The Extremes of Good and Evil

Bachelor Thesis

presented by
Larissa Strauch
Matriculation Number 1518629

submitted to the
Data and Web Science Group
Prof. Dr. Ponzetto
University of Mannheim

August 2014

Contents

1 Introduction **1**

1.1 Problem Statement 1

1.2 Contribution 1

1.3 Related Work 1

2 Data Analysis **2**

2.1 Data Description 2

2.1.1 Dataset labelled by Publisher 3

2.1.2 Dataset labelled by Article 4

2.2 Data Analysis 5

List of Algorithms

List of Figures

2.1 Distribution of as Hyperpartisan labelled articles by publisher . . . 3

List of Tables

Chapter 1

Introduction

1.1 Problem Statement

1.2 Contribution

1.3 Related Work

Chapter 2

Data Analysis

2.1 Data Description

The given data, on which we want to build our model on, was provided by zenodo.org as part of SemEvals Task 4 and consists of 2 independent datasets, which in turn have been divided into a GroundTruth-, Training- and Validation set.
The first dataset, recognizable by the term 'byPublisher', reflects the publisher's general bias set forth by BuzzFeed journalists or MediaBiasFastCheck.com beforehand. It consists of a total of 750,000 items, of which 600,000 belong to the Training- and 150,000 to the Validation set.
In return, the second dataset, recognizable by the term 'byArticle', was scrapped by crowdsourcing at hand and therefore consists of only 645 items without a Validation set.

The GroundTruth dataset was provided as an XML File and consists of the features 'article url', 'labeled-by', 'id' and 'hyperpartisan'. In addition, the GroundTruth dataset scrapped 'byPublisher' contains the feature 'bias'.

- Article-url: Contains the article's url.
- Labeled-by: Reflects whether the respective article was labeled 'byPublisher' or 'byArticle'.
- Id: Allocates each article a unique id.
- Hyperpartisan: Displays whether the particular article has been labeled as hyperpartisan or not.
- Bias: Divides the publisher's bias into 'left', 'left-center', 'least', 'right-center' and 'right'.

The Training dataset was as well provided as an XML file and contains the contents of the website of the respective article. In addition, it consists of the features 'article title' 'published-at' and 'id'.

- Article title: Represents the articles title.
- Published-at: Specifies the published date.
- Id: A unique id, which is the same as the corresponding entry in the GroundTruth dataset.

The given Data has been cleaned in advance, therefore no additional steps were necessary.

The main focus of the datasets is on the Hyperpartisan feature, on which we want to classify the articles as this thesis progresses.

2.1.1 Dataset labelled by Publisher

As mentioned above, this dataset consists of a total of 750,000 articles and is divided into a training record consisting of 600,000 articles and a validation set consisting of 150,000 articles. Summarizing these two sets of data, a total of 375,000 were labelled as 'Hyperpartisan' and 375,000 were not – which corresponds to a 50:50 distribution. But even individually, this distribution does not change.

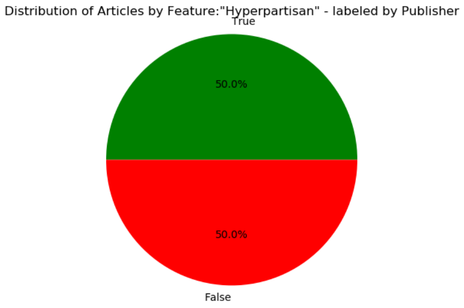
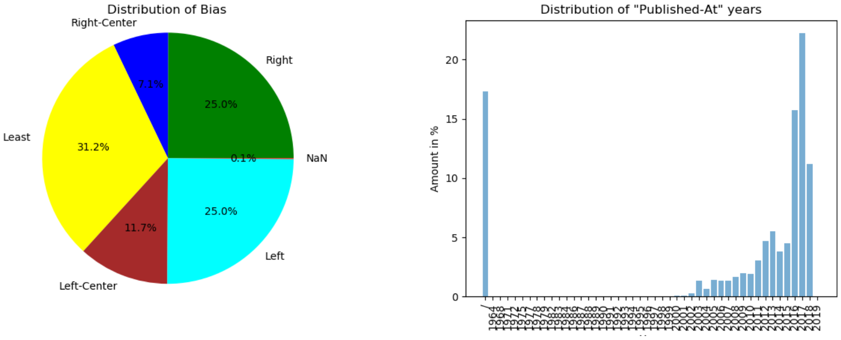


Figure 2.1: Distribution of as Hyperpartisan labelled articles by publisher

This dataset also includes the feature 'bias', which informs you about the general bias of the publisher. All 375,000 Hyperpartisan labelled all are assigned to either the left or right sectors, but none are right-centre, least or left-centre and are again 50:50 distributed.

The other 50% are split between the remaining bias, with 'Least' owning the largest share at 37%.
The publicity data is distributed over the years 1964-2018, with most of the data coming from 2012-2018.

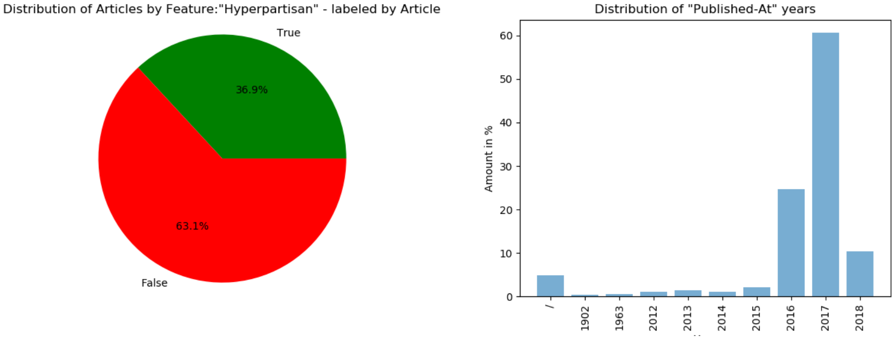


(a) Distribution of Bias

(b) Distribution of publishing years

2.1.2 Dataset labelled by Article

The dataset labelled by Article is a little different to the larger one labelled by publisher. Here the articles were individually labelled by hand. Accordingly, the distributions of this dataset are completely different. This becomes quiet obvious if we look at how the distribution of the Hyperpartisan labelled articles is here. Here we can see that there is no 50:50 distribution left. Only 36.9% were defined here as Hyperpartisan as shown in figure 2.4.



(c) Distribution of Hyperpartisan labelled articles

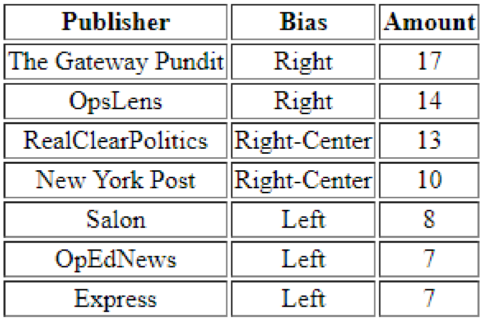
(d) Distribution of publishing years

Moreover, in this dataset, the distribution of publication data is not mainly from the years 2012-2018, but in 2016-2018, with the largest number of articles dating back to 2017 at just under 60% as we can see in figure 2.5. Altogether all 645 articles date from the years 1902-2018.

2.2 Data Analysis

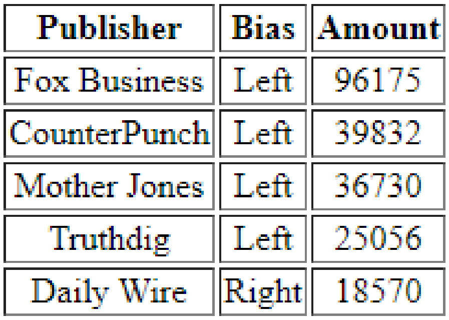
Hyperpartisan means ”extremely partisan; extremely biased in favor of a political party.” [definition]. This often materializes in relation to significant political events. As a result, in the following chapter, I will discuss in detail the direct link between the various features, especially the correlations between publication dates and label, as well as the connections between publisher and label.

The two figures e and f list the publisher who produced the highest amount of hyperpartisan articles in the respective dataset. To keep track, only those publishers who have publisher more than 7 hyperpartisan articles in the dataset ’labelled by article’ and who have published more than 10.000 hyperpartisan articles in the dataset ’labelled by publisher’ are included in the tables. However, a problem here is the dataset ’Labelled by Publisher’, since this record, as previously described, has been labelled by the overall bias of the publisher. Meaning, for the further development, we can not include the publishers, since each article of a publishing house has the same label. What I would like to discuss later, however, is the connection between whether or not one of the publishers listed in Table x published more articles in important political years



Publisher	Bias	Amount
The Gateway Pundit	Right	17
OpsLens	Right	14
RealClearPolitics	Right-Center	13
New York Post	Right-Center	10
Salon	Left	8
OpEdNews	Left	7
Express	Left	7

(e) Publisher with the highest amounts of Hyperpartisan Articles labelled by Article



Publisher	Bias	Amount
Fox Business	Left	96175
CounterPunch	Left	39832
Mother Jones	Left	36730
Truthdig	Left	25056
Daily Wire	Right	18570

(f) Publisher with the highest amounts of Hyperpartisan Articles labelled by Publisher

Ehrenwörtliche Erklärung

Ich versichere, dass ich die beiliegende Master-/Bachelorarbeit ohne Hilfe Dritter und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen. Ich bin mir bewusst, dass eine falsche Er- klärung rechtliche Folgen haben wird.

Mannheim, den 31.08.2014 Unterschrift