Using Scattertext for visually analysing Hyperpartisan News Detection data.

Mikel Salvoch Vilches and **Julen Rodríguez Meneses**
University of the Basque Country
Donostia, Gipuzkoa

1 Introduction

In the following document we detail the process and the results obtained from applying Scattertext (Kessler, 2017) to a dataset that we will detail later on in section 3. This is developed in the context of the final assignment of the course Natural Language Processing.

The paper cited above in which Scattertext is presented defines the tool as an open source tool for visualizing linguistic variation between document categories in a language-independent way.

The main objective of the project will be to perform a qualitative analysis of the obtained results. The application of Scattertext on our dataset will be done following the tutorial proposed in its documentation. In this sense, and once the process has been applied, we will obtain an interactive HTML document in which we will be able to visualize the obtained results.

2 Related Work

In terms of related work, it does not make sense to evaluate the existing use of this tool in any particular dataset. Therefore, in this section we focus on similar tools that could be equally applied to achieve the stated objective in section 1.

In the original document presenting the tool, previous text visualization methods are compared, like scatterplots and word clouds, which struggle with readability and word comparison. Scatterplots (Monroe et al., 2017) suffer from overlapping labels. On the other hand, word clouds (Schwartz et al., 2013) distort word importance. This merely scale words by frequency, failing to capture patterns that reveal deeper linguistic information.

Scattertext, which can be seen as a natural evolution of scatterplots, show how words relate to each other in the text, helping to uncover hidden connections and details. It optimizes word placement and improves interactivity visualizations (Bostock et al., 2018), but require manual selection.

3 Data

The dataset we are going to use is the Hyperpartisan News Dataset from SemEval-2019 (Kiesel et al., 2019).

Hyperpartisan news is refered to news which take extremist political positions, both to the right and the left.

The dataset is composed of 1273 news manually tagged by 3 annotators. Additionally, it has 754000 semi-automatically tagged news items but this last block is not of interest to us.

Of the 1273 manually labeled articles, we have available $\sim 50\%$ of the data (645 articles) while the remaining (628) were reserved. It should be remembered that the purpose of publishing this dataset was the construction of a classifier, thus having data not seen by the participants' models was required. As a curiosity, the best model achieved an accuracy higher than 82%. Also, the authors highlight the success of participation.

The annotation process consisted of asking each annotator to determine, for each new, a number on a scale of 1 to 5 to determine whether the new was hyperpartisan content or not. In particular, the proposed scale was as follows:

1. No hyperpartisan content
2. Mostly unbiased, non-hyperpartisan content
3. Not sure
4. Fair amount of hyperpartisan content
5. Extreme hyperpartisan content

Finally, the authors removed the news for which there was not much consensus and for which the result was the third option ("Not sure"). For the remaining data, they classified in a binary way between hyperpartisan for scores above 4 and not

hyperpartisan for scores below 2. For further detail, check the original paper which describes the dataset in a very precise way.

4 Results

During the experimentation, we have generated several graphs to determine which ones are of interest to us and allow us to obtain useful graphical information about our dataset. Scattertext allows us to create many types of graphs as it is shown in the proposed tutorial.

In our case, we have decided to keep 4 plots. For some plots, we had to discard them because of the slow loading of the HTML document (mainly because of the high volume of data).

Next, we proceed to show a static image of each of the plots. It should be remembered that one of the main advantages of the tool we are using is the generation of interactive graphics. Therefore, in addition, we attach to each of the plots a link to the HTML document hosted on GitHub and visible directly from the browser by clicking on the link.

Result analysis will be carried out in section 5.

4.1 Word clouds

Just as a previous, we can see how word clouds[1] used to be.

Figure 1: Word clouds for each class

As it is really difficult to differentiate which term is from which class, we needed to create two separate plots for each one.

4.2 Visualizing term associations

As a first scattertext plot we are going to show terms *per se*. In particular, we show the terms according to the degree of association they have with Hyperpartisan News.

Remember that for all the interactive graphs shown you can click on a point/term and observe the frequency for each class, this is shown in fig. 3.

[1]Created with the package managed by (Müller, 2016)

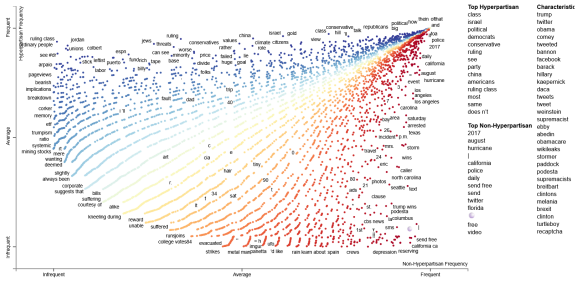

Figure 2: Full interactive plot available here (it may take a couple of minutes to load).

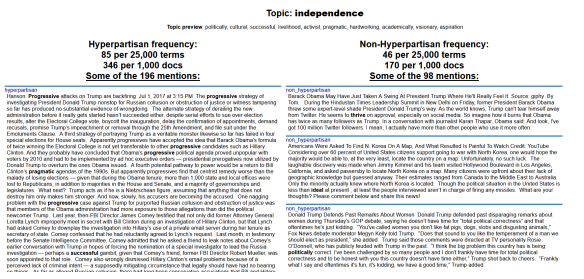

Figure 3: Information available when clicking a data point.

4.3 Empath

As we were saying, one of the limitations, not only in terms of loading the graph, but also in terms of interpretability, is the large number of terms that appear in our dataset.

An interesting alternative is to display Empath (Fast et al., 2016), a categorization of terms. In very general terms Empath is a word categorization tool, as exemplified in their paper, from small set of seed terms like “bleed” and “punch” they generate the category “violence”.

Figure 4: Empath full interactive plot available here.

4.4 Ordering Terms by Corpus Characteristicness

The third graph we will analyze is the one in which we order the terms by corps characteristicness. Let us first stop to have a look to what this means.

Basically, when performing this procedure, we are identifying terms that are frequent within the studied documents but less common in general language. The characteristic score compares these terms against a general English frequency list. After this is performed, we are able to get the following plot:

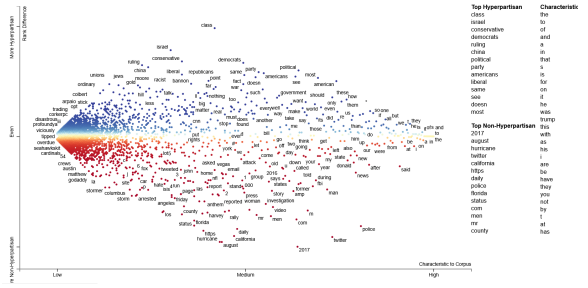

Figure 5: Full interactive plot available here (it may take a couple of minutes to load).

4.5 Embedding the data and plotting in the same representation space

The fourth and last graph is a projection of word embeddings (made with a word2vec method (Mikolov et al., 2013) which is a non-contexutal construction) and dimensionality reduction, utilizing the UMAP method (McInnes et al., 2018), to map words into a semantic space.

This makes us able to revealing relationships among words that go beyond mere frequency counts.

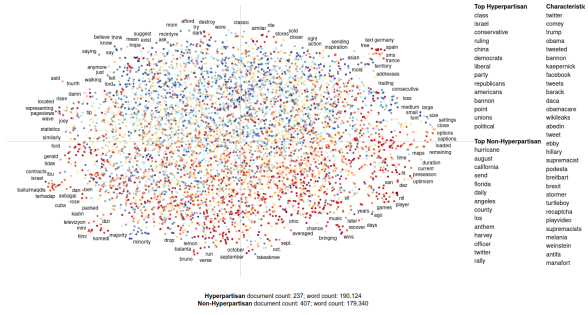

Figure 6: Full interactive plot available here (it may take a couple of minutes to load).

5 Analysis

5.1 Word clouds

Word clouds provide a quick overview of the most frequently used words in hyperpartisan and non-hyperpartisan news articles. In hyperpartisan news, words with strong ideological connotations such as "corrupt," "radical," "elite," "scandal," and "threat" appeared with high frequency. This suggests that hyperpartisan articles tend to focus on polarizing topics, specifically into conflicts and political entities. Conversely, non-hyperpartisan news featured words like "report", "official", "statement", "policy", "analysis" and "fraud", which indicate a more neutral, information-driven approach.

Althoug easy to interpret, this method lacks context and does not tell us how words are being used within sentences, e.g. whether they are used positively or negatively. This makes it a limited tool for in-depth analysis. Thats why the following scattertext plots will gave us more information of the hyperpartisan dataset.

5.2 Visualizing term associations

This visualization method displays how strongly certain words are associated with the hyperpartisan or non-hyperpartisan class. For instance, words like "deep state", "mainstream media", "rigged" and "traitor" were strongly associated with hyperpartisan articles. Meanwhile, words such as "fact-check", "research", "confirmed" and "expert" were more frequent in non-hyperpartisan texts. It is interesting to see how the same term, in this case "climate change", can be next to others words like "hoax" or "exaggerated", while in non-hyperpartisan articles, it was linked to terms like "scientific consensus" and "report".

This visualization is useful for identifying key words between hyperpartisan and non-hyperpartisan news. It provides more insight than simple word clouds or frequency-based metrics. However, this plot do not capture the full usage of a word. A word may appear frequently in one category but be used in different ways.

5.3 Empath

Empath revealed clear thematic differences between hyperpartisan and non-hyperpartisan news articles. Categories such as violence, fear, war, and anger were significantly more prevalent in hyperpartisan articles. For instance, words from the violence category, including "attack", "destroy" and "kill", appeared frequently, suggesting a focus on danger and conflict reports. In contrast, non-hyperpartisan articles showed higher frequencies in categories like science, business, and law, indicating an emphasis on factual reports.

Empath categorization allows us to move beyond raw frequency and identify thematic trends, making

it more informative than previous plots. However, it has limitations. Since it relies on predefined dictionaries, it may miss relevant terms that are relevant to the hyperpartisan artcles that may not appear there. Additionally, the same word can fit into multiple categories depending on the context. While useful for detecting general trends, it cannot replace manual linguistic analysis.

5.4 Ordering Terms by Corpus Characteristicness

This method identifies words that are frequent in hyperpartisan or non-hyperpartisan texts but rare in general English. This approach revealed that hyperpartisan news relies on a very emotional vocabulary, using charged or conspiratorial terms. Words like "cabal", "establishment", "globalists" and "patriot" are common in hyperpartisan news but rarely appeared in other texts.

While this visualization helps differentiate non-common terms, it does not directly explain why certain words are used more frequently.

5.5 Embedding the data and plotting in the same representation space

As previously explained, the embeddings are made in a non-contextual method and only captures semantic relationships based on word co-occurrence. The UMAP visualization showed clear clustering of hyperpartisan and non-hyperpartisan words. For example, words associated with conspiracy theories, like "hoax", "deep state" and "cover-up", formed a distinct cluster. On the contrary words related to scientific discourse, like "evidence", "study" and "peer-reviewed", were more dispersed in the non-hyperpartisan section. This suggests that hyperpartisan words are more semantically joined.

This technique can be difficult to interpret if you compare with the other plots and small changes in dataset composition can significantly alter the output. This serves as a good starting point for deep learning analysis, where more complex contextual models could further refine the insights.

6 Conclusions

To sum up, this has demonstrated how the ussage of scattertext can leverage our understanding in analyzing hyperpartisan articles in comparison with other methodologies. For a future research it would be interesting if we compare the results with a more in deep analysis and not just qualitative, and maybe use other tools.

References

Mike Bostock, Shan Carter, and Matthew Ericson. 2018. At the national conventions, the words they used - interactive feature - nytimes.com.

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*, CHI'16, page 4647–4657. ACM.

Jason S. Kessler. 2017. Scattertext: a browser-based tool for visualizing how corpora differ.

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*, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2017. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Andreas Müller. 2016. word_cloud: A python package for generating word clouds. `https://github.com/amueller/word_cloud`. Version 1.9.4.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.