Team Ned Leeds at SemEval-2019 Task 4: Exploring Language Indicators of Hyperpartisan Reporting

Bozhidar Stevanoski, Sonja Gievska
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Rugjer Boshkovikj 16, Skopje, Republic of North Macedonia
bozidar.stevanoski@students.finki.ukim.mk
sonja.gievska@finki.ukim.mk

Abstract

This paper reports an experiment carried out to investigate the relevance of several syntactic, stylistic and pragmatic features on the task of distinguishing between mainstream and partisan news articles. The results of the evaluation of different feature sets and the extent to which various feature categories could affect the performance metrics are discussed and compared. Among different combinations of features and classifiers, Random Forest classifier using vector representations of the headline and the text of the report, with the inclusion of 8 readability scores and few stylistic features yielded best result, ranking our team at the 9^{th} place at the SemEval 2019 Hyperpartisan News Detection challenge.

1 Introduction

Current influential technological megatrends, such as, smart phones and social networking often come with some unwanted side effects - prolific spread of false, biased and misleading information, for instance. In the past few years, the potential threats and consequences of disseminating fake news has reinforced the discussion on the responsibility of social media and governments to tackle the issue, sooner rather than later. Enabling users to report on and be informed of untruthful, deceitful and fraudulent content and sources is expected to become a type of guiding principle for those involved in publishing and disseminating content online.

There is a blurred line between deceptive writing and hyperpartisan reporting, producing extremely biased articles in favor of one political party, cause or individual, while preserving the format and appearance of professional articles. Adherence to the ethics and rules of objective reporting is frequently debatable when it comes to political analysis in media articles. While certain truthful facts are present, they are carefully entangled in a narrative package with biased views, populistic messages and divisive topics, using language that polarizes and flares emotions. Rather than labeling and grading news articles on the truth continuum, researchers usually opt for identification of the phenomenological and contextual features of distinguishing hyperpartisanship in online news articles.

People use diverse set of cues extrapolated from published text and external knowledge and sources, when verifying the veracity of information imparted by others. A large body of evidence documents the impact of deception has on language choices people make. A notable body of work exists revealing insights into the language of deceit in interrogation context (Porter and Yuille, 1996), court hearings (Coulthard et al., 2016), or personal relationships (Miller et al., 1986). Empirical studies still remain the primary manner in which manifestation of deceptive human behavior online is studied. Analysis of political language (Rashkin et al., 2017), partisan media (Gervais, 2014), and news publishing (Rubin et al., 2015) were also guided broadly by the questions pertaining to detecting deception in written language.

In what follows, we highlight the primary findings of our empirical research in identifying tangible verbal indicators as they relate to our central commitment of detecting deception in text. In this paper we examine the impact of grammar and psycho-linguistic word categories, syntactic word connotations and text complexity metrics on the task of distinguishing hyperpartisanship in real news articles.

2 Related Work

Given how prolific fake content has become, the phenomenon has challenged the interdisciplinary

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 1026–1031
Minneapolis, Minnesota, USA, June 6–7, 2019.

research community and has been the focus of notable research studies, especially in the field of natural language processing (NLP) and social network analysis.

While it is beyond the scope of this paper to exhaust a review on the topic, of particular relevance to the authors of this paper are the works in monitoring and detecting what is considered untruthful and deceitful content. The differences in the type of conveyed text and the underlying context are likely to afford contrasting models of deception i.e., combination of linguistic features and selection of classification algorithm they rely upon.

It is interesting to note that rather simple linguistic analysis could be successful on a number of NLP tasks relating to detection of deceptive text, such as fake news, opinions, trolling, hate and abusive language, including hyperpartisan reporting. This indicates that it is not semantics, but rather syntactic and pragmatics of the language style of the author that give clues of the underlying cognitive states relating to deception.

Low-level linguistic features such us word counts and frequencies (Horne and Adali, 2017), language modeling (Conroy et al., 2015; Potthast et al., 2018; Pérez-Rosas et al., 2018), part-of-speech tags (POS) (Lim et al., 2018; Conroy et al., 2015), Probabilistic Context Free Grammar (Feng et al., 2012), readability scores (Potthast et al., 2018), and their combinations have proved to be successful with varying performance and generalization power, especially for testing on cross-domain datasets. The research study most closely related to ours, proposes a model for hyperpartisan classification that yielded accuracy of 0.75 (Potthast et al., 2018), which will be used as a baseline accuracy against which our model will be compared.

The use of deep learning architectures (Wang, 2017) have complemented the list of traditional machine learning algorithms (ML), such as SVM (Yang et al., 2017; Lim et al., 2018), logistic regression, discriminant analysis, decision trees (Potthast et al., 2018) and neural networks (Vuković et al., 2009), used in the field of deceptive detection. An unavoidable discussion on the trade-offs between generality and specificity of the models has never ceased to flavor the interpretation of results and point out directions for future improvements.

3 Dataset

Two datasets of news articles were available for the SemEval 2019 Task 4: "Hyperpartisan news detection" (Kiesel et al., 2019), one labeled *"by-article"* by professional journalists, and the other labeled *"by-publisher"*.

Our empirical study was focused on the former one, whose training dataset consists of 645 articles. The testing dataset, which is not publicly released, are made available via TIRA (Potthast et al., 2019), and it contains 628 *by-article* articles. It is balanced and consists of articles from previously unseen publishers in the training sets. For evaluation purposes, we randomly choose 80% of the *by-article* data for training, and the remaining 20% for validation.

4 Our Methodology

In this paper, we further enhance the feature set explored by related research, and explore few features that appeared to be promising to capture syntactic and pragmatic aspect of hyperpartisan reporting.

Word vector representations: Though previous research studies on this topic use language modelling i.e., frequencies of n-grams in an article to unmask the style of hyperpartisan reporting, our view is that it is distributed word vector representations might augment the model in capturing the style of deceptive and biased political reporting.

Word2Vec has been emphasized as providing better performance, generalizability and transfer of knowledge on a number of related NLP problems. In consequence, word2vec, pre-trained on part of the Google News dataset consisting of cca 100 billion words (Mikolov et al., 2013) was utilized in our model.

Indication of hyperpartisan language and style could be found in various parts of a journal article - article headline and individual sentences could be indicative of biased and partisan language. Transmission of context, set by a sentence that is entailed in the consecutive sentences in a document, is the core idea underlying the proposed word vector representations on two different levels, one on a sentence level and another on a document/article level. Consequently, three word embeddings representing the headline, the sentences and the entire document text were concatenated creating the final word2vec vector.

For the word and sentence tokenization we use the Natural Language Toolkit (NLTK)[1].

Readability scores: Readability scores measure the ease of comprehension of a particular style of writing based on metrics such as, word and sentence length and various weighting factors and ratios, making them closely related to the quantitative aspect of text complexity. In accordance with successful practices reported in previous research in text deception detection (Pérez-Rosas et al., 2018; Yang et al., 2017), we use eight such scores[2], namely Flesch Reading Ease (Flesch, 1948), Flesch Kincaid Grade Level (Kincaid et al., 1975), Coleman Liau Index (Coleman and Liau, 1975), Gunning Fog Index (Gunning, 1952), SMOG Index (Harry and Laughlin, 1969), ARI Index (Senter and Smith, 1967), LIX Index (Björnsson, 1968) and Dale-Chall Score (Chall and Dale, 1995).

General stylistic measures: We also employ elementary measures - number of characters, total words, different words, sentences, syllables, polysyllable words, difficult words (as defined by (Dale and Chall, 1948)), and words longer than 4, 6, 10 and 13 characters.

Psycho-linguistic features: Motivated by previous studies in the field of deceptive text analysis, including fake news examination (Cunha et al., 2018), exploring fraudulent hotel reviews (Fast et al., 2016), characterizing and detecting hateful Twitter users (Ribeiro et al., 2018), we explore the effect of all 194 types of features from the Empath (Fast et al., 2016) lexicon on the task of hyperpartisan news detection.

Part-of-speech tagging: The frequencies of part-of-speech (POS) categories of the words in text, in particular frequencies of nouns, proper singular nouns, personal and possesive pronouns, wh-pronouns, determiners, wh-determiners, cardinal digits, particles, interjections, adjectives, verbs in base form, past tense, gerund, past participle, 3rd and non-3rd person singular present, were added to our model.

Augmented stylistic feature set: Instead of eliminating stop-words, we take the number of their occurrences as a feature. We use the corpus made available by NLTK. Frequencies of interrogative (how, when, what, why) and all-caps words, negations (not, never, no) and punctuation marks are as stylistic features. The stylistic features were normalized by article length.

Bag-of-words of hyperlinks: The links in each article are abbreviated to their base URL form, using Python's Urllib[3], and further transformed into a bag-of-words (BoW) representation. Both internal (anchor links) and external links in respect to the articles, are taken into account for the BoW representation.

5 Results and Discussion

The relationship between various predictive models and evaluation metrics has always been a topic of interest in machine learning and NLP, and this section describes the performance of the feature sets we have experimented with. It is important to note that since the features we test can take values from different ranges, we perform min-max normalization on all of them to bring them in the $[0, 1]$ interval.

We have experimented with various classifiers, such as Logistic Regression, Multilayer Perceptron and Extra Trees, although the most successful one was Random Forest (RF) classifier with 100 trees, which is in line with the findings of the baseline model (Potthast et al., 2018). We use the Python implementation of the classifiers from the Scikit-learn library. [4]

While aiming for achieving high accuracy, avoiding overfitting was also an objective to ensure the model is robust enough to handle previously unseen data. The evaluation results obtained by the models on *by-article* validation and test datasets are presented in Table 1. A short description of the evaluated models follows:

- **Model 1** - A model that incorporates three concatenated word representation vectors, eight readability scores and the general stylistic features
- **Model 2** - The set of features of Model 1 augmented with psycho-linguistic features
- **Model 3** - Frequencies of the POS tags and additional stylistic features were added to the set of features included in Model 2
- **Model 4** - An extension of Model 1 feature set that included hyperlink features

[1]https://www.nltk.org/ Last accessed: 23 February 2019.

[2]https://pypi.org/project/ReadabilityCalculator. Last Accessed: 20 February 2019

[3]https://docs.python.org/3/library/urllib.html Last Accessed: 23 February 2019.

[4]https://scikit-learn.org/ Last accessed: 23 February 2019.

Models	*By-article* test dataset				*By-article* validation dataset			
	accuracy	*precision*	*recall*	*F1 score*	*accuracy*	*precision*	*recall*	*F1 score*
1	0.775	0.865	0.653	0.744	0.837	0.857	0.652	0.741
2	0.769	0.860	0.643	0.736	0.798	0.833	0.543	0.658
3	0.760	0.844	0.637	0.726	0.814	0.844	0.587	0.692
4	0.763	0.851	0.637	0.729	0.837	0.903	0.609	0.727
5	0.710	0.784	0.580	0.667	0.806	0.784	0.630	0.699

Table 1: Performance comparison of models trained on the *by-article* dataset.

- **Model 5** - Principal Component Analysis (PCA) was used to reduce dimensionality of Model 3 to 50 features

The model i.e., feature set that exhibits the best performance is Model 1, that outperforms the other models on the validation as well as on the *by-article* test dataset, but also outperforms the baseline accuracy results presented in (Potthast et al., 2018) by 2.5%.

The attempts to improve the performance on the same dataset by augmenting the set of features with new ones were dissatisfactory and did not lead to any performance advantage. Augmenting the feature set with psycho-linguistic features or POS tags in Model 2 and Model 3 respectively, failed to gain any performance advantage compared to Model 1 Model 4 yielded the worst results. Reducing the dimensionality of the feature space of Model 3 to a 50-dimensional one by using PCA in Model 5, led to even greater degradation of performance metrics. When testing the predicting power of the hyperlink features independently from all other features, the results were significantly better than chance.

The weakness of the models can be explained by the difficulty in defining general heuristics with which to detect biased and deceptive reports. Much of this research represents an effort to understand the clues which give insight into the underlying conditions pertaining to such reporting in news articles. Close inspection of data and comparative analysis with the models participating on the same SemEval task could better support the interpretation of our results. In addition, not having information on the cases that were misclassified by our models, makes it difficult to speculate and offer solutions for proper treatment and improvement of the limitations of our model.

6 Conclusion

In this paper, we report on an experiment that examines the predictive effect of the different feature sets on automatic detection of hyperpartisan articles. Results implicate that the features examined in this research, to varying degree, capture the syntactic and pragmatic aspects of hyperpartisan style, and generalize well to a set of previously unseen articles by unseen publishers. The findings provide evidence of strong modeling capability of word vector embeddings combined with text complexity metrics of the reports and psycholinguistic features, demonstrating that the model accuracy rivals the performance of other teams participating in the SemEval 2019 hyperpartisan challenge, positioning our team at the 9^{th} place on the task's leaderboard.

References

Carl-Hugo Björnsson. 1968. *Läsbarhet: hur skall man som författare nå fram till läsarna?* Bokförlaget Liber.

Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Niall J Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, page 82. American Society for Information Science.

Malcolm Coulthard, Alison Johnson, and David Wright. 2016. *An introduction to forensic linguistics: Language in evidence*. Routledge.

Evandro Cunha, Gabriel Magno, Josemar Caetano, Douglas Teixeira, and Virgilio Almeida. 2018. Fake news as we feel it: perception and conceptualization

of the term fake news in the media. In *International Conference on Social Informatics*, pages 151–166. Springer.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4647–4657. ACM.

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 171–175. Association for Computational Linguistics.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Bryan T Gervais. 2014. Following the news? reception of uncivil partisan media and the use of incivility in political expression. *Political Communication*, 31(4):564–583.

Robert Gunning. 1952. *The technique of clear writing*. McGraw-Hill, New York.

McLaughlin G Harry and M Laughlin. 1969. Smog gradinga new readability formula. *Journal of Reading*, 12(8):639–646.

Benjamin D Horne and Sibel Adali. 2017. This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Eleventh International AAAI Conference on Web and Social Media*.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 Task 4: Hyperpartisan News Detection. In *Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval 2019)*. Association for Computational Linguistics.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Institute for Simulation and Training, University of Central Florida.

Sora Lim, Adam Jatowt, and Masatoshi Yoshikawa. 2018. Understanding characteristics of biased sentences in news. *INRA 2018, October 2018, Turin, Italy*.

Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013.

Gerald R Miller, Paul A Mongeau, and Carra Sleight. 1986. Invited article fudging with friends and lying to lovers: Deceptive communication in personal relationships. *Journal of Social and Personal Relationships*, 3(4):495–512.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. *Proceedings of the 27th International Conference on Computational Linguistics, pages 33913401 Santa Fe, New Mexico, USA, August 20-26, 2018*.

Stephen Porter and John C Yuille. 1996. The language of deceit: An investigation of the verbal clues to deception in the interrogation context. *Law and Human Behavior*, 20(4):443–458.

Martin Potthast, Tim Gollub, Matti Wiegmann, and Benno Stein. 2019. TIRA Integrated Research Architecture. In Nicola Ferro and Carol Peters, editors, *Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF*. Springer.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A Stylometric Inquiry into Hyperpartisan and Fake News. In *56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 231–240. Association for Computational Linguistics.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937. Association for Computational Linguistics.

Manoel Horta Ribeiro, Pedro H Calais, Yuri A Santos, Virgílio AF Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on twitter. In *Twelfth International AAAI Conference on Web and Social Media*.

Victoria L Rubin, Niall J Conroy, and Yimin Chen. 2015. Towards news verification: Deception detection methods for news discourse. In *Hawaii International Conference on System Sciences*.

RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report, CINCINNATI UNIV OH.

Marin Vuković, Krešimir Pripužić, and Hrvoje Belani. 2009. An intelligent automatic hoax detection system. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 318–325. Springer.

William Yang Wang. 2017. "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)*, pages 422–426. Association for Computational Linguistics.

Fan Yang, Arjun Mukherjee, and Eduard Dragut. 2017. Satirical news detection and analysis using attention mechanism and linguistic features. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1979–1989. Association for Computational Linguistics.