Cross-Domain Failures of Fake News Detection

Maria Janicka, Maria Pszona, Aleksander Wawer

Samsung R&D Institute Poland
pl. Europejski 1
00-844 Warsaw, Poland
{m.janicka, m.pszona, a.wawer}@samsung.com

Abstract. Fake news recognition has become a prominent research topic in natural language processing. Researchers reported significant successes when applying methods based on various stylometric and lexical features and machine learning, with accuracy reaching 90%. This article is focused on answering the question: are the fake news detection models universally applicable or limited to the domain they have been trained on? We used four different, freely available English language Fake News corpora and trained models in both in-domain and cross-domain setting. We also explored and compared features important in each domain. We found that the performance in cross-domain setting degrades by 20% and sets of features important to detect fake texts differ between domains. Our conclusions support the hypothesis that high accuracy of machine learning models applied to fake news detection may be related to over-fitting, and models need to be trained and evaluated on mixed types of texts.

Keywords: Fake news recognition · Machine Learning · Feature Analysis · Cross-domain analysis.

1 Introduction

Recognizing fake news is a problem of automatically detecting misleading news stories, ones that often come from non-reputable sources. The research on fake-news detection surged since the 2016 US presidential campaign. While the most reliable approach is human fact-checking, the one we focus on in this paper is the analysis of non-lexical properties, such as psycholinguistic and stylometric variables obtained from several available tools. Non-lexical analysis is interesting because, at least in theory, it should allow to abstract from topics and domains, resulting in a more universally applicable solution.

The paper is organized as follows: Section 2 describes previous studies on the topic of fake news detection, Section 3 describes fake news data sets, Section 4 outlines the input features and machine learning methods. Section 5 contains the results of experiments and Section 6 analysis of features.

2 Previous Work

As typically fake news is intentionally created to spread misinformation, their writing style slightly differ from that of reliable content. Therefore, the tradi-

tional approach to detect false content is based on linguistic features. Until now, one of the top classifiers relying on these features achieved the accuracy of up to 76% [15]. The studies revealed that the punctuation and factors related to the complexity of text - including a number of characters, words, syllables, complex words, long words and several readability metrics such as Flesch-Kincaid [6], Gunning Fog [7] and Automatic Readability Index [18], are of the highest importance. The semantic features, which can be extracted from Linguistic Inquiry and Word Count software (LIWC) [14] are also crucial. LIWC not only provides a number of words that fall into different meta-language categories such as positive emotions, analytical thinking, cognitive process but also carry out part-of-speech tagging. The successful usage of these features was demonstrated in the detection of falsified reviews [13] or prisoners lies [2].

Zheng et al. demonstrated that by using relationships between news article, its creator, subject and fake/true label, it is possible to achieve accuracy as of 0.63 [21]. They designed a diffusive network based on a set of explicit and latent features extracted exclusively from textual content.

A news article is often accompanied by visual materials like images or videos, which are rarely taken into account. Nevertheless, recent studies revealed that fake and real news exhibit different image distribution patterns. Jin et al. proposed several visual and statistical features supporting the detection of fake news [8].

Another common approach, adopted for example by B.S. Detector [1], a browser extension alerting users about unreliable news source, is simply based on a curated open-source database containing an assessment of online information sources. We observed that some of the already unmasked pages that deliberately publish fake content, use redirection to different URL, so there is an overwhelming need to update this database regularly. While detecting false content, the initial step should involve the assessment of source credibility. The authors of the already mentioned database suggest 6 practical steps. For instance, checking whether the title or domain is not just a slight variation on a well-known website, verification of mentioned links, referenced sources and quotation, both aesthetic and writing style analysis.

Some studies revealed that the fake and reliable news follow different patterns of propagation in social media [5]. Interestingly, this is noticeable even at early stages of spreading, which is extremely useful in preventing the negative impact of misinformation on society [22].

3 Fake News Data Sets

This section describes data sets used in our experiments.

[1] `https://bsdetector.tech/`

3.1 Kaggle

The data set [2] contains news collected with B.S. Detector, a browser extension, which provides a list of unreliable sources. It is the biggest data set used in this research containing 20800 texts falling into two categories: fake and true.

3.2 LIAR

The data set [20] contains short statments manually labeled by PolitiFact ([3]) fact-checkers. These short texts are categorized into 6 groups, but we included only texts labeled as true together with false and pants-fire category (the most obviously fake information) labeled as fake.

3.3 Pérez-Rosas et al.

Fake news data was generated using crowdsourcing via Amazon Mechanical Turk. The AMT workers were asked to generate fake versions of true news collected earlier in a corpus. Each of the fake news had to mimic a journalistic style [16].

3.4 Buzzfeed

This data set is created from top fake news on Facebook reported in years 2016 and 2017. The data was collected using BuzzSumo with the help of PolitiFact information and Buzzfeed own resources. Then complementary 91 real news was added.

3.5 Data Set Summary

The data sets are summarized in Table 1. They differ not only in origin and length, but also in topic (although politics somehow dominates). The proportions of fake vs true are somewhat balanced.

4 Machine Learning

To build a comprehensive set of features we discriminated four areas of linguistic investigation. In total, we collected 279 features. The biggest subset consists of 182 General Inquirer features plus two special features from purposely designed dictionaries. The second group is built out of 61 features containing POS tags (56) and syntactic information (3). We also add 35 psycholinguistic features connected with readability and one feature containing information about text subjectivity.

[2] `https://www.kaggle.com/c/fake-news/data`
[3] `https://www.politifact.com/`

Dataset	Size	Length	Comments
Kaggle	10 387 true 10 413 fake	medium	only politics
LIAR	2053 true 2454 mostly-true 2627 half-true 2103 barely-true 2507 false 1047 pants-fire	very short	only politics
Pérez-Rosas et al.	240 true 240 fake	medium	seven domains; for each legitimate news fake news generated via Amazon Mechanical Turk (AMT)
Buzzfeed	91 true 91 fake	medium	categories: politics, breaking news, business, local news, medicine, race

Table 1: Data set summary

In the first step, we carried out basic analysis including information about parts of speech and syntactic structure. We used Spacy library to count the percentage of an occurrence of a given POS tag in news text. CoreNLP dependency parser was employed to measure parse tree depth together with the depth of a noun phrase.

I the next step we use General Inquirer [19] – a tool for text content analysis which provides a wide range of categories. It helps to characterize text by defining words in terms of sentiment, intensity, varying social and cognitive contexts. Categories were collected from four different sources - the Harvard IV-4 dictionary, the Lasswell value dictionary [10] - several categories were constructed based on work of Semin and Fiedler on social cognition and language [17], finally, marker categories were adapted from Kelly and Stone work on word sense disambiguation [9]. In addition to enrich existing feature set with domain-relevant terms, we created two special dictionaries containing linguistic hedges and exclusion terms. We created features from General Inquirer in fake news classifier by measuring the ratio of words in a given category to all words in a text.

Further, we enriched feature space with readability indices[4]. We used popular measures which represent an approximation of the level of education needed to understand a text - Flesh-Kincaid [6], ARI [18], Coleman-Liau [4], Gunning Fog Index [7], LIX [1], SMOG Index [11], RIX [1], Dale-Chall Index [3]. Each metric uses different premises connected with word-level and sentence-level complexity - e.g., sentence length, word length, number of syllables per word, number of long words in a text or information about part-of-speech or sentence beginnings. We also use these indicators in a stand-alone manner as a set of psycholinguistic features.

[4] `https://github.com/andreasvc/readability/`

The last is a sentence-level feature – subjectivity. We used subjectivity classifier [5] based on bi-directional GRU to find subjective sentences in a text. Percentage of subjective sentences serves as a feature in fake news classifier.

Values of all features were normalised and four different approaches to classification task were tested. We trained support vector classifier with the linear kernel, support vector machine with stochastic gradient descent, extremely randomized decision trees and extreme gradient boosting.

5 Results

This section presents the results of machine learning experiments. For each data set, we treated it as a training data source, and performed a number of cross-domain experiments. We tested the obtained model on the same data set (in this case, we split data into random 80% train and 20% test subsets). We also applied it to all other data sets (in this case, we used the whole data set for training). We present each experiment in a separate table. Table 2 contains the results of models trained on Mihalcea data set [16], Table 3 illustrates the performance of models trained on Kaggle data, Table 4 shows the results of models trained on Politifact, and finally Table 5 shows the accuracy of Buzzfeed-trained models.

Classifier	Training set	Pérez-Rosas et al.	Kaggle	LIAR	Buzzfeed
LinearSVC	Pérez-Rosas et al.	0.667	0.506	0.549	0.626
SGDClassifier	Pérez-Rosas et al.	0.675	0.466	0.535	0.626
ExtraTreesClassifier	Pérez-Rosas et al.	0.733	0.517	0.535	0.495
XGBoost	Pérez-Rosas et al.	0.767	0.42	0.463	0.588

Table 2: Accuracy on Pérez-Rosas et al. data set

Classifier	Training set	Kaggle	LIAR	Pérez-Rosas et al.	Buzzfeed
LinearSVC	Kaggle	0.974	0.569	0.473	0.33
SGDClassifier	Kaggle	0.973	0.586	0.465	0.308
ExtraTreesClassifier	Kaggle	0.962	0.628	0.506	0.401
XGBoost	Kaggle	0.977	0.628	0.483	0.352

Table 3: Accuracy on Kaggle data set

The results reveal that fake news detection, once models are trained and tested within the same data set, appears to be a promising problem to solve. Here, Kaggle data set is a special one: when training and testing models on it, the accuracy is astonishing (near one). The best classifiers on Pérez-Rosas et al. and Buzzfeed data sets reach accuracy in the range of 0.76-0.78. LIAR set is the most difficult even for models trained on this data set, as the accuracy does not exceed 0.653.

[5] `https://github.com/fractalego/subjectivity_classifier`

Classifier	Training set	LIAR	Kaggle	Pérez-Rosas et al.	Buzzfeed
LinearSVC	LIAR	0.636	0.785	0.562	0.61
SGDClassifier	LIAR	0.533	0.468	0.517	0.555
ExtraTreesClassifier	LIAR	0.629	0.464	0.552	0.516
XGBoost	LIAR	0.653	0.486	0.496	0.511

Table 4: Accuracy on LIAR data set

Classifier	Training set	Buzzfeed	Kaggle	LIAR	Pérez-Rosas et al.
LinearSVC	Buzzfeed	0.674	0.625	0.519	0.59
SGDClassifier	Buzzfeed	0.674	0.604	0.498	0.59
ExtraTreesClassifier	Buzzfeed	0.739	0.547	0.522	0.535
XGBoost	Buzzfeed	0.783	0.554	0.586	0.567

Table 5: Accuracy on Buzzfeed data set

Among classification algorithms, the one that performs best within the same data set is XGBoost. However, cross-domain application reveals that it comes at the cost of overfitting: it does not generalize well to other types of fake news data. In most cross-domain settings it is significantly outperformed by LinearSVC.

However, the most interesting observation is that in all of the cases (datasets and classifiers), applying the models to other data sets yields sharp drops of accuracy, often down to values similar to random baselines. Kaggle-trained classifiers are not better in this respect, since the accuracy ranges between 0.62 when applied to LIAR to as low as 0.4 when applied to Buzzfeed.

Surprisingly, the LinearSVC classifier trained on LIAR data set, where it reached 0.636, managed to perform better when applied to the Kaggle data (0.785).

6 Feature Analysis

To gain more understanding of the data, we have performed an analysis of feature distribution in each of the data sets. We have selected four features that are both relevant and exhibit interesting patterns, and illustrated their occurrences as histograms. The features are as follows:

- Linguistic Category Model's Descriptive Action Verbs (DAV).
- Linguistic Category Model's State Verbs (SV)
- Verbs in Past Tense.
- Automated Readability Index (ARI).

Linguistic Category Model [17] is a framework for verb categorization according to their abstractness. DAV verbs correspond to the most concrete class, while SV verbs are the most abstract. According to Pennebaker et al. [12], the language of deception is linked to the higher levels of abstraction. This finding is reflected in the Kaggle data set distributions for DAV and SV verbs. True

texts contain more DAVs and less SVs (are less abstract), and vice versa. This conclusion can not be observed in other three data sets.

Verbs in the past tense can also help distinguish fake and true news on the Kaggle data. Fake news less often refer to past actions and events (contain less verbs in the past tense) than true news.

Automated Readability Index (ARI) is a tool to measure language complexity and understandability. On Kaggle news, it reveals that fake news are on average less readable (a high spike in distribution) than true news. ARI is used as an example, but we can observe similar differences with respect to all of the readability measures. Those features show the most distinct values discrepancy between fake and real content, which suggests that readability plays a major role in a good in-domain performance of a classifier trained on the Kaggel data set. Again, the observation does not hold for the remaining data sets.

In the LIAR data, larger than Pérez-Rosas et al. and Buzzfeed data, distribution of four features within fake news was similar to that within true news. No apparent patterns could be observed. Feature distributions in Pérez-Rosas et al. and Buzzfeed data are similar, expectedly the amount of noise increases with decreasing size of the corpora. None of the three data sets seems to be usable for high performance detection of fake vs true news using stylometric and psycholinguistic features.

Contradictory, in the Kaggle data set we can see some clear differences in feature distribution, which reflects in high accuracy of classifiers trained and tested on this data set. Our hypothesis is that those news articles are of special character, as they were not collected in a manual fact-checking procedure, but were added from sources marked as unreliable. This kind of web pages are created to manipulate audience and language may differ from those of legitimate journalism.

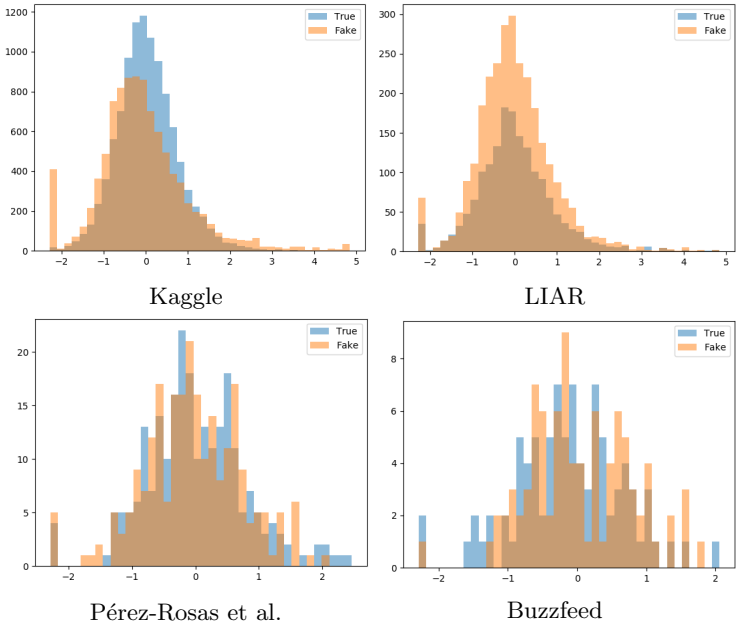

Fig. 2: Descriptive Action Verbs (DAV)

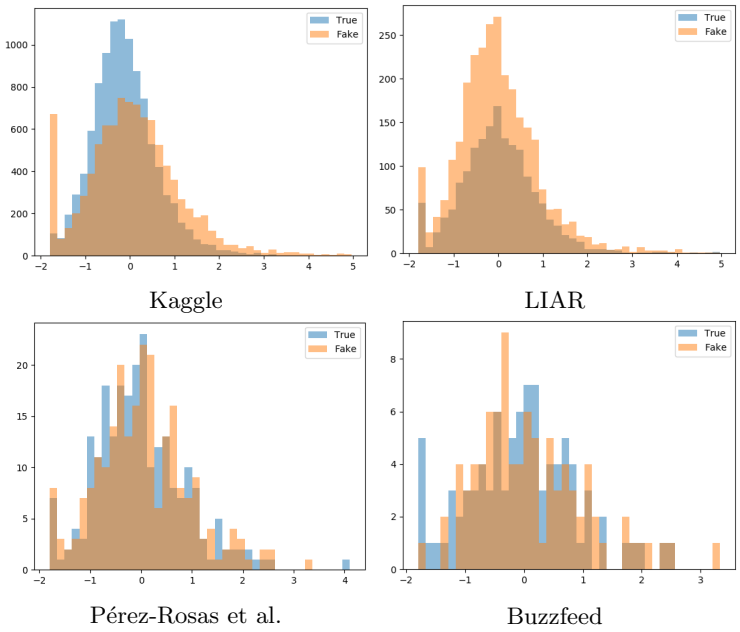

Fig. 4: State Verbs (SV)

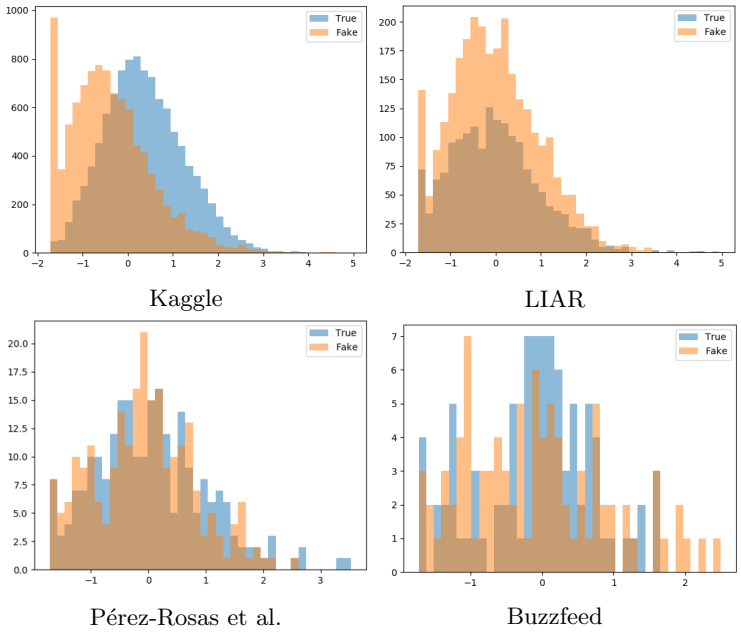

Fig. 6: Verb - Past Tense

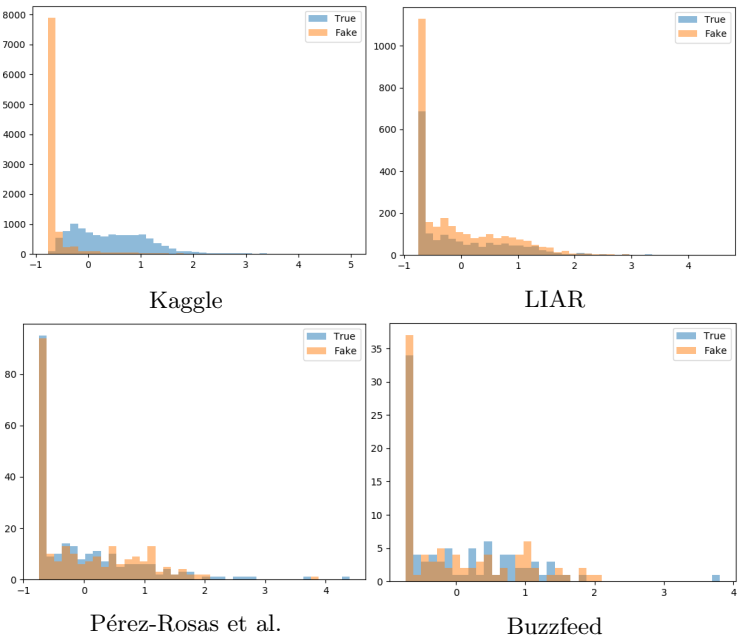

Fig. 8: Automated Readability Index

7 Conclusions

Stylometric and psycholinguistic features, such as those used in our paper, were hoped to introduce universal character to fake news recognition models, and to outperform traditional machine learning based on word occurrence vectors as features. This paper argues for the opposite: successes of machine learning, measured as high accuracy in recognizing fake news texts, are strongly linked and in fact constrained to types of texts on which the models have been trained. One may think about this problem as a form of over-fitting. Also the models are hardly usable for real-life fake news detection.

Therefore, our paper outlines an important future direction for studying fake news. Instead of preparing fake news data sets which consist of texts of similar structure and the same source researchers should attempt to compile more mixed corpora, gathering short and long texts on politics, economy and many other topics, from both social media and printed sources.

The design of machine learning models should take into account postulated corpora diversity. One of the observations made in this paper was over-fitting (domain dependency) of models based on gradient boosting (eg. XGBoost) and relatively better performance of Linear SVM.

In the future, we plan to experiment with training and testing corpora compiled from varied texts with the focus on machine learning methods that prevent over-fitting. We also intend to conduct similar studies using deep learning methods.

References

1. Anderson, J.: Lix and rix: Variations on a little-known readability index. Journal of Reading **26**(6), 490–496 (1983)
2. Bond, G.D., Lee, A.Y.: Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. Applied Cognitive Psychology **19**(3), 313–329 (2005)
3. Chall, J.S., Dale, E.: Readability revisited: The new Dale-Chall readability formula. Brookline Books (1995)
4. Coleman, M., Liau, T.L.: A computer readability formula designed for machine scoring. Journal of Applied Psychology **60**(2), 283 (1975)
5. Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H.E., Quattrociocchi, W.: The spreading of misinformation online. Proceedings of the National Academy of Sciences **113**(3), 554–559 (2016)
6. Flesch, R.: Flesch–kincaid readability test. Retrieved October **26**, 2007 (2007)
7. Gunning, R.: The fog index after twenty years. Journal of Business Communication **6**(2), 3–13 (1969)
8. Jin, Z., Cao, J., Zhang, Y., Zhou, J., Tian, Q.: Novel visual and statistical image features for microblogs news verification. IEEE transactions on multimedia **19**(3), 598–608 (2017)
9. Kelly, E.F., Stone, P.J.: Computer recognition of English word senses, vol. 13. North-Holland (1975)
10. Lasswell, H.D., Namenwirth, J.Z.: The lasswell value dictionary. New Haven (1969)

11. Mc Laughlin, G.H.: Smog grading-a new readability formula. Journal of reading **12**(8), 639–646 (1969)
12. Newman, M.L., Pennebaker, J.W., Berry, D.S., Richards, J.M.: Lying words: Predicting deception from linguistic styles. Personality and Social Psychology Bulletin **29**(5), 665–675 (2003)
13. Ott, M., Choi, Y., Cardie, C., Hancock, J.T.: 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. pp. 309–319. Association for Computational Linguistics (2011)
14. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic inquiry and word count: Liwc 2001. Mahway: Lawrence Erlbaum Associates **71**(2001), 2001 (2001)
15. Pérez-Rosas, V., Kleinberg, B., Lefevre, A., Mihalcea, R.: Automatic detection of fake news. arXiv preprint arXiv:1708.07104 (2017)
16. Pérez-Rosas, V., Kleinberg, B., Lefevre, A., Mihalcea, R.: Automatic detection of fake news. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 3391–3401. Association for Computational Linguistics (2018), http://aclweb.org/anthology/C18-1287
17. Semin, G.R., Fiedler, K.: The cognitive functions of linguistic categories in describing persons: Social cognition and language. Journal of personality and Social Psychology **54**(4), 558 (1988)
18. Senter, R., Smith, E.A.: Automated readability index. Tech. rep., CINCINNATI UNIV OH (1967)
19. Stone, P.J., Dunphy, D.C., Smith, M.S.: The general inquirer: A computer approach to content analysis. (1966)
20. Wang, W.Y.: "liar, liar pants on fire": A new benchmark dataset for fake news detection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 422–426. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/P17-2067, http://aclweb.org/anthology/P17-2067
21. Zhang, J., Cui, L., Fu, Y., Gouza, F.B.: Fake news detection with deep diffusive network model. arXiv preprint arXiv:1805.08751 (2018)
22. Zhao, Z., Zhao, J., Sano, Y., Levy, O., Takayasu, H., Takayasu, M., Li, D., Havlin, S.: Fake news propagate differently from real news even at early stages of spreading. arXiv preprint arXiv:1803.03443 (2018)