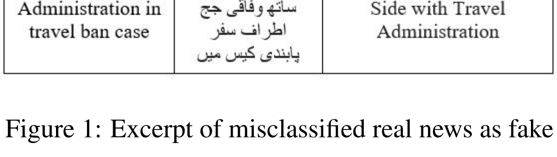
Feature Analysis. For the Original dataset, the combination of character and word unigrams (*1c-1w-0f*) achieves the highest results with 0.84 F1_{Fake} and 0.94 ROC-AUC scores. Subsequently, character bi-grams only (*2c-0w-0f*) present slightly lower ROC-AUC of 0.93. These feature sets show best results for all non-cross datasets, with the purely MT dataset achieving the highest of all (but the original) results of 0.83 F1_{Fake} and 0.91 ROC-AUC and the Augmented dataset achieving its highest ROC-AUC score of 0.90, F1_{Fake} of 0.79. The ROC curves for this feature set are shown in Figure 3.

We observe that solely word bi-grams (*0c-2w-0f*) gave the worst results for all datasets compared to other n-gram types. This means that relying solely on word phrases is insufficient for detecting fake news.

One of the potential reasons behind the character bi-grams performing generally better than word bi-grams is the relatively small size of the datasets. Training on smaller datasets is susceptible to the feature space dimension, which becomes quite large for word bi-grams as there is a larger variety of two word combinations compared to two character combinations.

However, concerning the datasets that include machine-translated text, another possible reason for bad performance is that during machine translation, a translation could change the natural word order, creating otherwise uncommon combinations of words, which do not provide useful input for fake news discrimination. Hence, the signal from word n-grams drops. On the contrary, the character bi-grams maintain strong signal because machine translation does not affect the order of characters.

Quality of Machine Translation. To investigate the low performance on the datasets containing machine-translated articles, we manually verified the quality of the Google Translate's English-Urdu machine translation. We conclude that the translation is fairly inaccurate. For example, in figure 1 legitimate political domain news was classified as fake by the classifier using the (*0c-2w-0f*) feature set potentially due to heavily distorted semantics.



English (Original)	Machine Translation	Translation of (Machine Translation)
Federal Judge sides with Trump Administration in travel ban case	ٹریول ایڈمنسٹریشن کے ساتھ وفاقی جج اطراف سفر پابندی کیس میں	In the Travel Bonding Case on Federal Judge Side with Travel Administration

Figure 1: Excerpt of misclassified real news as fake

In Figure 2 a fake news article on technology was classified as real using the more robust (*2c-0w-0f*) feature set. The misclassification could be due to the almost contrary mean-

Dataset	1c-1w-0f			2c-2w-2f			2c-1w-0f		
—	F1_{Fake}	F1_{legit}	ROC	F1_{Fake}	F1_{legit}	ROC	F1_{Fake}	F1_{legit}	ROC
Original	0.84	0.88	0.94	0.82	0.89	0.93	0.83	0.87	0.93
MT	0.65	0.72	0.70	0.80	0.77	0.88	0.78	0.76	0.83
Augm.	0.75	0.78	0.87	0.74	0.80	0.86	0.80	0.83	0.88
Augm.$_{Downsized}$	0.73	0.76	0.83	0.75	0.80	0.86	0.74	0.80	0.85
MT to Original	0.36	0.79	0.54	0.19	0.72	0.58	0.44	0.66	0.62
Original to MT	0.34	0.53	0.47	0.08	0.63	0.41	0.37	0.61	0.51

Table 2: Classification results of fake news detection in terms of F1_{Real}, F1_{Fake} and ROC-AUC

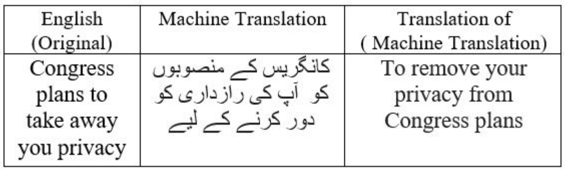
Dataset	2c-0w-0f			0c-2w-0f		
—	F1_{Fake}	F1_{legit}	ROC	F1_{Fake}	F1_{legit}	ROC
Original	0.84	0.88	0.93	0.54	0.70	0.69
MT	0.83	0.83	0.91	0.63	0.60	0.62
Augm.	0.79	0.84	0.90	0.53	0.64	0.64
Augm.$_{Downsized}$	0.75	0.82	0.89	0.48	0.65	0.62
MT to Original	0.50	0.59	0.58	0.52	0.55	0.56
Original to MT	0.22	0.60	0.44	0.15	0.57	0.44

Table 3: Classification results of fake news detection in terms of F1_{Real}, F1_{Fake} and ROC-AUC

ing in the translation as well as the unnatural word order.

So, data augmentation using machine translation does not provide any improvement for fake news detection in Urdu. The drastically low classification results for cross datasets confirm that the language of the original and MT articles is substantially different in word order, syntax, and semantics.

Our conclusion is that the present quality of publicly available MT systems for the English-Urdu language pair does not enable considering MT as a feasible technique for data augmentation particularly in case of fake news detection.



English (Original)	Machine Translation	Translation of (Machine Translation)
Congress plans to take away you privacy	کانگریس کے منصوبوں کو آپ کی رازداری کو دور کرنے کے لیے	To remove your privacy from Congress plans

Figure 2: Excerpt of misclassified fake news as real

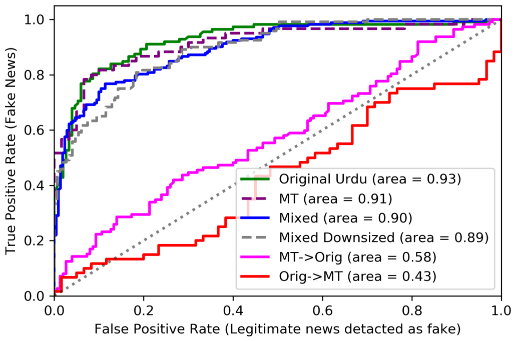


Figure 3: ROC curves for the *2c-0w-0f* feature set

5. Conclusion

We investigated whether machine translation from English to Urdu can be applied as a text data augmentation technique to expand the limited annotated resources for Urdu. Yet the empirical results show that at its current stage, the machine translation quality for this language pair does not enable efficient automated data augmentation, in particular, for such high level task as fake news detection which relies on linguistically more precise data. A lot of misclassification errors stem from the poor quality of translation and unnatural sentences generated in the target language (Urdu).

Additionally, our findings indicate that character bi-grams are the most robust and highest performing feature for the datasets augmented with machine-translated texts. Whereas the solely word bi-gram features performed poorly in general for all datasets, potentially due to the high dimensionality of the word bi-gram space w.r.t. the sizes of the available datasets.

In future, we intend to explore other text data augmentation techniques, in particular, the synonym replacement, and its application to fake news detection. We also plan to investigate whether text data augmentation through MT for other language pairs with larger parallel resources will provide more promising results. We plan to start with Spanish-English based on the availability of the annotated benchmarking fake news datasets in these languages. We will also flip the direction of machine translation from Urdu to English to see whether other features and systems available for the English language can perform better on the machine-translated texts from Urdu to English.

6. Acknowledgements

The work was done with partial support of CONACYT project 240844 and SIP-IPN projects 20195719.

7. Bibliographical References

Amjad, M., Sidorov, G., Zhila, A., Gómez-Adorno, H., Voronkov, I., and Gelbukh, A. (2020). Bend the Truth:

A Benchmark Dataset for Fake News Detection in Urdu Language and Its Evaluation. *Journal of Intelligent & Fuzzy Systems*. In press.

Banea, C. and Mihalcea, R. (2011). Word Sense Disambiguation with Multilingual Features. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*, pages 25–34.

Bergmanis, T., Kann, K., Schütze, H., and Goldwater, S. (2017). Training Data Augmentation for Low-Resource Morphological Inflection. *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 31–39.

Cui, X., Goel, V., and Kingsbury, B. (2015). Data Augmentation for Deep Neural Network Acoustic Modeling. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(9):1469–1477.

Gómez-Adorno, H., Martín-del Campo-Rodríguez, C., Sidorov, G., Alemán, Y., Vilariño, D., and Pinto, D. (2018a). Hierarchical Clustering Analysis: The Best-Performing Approach at PAN 2017 Author Clustering Task. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, CLEF'2018, pages 216–223. Springer.

Gómez-Adorno, H., Ríos-Toledo, G., Posadas-Durán, J. P., Sidorov, G., and Sierra, G. (2018b). Stylometry-based Approach for Detecting Writing Style Changes in Literary Texts. *Computación y Sistemas*, 22(1):47–53.

Horn, R. A. and Johnson, C. R., (1990). *Norms for Vectors and Matrices*, pages 313–386. Cambridge University Press, Cambridge, England.

Ireland, M. E. and Pennebaker, J. W. (2010). Language Style Matching in Writing: Synchrony in Essays, Correspondence, and Poetry. *Journal of Personality and Social Psychology*, 99(3):549–571.

Klein, D. and Manning, C. D. (2003). Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio Augmentation for Speech Recognition. In *INTERSPEECH*, pages 3586–3589. ISCA.

Kobayashi, S. (2018). Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457. Association for Computational Linguistics.

Loukachevitch, N., Blinov, P., Kotelnikov, E., Rubtsova, Y., Ivanov, V., and Tutubalina, E. (2015). SentiRuEval: Testing Object-Oriented Sentiment Analysis Systems in Russian. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue 2015*, 2:3–13.

Markov, I., Gómez-Adorno, H., Sidorov, G., and Gelbukh, A. (2017). The Winning Approach to Cross-Genre Gender Identification in Russian at Rusprofiling 2017. In *Working notes of FIRE 2017 - Forum for Information Retrieval Evaluation*, pages 1–216.

Ott, M., Choi, Y., Cardie, C., and Hancock, J. T. (2011). Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 309–319. Association for Computational Linguistics.

Pang, B. and Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2018). Automatic Detection of Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401.

Posadas-Durán, J. P., Gómez-Adorno, H., Sidorov, G., and Moreno Escobar, J. (2018). Detection of Fake News in a New Corpus for the Spanish Language. *Journal of Intelligent & Fuzzy Systems*, 36(5):4868–4876.

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., and Stein, B. (2018). A Stylometric Inquiry into Hyperpartisan and Fake News. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240. Association for Computational Linguistics.

Rubin, V., Conroy, N., Chen, Y., and Cornwell, S. (2016). Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17.

Sanchez-Perez, M. A., Markov, I., Gómez-Adorno, H., and Sidorov, G. (2017). Comparison of Character N-grams and Lexical Features on Author, Gender, and Language Variety Identification on the Same Spanish News Corpus. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, CLEF'2017, pages 145–151. Springer.

Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., Díaz-Rangel, I., Suárez-Guerra, S., Trevino, A., and Gordon, J. (2012). Empirical Study of Machine Learning based Approach for Opinion Mining in Tweets. In *Mexican International Conference on Artificial intelligence*, MICAI'2012, pages 1–14. Springer.

Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Wang, W. Y. and Yang, D. (2015). That's So Annoying!!!: A Lexical and Frame-Semantic Embedding Based Data Augmentation Approach to Automatic Categorization of Annoying Behaviors using #petpeeve Tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2557–2563. Association for Computational Linguistics.

Yu, A. W., Dohan, D., Luong, M.-T., Zhao, R., Chen, K., Norouzi, M., and Le, Q. V. (2018). QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. *CoRR*.

Ziang, X., Wang, S. I., Li, J., Levy, D., Nie, A., Jurafsky, D., and Ng, A. Y. (2017). Data Noising as Smoothing in Neural Network Language Models. In *5th International Conference on Learning Representations, ICLR 2017*.

8. Language Resource References

Amjad, M., Sidorov, G., Zhila, A., Gómez-Adorno, H., Voronkov, I., and Gelbukh, A. (2019). *Bend The Truth: Benchmark Dataset for Fake News Detection in Urdu*. M. Amjad, distributed via GitHub, 1.0, https://github.com/MaazAmjad/Datasets-for-Urdu-news.git.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2018). *Fake News*. R. Mihalcea, https://web.eecs.umich.edu/˜mihalcea/downloads.html #FakeNews.