Master thesis

Simon Lorent

July 21, 2019

Contents

Master thesis

Fake news detection using machine learning

Simon Lorent

Acknowledgement

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Master thesis

Fake news detection using machine learning

Simon Lorent

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Chapter 1

Introduction

1.1 What are fake news?

1.1.1 Definition

Fake news have quickly become a society problem, being used to propagate false or rumorus informations in order to change behaviors of peoples. Before stating to work on detecting fake news, it is needed to first understand what they are. It have been show that propagation of fake news have had a non negligable influence of 2016 US presidential elections[1]. A few facts on fake news in the United States:

- 62% of US citizen get there news for social medias[2]

- Fake news had more share on facebook than mainstream news[3].

Fake news have also been used in order to influence the referendum in the United Kingdom for the "Brexit".

There are two aspects of fake news detection that need to be taken into account according to Shu et al[4]. The first is characterization or what are fake news and the second is detection. In order to build detection models, it is need to start by charaterization, indeed, it is need to understand what are fake news before trying to detect them.

Fake news definition is made of two part: authenticity and intent. Authenticity means that fake news content fale information that can be verified as such, which means that concpiracy theory is not included in fake news as there are difficult to be proven true or false in most cases. The second part, intent, means that the false information have been written with the goal of missleading the reader.

1.1.2 Fake News Characterization

Definition 1 *Fake news is a news article that is intentionally and verifiably false*

The part of the definition introducing the intent of missleading the reader automaticaly discard satire news medias, that is why this works will focus on the first part, the fact that the piece of information is verifiably false or true. Indded, even if satire news medias does not have the intent to misslead the readers, not all of them have the ability of making
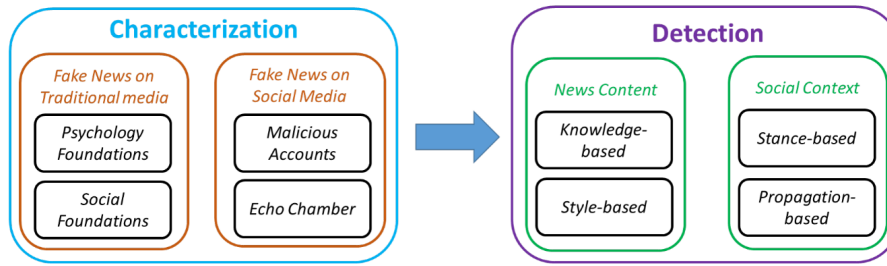
Figure 1.1: Fake news on social media: from characterization to detection.[4]

criticisme and not taking it to the first degree. On the other hand, in the case of political media, even if it clearly try to influence the consumer, verifying the authenticity of there claims is usualy harder as, in most of the cases, openly lies.

1.2 Feature Extraction

1.2.1 News Content Features

Now that fake news have been defined and the target have been set, it is needed to analize what features can be used in order to classify fake news. Starting by looking at news content, it can be seen that it is made of four principal raw components:

- **Source**: Where does the news come from, who wrote it, is this source reliable or not.

- **Headline**: Short summary of the news content that try to attract the reader.

- **Body Text**: The actual text content of the news.

- **Image/Video**: Usualy, textual information is agremented with visual information such as images, videos or audio.

Features will be extracted from these four basic components, with the mains features being linguistic-based and visual-based. As explained before, fake news are used to influence the consumer, and in order to do that, they often use a specific language in order to attract the readers. On the other hand, non fake news will mostly stick to a different language register, being more formal. This are linguistic-based features, to which can be added lexical features such as total number of words, frequency of large words or unique words.

The second features that need to be taken into account are visual features. Indeed, modified images are often used to add more weight to the textual information. For example, the **Figure 1.2** is supposed to show the progress of deforestation, but the two images are acutaly from the same original one, and in addition the WWF logo make it looks like to be from a trusted source.

1.2.2 Social Context Features

In the context of news sharing on social medias, multiples aspect can be taken into account, such as user aspect, post aspect and group aspect. For instance, it is possible to

Figure 1.2: The two images provided to show deforestation between two date are from the same image taken at the same time.

analyze the behavior of specific users and use their metadata in order to find if a user is at risk of trusting or sharing false information. For instance, those metadata can be its center of intereset, its number of followers, or anything that relates to it.

Post-based aspect is in a sense similar to user based: it can use post metadata in order to provide usefull informations, but in addition to metadata, the actual content can be used. It is also possible to extract features from the content using latent Dirichlet allocation (LDA)[5].

1.3 News Content Models

1.3.1 Knowledge-based models

Now that the differents kinds of features available for the news have been defined, it is possible to start to explain what kind of models can be built using these features. The first models that relates to the news content is based on knowledge: the goal of this model is to check the truthfulness of the news content and can be acheived in tree differents way (or a mixture of them):

- **Expert-oriented**: relies on expert, such as journalist or scientist, to asses the news content.

- **Crowdsourcing-oriented**: relies on the wisdom of crowd that says that if a suficiently large amount of persons says that something is false or true then it should be.

- **Computational-oriented**: relies on automatic fact checking, that could be based on external resources shuch as DBpedia.

These methods all have pros and cons, hiring expert might be constly, and expert are limited in number and might not be able to treat all the news that are produced. In the case of crowdsourcing it can easily be fooled if enought bad annotators breaks the system and automatic fact checking might not have the necessary accuracy.

1.3.2 Style-based model

As explained earlier, fake news usualy tries to influence consumier behavior, and thus generaly use a specific style in order to play on the emotion. These methods are called deception-oriented stylometric methods.

The second method is called objectivity-oriented approaches and tries to capture the objectivity of the texts or headlines. These kind of style is mostly used by partisan article or yellow-journalism, that is, website that relies on eye-catching headline without reporting any useful information. An example of these kind of headline could be:

You will never beleive what he did !!!!!!

This kind of headline plays on the curiosity of the reader that would click to read the news.

1.4 Social Context Models

The last features that have not been used yet are social media features. There are two approaches to use these features: stance-based and propagation-based.

Stence-based approaches use implicit or explicit representation. For instance, explicit representation might be positive or negative votes on social medias. Implicit representation need to be extracted from the post itself.

Propagation-based approaches use features related to sharing such as the number of retweet on twitter.

1.5 Datasets

1.5.1 Fake News Corpus

This works uses multiples corpus in order to train and test different models. The main corpus used for training is called Fake News Corpus[6]. This corpus have been automatically crawled using `opensources.co` labels. In other words, domains have been labeled with one or more labels in

- Fake News
- Satire
- Extreme Bias
- Conspiracy Theory
- Junk Science
- Hate News
- Clickbait
- Proceed With Caution
- Political
- Credible

These annotations have been provided by crowdsourcing, which means that they might not be exactly accurate, but are expected to be close to the reality. Because this works focus on fake news detection against reliable news, only the news labels as fake and credible have been used.

1.5.2 Fake News Net

The second dataset used is fake news net[7, 8, 4]. This corpus is made of news from two different sources, PolitiFact and GossipCop. An older version also provide news from BuzzFeed. News are categorized in two classes: fake and non fake. Being quite smaller than fake news corpus, this dataset will be used as a test dataset.

1.5.3   Liar, Liar Pants on Fire

The third and last dataset is **Liar, Liar Pants on Fire** dataset[9], which is a collection of twelve thousand small sentences collected from various sources and hand labeled. They are devided in six classes:

- pants-fire

- false

- barely-true

- half-true

- mostly-true

- true

This set will be used a second test set. Because in this case there are six classes againt two in the other cases, a threshould should be used in order to fix which one will be concidered as true or false.

It should be noted that this one differ from the two other dataset is it is composed only on short sentences, and thus it should not be expected to have very good results on this dataset for models trained on Fake News Corpus which is made of full texts.

1.6   State of the Art

There are two mains categories of state of the art that are interesting for this work: previous work on fake news detection and on general text classification. Works on fake news detection is almost inexistant and mainly focus on 2016 US presidential elections or does not uses the sames features. That is, when this work focus on automatic features extraction using machine learning and deep learning, other works make use of hand crafted features[10, 11] such as psycholinguistic features[12] which is not the goal here.

When it comes to state of the art for text classifiction, it includes Long short-term memory (LSTM)[13], Attention Mechanism[14], IndRNN[15], Attention-Based Bidirection LSTM[16], Hierarchical Attention Networks for Text Classification[17], Adversarial Training Methods For Supervised Text Classification[18], Convolutional Neural Networks for Sentence Classification[19] and RMDL: Random Multimodel Deep Learning for Classification[20]. All of these models have comparable performances.

1.7   Conclusion

As it have been shown in **Section 1.2** and **Section 1.3** mutliples approaches can be used in order to extract features and use them in models. This works focus on textual news content features. Indeed, other features related to social medias are difficult to aquire. For example, users information are difficult to obtain on Facebook, as well as post information. In addition, the different datasets that have been presented at **Section 1.5** does not provides any other information than textual ones.
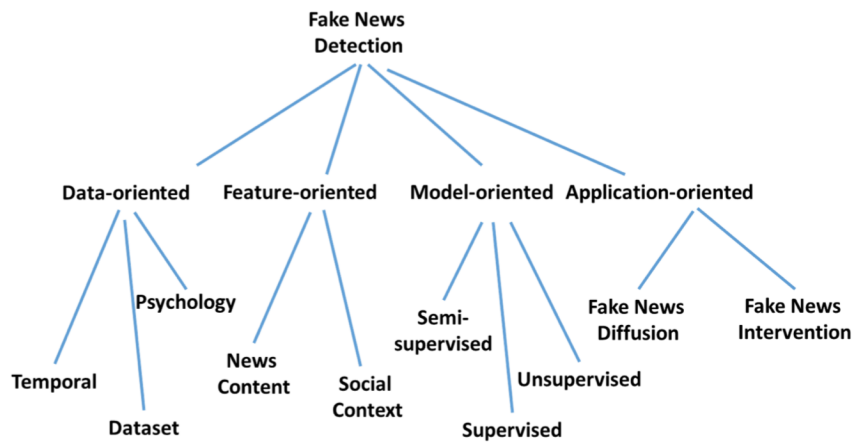
Figure 1.3: Differents approaches to fake news detection.

Looking at **Figure 1.3** it can be seen that the main focus will be made on unsupervised and supervised learning models using textual news content. It should be noted that machine learning models usualy comes with a trade-off between precision and recall and thus that a model which is very good at detected fake news might have a high false positive rate as opposite to a model with a low false positive rate which might not be good at detecting them. This cause ethical questions such as automatic censorship that will not be discused here.

Chapter 2

Data Exploration

2.1 Introduction

A good starting point for the analysis is to make some data exploration of the data set. The first thing to be done is statistical analysis such as counting the number of text per class or counting the number of words per sentence. The second step consit of doing Latent Dirichlet Allocation[5] in order to make unsupervised clustering of the text and see if there is some kind of correlation between the clusters to which a text belongs and its labels.

2.2 Dataset statistics

2.2.1 Fake News Corpus

General Analysis

Because **Fake News Corpus** is the main dataset, the data exploration will start with this dataset. And the first thing is to count the number of items per class. Before starting the analysis, it is needed to cleanup the dataset. As it is originaly given in a large 30GB CSV file, the first step is to put everything in a database in order to be able to retreive only wanted peice of information. In order to do so, the file have been red line by line. It appears that some of the lines are baddly formated, preventing them to be red correctly, in this case they are dropped without being put in the database. Also, each line that is a duplicate of a line already red is also dropped. The second step in cleaning the set consist of some more duplicate removal. Indeed, dropping same lines remove only exact duplicate. It appears that some news does have the same content, with slight variation in the title, or a different author. In order to remove the duplicate, each text is hashed using SHA256 and those hash a compared, removing duplicates and keeping only one.

Because the dataset have been cleaned, numbers provided by the dataset creators and number computed after cleaning will be provided. We found the values given at **Table 2.1**. It shows that the number of fake news is smaller by a small factors with respect to the number of reliable news, but given the total number of items it should not cause any problems. But it will still be taken into account later on.

In addition to the numbers provided at **Table 2.1**, there are also two more categories that are in the dataset but for which no description is provided:

Type	Provided	Computed
Fake News	928, 083	770, 287
Satire	146, 080	855, 23
Extreme Bias	1, 300, 444	771, 407
Conspiracy Theory	905, 981	495, 402
Junk Science	144, 939	79, 342
Hate News	117, 374	65, 264
Clickbait	292, 201	176, 403
Proceed With Caution	319, 830	104, 657
Political	2, 435, 471	972, 283
Credible	1, 920, 139	1, 811, 644

Table 2.1: Number of texts per categories

- Unknwon: 231301

- Rumor: 376815

To have a better view of the distribution of categories, an histogram is provided at **Figure 2.1**.
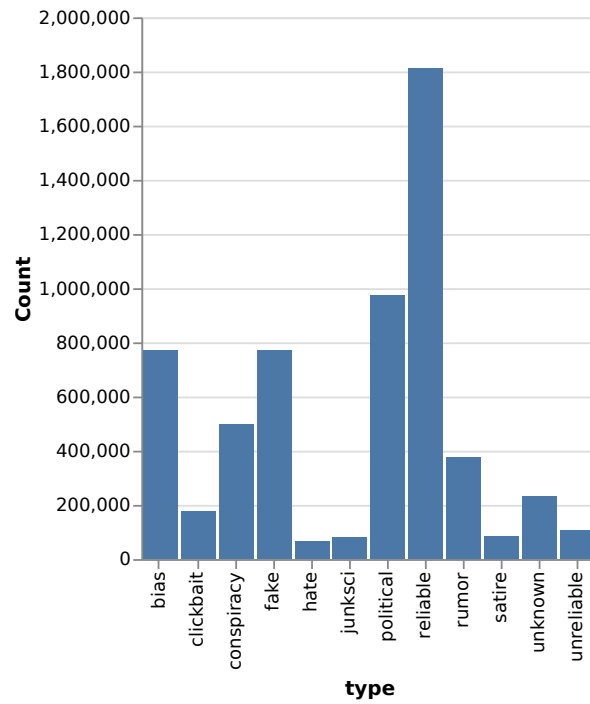
Figure 2.1: Histogram of text distribution along their categories on the computed numbers.

In addition, the number of words per text and the average number of words per sentences have been computed for each text categories. **Figure 2.2** shows the boxplots for these values. It can be seen that there is no significative difference that might be used in order to make class prediction.

Before counting the number of words and sentences, the texts are preprocessed using gensim[21] and NLTK[22]. The first step consist of spliting text into an array of sentences
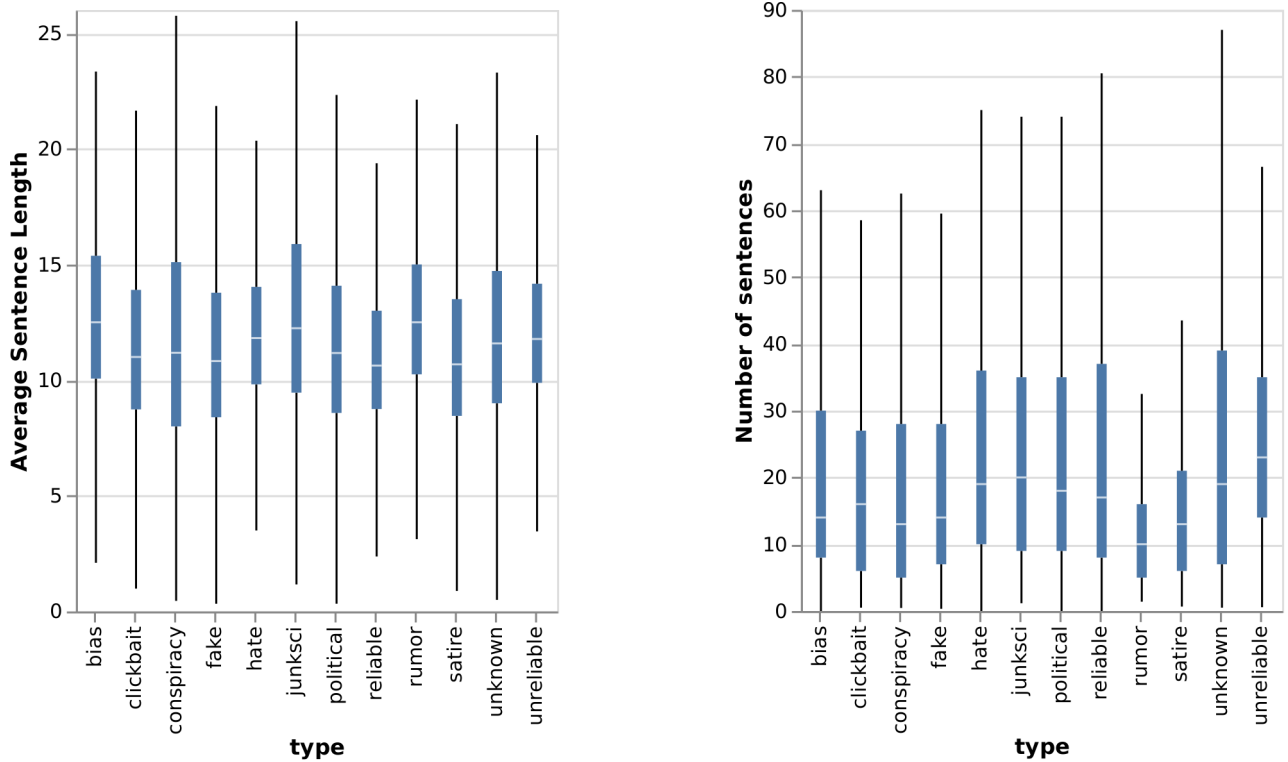
(a) Boxplot of average sentence length for each category.

(b) Boxplot of number of sentences for each category.

Figure 2.2: Summary statistics

on stop punctuation such as dots or questions mark, but not on commas. The second step consist on filttering words that are contained in these sentences, to do so, stopwords (words such as 'a', 'an', 'the'), punctuation, words or size less or equal to tree, non alphanumeric words, numeric values and tags (such as html tags) are removed. Finnaly, the number of word still present is used.

An interesting features to look at is the distrubtion of news sources with respect to their categories. It shows that in some case some source a predominant. For instance, looking at **Figure 2.3** shows that most of the reliable news are from *nytimes.com* and in the same way, most of the fake news are coming from *beforeitsnews.com*. That have to be taken into account when training and testing models as the goal is not to distinguish between these two sources but between fake news and reliable news.

Fake News analysis

In this section, the analysis will focus on fake news and reliable news. Because of what shows **Figure 2.3**, that is, some categories are almost all from the same source, an analysis of what happens when dropping these sources. First, lets compare the number of news while and while not taking into account majors sources.
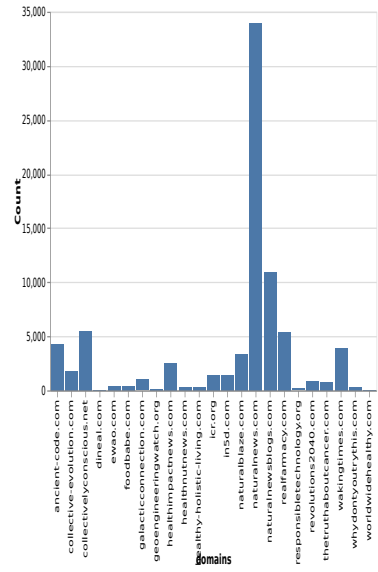
Comparing **Figure 2.4** and **Figure 2.5** shows that even by removing *nytimes.com* and *beforeitsnews.com* news from the dataset still leave enought texts to train the different models without the risk of learning something unwanted such as only separating these two sources. But one drawback is that the ratio between fake news and reliable news is
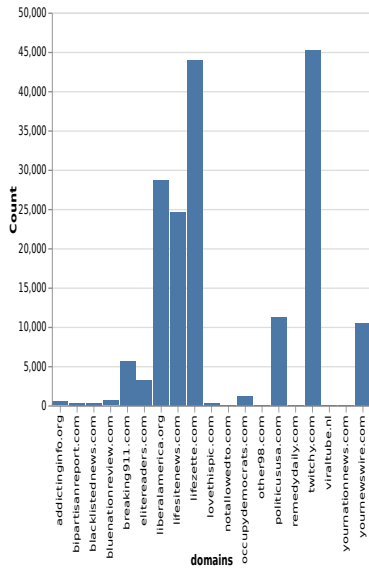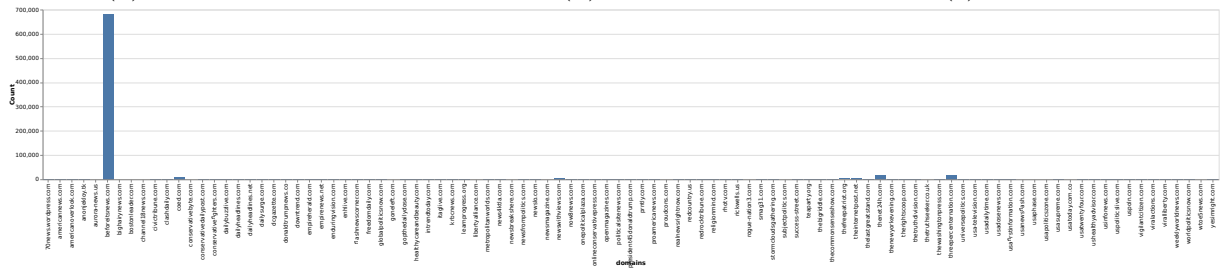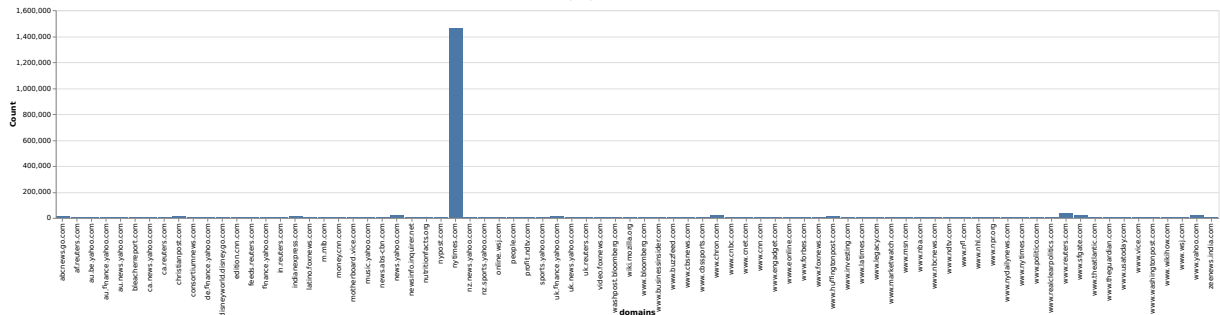
(a) Clickbaits
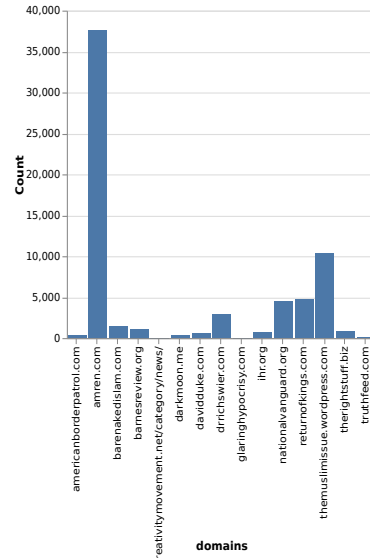
(b) Hate

(c) Junk sience

(d) Fake

(e) Reliable

going from around one half to around one fourth.

2.2.2   Liar-Liar corpus

As said in **Chapter 1**, this dataset is made of small text of one or two sentences at most. Which means that they are smaller than than the ones in *fake news corpus*. The distribution of words length can be seen at **Figure 2.6**. In addition, this dataset is not unbalanced as the other corpus is, which means that precautions does not have to be taken.

Visualisation with t-SNE[23]

In order to visualize the data it is needed to transform text in a numerical way and eventualy reduce the dimension in order to to allow it to be plotted on a 2D or 3D plot.

(f) Political

(g) Satire

(h) Unknown

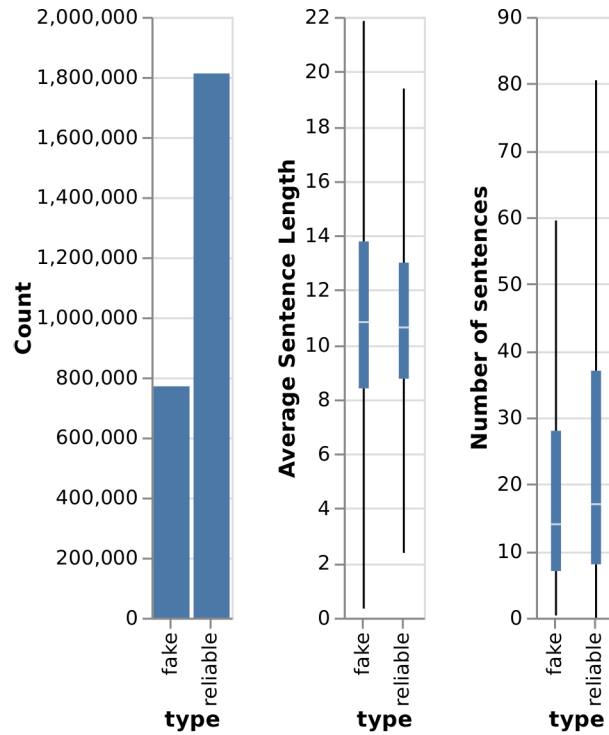Figure 2.3: Histogram of news origin for each categories.

Figure 2.4: Summary statistics for not downsampled fake and reliable news.

Here TF-IDF (term frequency, inverse document frequency) is used. How it works will be details later on. This produce a sparse matrix with each document being represented as an array, each value of the array being a value for one term. That is, the more term in the corpus, the longer the array becomes. For example, a corpus of 10.000 text with 20.000 uniques words would be represented as a 10000×20000 sparse matrix. As said before, plotting in 20000 dimension is not possible. In order to do so, the number of dimension needs to be reduced. Here, principal component analysis on t-SNE will be used together.

What is t-SNE? It stands for **t-distributed stochastic neighbor embedding**. The goal of this method is that two points to are closed in $\mathcal{R}^Q$ should be close in $\mathcal{R}^S, S \ll Q$. In order to do so, the algorithm starts by fitting a probability distribution on the input points. It starts by computing

$$p_{i|j} = \frac{\frac{exp(-||x_i - x_j||^2)}{2\sigma_i^2}}{\sum_{k \neq i} \frac{exp(-||x_i - x_k||^2)}{2\sigma_i^2}}$$

Each of the $y_i, i \in 0 \ldots N$ being a point of dimension Q.
Then

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N}$$

is computed.

The probability distribution for the low density map is given by

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l}(1 + ||y_k - x_l||^2)^{-1}}$$
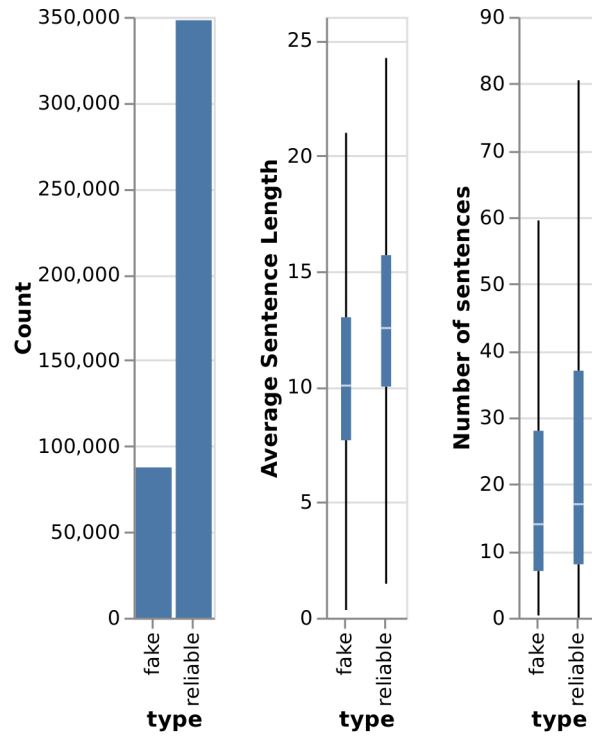
Figure 2.5: Summary statistics for downsampled dataset on fake and reliable news

And in order to have a low dimension probability distribution as close as possible to the high dimension one, it minimize the KullbackLeibler divergence of the two distributions.

$$KL(P||Q) = \sum_{i \neq j} p_{ij} * \log(\frac{p_{ij}}{q_{ij}})$$

The minimization of KL divergence is acheived by gradient descent.

The documentation of the module of scikit-learn[24] used for computing this value recomand to apply PCA first in order to reduce the original dimension and speed-up computation. The result of these computation can be seen at **Figure 2.7**.
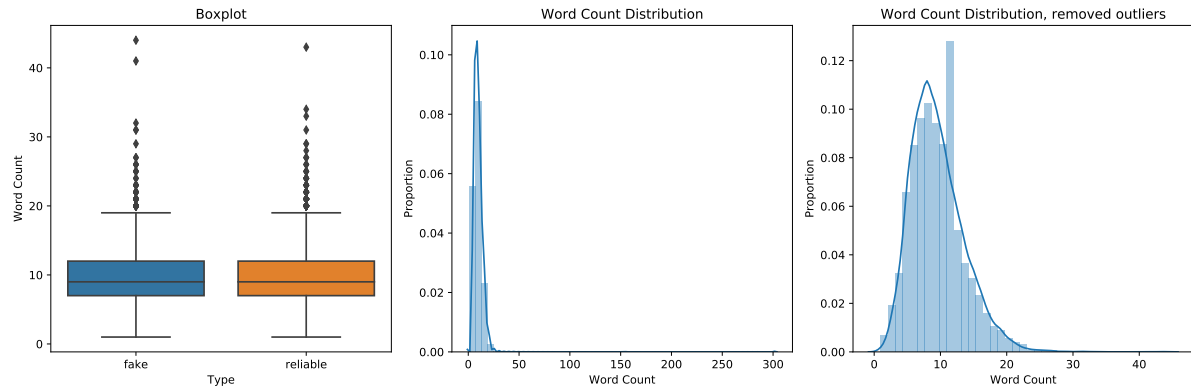
Figure 2.6: Number of words distributions for liar-liar dataset. On the first and the third plots, a few outliers with length greater than 50 have been removed in order to make the plots more readables.
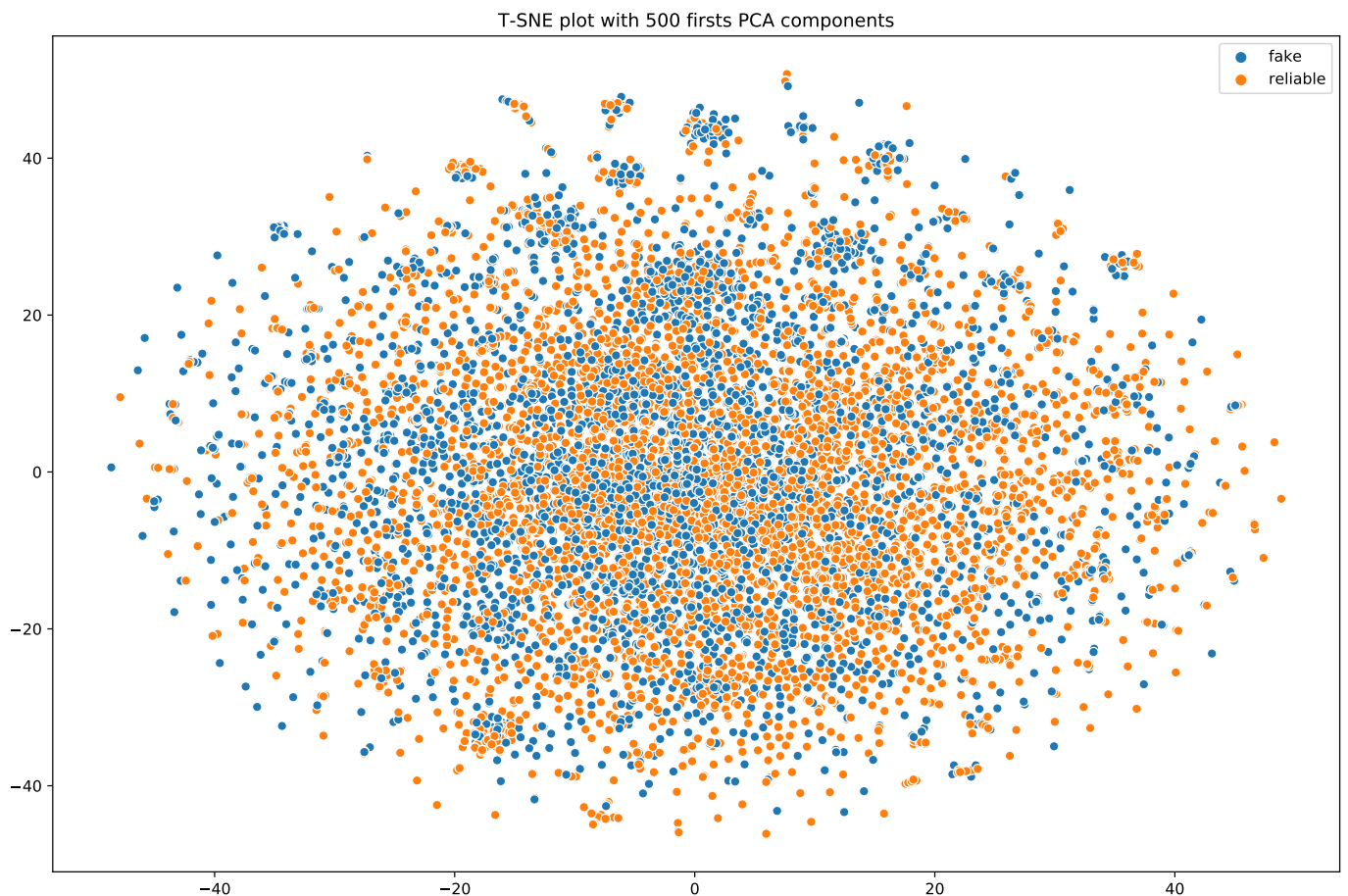

Figure 2.7: t-SNE plot for liar-liar dataset.

Bibliography

[1] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. In *Journal of Economic Perspective*, volume 31, 2017.

[2] Jeffrey Gottfried and Elisa Shearer. *News Use Across Social Medial Platforms 2016*. Pew Research Center, 2016.

[3] Craig Silverman and Lawrence Alexander. How teens in the balkans are duping trump supporters with fake news. *Buzzfeed News*, 3, 2016.

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

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[6] Maciej Szpakowski. Fake news corpus. `https://github.com/several27/FakeNewsCorpus`. Accessed: 2018-10.

[7] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*, 2018.

[8] Kai Shu, Suhang Wang, and Huan Liu. Exploiting tri-relationship for fake news detection. *arXiv preprint arXiv:1712.07709*, 2017.

[9] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection.

[10] Julio CS Reis, André Correia, Fabrício Murai, Adriano Veloso, Fabrício Benevenuto, and Erik Cambria. Supervised learning for fake news detection. *IEEE Intelligent Systems*, 34(2):76–81, 2019.

[11] Vernica Prez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news.

[12] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 1997.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[15] Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. Independently recurrent neural network (indrnn): Building a longer and deeper rnn.

[16] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212, Berlin, Germany, August 2016. Association for Computational Linguistics.

[17] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California, 2016. Association for Computational Linguistics.

[18] Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. Adversarial Training Methods for Semi-Supervised Text Classification. *arXiv:1605.07725 [cs, stat]*, May 2016. arXiv: 1605.07725.

[19] Yoon Kim. Convolutional Neural Networks for Sentence Classification. *arXiv:1408.5882 [cs]*, August 2014. arXiv: 1408.5882.

[20] Kamran Kowsari, Mojtaba Heidarysafa, Donald E. Brown, Kiana Jafari Meimandi, and Laura E. Barnes. RMDL: Random Multimodel Deep Learning for Classification. *Proceedings of the 2nd International Conference on Information System and Data Mining - ICISDM '18*, pages 19–28, 2018. arXiv: 1805.01890.

[21] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

[22] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.

[23] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, (9):2579–2605, 2008.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.