Data exploration for transparency

Rishabh Shukla, Magdalena Lis
Factmata Ltd., Mindspace
114, Relay building, High Street
Whitechapel, London
rishabh.shukla@factmata.com, magdalena.lis@factmata.com

Abstract

Transparency of Machine Learning systems is necessary for building trust with their users. For interpreting and explaining model predictions, a good understanding of data is crucial. In this paper, we present results of applying data exploration methods in real-world industry scenario. We then discuss our approach to improving the data and models quality based on the resulting insights. Finally, we show how we transform these insights into publicly available information to increase transparency.

1 Introduction

With the increasing presence of algorithmic decision making, transparency in Machine Learning has become a topic of major social interest. Insight into models decisions is vital for preventing algorithmic bias (Gilpin et al., 2018) that can result in discrimination. For example, the prediction of a criminal risk assessment tool was found to be racially biased (Angwin et al., 2016). Furthermore, in the industry, unexpected results and seemingly black box predictions from ML systems might result in losing the trust of potential clients and users. Therefore, a clear understanding of the limitations of the ML models, and being transparent about them, is crucial for businesses.

The issue of algorithmic fairness is strongly related to the data used for model training (Kamishima et al., 2018). Therefore, data exploration plays an important role in understanding models' predictions and biases.

In this paper, we describe our application of data exploration methods to hate speech and hyperpartisan datasets. Firstly, we introduce the datasets. Next, we outline a set of well-known data exploration methods and how they can be employed to get valuable model insights. We then present the results and insights obtained from applying those methods to our datasets. Finally, we discuss how we utilise those results to improve the models as well as to provide users of our products with an understanding of our technology.

2 Datasets

The current study concerns datasets used for training our models for hate speech and hyperpartisanship detection. Despite the focus on a limited set of datasets, the data exploration methods described in the following section are equally applicable to any dataset for training an ML systems.

Hate speech refers here to derogatory statements based on the individual or groups identity. We use a public hate speech dataset (Waseem and Hovy, 2016), which consists of 16,907 manually annotated tweets. The samples were initially bootstrapped using hashtags that occur in hateful tweets. The tweets were then annotated manually with one the following labels: 'racist', 'sexist' and 'none.' For the current study, we aggregate the former two labels under an umbrella label 'hate speech.'

Hyperpartisanship indicates content which is politically biased (Entman, 2007). Our proprietary hyperpartisanship dataset consists of articles gathered using seed domains from www.mediabiasfactcheck.com. Articles from the extreme right and extreme left websites were used as positive samples, whereas the negative samples consist of articles from reputed news websites (NewYork Times, Guardian, etc.). The final dataset encompasses 40k articles labelled as either 'hyperpartisan' or 'not hyperpartisan.'

3 Data exploration methods

To obtain a better understanding of the datasets to be used for modelling, we employ a number of

well-known data exploration methods. The triangulation of these methods leads to a more comprehensive understanding of the data, as each method provides additional insights.

Manual analysis: Qualitative analysis of the dataset and predictions helps in getting a general idea of the data and the model, including an approximation of the quality as well as potential explanations of false positives and negatives. Manual analysis is complemented by statistical methods, which provide robust results in an efficient manner.

Latent Semantic Indexing: LSI is a method for extracting the context-dependent meaning of words using statistical analysis of corpora (Landauer et al., 1998). LSI provides the most frequently co-occurring word groups in a dataset. This method enables us to obtain insights beyond basic word frequencies and individual tokens by providing co-occurrences of words in similar contexts.

Local Interpretable Model-Agnostic Explanations: LIME trains a linear model on locally modified data points by perturbing the original dataset (Tulio Ribeiro et al., 2016). It helps in understanding the contribution weights of individual tokens to a prediction. By exploring the predictions instead of the datasets directly, LIME complements the previous methods, while providing latent insights about the overfitting of the models towards certain features.

4 Results

We applied the above-mentioned methods to the hate speech and hyperpartisanship datasets and the trained models. The data exploration methods provided us with the following insights:

Reporting vs expressing content: Some articles do not express hateful/hyperpartisan views but merely report on them. That is, they report on a hateful incident or quote biased language, often to rebuke them. Manual analysis of the models' predictions have indicated that the models can't differentiate between reported and expressed content. This results in false positives for texts reporting on hate speech/hyperpartisanship.

Targets: LSI have helped us in identifying demographic bias in the hyperpartisan dataset. The results demonstrated a bias towards American context, i.e. the dominance of political terms from US elections. In the hate speech dataset, most of the data samples focused on explicit Islamophobic remarks and sexist language against women. As a result, the models trained on these datasets couldn't generalise well to other demographies (e.g. UK politics) and other targets of hate speech (e.g. LGBT community). Furthermore, LSI showed frequent co-occurrences of sexist and profane language in the same contexts. In the hyperpartisanship dataset, we have found a co-occurrence of token Trump with fraud and Muslims in as many as 10% of the clusters, which suggests a lack of diversity in the training data.

Overfitting: LIME results have indicated overfitting of both hate speech and hyperpartisanship models. Bias towards certain words like muslims, girl, Trump, conservatives, etc. always resulted into a positive class prediction.

5 Discussion

To account for the difference between expressing and reporting hateful/hyperpartisan content, we included a quote detection feature as a filter in our models[1] as a result of which we could differentiate between expressed politically bias[2] and a fact-check analysis of the same incident.[3]

To counter the geographical bias, we have started collecting hyperpartisan data from English news outlets around the world. For hate speech, we gathered anti-semitic comments, hateful samples against LGBT community and black people.

Finally, we have transformed the insights from the above analyses into information suitable for non-experts, providing transparency to users, clients and the general public. Factmata's moderation API (`https://try.factmata.com/`) is one such example where we clearly communicate the strengths and weaknesses of our ML models. We explicitly mention the distribution of the datasets and limitations of the models on our website and in our public technical documentation available to any external party.

[1] Results based on our API in July 2019. https://try.factmata.com

[2] https://crooksandliars.com/2019/03/trump-screams-democrats-during-psychotic

[3] https://www.politifact.com/truth-o-meter/article/2019/mar/29/fact-checking-donald-trumps-grand-rapids-rally-aft/

References

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias.

Robert M. Entman. 2007. Framing Bias: Media in the Distribution of Power. *Journal of Communication*, 57(1):163–173.

Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. *arXiv e-prints*, page arXiv:1806.00069.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2018. Model-based and actual independence for fairness-aware classification. *Data Mining and Knowledge Discovery*, 32(1):258–286.

Thomas K Landauer, Peter W. Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv e-prints*, page arXiv:1602.04938.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.