Fermi at SemEval-2019 Task 4: The sarah-jane-smith Hyperpartisan News Detector

Nikhil Chakravartula[1,3] Vijayasaradhi Indurthi[1,2], Bakhtiyar Syed[2],
[1] Teradata, [2] IIIT Hyderabad
[1]{nikhil.chakravartula,vijayasaradhi.indurthi}@teradata.com
[2]{vijaya.saradhi, syed.b}@research.iiit.ac.in
[3]{nikhil.chakravartula}@gmail.com

Abstract

This paper describes our system (Fermi) for Task 4: Hyper-partisan News detection of SemEval-2019. We use simple text classification algorithms by transforming the input features to a reduced feature set. We aim to find the right number of features useful for efficient classification and explore multiple training models to evaluate the performance of these text classification algorithms. Our team - **Fermi**'s model achieved an accuracy of 59.10% and an F1 score of 69.5% on the official test data set.

In this paper, we provide a detailed description of the approach as well as the results obtained in the task.

1 Introduction

Hyper-partisan refers to a person or a group's tendency to be extremely partisan or biased towards a person or a group and specifically towards a political person or a political party. With the tremendous increase in citizen-based journalism, where anyone can create a website and post his (biased) views, there is a new phenomenon called fake news and it's potential role in affecting the election results, and has the ability to modify and impact the public's perception towards various people, companies and political parties. These kind of 'news' are usually one-sided, inflammatory, emotional and mostly woven around untruths. Combined with the proliferation of social media platforms, these 'fake news' signals get amplified and may potentially mask the signal of the real news. The fake news phenomenon hype has caused irreparable loss to many politicians, companies and in some cases involved the death of fellow citizens.

While Social media platforms can be used for constructive ideas, a small group of people can propagate their notions including hatred or affinity towards or against an individual, or a group or a race to the entire world in a few seconds. This necessitates the need to come up with computational methods to identify hyper-partisan news in user generated content.

Using computational methods to identify hyper-partisan news has been gaining attention in recent years as evidenced in (Potthast et al., 2018).

2 Related Work

In this section, we briefly describe other work in this area.

Hyper-partisan news detection is a new area and to the best of the knowledge of the authors, not much work has been done in this area. However, a close and related task is that of fake news detection. (Pérez-Rosas et al., 2017) use linguistic features to distinguish between fake and legitimate news content. (Wang, 2017) collect a decade long manually labelled sor statements in various context from a political fact checking website and create fake news classifiers using surface level linguistic patterns. (Tschiatschek et al., 2018) leverage crowd signals for detecting fake news. (Long et al., 2017) tackles the problem of fake news through multi-perspective speaker profiles.

Papers published in the last two years include the surveys by (Zhou and Zafarani, 2018), (Zhou et al., 2019) and (Shu et al., 2017), the paper by (Kumar and Shah, 2018).

A shared task on Hyper-partisan News detection(Kiesel et al., 2019) was announced as part of the annual workshop SemEval 2019. The task was to find if the given news article text and classify if it follows a hyper-partisan argumentation, i.e., whether it exhibits blind, prejudiced, or unreasoning allegiance to one party, faction, cause, or person.

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 954–956
Minneapolis, Minnesota, USA, June 6–7, 2019.

3 Methodology and Data

The data collection methods used to compile the data set in Hyperpartisan news detection is described in (Kiesel et al., 2019). We tackle the problem of identifying a piece of news as hyperpartisan or not by formulating it as a text classification problem. We use bag of words representation to transform the individual documents into vectors. After the transformation, we reduce the number of dimensions by using chi-square feature selection technique. In this method, the chi-square statistics between every feature variable and the target variable are computed, and then the existence of a relationship between the variables and the target is calculated. If the target variable is independent of the feature variable, that feature variable is not useful for prediction. If the two are dependent, then that feature variable is very important. In text classification, the feature selection is the process of selecting a specific subset of the terms of the training set and using only them in the classification algorithm. The feature selection process takes place before the training of the classifier. We use Random Forest Classifier from scikit-learn[1] machine learning library to generate models on these reduced features. The number of estimators in all the experiments is 20. All other parameters are default.

Our results on the different number of important features have been mentioned and described in the results section.

We haven't used any external datasets to augment the data for training our models.

No of features	F1 (macro)	Accuracy
200	0.39	0.51
400	0.40	0.51
600	0.42	0.51
800	0.44	0.52
1000	**0.46**	**0.52**

Table 1: *Dev* set Accuracy and Macro-F1 scores on labels by publisher dataset.

Dataset	F1 (macro)	Accuracy
Labels-by-article	0.69	0.59
Labels-by-publisher	0.66	0.61

Table 2: *Test set* Accuracy and Macro-F1 scores.

[1] https://scikit-learn.org/

4 Results and Analysis

Table 2 shows the dev set macro-averaged F-1 and accuracy for different number of important features.

We notice that the best performance was bagged by the model which uses 1000 features with Random Forest. We submitted this best model for evaluation on the test data and Table 4 shows the results.

The potential applications of this work show how different number of important features affect the performance of the classification task.

5 Future Work

Due to some constraints on the TIRA[2] platform, we were unable to use state-of-the-art deep learning techniques for text classification, which gained immense popularity in the past few years. In the future, we would like to explore transfer learning and deep learning algorithms to create models for and evaluate their performance for this task.

References

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.

Srijan Kumar and Neil Shah. 2018. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559*.

Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. 2017. Fake news detection through multi-perspective speaker profiles. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 252–256.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*.

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.

[2] https://tira.io

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.

Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. 2018. Fake news detection in social networks via crowd signals. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pages 517–524. International World Wide Web Conferences Steering Committee.

William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Xinyi Zhou and Reza Zafarani. 2018. Fake news: A survey of research, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315*.

Xinyi Zhou, Reza Zafarani, Kai Shu, and Huan Liu. 2019. Fake news: Fundamental theories, detection strategies and challenges. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 836–837. ACM.