Augmenting Stylometric Features to Improve Detection of Propaganda and Manipulation

Radoslav Sabol and Aleš Horák

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00, Brno, Czech republic
xsabol@fi.muni.cz, hales@fi.muni.cz

Abstract. Identification of manipulative techniques in newspaper texts allows an informed reader to cope with the text content without being negatively influenced.

In this paper, we present new developments in using stylometry to support a deep learning neural network model in labelling newspaper articles for the presence of specific manipulative techniques. We also evaluate all stylometric features in 16 groups and improve the manipulation detection results in 15 of 17 techniques.

Keywords: propaganda detection, manipulative techniques, Propaganda dataset, stylometry

1 Introduction

Propaganda newspaper articles employ specific rhetorical figures to drive the readers opinion or to manipulate them, for example fabulation, labelling or demonization [10]. Detecting a presence of these devices in the text can be a strong indication that the text embodies malicious or ulterior motives.

In the previous work [12,11], a new deep learning approach that combines transformer-based large language model analysis with stylometric features has been introduced. The combination allowed to improve state-of-the-art results with the Propaganda benchmark dataset [1] for 14 of 17 manipulative techniques.

In the current paper, we present the new developments in the stylometric features set and evaluate feature group assets by a series of ablative sets. The final results reveal further enhancement of the results for 15 techniques.

1.1 Related Works

There has yet to be a genuine consensus within the scientific community on the optimal and universal set of stylometric features to be used in style analysis tasks. The choice usually depends on the currently solved task and applied

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2023, pp. 67–77, 2023. © Tribun EU 2023

classification algorithm [9]. For the best variety, numerical features that reflect the author's writing style are tailored from multiple levels of linguistic analysis.

Syntactic features attempt to exploit the sentence structure. Straightforward and common approaches make use of punctuation mark frequency, placement, and sentence lengths. More complex methods involve the extraction of information from the syntactic trees. Feng et al. [4] use two kinds of syntactic features for deception detection. Shallow syntactic features utilize the part of speech tags, while deep features encode the tree as a probabilistic context-free grammar. It was shown that the syntactic features do not outperform other feature types by themselves; they still carry viable information that can be utilized in conjunction with different feature types [7].

2 Stylometric Feature Set

The following section describes modifications to the previous stylometric feature set [12] to provide more detailed insight from various levels of linguistic analysis. Table 1 briefly explains the current state of implemented features, while other subsections will present the proposed changes in detail.

Table 1: Overview of the updated set of stylometric features. Features highlighted in bold are brand new additions to the old set. The feature highlighted in italic was significantly modified from the previous iteration.

Feature Type	# features	Language Independent
Word Length	137	✓
Sentence Length	177	✓
Word Repetition	140	
Word Class n-Grams	514	
Morphological Tags n-Grams	1,434	
Letter Casing	494	✓
Word Suffixes	425	✓
Word Richness	6	✓
Stopwords	600	✓
Punctuation	147	✓
Typography	111	✓
Character n-Gram Distribution	6,550	✓
Emoticons Presence	28	✓
Readability Metrics	4	
Structural Tree Characteristics	180	
Dependency relations n-Grams	3,208	
Total	14,155	

2.1 Syntactic Features

The following subsection covers new stylometric feature extractors that describe sentence structure from dependency trees. This information is subtly described using several existing features (sentence lengths, punctuation frequencies).

However, dependency trees allow for unique details that may improve the current feature set.

First, a dataset needs to be augmented by an additional support object. Before this work, the list of objects for each document was the following:

- `text`: the original plaintext document
- `lemmas`: a list of tokens and lemmatas
- `morphology`: morphological annotations from `majka` [14] and `desamb` [15]

To utilize the syntactic information, we create a new support object called `syntax`, which contains dependency trees for each document sentence. The trees are extracted using **UDPipe** [13], allowing a straightforward switch to other languages when necessary.

Structural Tree Characteristics The first group of features ignores all syntactic relations and observes only the structure of a tree. There are currently three feature extractors implemented:

1. **depth of the tree** (40 features)
   - for each tree, compute the longest path from the root node to any of the leaf nodes
2. **branching factors** (40 features)
   - for each non-leaf node of every tree, the number of children
3. **tree width** (100 features)
   - for each tree, compute the number of leaf nodes

The resulting vectors correspond to the relative frequency distributions of depths/widths/branching factors. To better convey the notion of adjacency between individual values, additional bins of size 2-3 are added to capture close values. The bins are illustrated in Figure 1.

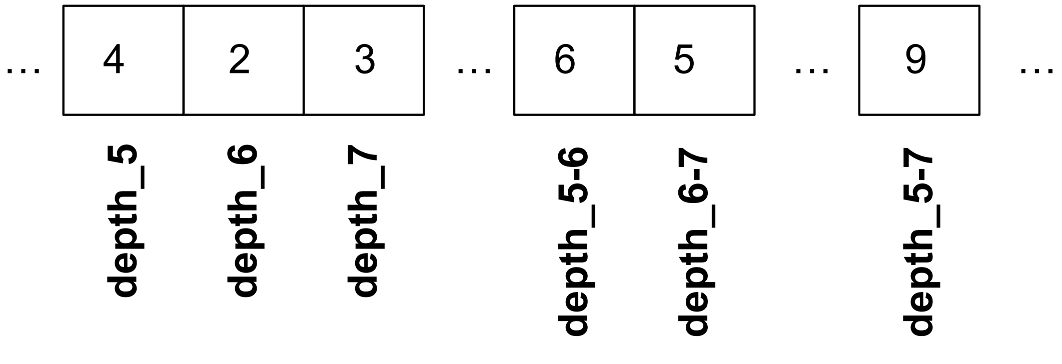


Fig. 1: The example of value binning for tree depth feature extractors. For illustrative purposes, the features are not normalized.

Relation-based Features The second set of features focuses on part-of-speech tags in the tree's nodes and the relations between them. We propose three types of characteristics, where each feature extractor reveals different information about the relationships in the dependency tree.

1. **Node *n*-grams** (735 features)
   - 2–4-grams, unigrams are essentially identical to word-class unigrams already present in the set
   - *n*-gram is constructed as an ascending path of node labels
2. **Relation *n*-grams** (1,415 features)
   - 1–4-grams using an ascending path of edge labels
3. **Complete *n*-grams** (1,058 features)
   - 2–4-grams, where *n*-gram is a path containing both node and edge labels (however, only node labels count towards an *n*-gram)

The training corpus calculates the list of allowed *n*-grams in advance. The preparation extracts all relevant *n*-grams from the documents and constructs the vector only using instances present in at least 1% of the documents. An example tree along with extracted *n*-gram is shown in Figure 2.

2.2 Readability Measures

The readability extractors present a group of four numerical features, where each element corresponds to a readability measure extracted from the input document. All of the metrics depend on the number of words, sentence lengths, and syllables, making them straightforward to adapt to other languages.

The **Flesch Reading Ease** [5] is considered to be one of the most commonly used and reliable readability metrics [8]. The values are scaled from 0 to 100, where higher values indicate that the text is easier to understand. The score is computed in the following way:

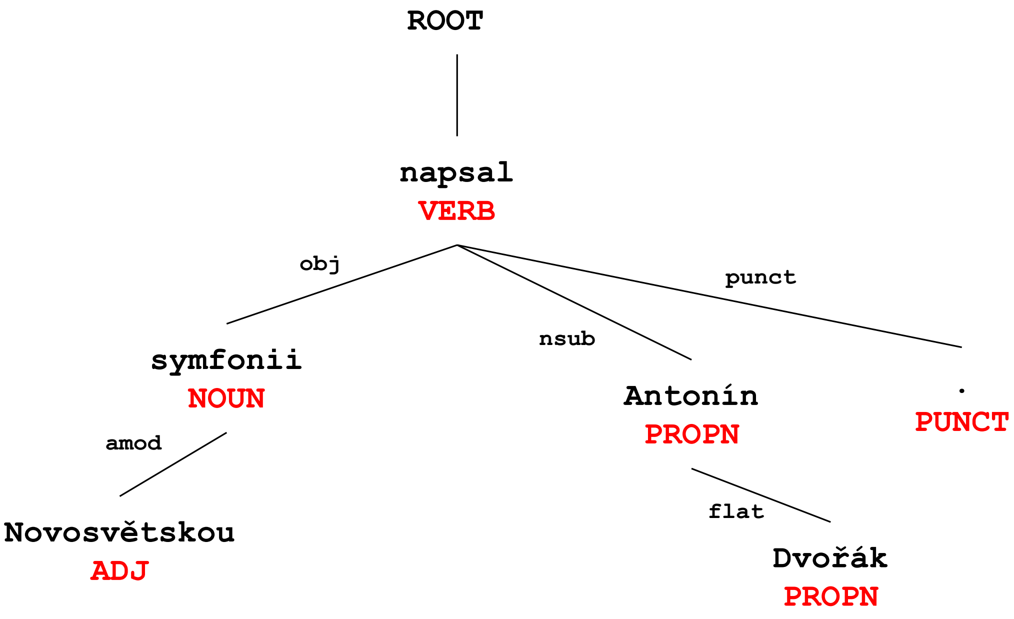
$$FLESCH = 206.835 - (1.015 * \overline{S}) - (84.6 * \overline{W})$$

Where $\overline{S}$ is the average number of syllables per word (or the total number of syllables divided by the number of words); similarly, $\overline{W}$ represents the average number of words per sentence. The final score is divided by 100 to make the domain consistent with other features.

The **Gunning Fog formula** is computed from a random sample of 100 sentences [2]. The resulting index approximates the years of formal education required to comprehend the text easily. The following formula computes it:

$$GUNNING\_FOG = 0.4 * \overline{S} * H$$

Where H stands for a percentage of complex words. We consider a word to be complex when its lemma is more than two syllables long. The final index is normalized according to the most difficult article within the training corpus.



Feature type	**Instances**
POS bigrams	[ADJ][NOUN], [NOUN][VERB], [VERB][ROOT], [PROPN][PROPN], [PROPN][VERB], [PUNCT][VERB]
Relation unigrams	(amod), (obj), (root), (nsub), (punct), (flat)
Relation bigrams	(amod)(obj), (obj)(root), (flat)(nsub), (nsub)(root), (punct)(root)
Complete bigrams	[ADJ](amod)[NOUN], [NOUN](obj)[VERB], [VERB](root)[ROOT], [PROPN](flat)[PROPN] [PROPN](nsub)[VERB], [PUNCT](punct)[VERB]

Fig. 2: An example dependency tree. The table below the tree lists the extracted *n*-grams from the sample tree.

The **McLaughlin's SMOG formula** is considered a more easily computed substitute for the Gunning Fog Index [6]. The interpretation of values remains the same, while the index is computed as:

$$SMOG = 3 + \sqrt{H}$$

Where H is the percentage of hard words in a random sample of 30 words.

Last but not least, the **FORCAST formula** uses an opposite approach where ″easy″ words are counted instead of the difficult ones [3]. It is computed as follows:

$$FORCAST = 20 - (E/10)$$

Where E is the number of single-syllable lemmas in a 150 word sample.

2.3 Limiting Morphological Tags

The original stylometric feature set included 10,000 features for various n-grams of morphological tags extracted from the training corpus. The feature amount was fixed to the most common morphological tags n-grams dependent only on n, and did not factor in the actual frequencies in the training corpus. This method led to the feature sets containing highly improbable n-grams where it is unclear whether the tags yield any significance or can be discarded as random noise. The noisiness can be observed from the previous works were the tags were frequently the least significant feature of the feature set via ablation tests [11].

The solution includes more strict limits for selecting morphological n-grams based on document frequencies. There is no strict limit on how many morphological n-grams need to be present in the feature set; however, it is required that the n-gram is present in at least 2% of the training documents. This method ensures that the feature vector will not contain improbable phenomena.

3 Experiments

The performed experiments focus on two aspects of the stylometric feature set: the importance of individual feature extractors for the current task and the overall performance of the modified stylometric feature set against the previous one. Both goals are benchmarked using the Propaganda dataset [1].

The Propaganda dataset includes 17 attributes, where 8 of them are manipulative techniques commonly used in misinformative news domains and thus are sensitive to style analysis. The remaining attributes focus on the properties or specific phenomena of the article, like genre, topic, or the writer's opinion of Russia. The dataset is split into train and test partitions, where the test partition contains a balanced sample of approximately 1000 documents as in [11].

The benchmarking uses gradient-boosted decision trees (GBDT) due to their reasonable performance and running times. Each experiment is repeated three times (indexed as i), with seeds fixed to $40 + i$ for results to be reproducible. The size of the ensemble is limited to 100 trees. The Weighted F1 is used as the performance metric to factor in the imbalance in the dataset.

3.1 Feature Selection

This experiment aims to measure the importance of individual feature extractors and select the most appropriate feature for each extractor. For this purpose, all feature extractors are grouped into 16 categories. First, the performance on the full feature set is measured as a base. Then, one of the categories is selected and removed from the complete feature set. Finally, a new model is trained on the reduced feature set, and a difference in weighted F1 from the complete feature set is measured. This process is applied to all feature categories. Similarly to the base experiments, the ablation tests are repeated three times on different seeds to estimate the difference in performance better.

For the fairness of comparison, a development set of approximately 20% of instances from the original training set is created. After the least significant features are determined, new results will be computed using the refined feature set with the feature of least significance removed.

4 Results and Discussion

In this section, the comparison of results on the benchmarking dataset Propaganda is performed. Detailed results of the ablation tests are discussed to better understand the interaction of stylometric features with the classes in the dataset.

Table 2: Summary of results for all attributes of the Propaganda dataset. The first and second row compares the old and new features sets. The third row describes the weighted F1 after the removal of the least significant feature group.

	argumentation	blaming	demonization	emotions	fabulation	fear-mongering	labelling	relativization	genre	location	sentiment	scope	topic	expert	opinion	Russia	source
Old Features	68.54	71.67	95.60	**80.69**	79.72	90.02	82.13	92.48	95.47	69.42	79.45	**86.65**	58.21	71.44	87.41	80.61	67.07
New Features	**69.04**	**71.72**	**95.75**	80.56	**80.21**	**90.73**	**82.78**	**92.56**	**95.71**	**70.47**	**81.19**	85.51	**59.37**	**73.75**	**87.74**	**81.11**	**67.54**

4.1 Comparison with the Previous Feature Set

A comparison of results between old and new feature sets can be seen in Table 2. The table also contains a third row with weighted F1 corresponding to the model trained on new features with the least important feature category removed.

Overall, the weighted F1 was improved by the refined features for almost every attribute of the Propaganda dataset. The only exceptions are emotions, where the new features perform worse by 0.13%, which is a difference that can be attributed to random error in measurement. A much more significant loss of 1.14% can be observed with the *scope* attribute. Even when removing the least significant feature (word class *n*-grams), the performance cannot be matched with the old feature set. In this instance, changing the morphological features could harm this attribute's performance.

The greatest improvement was achieved for the *expert* attribute (2.33%) that could be explained by proper noun relations that were extracted as part of dependency tree analysis. A similar argument could be used to describe the enhancement for the *labelling* attribute, as propaganda labels usually consist of literal or metaphorical comparisons that can be extracted using relations in the dependency trees.

4.2 Feature Selection Results

The heatmap in Figure 3 shows the results for the performed ablation tests. Brighter colors indicate that the selected feature subset has higher importance, while the dark ones suggest that the feature is either not essential or even performance-degrading.

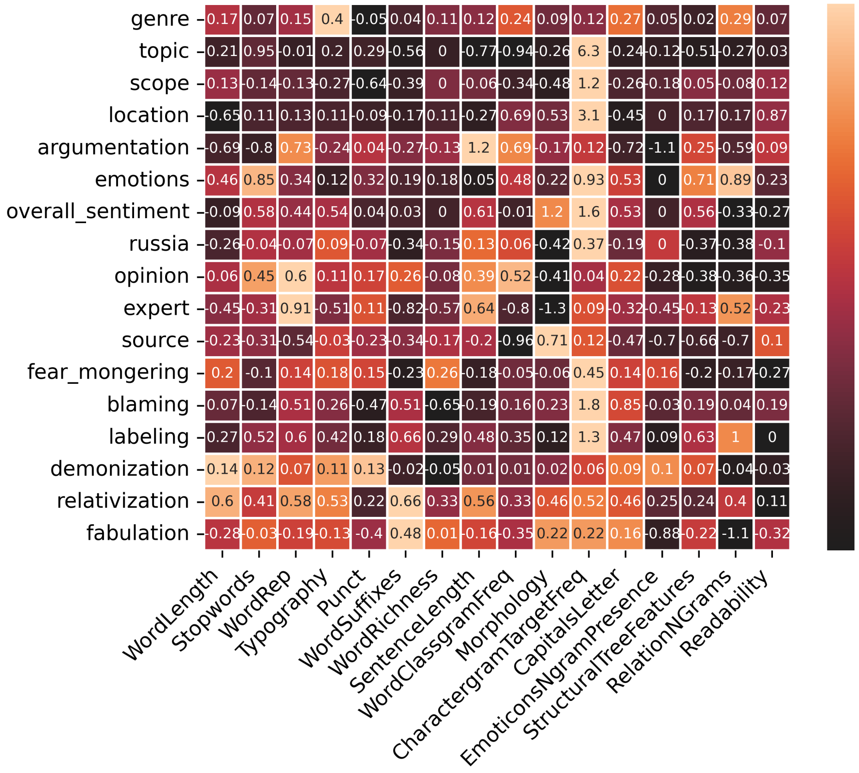


Fig. 3: Heatmap of feature group importances. The color map is normalized in a row-wise fashion to highlight notable importances.

The highest observed feature importances are 6.3% and 3.1% for Character *n*-Grams in *location* and *topic* attributes. These two attributes are heavily tied with the semantics of the text, and sufficiently large character *n*-grams (in this case, 5-grams) can capture keywords that are tied with these attributes. As *location* and *topic* should not be affected by the writing style, having semantic cues captured within character *n*-grams is vital for improving accuracy for such attributes.

Another notable importance is 1.2% for sentence lengths in *argumentation* detection. A possible explanation is that when the author uses more complex reasoning, his sentences are usually longer as they contain sub-sentences which logically follow-up the argument.

The new features presented in this paper positively contribute to *labelling* detection (where relation *n*-grams are the second most important features), *expert*, and *emotions*.

To finish off the experiment, we remove the least important feature for each attribute, and evaluate a new model. The results can be observed on Table 3.

Table 3: Re-evaluation of GBDT with the least significant feature removed. **Importance** column refers to the difference in weighted F1 as shown on the heatmap. **F1** refers to the Weighted F1 performance metric of the new model with **bold** values referring to the best numbers for each attribute. **Diff** is a difference against the model trained on a complete feature set.

Attribute	**Removed Feature**	**Importance**	**F1 (%)**	**Diff**
Argumentation	Emojis	-1.1	**69.34**	0.30
Blaming	Word Richness	-0.65	71.33	-0.39
Demonization	Word Richness	-0.05	**95.75**	0.00
Emotions	Emojis	0	80.40	-0.16
Fabulation	Relation *n*-grams	-1.1	**80.52**	0.31
Fear-mongering	Readability	-0.27	90.63	-0.10
Labeling	Readability	0	82.09	-0.69
Relativization	Readability	0.11	**92.64**	0.08
Genre	Punctuation	-0.05	95.52	-0.19
Location	Word Lengths	-0.65	69.75	-0.72
Sentiment	Relation *n*-grams	-0.33	80.60	-0.59
Scope	Punctuation	-0.64	**86.04**	0.53
Topic	Word Classes	-0.94	**60.05**	0.68
Expert	Morphology	-1.3	**73.83**	0.08
Opinion	Morphology	-0.41	**87.95**	0.21
Russia	Morphology	-0.42	80.87	-0.24
Source	Word Classes	-0.96	67.07	-0.47

If the importance is low enough, it is possible to further improve the performance by removing the feature extractor. The improvement holds mainly for the *Topic* and *Scope* attributes. However, this is not guaranteed, as the features removed for the *Source* attribute also have a low importance, but the performance degrades significantly.

There is no significant correlation between the feature importance and the difference in performance. There may be more variables involved, like the actual number of removed numerical features and the choice of the learning algorithm.

5 Conclusion and Future Work

We have presented an extension of stylometry features used, in conjunction with large language models, for identification of 17 different manipulative techniques and propaganda reflection techniques as employed in newspaper texts. By adding syntactic features, readability metrics and by adjusting the previous morphological features, the manipulation detection models are improved with 15 of the 17 text attributes.

In the future work, we plan to accomplish further steps in adjusting the detailed best sets of features for each attribute and to tune the feature weights by comparing their values in propagandistic and standard newspaper texts.

Acknowledgements. This work has been partly supported by the Ministry of Education of CR within the LINDAT-CLARIAH-CZ project LM2023062. Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

References

1. Baisa, V., Herman, O., Horak, A.: Benchmark dataset for propaganda detection in Czech newspaper texts. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). pp. 77–83. INCOMA Ltd., Varna, Bulgaria (Sep 2019). https://doi.org/10.26615/978-954-452-056-4_010, https://aclanthology.org/R19-1010
2. Bogert, J.: In defense of the fog index. The Bulletin of the Association for Business Communication **48**(2), 9–12 (1985). https://doi.org/10.1177/108056998504800203, https://doi.org/10.1177/108056998504800203
3. Caylor, J.S., Others: Methodologies for determining reading requirements of military occupational specialties (1973)
4. Feng, S., Banerjee, R., Choi, Y.: Syntactic stylometry for deception detection. In: Li, H., Lin, C.Y., Osborne, M., Lee, G.G., Park, J.C. (eds.) Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 171–175. Association for Computational Linguistics, Jeju Island, Korea (Jul 2012), https://aclanthology.org/P12-2034
5. Flesch, R.: A new readability yardstick. Journal of applied psychology **32**(3), 221 (1948)
6. Harry, G., Laughlin, M.: SMOG Grading – a new readability formula. The Journal of Reading (1969), https://api.semanticscholar.org/CorpusID:9571753
7. Hollingsworth, C.: Syntactic stylometry: Using sentence structure for authorship attribution (2012), https://api.semanticscholar.org/CorpusID:15376304
8. Klare, G.R.: The measurement of readability: useful information for communicators. ACM Journal of Computer Documentation (JCD) **24**(3), 107–121 (2000)
9. Lagutina, K., Lagutina, N., Boychuk, E., Vorontsova, I., Shliakhtina, E., Belyaeva, O., Paramonov, I., Demidov, P.: A survey on stylometric text features. In: 2019 25th Conference of Open Innovations Association (FRUCT). pp. 184–195 (2019). https://doi.org/10.23919/FRUCT48121.2019.8981504

10. Miles, C.: Rhetorical Methods and Metaphor in Viral Propaganda. In: Bains, P., O'Shaughnessy, N., Snow, N. (eds.) The SAGE Handbook of Propaganda, pp. 155–170. SAGE Publications (2019)
11. Sabol, R.: Propaganda Detection using Stylometric Text Analysis. Master thesis, Masaryk University, Faculty of Informatics, Brno (2023), `https://is.muni.cz/th/b83ai/`
12. Sabol, R., Horák, A.: Manipulative Style Recognition of Czech News Texts using Stylometric Text Analysis. In: Recent Advances in Slavonic Natural Language Processing, RASLAN 2022. pp. 191–199 (2022)
13. Straka, M., Straková, J.: Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. pp. 88–99. Association for Computational Linguistics, Vancouver, Canada (August 2017), `http://www.aclweb.org/anthology/K/K17/K17-3009.pdf`
14. Šmerk, P.: Fast Morphological Analysis of Czech. In: Sojka, P., Horák, A. (eds.) Third Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2009. pp. 13–16. Masaryk University (2009)
15. Šmerk, P.: K počítačové morfologické analýze češtiny (in Czech, Towards Computational Morphological Analysis of Czech). Ph.D. thesis, Faculty of Informatics, Masaryk University (2010)