

Master thesis

Simon Lorent

August 4, 2019

Contents

1	Introduction	5
1.1	What are fake news?	5
1.1.1	Definition	5
1.1.2	Fake News Characterization	5
1.2	Feature Extraction	6
1.2.1	News Content Features	6
1.2.2	Social Context Features	6
1.3	News Content Models	8
1.3.1	Knowledge-based models	8
1.3.2	Style-based model	8
1.4	Social Context Models	9
1.5	Datasets	9
1.5.1	Fake News Corpus	9
1.5.2	Fake News Net	9
1.5.3	Liar, Liar Pants on Fire	10
1.6	State of the Art	10
1.7	Conclusion	10
2	Data Exploration	12
2.1	Introduction	12
2.2	Dataset statistics	12
2.2.1	Fake News Corpus	12
2.2.2	Liar-Liar corpus	17
2.3	Visualisation with t-SNE	17
2.4	Conclusion	19
3	TODO: find a title	24
3.1	Introduction	24
3.2	Words and text embedding	24
3.3	Methodology	25
3.3.1	Evaluation metrics	25
3.4	Models	26
3.4.1	Naïve-Bayes[1]	26
3.4.2	LIBLINEAR	26
3.5	Models on liar-liar dataset	27
3.5.1	Linear SVC	27
3.5.2	Decision Tree	27
3.5.3	Max Feature Number	27

3.6	Models on fake corpus dataset	30
3.6.1	SMOTE: Synthetic Minority Over-sampling Technique[2]	30
A	Appendix to Chapter 3	34
A.1	TF-IDF max features row results	34

Master thesis

Fake news detection using machine learning

Simon Lorent

Acknowledgement

Lorem ipsum dolor sit amet, consectetur 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, consectetur 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, consectetur 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, consectetur 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 rumorous 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 negligible influence of 2016 US presidential elections[3]. A few facts on fake news in the United States:

- 62% of US citizen get there news for social medias[4]
- Fake news had more share on facebook than mainstream news[5].

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[6]. 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 conspiracy 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 automatically 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

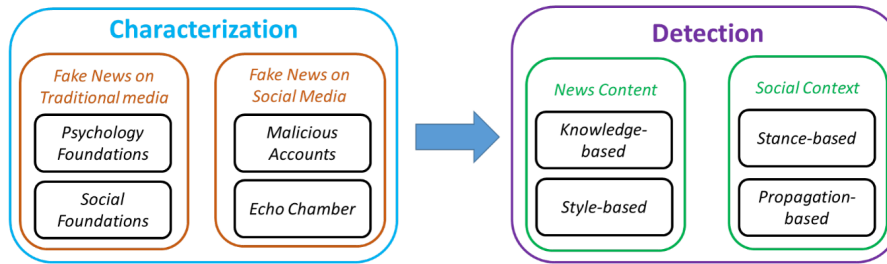


Figure 1.1: Fake news on social media: from characterization to detection.[6]

criticism 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 usually 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 analyze 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:** Usually, 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 interest, 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 useful 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)[7].

1.3 News Content Models

1.3.1 Knowledge-based models

Now that the different 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 achieved in three different ways (or a mixture of them):

- **Expert-oriented:** relies on expert, such as journalist or scientist, to assess the news content.
- **Crowdsourcing-oriented:** relies on the wisdom of crowd that says that if a sufficiently 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 such as DBpedia.

These methods all have pros and cons, hiring expert might be costly, 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 enough bad annotators break the system and automatic fact checking might not have the necessary accuracy.

1.3.2 Style-based model

As explained earlier, fake news usually tries to influence consumer behavior, and thus generally use a specific style in order to play on the emotion. These methods are called deception-oriented stylistic 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 believe 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[8]. 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[9, 10, 6]. 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[11], which is a collection of twelve thousand small sentences collected from various sources and hand labeled. They are divided 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 against two in the other cases, a threshold should be used in order to fix which one will be considered as true or false in order to be compared with the other dataset.

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. In addition, the texts from the latest dataset are more politically oriented then the ones from the first one.

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 same 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[12, 13] such as psycholinguistic features[14] which is not the goal here.

When it comes to state of the art for text classification, it includes Long short-term memory (LSTM)[15], Attention Mechanism[16], IndRNN[17], Attention-Based Bidirectional LSTM[18], Hierarchical Attention Networks for Text Classification[19], Adversarial Training Methods For Supervised Text Classification[20], Convolutional Neural Networks for Sentence Classification[21] and RMDL: Random Multimodel Deep Learning for Classification[22]. All of these models have comparable performances.

1.7 Conclusion

As it have been shown in **Section 1.2** and **Section 1.3** multiples 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

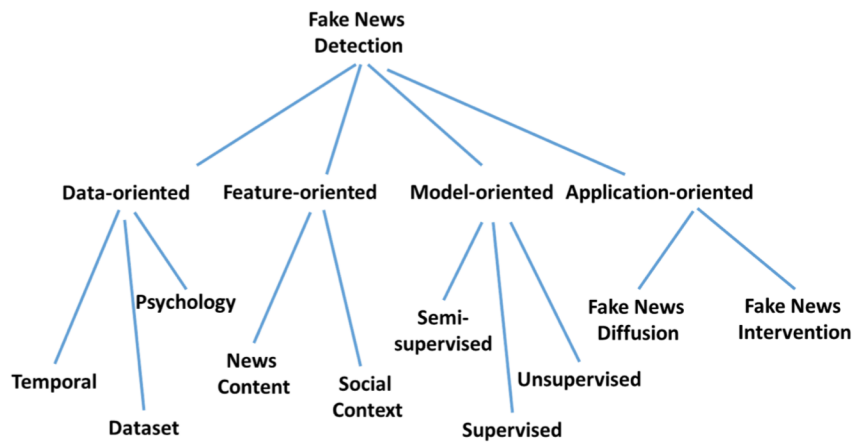


Figure 1.3: Different approaches to fake news detection.

acquire. 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.

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 usually 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 discussed 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 consist of doing Latent Dirichlet Allocation[7] 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 originally given in a large 30GB CSV file, the first step is to put everything in a database in order to be able to retrieve 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**.

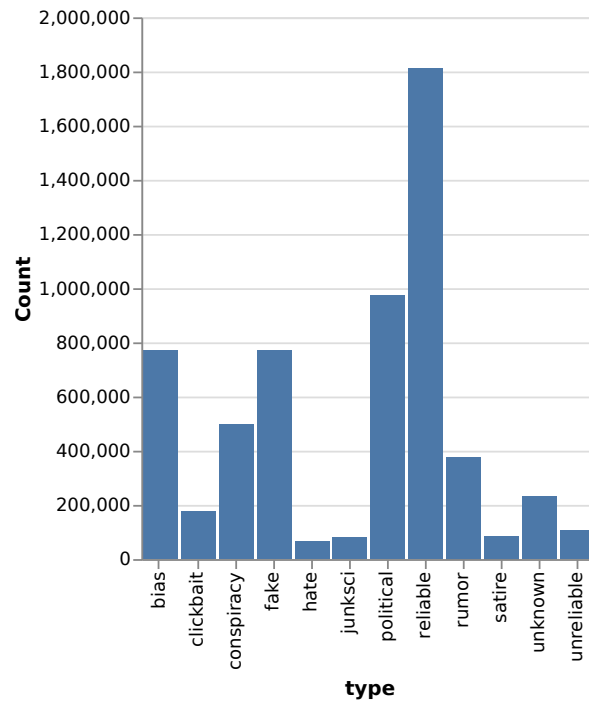
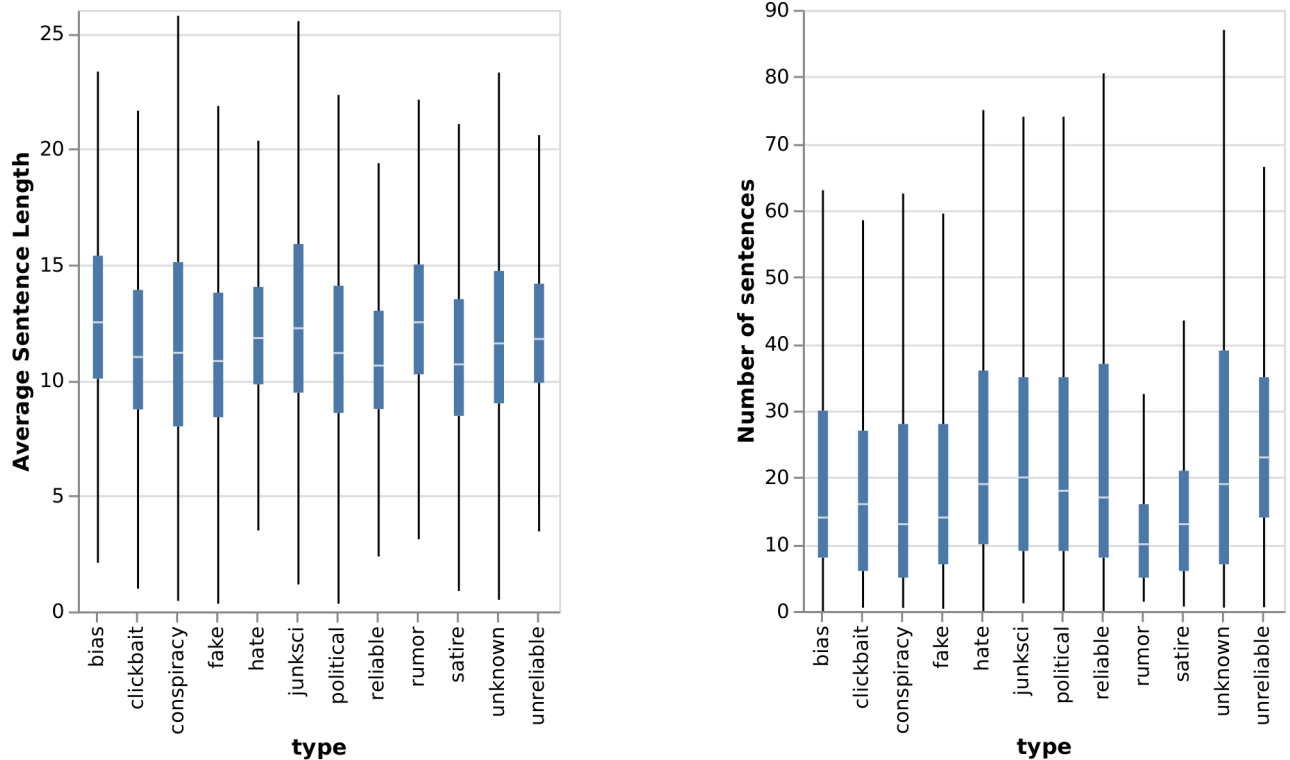


Figure 2.1: Histogram of text distribution along their categories on the computed numbers.

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[23] and NLTK[24]. The first step consist of splitting text into an array of sentences



(a) Boxplot of average sentence length for each category.

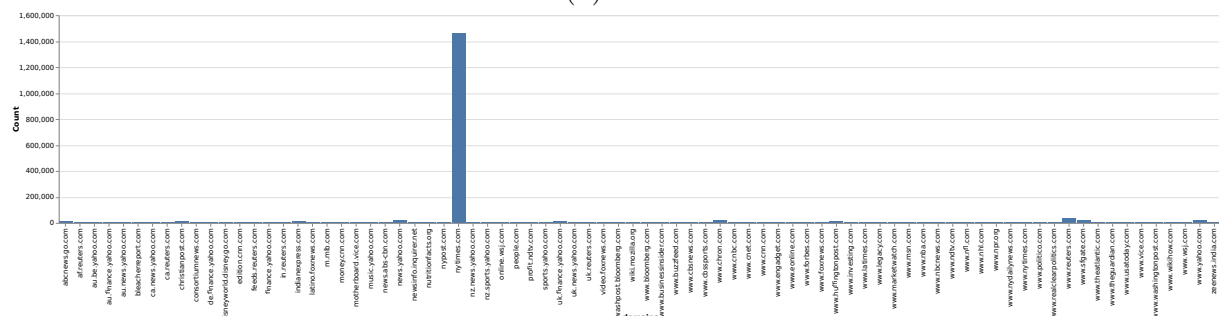
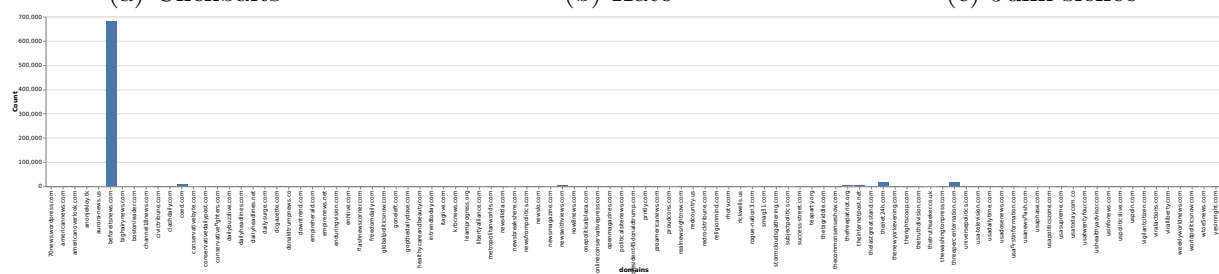
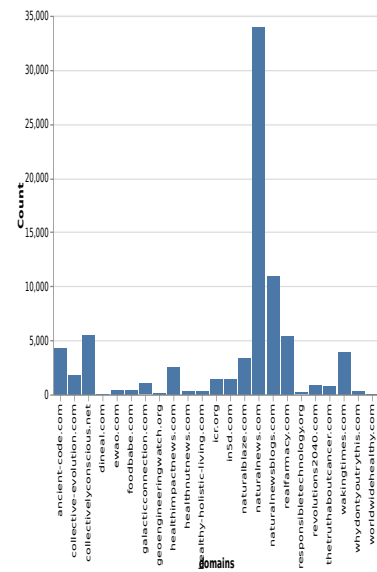
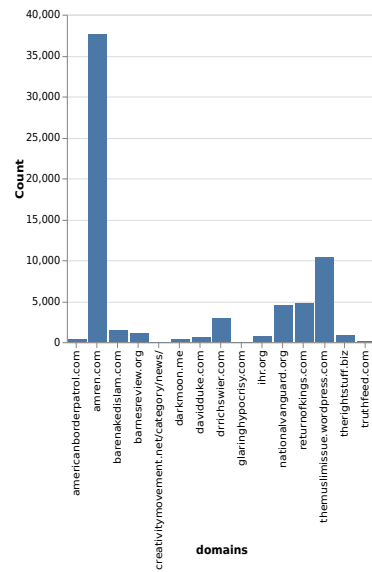
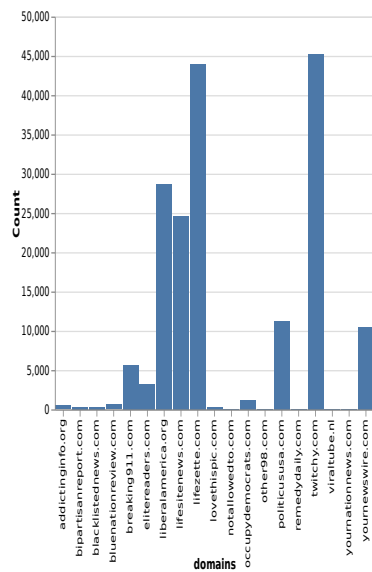
(b) Boxplot of number of sentences for each category.

Figure 2.2: Summary statistics

on stop punctuation such as dots or questions mark, but not on commas. The second step consist on filtering 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 alpha-numeric 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.

An other import feature to look at is the distribution of the number of words in the text. Indded, at some point it will be needed to fix a constant length for the texts and using to small length would mean a lot of cutting and using too long size would mean too much padding. It is thus needed to investigate the length of the texts in order to choose the right one. It can be seen at **Figure 2.4** that reliable news have slightly more words than fake news, but the difference is minimal.



Comparing **Figure 2.5** and **Figure 2.6** shows that even by removing *nytimes.com* and *beforeitsnews.com* news from the dataset still leave enough 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 going from around one half to around one fourth.



Figure 2.3: Histogram of news origin for each categories.

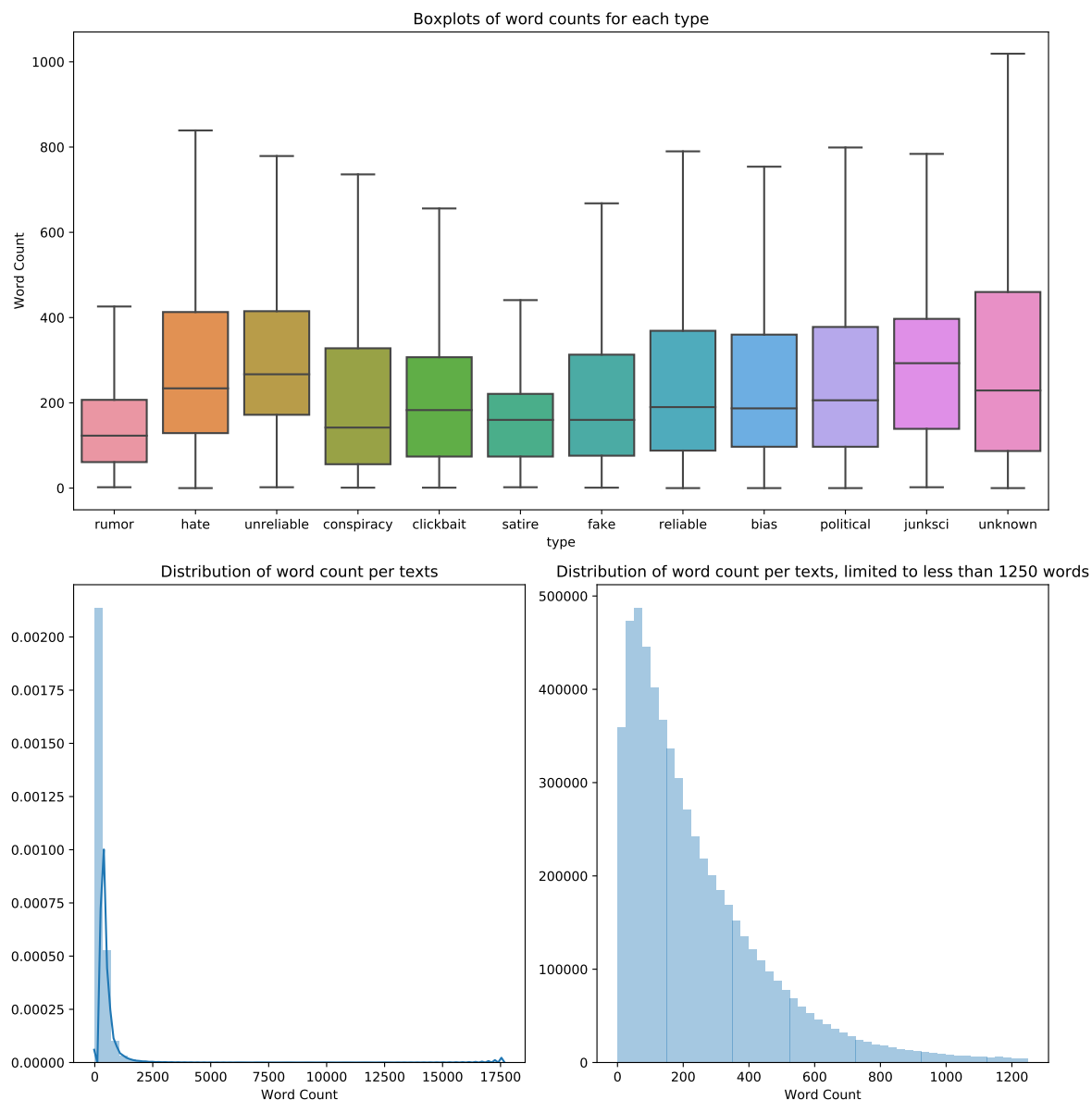


Figure 2.4: Distribution of the number of words per text

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 the ones in *fake news corpus*. The distribution of words length can be seen at **Figure 2.7**. In addition, this dataset is not unbalanced as the other corpus is, which means that precautions does not have to be taken.

2.3 Visualisation with t-SNE

In order to visualize the data it is needed to transform text in a numerical way and eventually reduce the dimension in order to allow it to be plotted on a 2D or 3D plot. Here TF-IDF (term frequency, inverse document frequency[25, 26]) is used. How it works will be details later on. This produce a sparse matrix with each document being

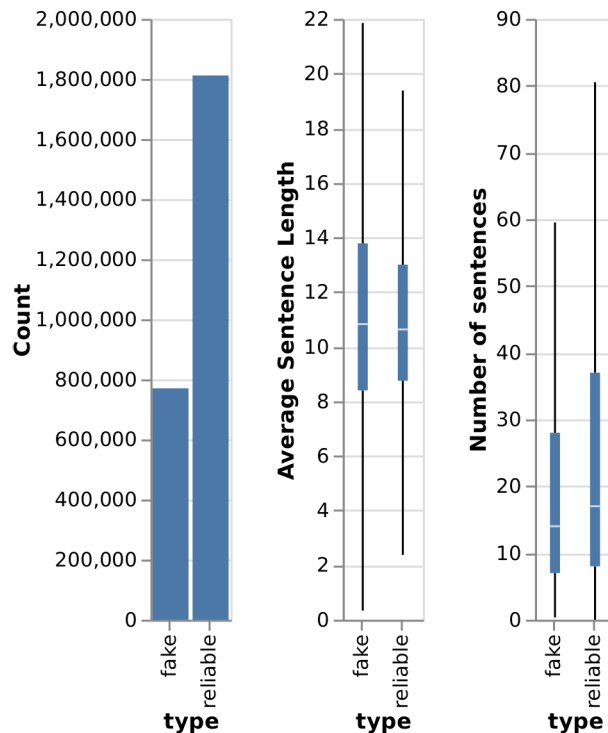


Figure 2.5: Summary statistics for not downsampled fake and reliable news.

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 before t-SNE[27] 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 $y_i, i \in 0 \dots N$ 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

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l} (1 + ||y_k - y_l||^2)^{-1}}$$

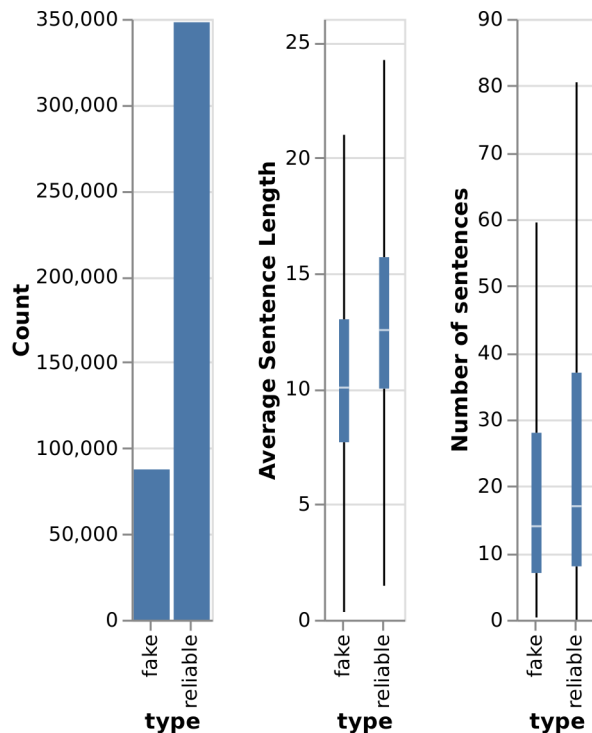


Figure 2.6: Summary statistics for downsampled dataset on fake and reliable news

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.

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

The minimization of KL divergence is acheived by gradient descent.

The documentation of the module of scikit-learn[28] used for computing this value recommand 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 2.8**. These 500 firsts PCA components explains around 47% of the total variance. The figure shows that there is no clear clustering of the classes.

Increasing the number of PCA components to 1750 gives the results at **Figure 2.9** and does not shows more clustering, even if it explains 75% of the variance. This shows that classifying the dots might not be easy, but it should be reminded that it is a dimensionality reduction and that there is a loss of information. Some of the original data dimension can have a better separation of the classes.

It is not possible to use t-SNE on the *Fake News Corpus* because the algorithm is quadratic with respect to the number of samples. Which make it impossible to compute for that corpus which is larger than the *liar-liar* corpus. But it is still possible to try to make some vizualisation using truncated singular v alues decomposition.

2.4 Conclusion

Data exploration have shown that there is no real statistical differences between text metadata for fake and reliable news, and thus make it not interesting for using it for

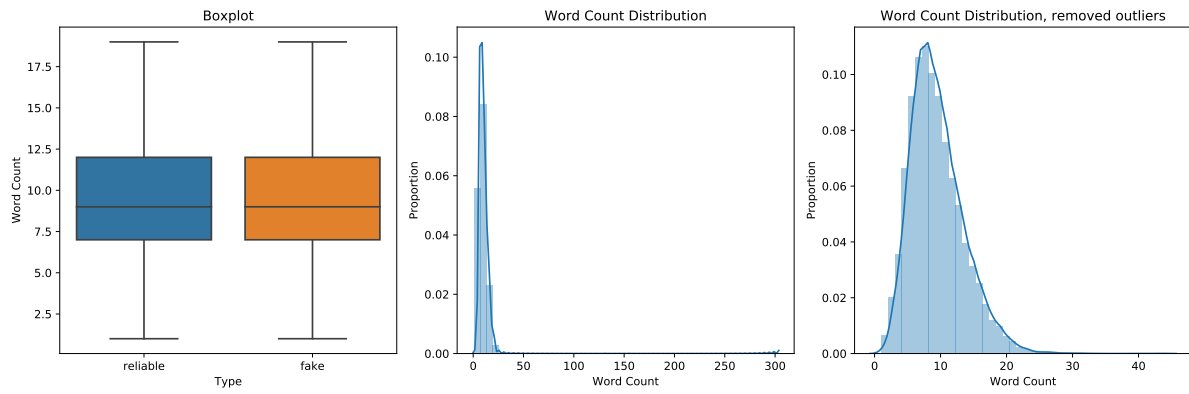


Figure 2.7: Number of words distributions for liar-liar dataset. On the first and the third plots, a few outliers with length greater than 50 have been removed in order to make the plots more readable.

classifying new texts. In addition, dimensionality reduction does not show any sign of helpfulness for the classification.

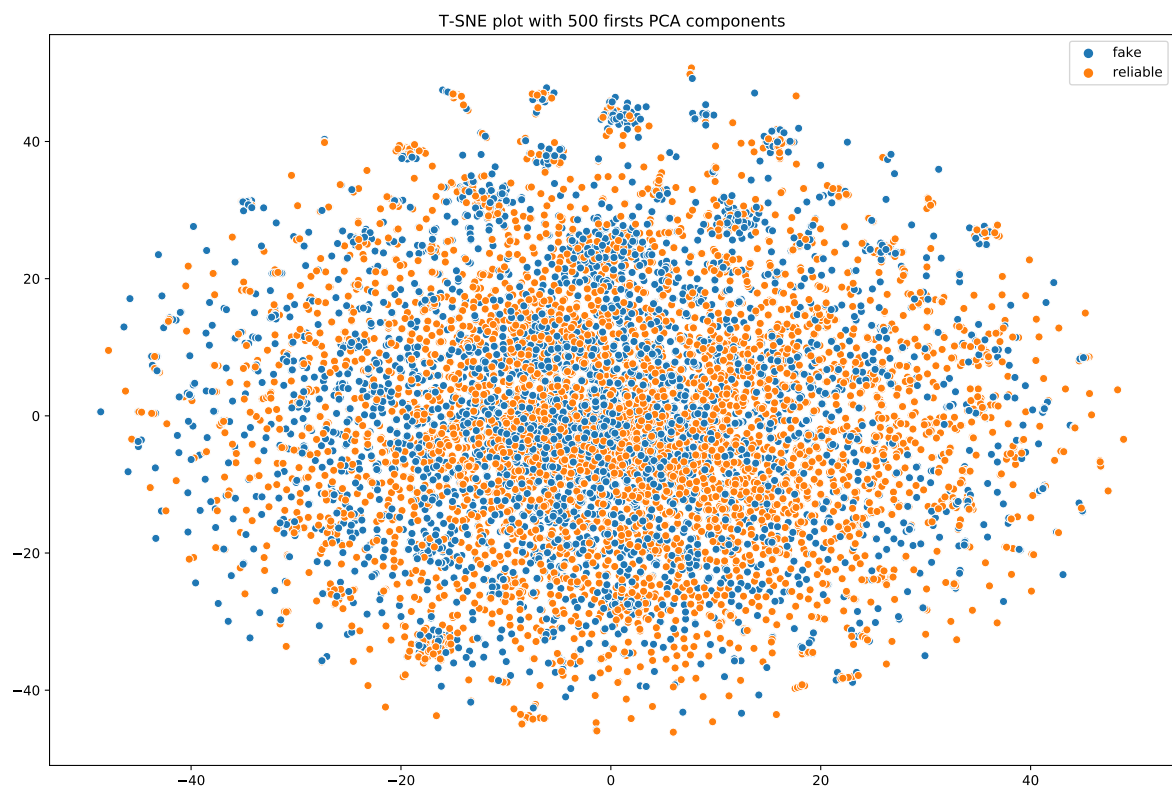


Figure 2.8: t-SNE plot for liar-liar dataset.

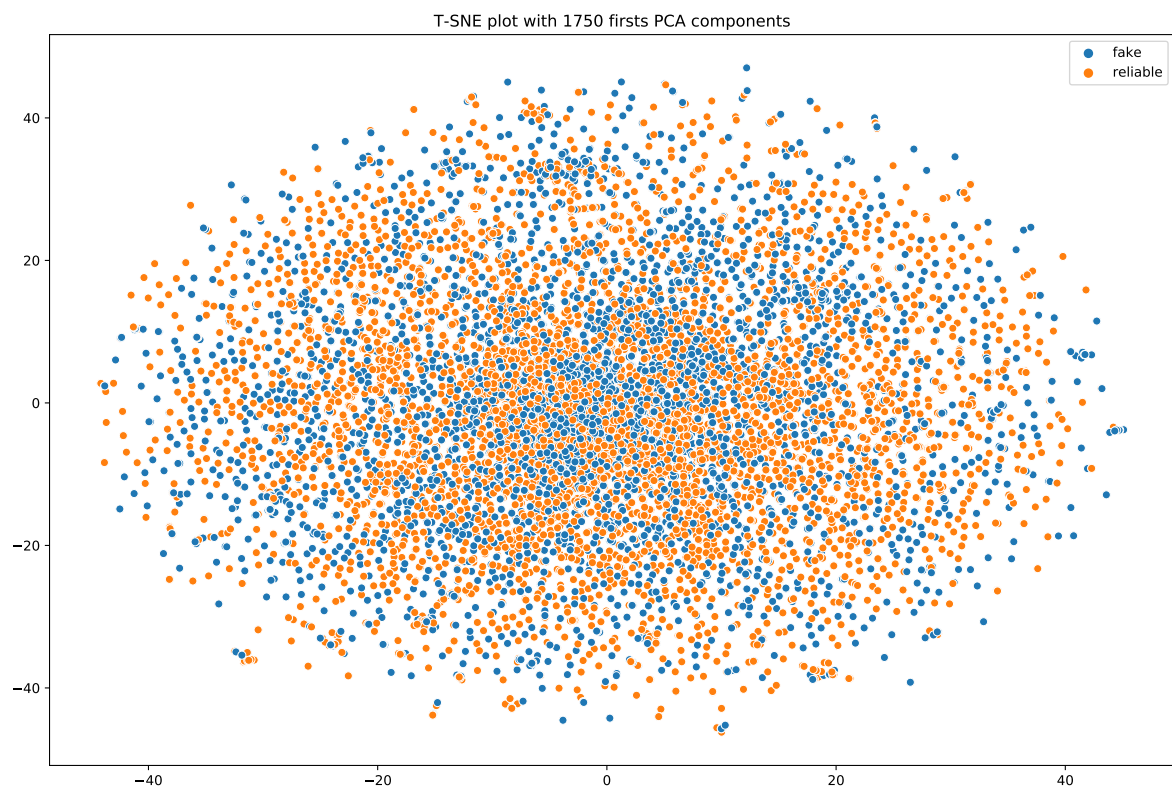


Figure 2.9: t-SNE plot for liar-liar dataset.

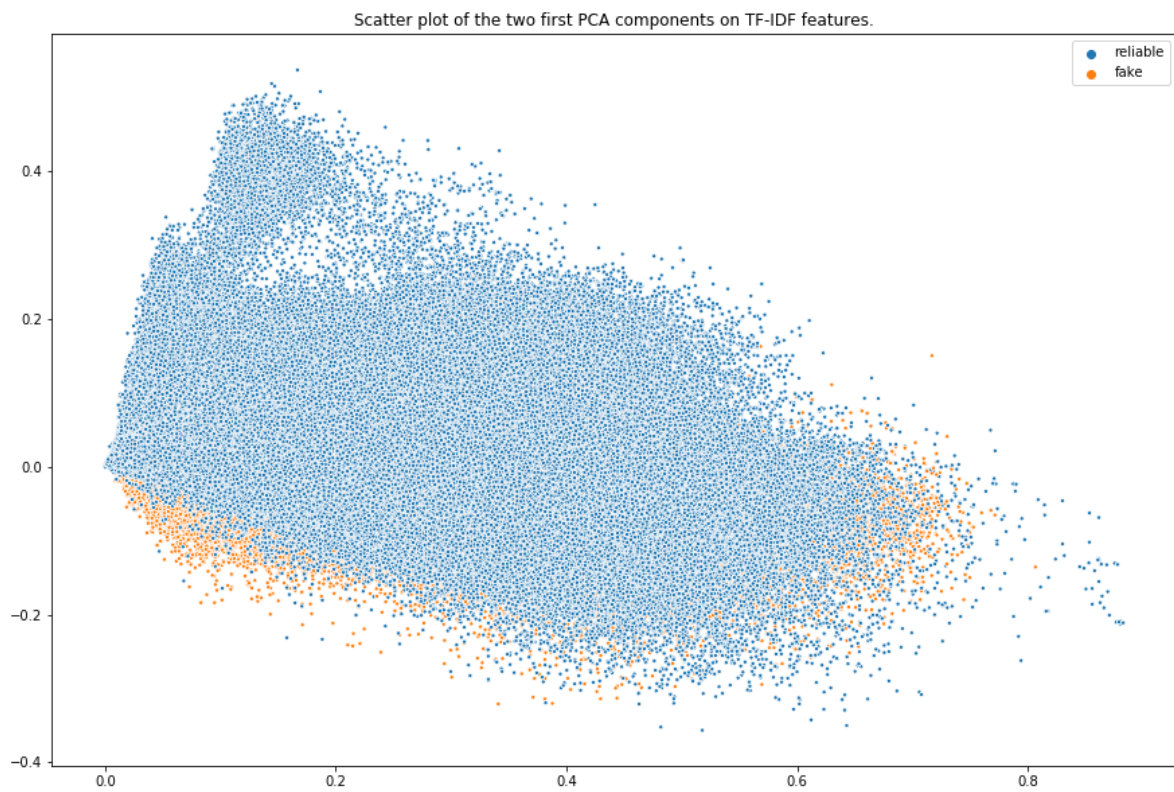


Figure 2.10: First two LSA compnoents for fake news corpus

Chapter 3

TODO: find a title

3.1 Introduction

In this chapter, we will focus on the more traditional methods used in natural language processing such as Nave-Bayes, decision trees, linear SVM and others. These will serve as baseline for comparing the performances of the more two advanced models that will be analyse later on: LSTM and Attention Mechanism. The first thing to do when working with text is the do words and texts embedding, indeed, in order to use machine learning algorithms on texts, a mathematical representation of these texts is required.

3.2 Words and text embedding

As explained before, text need to be represented in a way that gives more meaningful information than a simple sequence of bits, which have additional drawbacks such that for a given word, the sequence of bits representing it depends on the coding. The first and simplest coding that come to mind is a one-hot encoding: a matrix M of size number of texts \times number of words where $M_{ij} = 1$ if the word j is present in the text i and 0 in the other case. But this is till not enough as each word is given the same weight, no matter how often it appears in the text.

In order to overcome this problem, term-frequency might be used, that is, rather than setting M_{ij} to 0 or 1 we set it to the number of time it appears in the text.

It is possible to use even better text embedding. It is called term-frequency, inverse document frequency. The main idea is that a word that appears often in all the documents is not helpful in order to classify the documents. For example, if the task is to classify books of biology and physics, words atom, cell or light are more useful than today or tomorrow.

In order to compute tf-idf, it is separated in two part, the first one being term-frequency and the second one inverse document frequency. We have that

$$tf_{ij} = \#(W_j | W_j \in D_i) \quad (3.1)$$

That is, tf_{ij} is the number of time the word j appears in the document i . Secondly, we have that

$$idf_j = \log\left(\frac{\#D}{\#(D_i | W_j \in D_i)}\right)$$

This is the log of the total number of documents, over the number of documents that contains the word j . Finally, the value $tfidf$ value is computed by

$$tf - idf_{ij} = tf_{ij} * idf_j \quad (3.2)$$

This the text embedding methods that will be used in this section.

3.3 Methodology

All the methods presented will be tested in three different ways:

- On the liar-liar dataset
- On the fake corpus dataset, excluding the news from *beforeitsnews.com* and *ny-times.com*
- On the fake corpus dataset, including the two previous domains.

To be more precise, in the first case, the models will be trained on a training set, tuned using validation set and finally tested using test set. In the second case, the same methodology will be used, the dataset have been split by choosing 60% of the text from each domain for training, and 20% for validation and testing. This way of splitting has been chosen because of the uneven representation of each domain in the dataset in order to ensure representation of all the domains in the three subsets.

In the last case, the model is trained on all the news but the two domains previously cited and tested on these two domains.

3.3.1 Evaluation metrics

In order to evaluate each model, multiple evaluation metrics have been used. There are recall, precision and f1-score. It is needed to use multiple metrics because they don't all account for the same values. For instance, it is possible to have a model with a recall of 1 that behaves extremely bad because it simply classifies all the inputs in the same single class.

Remember that precision is defined as

$$Precision = \frac{TP}{TP + FP} \quad (3.3)$$

Which means that we can have two different precisions, depending on which classes are considered as being positive. This is the proportion of correctly classified positive elements over the number of elements classified as positive. It equals to 1 when there is no false positive, but it does not mean that all the positive elements are correctly classified as they might be some false negatives. The recall helps to solve these problems.

It is defined as

$$recall = \frac{TP}{TP + FN} \quad (3.4)$$

The f1-score combines the recall and the precision. It is defined by

$$f1 - score = \frac{2 * precision * recall}{precision + recall} \quad (3.5)$$

It is also possible to look at the weighted average of all these values. For instance, it is possible to compute a global recall by averaging the recall for both classes by the respective class ratio.

Finally, raw output can be used by looking at the confusion matrix.

The first parameters to tune is the max number of features used by tf-idf. This is the maximum number of words that will be kept to create the text encoding. The words that are kept are the most frequent words.

3.4 Models

Four models have been used in order to classify texts represented as a TF-IDF matrix. These are Multinomial Nave-Bayes, Linear SVM, Ridge Classifier and Decision Tree.

3.4.1 Naïve-Bayes[1]

The basic idea of Nave-Bayes model is that all features are independent of each other. This is a particularly strong hypothesis in the case of text classification because it supposes that words are not related to each other. But it works well given this hypothesis. Given an element of class y and vector of feature $\mathbf{X} = (x_1, \dots, x_n)$. The probability of the class given that vector is defined as

$$P(y|\mathbf{X}) = \frac{P(y) * P(\mathbf{X}|y)}{P(\mathbf{X})} \quad (3.6)$$

Thanks to the assumption of conditional independence we have that

$$P(x_i|y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i|y) \quad (3.7)$$

Using Bayes rules we have that

$$P(y|x_1, \dots, x_n) = \frac{P(y) \prod_{i=1}^n P(x_i|y)}{P(x_1, \dots, x_n)} \quad (3.8)$$

Because $P(x_1, \dots, x_n)$ is constant, we have the classification rule

$$\hat{y} = \underset{y}{\operatorname{argmax}} P(y) \prod_{i=1}^n P(x_i|y) \quad (3.9)$$

3.4.2 LIBLINEAR

Liblinear is a method for large linear classification. Given pairs of features-label (\mathbf{x}_i, y_i) , $y_i \in \{-1, 1\}$, it solves the following unconstrained optimization problem.

$$\min_w \frac{1}{2} \mathbf{w}^T \mathbf{w} + \mathbf{C} \sum_{i=1}^l \xi(\mathbf{w}; \mathbf{x}_i, y_i) \quad (3.10)$$

Where ξ is a loss function, in this case L2 loss function has been used, and $\mathbf{C} > 0$ a penalty parameter.

Class of new examples are assigned by looking at the value of $\mathbf{w}^T \mathbf{w}$. The class 1 is assigned if $\mathbf{w}^T \mathbf{w} \geq 0$ and the class -1 if $\mathbf{w}^T \mathbf{w} < 0$.

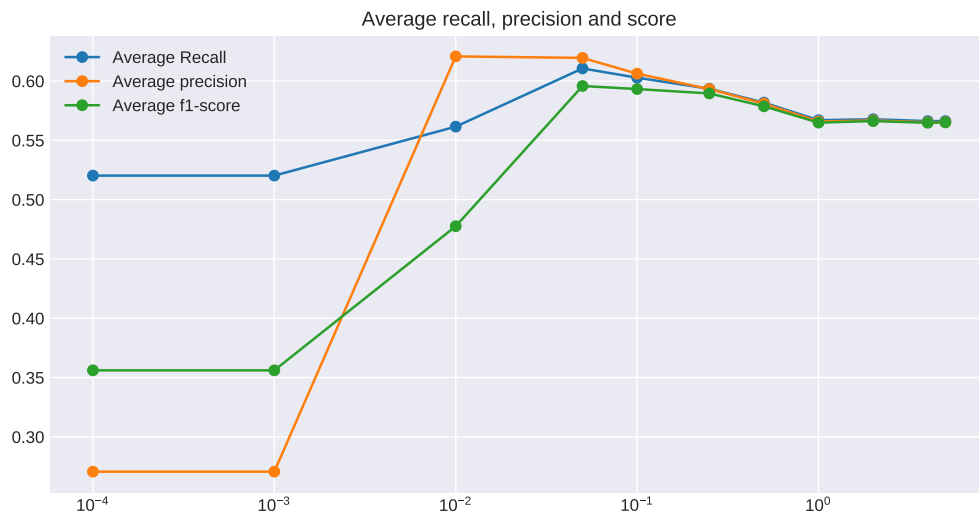


Figure 3.1: Tuning linearSVC parameters

3.5 Models on liar-liar dataset

3.5.1 Linear SVC

In the case of linear SVC there is one parameters to tune up, which is the penalty parameters for the error term. **Figure 3.1** show the tree main metrics with respect to the panlity parameter. This show that a parameters of around 0.1 is the best one.

3.5.2 Decision Tree

With decision tree, it is possible to reduce overfitting by pruning the tree. It is possible to do preprunning or post pruning. Pre pre preuning means that a stoping criterion is used to stop tree growth earlier and post pruning cut the tree once it have been fully grown. It this case pre pre pruning is done by limiting the maximum depth of the tree.

Figure 3.2 shows metrics values for different depths. It seems that tree of depth 1000 are the best ones.

3.5.3 Max Feature Number

Starting with the maximum number of features, the precision of each model can be analyzed when limiting the maxium number of words. The results for each model can be seen at **Figure 3.3, 3.4** and **3.5**. These Figures shows that in the case of Nave-Bayes it is better to use the maximum number of features (which is the total number of words in the training set) when it is better to limit it to 500 or 1000 for ridge classifier and linear SVC.

Based on **Figure 3.5** we can say that when it comes to precision and recall, Nave-Bayes is the one that perform the best.

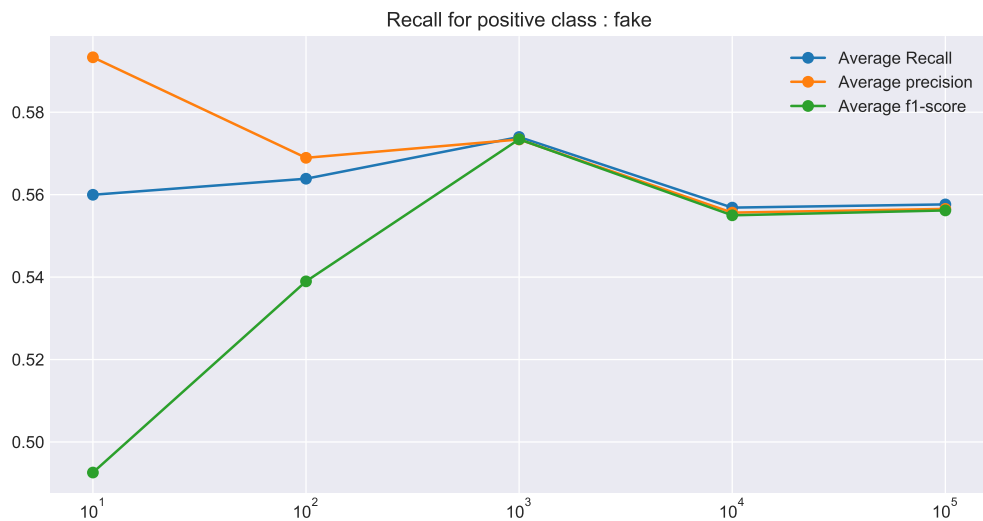


Figure 3.2: Tuning decision tree parameters

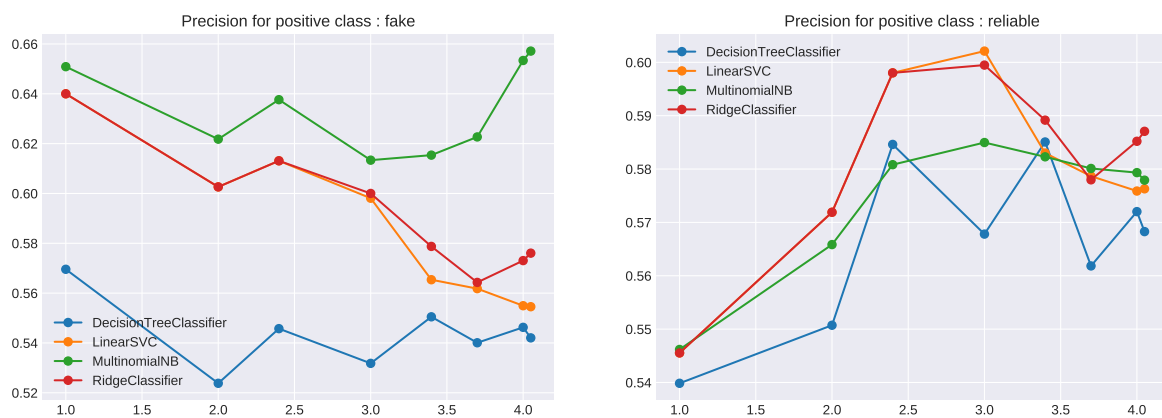


Figure 3.3: Precision of the model for each classes, the x axes is log scale of the number of features

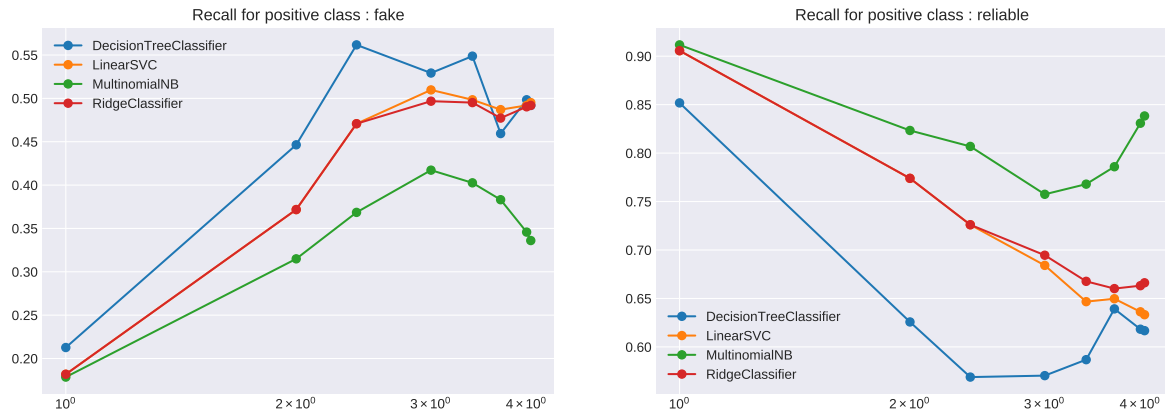


Figure 3.4: Precision of the model for each classes, the x axes is log scale of the number of features

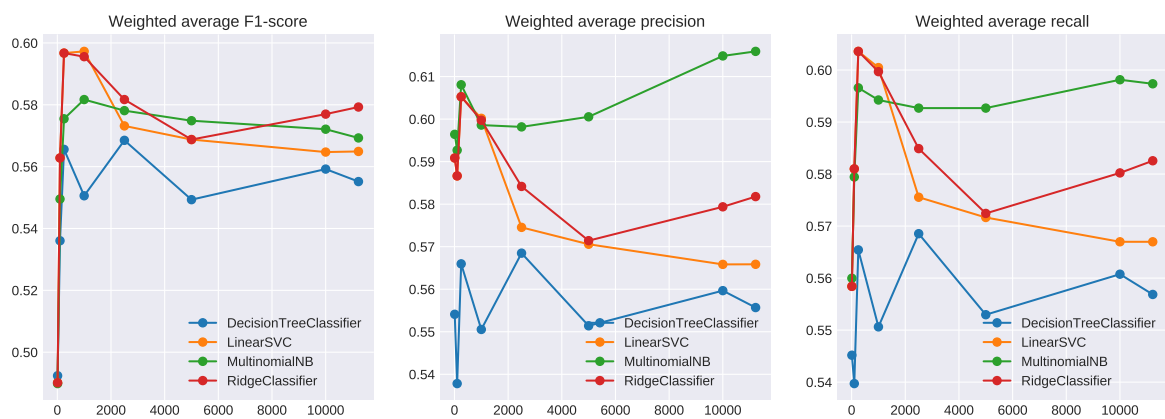


Figure 3.5: Weighted average of f1-score, precision and recall of each classes.

3.6 Models on fake corpus dataset

3.6.1 SMOTE: Synthetic Minority Over-sampling Technique[2]

As it have been show in **Chapter 2**, the fake news corpus is inbalanced. Synthetic minority over-sampling is a technique that allows to generate fake samples from the minor class. It works by randomly choosing one or many neirest neighbor in the minority class. For instance, if the algorithm is set to use 5 nearest neighbours, for each samples it will choose one of its nearest neighbours, and generate a new sample on the segment joining the sample and its neighbour.

Data: k = Number of nearest neighbors
Data: T = number of minority class samples
for $i \leftarrow 1 \dots T$ **do**
 | Compute k -nearest neighbors of sample i ;
 | Populate(knn , i);
end

Algorithm 1: SMOTE

Data: knn = the k nearest neighbour of sample i
Data: s = i th sample
 $nn = \text{random_choice}(knn)$;
 $\text{newSample} = s + \text{rand}(0, 1) * (nn - s)$;

Algorithm 2: Populate

Algorithm 1 and **2** shows how it works. The first one compute the k -nearest neighbors and the second one compute a new element by randomly choosing one of this neighbors.

Bibliography

- [1] Harry Zhang. The Optimality of Naive Bayes. page 6.
- [2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique.
- [3] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. In *Journal of Economic Perspective*, volume 31, 2017.
- [4] Jeffrey Gottfried and Elisa Shearer. *News Use Across Social Medial Platforms 2016*. Pew Research Center, 2016.
- [5] Craig Silverman and Lawrence Alexander. How teens in the balkans are duping trump supporters with fake news. *Buzzfeed News*, 3, 2016.
- [6] 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.
- [7] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
- [8] Maciej Szpakowski. Fake news corpus. <https://github.com/several27/FakeNewsCorpus>. Accessed: 2018-10.
- [9] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fak- enewsnet: A data repository with news content, social context and dynamic infor- mation for studying fake news on social media. *arXiv preprint arXiv:1809.01286*, 2018.
- [10] Kai Shu, Suhang Wang, and Huan Liu. Exploiting tri-relationship for fake news detection. *arXiv preprint arXiv:1712.07709*, 2017.
- [11] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection.
- [12] 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.
- [13] Vernica Prez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Au- tomatic detection of fake news.

- [14] 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.
- [15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 1997.
- [16] 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.
- [17] Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. Independently recurrent neural network (indrnn): Building a longer and deeper rnn.
- [18] 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.
- [19] 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.
- [20] 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.
- [21] Yoon Kim. Convolutional Neural Networks for Sentence Classification. *arXiv:1408.5882 [cs]*, August 2014. arXiv: 1408.5882.
- [22] 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.
- [23] 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>.
- [24] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.
- [25] Stephen Robertson. Understanding inverse document frequency: On theoretical arguments for idf, 2004.
- [26] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 2004.
- [27] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, (9):2579–2605, 2008.

- [28] 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.

Appendix A

Appendix to Chapter 3

A.1 TF-IDF max features row results

	model	max_features	recall	precision	f1
0	LinearSVC	10	0.558411	0.590855	0.490108
1	MultinomialNB	10	0.559969	0.596418	0.489844
2	DecisionTreeClassifier	10	0.545171	0.554105	0.492387
3	RidgeClassifier	10	0.558411	0.590855	0.490108
4	LinearSVC	100	0.580997	0.586645	0.562808
5	MultinomialNB	100	0.579439	0.592686	0.549533
6	DecisionTreeClassifier	100	0.539720	0.537812	0.536041
7	RidgeClassifier	100	0.580997	0.586645	0.562808
8	LinearSVC	250	0.603583	0.605262	0.596719
9	MultinomialNB	250	0.596573	0.608079	0.575478
10	DecisionTreeClassifier	250	0.565421	0.565966	0.565582
11	RidgeClassifier	250	0.603583	0.605262	0.596719
12	LinearSVC	1000	0.600467	0.600183	0.597274
13	MultinomialNB	1000	0.594237	0.598593	0.581692
14	DecisionTreeClassifier	1000	0.550623	0.550539	0.550578
15	RidgeClassifier	1000	0.599688	0.599731	0.595559
16	LinearSVC	2500	0.575545	0.574544	0.573173
17	MultinomialNB	2500	0.592679	0.598169	0.578113
18	DecisionTreeClassifier	2500	0.568536	0.568482	0.568508
19	RidgeClassifier	2500	0.584891	0.584169	0.581693
20	LinearSVC	5000	0.571651	0.570574	0.568765
21	MultinomialNB	5000	0.592679	0.600539	0.574854
22	DecisionTreeClassifier	5000	0.552960	0.551400	0.549321
23	RidgeClassifier	5000	0.572430	0.571418	0.568760
24	LinearSVC	10000	0.566978	0.565837	0.564715
25	MultinomialNB	10000	0.598131	0.614854	0.572109
26	DecisionTreeClassifier	10000	0.560748	0.559664	0.559212
27	RidgeClassifier	10000	0.580218	0.579376	0.576984
28	LinearSVC	11222	0.566978	0.565860	0.564915
29	MultinomialNB	11222	0.597352	0.615925	0.569280
30	DecisionTreeClassifier	11222	0.556854	0.555689	0.555171
31	RidgeClassifier	11222	0.582555	0.581782	0.579279