SZAKDOLGOZAT

Vall Petra

2019
Analysis of automated fake news detection

Készítette: Vall Petra
Gazdaságinformatikus alapszak
2019

Szakszeminárium vezető: Dr. Racskó Péter
# Table of Contents

1  Introduction .................................................................................................................. 5

2  Generic Perspective .................................................................................................... 6

   2.1  The Definition of fake news ................................................................................. 6

   2.2  Fake News’ Importance today ............................................................................. 7

   2.3  Key Players .......................................................................................................... 8

       2.3.1  Twitter ......................................................................................................... 8

       2.3.2  Facebook ..................................................................................................... 8

   2.4  Known Motifs ....................................................................................................... 9

       2.4.1  Political interest ............................................................................................ 9

       2.4.2  Commercial Interest ..................................................................................... 10

       2.4.3  „Fun” in misleading others ........................................................................ 10

3  Strategies for disseminating fake news ...................................................................... 11

   3.1  Messaging and Valence Strategies ...................................................................... 11

   3.2  Communication Strategies .................................................................................. 13

4  Algorithmic Perspective ............................................................................................ 15

   4.1  Fake News as a Machine Learning problem ....................................................... 15

   4.2  Machine Learning ............................................................................................... 15

       4.2.1  Supervised Learning .................................................................................... 16

       4.2.2  Unsupervised Learning ............................................................................... 17

       4.2.3  Reinforcement Learning ............................................................................. 17

       4.2.4  Deep Learning ............................................................................................ 17

   4.3  Natural Language Processing ............................................................................... 18

5  Fake News Detecting algorithms ................................................................................. 19

   5.1  Fake News Detecting Algorithm Categories ...................................................... 19
5.1.1 Content-based filtering tools..............................................................19
5.1.2 Social Context-based filtering tools......................................................25
5.1.3 Content and Social Context ...............................................................30
5.2 Conclusion............................................................................................35
6 Strategies for fake news detection today .....................................................36
   6.1 Government level ............................................................................37
   6.2 Social media platform level .............................................................38
7 Summary..................................................................................................42
References...................................................................................................45
Table of Figures ........................................................................................48
1 Introduction

Misinformation is not a new phenomenon. The manipulation of public opinion has been around for a long time, just like low trust in politicians. However, what puts this topic into a new perspective is the speed and scale of spreading misinformation. The internet makes it possible to disseminate information and misinformation, propaganda, fake news and hate speeches with such velocity that was till now unimaginable.

Fake news and their impact on society has grown greatly throughout the last couple of years. The information and misinformation that surrounds us on the internet, play lately a major role in how we see the world, it forms our opinion, our political views, it forms our way of acting upon our democratic rights. Social platforms like Twitter, Facebook, Instagram – just to name a few - make it possible to spread news (including junk and fake news) faster and easier than ever.

The spread of the fake news has a potential threat on the individuals as well as on our society and our democracy, which places the topic to the centre of interest lately. The 2 biggest events that drew attention to this problem were the 2016 US Election and 2016 Brexit. Most of the studies found in this topic were also prompted by the aforementioned political actions.

The amount of information – and misinformation - and the rapidity of its dissemination is asking for new perspectives, asking for new solutions, maybe a new automated filtering system - just like spam filtering for emails - to make the public’s life easier on being able to differentiate fake and true news, fake and trustworthy sources.

The purpose of this paper is to see if there are already solutions for this topic due to its freshness and if there are, to provide a brief introduction to each of these tools. I am mainly investigating tools which use machine learning algorithms in order to filter misinformation.

To understand the complexity of this topic, I start with providing an overall picture on fake news focusing on its presence on social media platforms, the reasons for creating fake news, how and why it became a pressing issue lately. I will then take a brief look at what machine learning (ML) means and how we can categorize the different machine learning techniques so that later we can understand how we can use ML for fake news detection. I continue with
presenting my research results for existing fake news detection algorithms, including a short summary of the dataset used, a brief introduction to the algorithm and a test if possible.

In the later chapters, I take a look at the current situation in tackling misinformation. Following that, I summarize what in my opinion are the realistic ideas about fighting against misinformation considering my research results.

Due to the freshness of this topic, my thesis relies mostly on studies, news articles, and podcasts found online.

### 2 Generic Perspective

#### 2.1 The Definition of fake news

There is an ever growing threat and anxiety over the topic of fake news, disinformation and misinformation in today’s society. There is an endless flow of information on the internet and many times it’s hard to decide if the information is one that we should rely on or not. When trying to find a algorithmized solution to the problem of fake news, one should first start to think about defining what fake news really is.

Fake news is rather complicated to define. It is a huge umbrella problem including different types of issues and terms such as the ones mentioned above - misinformation, disinformation. Based on my research, I came across the below definitions:

Misinformation: the unintended spread of false or misleading information (Staines - Moy, 2018)

Disinformation:”intentional actions by individuals and groups that – either knowingly or unknowingly- result in the spread of false or misleading information” (Hwang, 2017)

Fake News: “a made-up story with an intention to deceive, often geared toward getting clicks.” (Tavernise, 2016)
Each of these terms have also different taxonomies and harms that they can cause. Therefore, when working on the problem of fake news, it’s extremely important to first try to narrow down the issue and try to define the exact problem that we would like to work on and solve eventually.

Coming from the above, the problem of fake news is formulated in different ways in almost each cases and solutions that I studied. Some might see it as a binary classification issue – meaning that an article is either considered true or false – some might see it as a multiclassification issue – listing different categories that one can use for classifying the data analysed - whereas some might focus on the questionable sources more than on the fake news itself and try to find a way to eliminate those.

They all however agree on the fact, that it is a pressing issue and requires attention and an action plan in order to protect our democracy.

2.2 Fake News’ Importance today

The need for fact checking and filtering misinformation became the centre of attention with the fast development of the internet and social media platforms and artificial intelligence. Fake news has existed in multiple forms for a long time but the rapid changes in Information Technologies in the past few years added 2 arguments that humanity might not have counted with, namely velocity and quantity.

Facebook, Twitter and Google play now such a major role in how people form their views about politics, economics, global warming and the world, that having unfiltered information available on these pages could be a threat to our democratic processes.

The term „Fake news” became widely used in 2016, at the time of the US presidential election and Brexit going on. The word „Fake news” was even chosen the word of the year by the Macquarie dictionary in 2016 (Shu et al.,2017).

2016 was the year, when people had to face with the threat of large volumes of fake news with the aim of shifting their beliefs from one direction to another. Some argue that the results of the above mentioned election and referendum mirror these changes in opinion, however there is no solid proof for such statement. One thing however became evident; there is a need
for a protection mechanism in order to mitigate the threats caused by fake news and to assure that people will still have a choice to form their opinion based on valuable and trustworthy resources. With the European elections in sight in May 2019, the European Union already created an action plan against disinformation which will be analyzed in a later chapter.

2.3 Key Players

Although misinformation is not only present in the online media, this paper focuses mainly on the filtering techniques that could be or are applied already on the internet. Therefore, only those platforms will be shortly introduced here, which have been identified as key players in spreading (mis)information online.

2.3.1 Twitter

One of the most influential social media platforms enabling us to post almost anything we wish, is Twitter. Twitter’s co-founder Evan Williams’s original idea about creating the platform was to enable everybody to speak up. According to him, „once everybody could speak freely and exchange information and ideas, the world is automatically going to be a better place.” (Streitfeld, 2017, id.:Hwang, 2017). Unfortunately, his idealistic view about society and social platforms might have just failed with fake news being spread in millions of tweets. Twitter is an online news and social networking platform, where post with a restricted length of 140 characters could be shared by any registered users.

2.3.2 Facebook

Facebook, celebrating its 15th birthday this year, grew into a company with reaching over 2.32 billion monthly active users by Q4 in 2018 (Statista, 2019). Its users are spending hours daily on the News Feed - containing the activity of a user’s network of friends. Facebook’s algorithm running behind what we see on News Feed has a huge impact on what reaches us content-wise and what does not.
Other platforms that are worth mentioning are Google, Reddit, Instagram, Mozilla, Youtube and Snapchat however, since the algorithms discussed later in this paper work mostly with Twitter and Facebook data, I will not go into details on the above listed sites.

2.4 Known Motifs

Tim Hwang discusses that the there are three major categories of actors and via that, three main intentions of spreading misinformation (Hwang, 2017). These are: political interest, commercial interest and merely for fun.

2.4.1 Political interest

As mentioned earlier, Fake News as a leading topic, came into picture with the 2016 US presidential election in the United States and with Brexit in Europe. Ever since these events, whenever there is a politically important action such as an election or a referendum, there is a spike in sharing political content, propaganda articles and fake news – with fake news winning the battle. If we check on a Facebook data analysis by BuzzFeed between February 2016 and the month of the US election, we can see that fake news (here the term is used for false election stories generated by hoax sites) outperformed thrustworthy mainstream media. On figure 1, 19 of the major news channels – listing New York Times, Washington Post, The Guardian, BuzzFeed and Huffington Post among others - supply dataset for Mainstream news. The analysis checked on 20 of the best-performing election stories from both sides – mainstream and fake news - in terms of Facebook engagement covering shares, reactions and comments.

Misleading the public can be beneficial for parties before elections. With the development of the internet and the techniques of spreading information, political parties also learned how to use the available technical tools to promote their ideas and/or suppress the opposition. Computational propaganda became a term lately, often being used when reading about fake news – it is the „use of automation, algorithms and big-data analytics to manipulate public life” (Howard and Wolley, 2016, id.: Bradsha & Howard, 2018).
2.4.2 Commercial Interest

The business model of the majority of the platforms mentioned above relies on advertisement. This means that they create content – many times false information – that attracts attention in order to boost traffic of the websites paying for the advertisements. Since this paper will mainly analyze studies that were prompted by politically driven social media manipulation, I will not go into further details on the tools and techniques, however it is important to understand, why misinformation is partly beneficial for the social media platforms.

2.4.3 „Fun” in misleading others

Trolling campaigns are merely about fabricating misleading information many times only for seeking entertainment. However, there has been evidence that in some cases, these groups were contacted to cooperate with spreading misinformation during election season (Hwang, 2017). These groups are mainly informal and there is usually no coordinated operation behind them.
3 Strategies for disseminating fake news

In the following chapter, I will analyse the different strategies for spreading misinformation out of political interest.

In Bradshaw and Howard’s paper (2018), after studying thoroughly how state actors use social media to influence the opinion of the public in 48 countries – including Hungary – they came up with the below categorization of cyber troop strategies:

3.1 Messaging and Valence Strategies

Use of online commentators

One form of social media manipulation is to use online commentators to get in touch with genuine users in a variety of ways such as using direct chat or leaving reactions under articles. The aim of these communications could be as follows:

- Spreading pro-government messages
- Attacking the opposition
- Neutral strategies or distracting

Use of trolls

Trolls usually target specific individuals, communities, or parties with various forms of hate speech. There is an evidence of state sponsored trolling campaign also in Hungary based on table 1. This table lists the analyzed 48 countries and their ways of political proaganda.
Table 1. Social Media Manipulation Strategies: Messaging and Valence (Bradshaw - Howard, 2018)

<table>
<thead>
<tr>
<th>Country</th>
<th>Fake Account Type</th>
<th>Pro-Government or Party Messages</th>
<th>Attacks on the Opposition</th>
<th>Distracting or Neutral Messages</th>
<th>Trolling or Harassment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angola</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argentina</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Armenia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Australia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Austria</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Azerbaijan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bahrain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brazil</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cambodia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>China</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colombia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cuba</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Czech Republic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ecuador</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Egypt</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Germany</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hungary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>India</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iran</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Israel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Italy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kenya</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kyrgyzstan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malaysia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mexico</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Myanmar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Netherlands</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nigeria</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>North Korea</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pakistan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Philippines</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poland</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saudi Arabia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Serbia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>South Africa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>South Korea</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Syria</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taiwan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thailand</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Turkey</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ukraine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UAE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>United Kingdom</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>United States</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Venezuela</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vietnam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zimbabwe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Authors’ evaluations based on data collected. Note: This table reports on the messaging and valence strategies of cyber troops. A filled box indicates evidence found. For fake account types: ● human accounts, ●● automated accounts ●●● cyborg accounts, ●●●● - no evidence found.
The accounts misleading the public and carrying out the above mentioned messaging and valence strategies could be further categorized as follows (marked accordingly in table 1.):

- Automated accounts
- Human accounts
- Hybrid or cyborg accounts

Automated accounts, also called as „social bots”, could be defined as pieces of software or code designed to mimic human behaviour online (Bradshaw - Howard, 2018). Davis (2016) defines it as „(…) a computer algorithm that automatically produces content and interacts with humans on social media”. They could be used for several purposes like spreading junk news or by „Astroturfing” – a term which boosting someone’s image by fake comments and thus fake popularity. Automatic content creation can be achieved by using different language models – one such language model is the GPT-2 (Radford et al., 2019) which is capable of generating coherent texts after being trained with unsupervised learning techniques (more on this under 4.2.2 Unsupervised Learning). The bigger the training data on a certain topic is, the more coherent and comprehensive text/article it can create.

Another term above that needs some explanation is the so-called Hybrid or cyborg account. These accounts are usually run by humans, many times in coordinated teams, where they use automation only partly in order to be more efficient in misleading the public.

### 3.2 Communication Strategies

Many times, instead of spreading fake news, cyber troops are attacking the opposition by reporting legitimate users falsely to have the accounts or portals disabled. Furthermore, government actors, after realizing the power of social media platforms, started to create their own sites and portals to either combat fake news or to spread them themselves depending on what is more beneficial for them or how „democratic” is democratic leading in the given country. The communication strategies for social media manipulation could be categorized as follows (Bradshaw - Howard, 2018):

- Targeted ads
- Task Forces, Portals or Applications
Table 2. gives us a presentation on whether these strategies are present in the country and if so, how these different communication strategies distribute.

Table 2: Communication Strategies for Social Media Manipulation (Bradshaw - Howard, 2018):

<table>
<thead>
<tr>
<th>Country</th>
<th>Content Strategies</th>
<th>Targeted Ads</th>
<th>Task Forces, Portals or Applications</th>
<th>Chat Apps &amp; Other Platforms</th>
<th>SEO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angola</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argentina</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Armenia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Australia</td>
<td></td>
<td></td>
<td>Counter Info Ops</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Austria</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>WeChat</td>
</tr>
<tr>
<td>Azerbaijan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bahrain</td>
<td></td>
<td></td>
<td>Astroturf</td>
<td>WhatsApp</td>
<td></td>
</tr>
<tr>
<td>Brazil</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cambodia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>China</td>
<td></td>
<td></td>
<td>Fact Checking</td>
<td>WeChat</td>
<td></td>
</tr>
<tr>
<td>Colombia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cuba</td>
<td></td>
<td></td>
<td>Fact Checking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Czech Republic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ecuador</td>
<td></td>
<td></td>
<td>Astroturf</td>
<td>WhatsApp</td>
<td></td>
</tr>
<tr>
<td>Egypt</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Germany</td>
<td></td>
<td></td>
<td>Counter Info Ops</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hungary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>India</td>
<td></td>
<td></td>
<td>Astroturf</td>
<td>WhatsApp, Telegram</td>
<td></td>
</tr>
<tr>
<td>Iran</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Israel</td>
<td></td>
<td></td>
<td>Astroturf</td>
<td>Telegram</td>
<td></td>
</tr>
<tr>
<td>Italy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kenya</td>
<td></td>
<td></td>
<td>Astroturf</td>
<td>WhatsApp</td>
<td></td>
</tr>
<tr>
<td>Kyrgyzstan</td>
<td></td>
<td></td>
<td>Fact Checking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malaysia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mexico</td>
<td></td>
<td></td>
<td></td>
<td>WhatsApp, Snapchat</td>
<td></td>
</tr>
<tr>
<td>Myanmar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Netherlands</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nigeria</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>North Korea</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pakistan</td>
<td></td>
<td></td>
<td>Astroturf</td>
<td>WhatsApp</td>
<td></td>
</tr>
<tr>
<td>Philippines</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poland</td>
<td></td>
<td></td>
<td>Astroturf</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russian</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saudi Arabia</td>
<td></td>
<td></td>
<td>Astroturf</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Serbia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>South Africa</td>
<td></td>
<td></td>
<td>Astroturf</td>
<td></td>
<td></td>
</tr>
<tr>
<td>South Korea</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Syria</td>
<td></td>
<td></td>
<td>Reporting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taiwan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thailand</td>
<td></td>
<td></td>
<td>Reporting</td>
<td>Line, WeChat</td>
<td></td>
</tr>
<tr>
<td>Turkey</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ukraine</td>
<td></td>
<td></td>
<td>Fact Checking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UAE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>United Kingdom</td>
<td></td>
<td></td>
<td>Counter Info Ops</td>
<td></td>
<td></td>
</tr>
<tr>
<td>United States</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Venezuela</td>
<td></td>
<td></td>
<td>Counter Info Ops</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vietnam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zimbabwe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Authors’ evaluations based on data collected. Note: This table reports on the observed strategies of cyber troops. A filled box indicates evidence found. For content: ◼ = content creation; ◼ = malevolent content takedown, ◼/◼ = no evidence found.
4 Algorithmic Perspective

4.1 Fake News as a Machine Learning problem

In liberal democracies, freedom of expression is part of the fundamental rights of people. The emerging technologies enable us to share content – including false stories, fabrications and strong political views - and there is currently no widespread censorship or filtering method that would prevent us from fake news.

There are several machine learning techniques that are in place and are used in our everyday life – like spam detection or identifying harsh, violent images – but there is currently no such obvious method that could be used against fake news. That raises the question whether fake news detection could really be a machine learning problem or if it is rather a complex issue that could not be handled like the aforementioned two other problems. In the following chapters, I will explain what machine learning means and will introduce one way of categorizing Machine Learning techniques.

4.2 Machine Learning

As most of the algorithms and studies analyzed in this paper include Machine learning algorithms, it is important to discuss this topic in general and see the different machine learning techniques for a better understanding of the vocabulary used in the later parts.

Based on a Udemy course from Portilla (2018), Machine learning (ML) could be explained and categorized the below described way : ML is a data analysis method for automating analytical model building. The algorithms learn from data, which then allows automatic search for hidden patterns without being said where to look. ML is already a widely used tool to filter data that surrounds us. As an example, it helps to detect violent images on social media platforms, it helps to filter spam in our mailboxes. Using machine learning in order to detect fake news sounds great however it is a much harder task as it was for the previously mentioned topics.
Machine learning process consists of the following steps (figure 2.):

Firstly, we need to **acquire** data relevant for the topic.

Secondly, we need to **clean** data – deleting unnecessary elements, standardizing data.

Thirdly, we need to **split** the data into a Training and a Test dataset. Training Data is used to train the algorithm on what to search for and Test Data is used to test how well our machine learning algorithm performs a task.

![Machine Learning Process](image)

Figure 2: Machine Learning Process (Portilla, 2018)

We can categorize Machine learning algorithms the following way: Supervised Learning, Unsupervised Learning and Reinforcement Learning Algorithms.

### 4.2.1 Supervised Learning

Supervised Learning algorithms work with labelled training data to predict a label of a test dataset element based off the training dataset.

The learning algorithm is first fed with a set of inputs along with the corresponding correct outputs. Then the algorithm try to map the input and the output values accordingly, finds the error and modifies the model accordingly.

Supervised learning uses patterns to predict the values of the label on the Test data through methods like classification, regression, prediction. This type of algorithm is typically used for analysis where there is historical data which can predict likely future events.
4.2.2 Unsupervised Learning

In the case of unsupervised learning algorithms, the training set contains unlabeled data and the algorithms’ task is to try to group together similar data points based off of different features. It is typically used against data that has no historical labels. The system does not know the „right answer” , the algorithm has to find the pattern by itself. The overall goal of unsupervised learning algorithms is therefore to find some structure within the data. Popular techniques include self-organising maps, nearest-neighbor mapping, k-means clustering and singular value decomposition.

4.2.3 Reinforcement Learning

Reinforcement learning algorithms learn to perform an action from experience and are often used for robotics, gaming and navigation. It discovers through trial and error which actions have the greatest reward. It has 3 primary components:

- The agent (the learner or decision maker)
- The environment (everything the agent interacts with)
- Actions (what the agent can do)

The agent tried to choose actions that maximize the expected reward over a given amount of time which can be reached the fastest if finding and following a good policy. When using reinforcement learning algorithms, the goal is to find the best policy out of all.

4.2.4 Deep Learning

Deep Learning is a subset of machine learning. It is a Neural Network, which is modeled after the biological neural networks, mimicking the network of neurons in our brain, attempting to enable computers to learn the same way as humans do. Neural Networks try to solve problems that would normally be easy for humans but hard for computers, such as natural language understanding. Neural Networks comprises of several layers, each of them containing nodes – or others, artificial neurons (Portilla, 2018).
Deep learning (DL) is an autonomous, self-teaching system that uses the training (input) data for finding patterns and then uses those patterns for making predictions (output) about new data. The simplest Neural Network consists of 3 layers: the input layer – accepting input data. The second layer, called hidden layer – performing calculations such as assigning weight to the nodes and then summing up the result. The third layer, the output layer, contains the predicted output. „Deep” in Deep learning refers to the many numbers of hidden layers inbetween the input and the output layer. DL algorithms are able to determine if their predictions are accurate or not and then make adjustments accordingly.

Some fields where Deep learning methods are broadly used are Pattern Recognition, Time Series Prediction or Anomaly Detection.

4.3 Natural Language Processing

In order to understand the below discussed algorithms, it is important to talk about Natural Language Processing. Natural Language Processing (NLP) covers the different approaches of how to process the human language. The term lately has been also widely used to refer to the study of computer systems that work on developing an interpretation of the naturally spoken languages. NLP machine learning algorithms can both be supervised and unsupervised learning.

Language processing is not a deterministic science as for example mathematics; the same language, same sentence or even the same word does not always have the same meaning, same interpretation. Due to the undeterministic nature of NLP, translating a language to a computer is rather a hard task, however, there has been a great development in both NLP techniques and machine learning algorithms lately to get closer to the solution to this problem.

Machine learning for NLP and text analytics involves a set of statistical techniques in order to identify parts of speech, sentiments and other aspects of the text. When talking about supervised NLP machine learning algorithms, they are usually used for categorization and classification of texts or parts of texts. Unsupervised NLP learning methods are used for clustering similar documents together.
5 Fake News Detecting algorithms

In the below chapters, I will present my research results via looking at different areas of fake news detection that were already algorithmized. I will investigate how the data and the labels were collected and created, what aspects of this problem were considered as research questions and provide a brief introduction to the algorithm used. Additionally, I will also present test results if possible.

5.1 Fake News Detecting Algorithm Categories

During researching this topic, I found that the fake news detecting algorithms not only focus on the article’s content but also on the article’s dissemination network as well. One can learn a great deal about the nature of an article only by looking at the accounts and their networks spreading the article. The tools described rely on a combination of content-based algorithms and social context-based algorithms - having an emphasis either on the prior or on the latter. In the following chapters, I will introduce examples to how machine learning could be used in filtering fake news or spreading misinformation - grouped in the earlier described categories, discussing the data that was used to train and test the algorithm and the simplified theory behind the algorithm. Lastly, I will provide test results if possible for evaluation purposes.

5.1.1 Content-based filtering tools

5.1.1.1 Full Fact

Currently, there are a number of organizations that – as a solution to stop spreading fake news – offer fact checking. These organizations mostly help the work of journalists but there are some which let access to their research to the public as well. One such organization is UK’s independent fact checking charity called „Full Fact“ (Polich, 2018a). The charity operates since 2010, to provide information as close to real time as possible in order to give the people the possibility to make up their own minds on problems that they care about. Full
Fact is able to do real time fact checking at the time of elections (during speeches), however, their every day operation is rather to look at the most important news, the most influential claims that is around us and trying to take it back to the primary sources to present the news from there.

Full Fact currently chooses their topics based on what is trending on social media, political TV shows, press releases in the UK. Since they aim to be as unbiased as they can, they also monitor what topics they selected to work on, how many topics they picked from one party and how many from the other and much more to have in depth metrics to keep the balance between the different wings.

To check closer how Full Fact is using automation, we need to break down their day to day process into 5 major parts. These are:

- Monitoring daily media
- Spotting claims in the daily media
- Checking on the claim by calling people up
- Publishing
- Intervention via asking people to correct the records needed

Full Fact uses automation for the first 2 steps: monitoring and spotting claims plus one extra field: spotting repeated instances that have been already checked.

**Dataset:**

Their model is based on 25 000 annotations (labels) from 80 volunteers. The volunteers were asked to take sentences from political TV shows and label the sentences according to a 7 type taxonomy that was created by the Full Fact team. The taxonomy they came up with is a result of their 8-year long research in the field, by looking at a great amount of examples and then trying to classify them into supergroups and subgroups, going through a lot of trial and error. Once the volunteers labeled the data, they could begin with building their algorithm.
**Algorithm:**

Full Fact uses ML and NLP for detecting claims by extracting them from the news being monitored. In the first step of the factchecking, the issue is being treated as a binary classification issue. The supervised ML algorithm is trained on a dataset that has sentences labeled as „claim” and „not claim” and then decides further on the algorithm to learn the characteristics of both types of the labelled sentences. The organization has a 7-type taxonomy their ML algorithm is based upon to define what is a claim and what is not – some of these types are *Quantity in past or present*, *Legal claim*, *Prediction*, *Causation or Correlation*. The company currently has 2 ML-based products to help their everyday life: Live and Trends.

**Live**

Live is a live factchecking tool that does speech detection in real time and generates a transcript of it. The tool’s task can be split into 3 categories:

- *First*, it checks whether there is a fact check already for the claim spotted in the speech. If yes, it will surface that fact.
- *Secondly*, in case there is no fact check on the topic yet, it will search for existing data on the claim. If Live then finds data, it generates a graph of the latest data. A good example would be a claim on employment. A claim, such as *employment has fallen with 5 % in the past x years* could be easily checked by data from national statistical data.
- *Thirdly*, if none of the above is true, the tool will check if there is a claim at all in the speech for further checking.

Full Fact’s approach to the topic of misinformation is that instead of restricting information spread, the focus point should be on how to reduce the time that it takes to respond to misinformation by helping real journalism.
**Trends**

Trends is a tool that was built to find repeated instances of fake news over time. It checks the repetition of claims that are incorrect, checking the claims said in the Parliament, in public political facebook groups, in the newspapers and then tries to map the spread of misinformation that were already checked. Its purpose is to be able to go back to the originator and then try to take out the fake news from the circulation to try to minimize the repetition.

**Test:**

Live and Trends are being built internally and other than Full Fact’s team, only fact checkers and journalist can access them. However, Full Fact’s team is continuously publishing studies in the topic which are available on their website.

**5.1.1.2 FakerFact**

The below described study considers the fake news problem as a multiclassification issue, meaning that it does not say whether the news in question is *fake* or *true* but rather distinguishes multiple categories and tries to fit each articles into the best category. In Mike Tamir’s project (Polich, 2018b), the researchers focus on identifying fake news at the early stage without the help of checking on the dissemination networks – that is, as soon as the article is spotted, possibly before spreading the article would start. Fakerfact is a website and also a Chrome/Firefox plugin which leverages ML techniques to predict the category of a previously unseen website/article with high level sentiment analysis (an NLP task) via classes like *opinion*, *wiki* and *fake news*. Their starting point is not trying to figure out whether something is true or false, since they do not expect to find a lot of data about „new news”. Instead of going straight to fact checker websites (like Full Fact), they take a look at how the article itself is written, such as what intention was the article written with, e.g. sharing information, manipulation and so on.
Once a link is entered on the website, the algorithm will go through every word and different natural language understanding techniques its model was trained on. Finally, it comes up with a score to sort the article into one of the below 6 categories:

- Journalism (news)
- Wiki (meant to communicate information but not necessarily news)
- Sensational
- Opinion
- Satire
- Unreliable

**Dataset:**

In the early phase, the research team worked with an available public dataset, which they then started to label and then ended up with the taxonomy above. The team allows users to leave feedback after running an article through the page since “wisdom of the crowd” is also essential in this project. “Wisdom of the crowd” is the notion that the aggregated observations of many users will help to weed out inaccuracies and falsehoods. Implies that with a sufficient number of users, the user-generated content on a platform will essentially be self-filtered for truthful information” (Hwang, 2017). The idea might sound familiar as this was the original thought behind two of the biggest social media platforms: Twitter and Reddit.

**Algorithm:**

Different strategies work well with the different categories, meaning that the 6 labels are trained on different datasets, on different ways. The algorithm that is used behind the website relies on Deep Learning algorithms such as Long Short Term Memory network (LSTM) and Attention Mechanism.

For describing these algorithms, I used the publications from Britz (2016), Blier-Ollion (2016), Olah (2015), Skymind (2018a) and Veen (2016). Both Attention Mechanism and LSTM are a variants of Recurrent Neural Networks (RNNs). RNNs are one of the many
neural network/Deep Learning architectures defined to map input sequences to output sequences for recognition or prediction problems using time-series information. RNNs are very useful when trying to solve problems like recognizing patterns in handwriting. Recurrent Networks not only take the current input but also have the ability to look back at previous inputs and come to a conclusion on what the output data should be combining past and present data. From this point of view, it seems like the RNNs have a memory. RNNs are able to find correlations between present and past events – these dependencies are called „long-term dependencies”. All these past events are stored in the hidden layers of the RNNs, with a certain weight assigned to each event. RNNs’ biggest obstacle is the vanishing gradient problem (also referred to as the exploding gradient problem) which refers to the problem of how information gets rapidly lost over time, independently of the weight assigned to the pieces of information.

In the mid 90s, a variation of RNNs was proposed which offered a solution to the aforementioned problem: the Long Short-Term Memory units (LSTM). LSTM helps to preserve the weight information for much longer than RNNs. LSTMs contain information outside of the normal flow of the RNN in a gated cell. „The cells learn when to allow data to enter, leave or be deleted through the iterative process of making guesses, backpropagating error, and adjusting weights via gradient descent” (Skymind, 2018b).

The next big step in RNN is Attention Mechanism. The benefit of Attention Mechanism is that unlike LSTMs, it does not try to encode the full input (combination of previous and new data). With Attention mechanism, we look at every current and previous input data before deciding what to focus on.

Based on the above, we can conclude that assigning weight to the input data is a crucial part of Neural Networks, however, the different NNs use different credit assignment logic to channel the input and output data.

Test:

For testing purposes, I ran into a problem that many of the researches also described – it is not easy to find fake news articles when one is specifically looking for them. Finally, I found the article that later proved to be false and was still available – an article that stated that the
Pope endorsed Donald Trump. However, when pasting the link into the website, it did not give the expected result – it did not see any warning signs that the article could be fake. I then re-run the algorithm for a second fake news article where it worked better, presenting the result shown on figure 3:

![Image of testing FakerFact as a fake news detecting website](image_url)

Figure 3: Result of testing FakerFact as a fake news detecting website (FakerFact, 2019)

5.1.2 Social Context-based filtering tools

5.1.2.1 Hoaxy

Hoaxy is a tool that shows us interactively how information spreads online providing information also about how likely an account is to be bot. Based on an interview with Filippo Menczer (Polich, 2018c), there are 2 ways of how information spreads across the social media platforms: they either spread organically or artificially. The organic way, is the normal spread, when people talk to each other and exchange information online. The other way is the artificial spread - Menczner defines artificial information spreads as the networks which root back to social bots.
Menczer works with information graphs and states that based on the characteristics of the information dissemination networks, one can see whether we are talking about organic or artificial information spread. The graphs are built up in a way in which the nodes are the people / accounts on social network, and the edges or links between these accounts are the spreads of these pieces of information between the accounts. A piece of information could be a username, an article and html link, a hashtag or basically any kind of new information. Studying the information shared and the structure of the graphs (like the number of connected components, how dense the network is, whether it looks like a tree or a star…) would then lead him to a conclusion whether the news and the accounts are fake or not.

**Dataset:**

Menczer uses Twitter data for his research; among others, he stores feature information about users such as the number of connections, retweets and mentions.

**Algorithm:**

ML algorithms can be built once we are able to distinguish between the organic patterns and the artificial ones. Based on Mentzer’s early research, when he started to do his research study on fake news dissemination, this accuracy was very high but over time, the techniques/algorithms used by the bots became more sophisticated and therefore harder to spot. The ML techniques that were used during the research were mostly supervised learning algorithms using random forest method and later started to analyze the content as well with NLP algorithms such as part-of-speech tagging or sentiment analysis for differentiating between social bot accounts and real accounts, thus between fake news and real news. Their algorithm checks the followers, the number of the followers, mentionings of the account, retweet networks of the account.
Test:

Hoaxy on figure 4. helps to see how these diffusional networks are built up based on an article. It shows how selected stories spread from one account to another via retweets, replies, mentions.

Gray links are for stories from low credibility sources, yellow for stories from fact-checkers. The circles represent the nodes, which are the Twitter users. The bigger a circle is, the more time its stories were retweeted. The color bar on the right symbolizes how likely the account is a bot account.

Figure 4: Displaying diffusional network with Hoaxy (Hoaxy, 2019)

5.1.2.2 Some Like it Hoax

In a 2017 study (Tacchini et al., 2017), the researchers were interested in whether hoax could be identified merely by looking at who „liked” an article. The research works with data collected from Facebook including posts from scientific and hoax pages, likes and user information.

There are 3 main observations that the research brings attention to:

- The distribution of the number of likes per post is exponential
- The distribution of the number of likes per user is exponential
- Hoax pages have more users in common with other hoax pages

Studying the data on figure 5., we can see that the majority of the posts have a small number of likes and the distribution of the likes is exponential, just like the distribution of posts per user: the majority of users have only a single like. We can also see that the number of likes for hoax posts is in average higher than the number of likes for scientific articles.

![Figure 5: Likes per post and Likes per user histograms for the dataset (Tacchini et al., 2017)](image)

The study categorizes the users into 3 groups:

- users who liked hoax posts only
- users who liked non-hoax posts only
- users who liked at least one post belonging to a hoax page and one belonging to a non-hoax page

When analysing the dataset on users’ behaviour on a heatmap (figure 6. a), we can see that even though there is a high polarization, there are many users who like hoax and non-hoax articles as well (more detailed on figure 6. b). Due to this phenomenon, the researchers decided to create a subdataset containing only information of the users with mixed likes to study how well their original algorithm performs with a not strongly polarized dataset.
Dataset:

The dataset consists of 15500 public posts and more than 900000 users with around 230000 likes – the articles were selected from pages either dealing with scientific topics or with conspiracy news collected via Facebook Graph API. The ratio of hoax and non-hoax posts is 42.4% (hoax) and 57.6% (non-hoax).

Algorithm:

The study treats the fake news problem as a binary classification relying on labeled data thus using supervised learning methods.

Test:

The study provides results on how well the different models did on a cross-validation analysis. Both models reached a 99% accuracy that by splitting the dataset into training (80 %) and test data (20%), After training the model, it was able to tell with 99 % certainty if the articles of the test data are hoax or non-hoax articles. The article goes further on trying to find out what is the minimum number of articles needed to be evaluated for the training data keeping the high accuracy level. The study suggests that for the harmonic BLC algorithm, keeping the accuracy level above 99 %, it is enough to train the model only on 80 post – that is 0.5 % of the posts altogether.
5.1.3 Content and Social Context

5.1.3.1 Botometer

Another interesting tool is Botometer (Polich, 2018c) which has been developed to help identifying whether an account is controlled by a human or a machine. It assigns a score to an account between 1 to 5 using a classification algorithm. Other characteristics that Botometer looks at are how long ago was the account in question created, whether it has a default email address, how the account name is built up – a useful information could be if the account has a lots of digits, many times it means that the account is not real. There are also temporal features that are being monitored, such as whether the account tweets very often (much more than a human would do on a daily basis) or on a regular basis (tweets that are generated always around the same time could be suspicious).

Together with the above mentioned features, there are around 1200 features that Botometer checks on once there is a submission made on the website.

Dataset:

The model behind Botometer was trained on a dataset including labeled bot and human accounts. The research team used Twitter’s API to collect tweets of the dataset’s accounts and the retweets or mentioning of those tweets. Altogether, the dataset contained 15 k manually verified social bots and around 16k human accounts along with more than 5.6 millions tweets (Davis, 2016).

Algorithm:

It is a classification system checking on more than 1000 features which can be grouped into 6 main categories:

- **Network**: information diffusion patterns
- **User**: language, geographic location, account creation time
- **Friends**: statistics relative to the account’s contacts
- **Temporal**: Timing patterns of posts
• **Content:** linguistic cues via part-of-speech tagging

• **Sentiment:** emoticon scores

Botometer uses a supervised ML method to categorize users based on the features listed above.

**Test:**

As a test, I ran the account @realDonaldTrump against Botometer’s *check account* feature and it provided me a 0.2 result which means that the account most probably is not a fake account (Seen on figure 7.).

![Botometer](image_url)

**Figure 7:** Botometer showing evaluation result on the Twitter account @realDonaldTrump (Botometer, 2019)
In the result window, other than the overall score, I can also check on the different feature evaluations split into 2 main categories: language–specific features and language-independent features. I can then run a test for the account’s followers and the friends that are being followed by the original account. This way, I also found an example for a possible bot account seen on figure 8, with a score of 4.7, giving over 90% probability that the account is a social bot.

![Figure 8: Botometer showing evaluation result on a @realDonaldTrump follower’s Twitter account (Botometer,2019)](image)

A big problem in this research field is that not only bots but also humans are spreading fake news – many times unintentionally. A good way to avoid such dissemination of misinformation is to check who the article originates from – Botometer could be very useful on doing this task. However, it’s extremely hard to track back the origins of fake news. Also, as a human being, it is hard to resist to watch something that goes viral. We have the assumption that if something is being watched and shared many times, it must be worth watching. Unfortunately, using others as signals can be bias. There are bots that are built for boosting accounts that would trick search algorithms or social media platforms’ News Feed algorithms and would bring certain content as first to the public search but there are also many humans as well participating in this boosting method.

If something is spread by one’s social media friends, it gets priority on one’s Facebook feed causing this to be a phenomenon of a reinforcing loop. The biases of the algorithm and the biases of the people reinforce each other which can be exploited when spreading misinformation. What we are unsure about is the impact of fake news and misinformation on
the society. We can only have estimates on what could have happened if there were no fake news spread before the US elections. However, everything is based on assumptions, such as what is the likelihood that the chances of one candidate grow or shrink given that there is one more fake news about the other candidate?

5.1.3.2 Fake News Tracker

Fake News Tracker is a tool analyzing many aspects of the fake news problem. One is to try to find trends in plotting time series information of spreading fake news (figure 9.). Fake News Tracker is a software at its early stage created by Kai Shu’s research team (2017) that is checking over time what is the number of fake news and real news shared on Twitter adding the comparison of accounts that shared these articles, checking the characteristics of these account (such as gender, age, location…).

Figure 9: Trends in Twitter Data between January 2015 and December 2019 (Fake News Tracker, 2019)
Dataset:

The dataset used for building Fake News Tracker is publicly available on github for free analysis. The dataset was collected from Twitter and contains information about the news articles in multiple dimensions. The 2 main categories are Content Features and Social Context Features.

Content Features are for example the source of the article, the headline itself, the body text and any Image/Video that may be included in the article.

Social Context Features can be further distributed into three aspects, namely users, generated posts and networks.

The user-based features have 2 major levels: individual level and group level. On the individual level, it stores data such as the gender, the age, the number of followers and followees. On the group level, the dataset has information about the communities the alanyzed user is part of – groups usually also have certain characteristics coming from aggregating feature components of the indivial accounts in that specific community.

Post-based features are could be stance features indicating whether the user agrees or denies the article in question, or a score that uses the wisdom of the crowds.

Network-based features are collected via creating different networks of users who published related posts.

Algorithm:

Fake News Tracker handles the fake news problem as a binary classification problem assigning either 0 or 1 to a news article depending on whether it is fake or not. The Deep Learning classification algorithm tries to learn the above described feature representation of the news and other entities and use these features to do classification.

The algorithm works with several models used for the different parts of the data analysis. One model that was mentioned as the basis of categorizing post–based features was the Latent Dirichlet allocation (LDA). LDA is a generative statistical model in Natural Language Processing, that is being used here for classifying articles into different topics.
The website itself has several tabs showing different statistics and aspects on Twitter data, providing us with different statistics over fake news versus real news. On figure 9, we can see the news trends, showing the times when the number of fake news shares, post, mentions on Twitter was extremiliy high. On figure 10, we can see the most used words in fake versus real news.

![Word cloud in fake news versus real news content](image)

Figure 10: Word cloud in fake news versus real news content (Fake News Tracker, 2019)

### 5.2 Conclusion

There are several studies ands tool built around the topic of fake news however, almost none of them approach this problem the same way. Many of them see it as a binary classification issue on the top layer but even the binary categories are different. Some use the categories of „true news” and „fake news”, some focus on claims in articles and break down the article to smaller pieces to be identified as „claims” or „no claims”. Others would rather concentrate on the dissemination network and the accounts participating in sharing fake news than the content-specific features. This shows that the problem has many aspects, and how we analyze it, only depends on where we want to start from. There does not seem to be a „true way” – at least at this point of time, which is proved by the success level factor of the introduced products, let it be a high accuracy number or just my personal tests.
Almost in each cases, the researchers created an algorithm using some combinations of supervised learning methods with labeled data. Data collection, data cleaning and labelling the data seemed like the biggest challenge and work in the studies. I believe, that if there was an initiative for a cooperative work between the researchers - who have the tools and ideas - and the social media platforms - who have the data - , there would be more success in a shorter time that could be straight away tested and if possible, implemented.

As of now, there is no common agreement on the solution just like there is no common agreement on the problem’s definition either. Each approaches seem promising however not perfect and due to the areas of life and questions that this topic has an impact on, there is a need for a better and more finely tuned algorithm. Even though some of them showed high level of accuracy at the final testings, it does not seem to be enough to be implemented at this point. A human inspector’s presence is still needed to judge the articles but maybe not for many longer.

6 Strategies for fake news detection today

Online misinformation is a hard problem and as introduced in this paper, it is a complex issue not just considering the infrastructure of democracy but also algorithmically. Thus the question, who should be responsible handling this issue? From one hand, it is about demolishing the public’s confidence in a country and public faith in the country’s institutions. All being possible because of one of our fundamental values as a democratic country’s citizen – namely free speech. On the other hand, the social media platforms provide and boost us with limitless options to share anything we want including our opinion, other’s opinion and fake news. Although these platforms were originally set up to connect us everywhere, it seems that it is responsible now for driving us apart. When considering the above, we can conclude that the issue could be controlled at 2 levels: government-level and/or social media platform level.
6.1 Government level

Europe

I was interested in what is protecting me from fake news as a European citizen. I found a great amount of information and techniques about how and what measures the European Union took and takes in this measure, keeping in mind that we are approaching the 2019 European elections. My main source on this was an Action Plan against Disinformation that was published in December 2018 by the European Commission. Based on this document, The EU’s response are as below:

- Enabling EU forces more on detecting, analysing and exposing disinformation
- Work on joint response
- Finding a way to prompt private sector to focus on the issue
- Raising awareness on fake news existence

To reach the above, the EU decided to more than double the 2019 budget for strategic communication. This means, that it will raise from 1.9 million (2018) to 5 million in 2019. The EU is planning to collect more data to analyze and employ trained people who can work with the data. Furthermore, they want media monitoring services – covering the languages – and analytical tools (like the ones or their customized versions described in the previous chapter) to process the data.

The EU also seems to take the issue into its own hands from private sector. In September 2018, it released the Code of Practice on Disinformation and urged the dominant platforms to sign that which Google, Facebook, Twitter and Mozilla all did. The platforms since then have been making continuous efforts in filtering and removing the fake accounts and limiting visibility on malicious websites. The aim of Code of Practice of Disinformation is to create a trustworthy online ecosystem which instead of confusing the users, would rather help them to the valid sources and information.

Finally, it is organizing campaigns for the public to raise awareness of the dangers of fake news and misinformation and trying to train its citizens on how to differentiate between valid information and misinformation.
6.2 Social media platform level

By accepting the Code of Practice on Disinformation, many Social Media platforms accepted to join the efforts to mitigate disinformation. Below I will analyze how Facebook and Twitter have done so far in this matter.

**Facebook**

During the 2016 US presidential elections, a company called Cambridge Analytica was easily able to get access to the data of around 87 million Facebook user and use that in order to help Donald Trump’s campaign and the Brexit-supporter’s parties propaganda (Solon, 2018). After that, Facebook had to face with its relevance in manipulating public opinion one way or another.

There were several approaches of how Facebook tried to deal with fake news so far – such as using human moderators to spot violent elements and try to remove those before they go viral. Without a higher level of automation, these tasks were very time consuming and so partly inefficient. There has been efforts to try to use Artificial Intelligence to filter out misinformation on Facebook however it often also removed opinions and not purely just articles that had violent content. One of the many reasons that fake news detection is still an unresolved problem is that Artificial Intelligence is just simply not at the level yet where it would fully understand human writing the way humans do. It can already do a number of things such as spotting claims and checking on them whether they are true or not, checking on the sources if they are reliable or if they seem to be automated bot accounts, it can use NLP algorithms, do a sentiment analysis - meaning that the algorithm can judge how factual the article is. It has many perspectives, many different approaches trying to solve different parts of this complex issue.

Facebook’s newest approach (Mosseri, 2017) is a combination of human and artificial intelligence resources. The social media platform is actively working on trying to remove fake accounts and giving a notice to its users when the resource of an article is questionable. It is working closely with third - party fact checkers on checking news articles’ reliability. The algorithmic part is responsible for collecting articles by using feedback data from people
on Facebook. The stories marked as unreliable will then be passed on to fact-checkers who will rate the accuracy of the article. Articles with lower accuracy will appear lower in the users’ feed reducing the possibility of malicious, untrustworthy content reaching great amount of people.

Those pages and websites that are continuously posting fake news will experience a reduced distribution of their content and will not be able to advertise in the future – meaning that even though they would pay for advertisement, it will not be granted for them to do so.

Moreover, when an article is verified as false by fact-checkers, other than marking it fake, the article will have related articles shown in order to check more information on why it’s been ranked low. Also, when a user will try to share a story that’s been identified as fake, the user will be prompted that the article has been marked as false in order to prevent the public from spreading fake news.
Facebook at its current state enables us to mark seemingly malicious content. Figure 11 shows us how Facebook is filtering messages based on the users’ interactions. We, as users, get already a taxonomy that we can select from. It uses a multiclassification algorithm categorizing news into ten different categories: such as spam, false news, hate speech and so on. It then also gives us the opportunity for calling immediately in case of immediate danger.

Figure 11: Facebook’s taxonomy on questionable content (Facebook, 2019)

**Twitter**

In a recent interview (TED, 2019) with Twitter CEO, Jack Dorsey, he stated that this year, about 38% of abusive tweets are now identified with machine learning algorithms, which means that there is no need for human interaction in the reporting part. The comments are
filtered automatically and then reviewed by humans – so nothing is removed completely automatically. This is a huge step compared to last year, when the same task was in 0 % automated. Every single person who received an abusive comment needed to report that and those needed to be reviewed.

Twitter is working on combatting misinformation by trying to measure the „health of a conversation“. Instead of identifying claims or news whether they are true or fake, Twitter came up with an indicator system that they believe would reflect the conversational health. They compare the conversational health to human health and think that the prior can be measured just like the latter. The solution is currently in its early stage and there was no Estimated Time of Arrival specified.
7 Summary

The aim of this paper was to introduce the topic „fake news”, to see what is behind the rapid growth of spreading misinformation, to find out if this issue can be handled from an algorithmic point of view and if so, how. As discussed in the first chapter, the internet enables us to share content in all quality and quantity with a velocity never experienced before. The focus in this paper was specifically on the social media platforms and why and how the manipulation of the public is present there. To the question „Why is fake news present on social media?” I found three main reasons and after a short introduction to the other two motifs, I decided to shift my attention on the third, namely political interest.

I then introduced the different strategies and tools on how spreading misinformation works, who are behind that. I identified social/political bots as one of the biggest threats in the online world and showed how these accounts are present and what tasks could be executed by them.

I took a look at how machine learning could be at help to this problem – after a brief overview of the different machine learning techniques. As most of the techniques I found in filtering news used either Supervised Learning methods or Deep Learning techniques, I felt the importance to see what are the major differences between these learning methods. Moreover, I touchbased on the topic of Natural Language Processing as due to the nature of this topic, I found many solutions to the problem including NLP techniques.

When researching for studies on fake news detection, I found that there is a wide range of approaches offering solutions to this umbrella problem. On the highest level, almost all the analyzed solutions see this as a binary classification issue but not all with the same two classes (fake news and real news). The features represented in the descriptions included either content-related or social context-related features so I decided to further analyze the tools according to these two classes.

One common issue in all the analyzed cases was the data collection, creating a valid dataset with „true fake” news and with all the other features that the researchers identified to be essential in order to filter misinformation and the time it took to label data. I also observed
while trying to test some of the solutions, that finding „true fake” news, namely news articles that are prooved to be misleading, is a rather hard task.

As most of the examples in this paper show, the development of different fake news detecting tools were either triggered or accelerarated by one of the two major political events in 2016: Brexit and the US Presidential election. In the reasoning of all the studies stands, that our democracy and democratic procceses are at stake if this problem is not being handled „correctly”. However, it is a hard task to identify who is responsible for solving this issue. I analyzed different approaches on government-level and on social media platform-level.

As a conclusion, I see the need for a joint solution with the government creating policies for regulating social media platforms. This would include steps to be taken from a democratic point of view – like identifying the true filters or features to look for in fake news. The technology should be then developed by the social media platforms using information and forces of Fact-checkers and enable researchers to access to their data to ease the data collection process and accelerare finding the solution.

After analyzing this topic, I found that there are many tools out there to help filtering fake news however, most of them are not yet used or not yet well-known. As introduced in the case studies, fake news is hard to define, to specify and to identify even for humans and therefore it is hard to translate the problem for machines. This is one reason why we do not see these tools everywhere yet – the filtering technique in most of the above listed cases is robust and many times it is unable to categorize the news/accounts accordingly.

Many also argue that fake news should be handled as a binary problem – enough to mention opinion articles/posts with a heavy right or left political direction. It is not an easy task to find the line between these posts and therefore it is a big expectation from artificial intelligence to solve the issue. We need a clear definition and clear features of fake news problem that we are struggling to come up with.

As a conclusion, I see that - as of today - machine learning alone is not able to solve this problem without human interaction. As long as we are unclear about the problem’s definition or about which aspects could be the best to handle this issue, our best way to fight misinformation is a joint effort of the government, fact-checker organizations and social media platforms with using automation only as a helping tool but leaving humans with the decisions of what content to label/remove.
After going through the case studies, I found that many of the datasets and also of the algorithms are publicly available. As a next research topic, I would be interested in experimenting with creating a new filtering technique and describe the steps taken during trying to build the algorithm, see if new features could be added to improve the accuracy of the filtering technique and if so, how that piece of information could be collected with the policies changing rapidly in accessing data all around the world.
References


Table of Figures

Figure 1: Total Facebook engagements for Top 20 Election Stories (BuzzFeed News, 2016) ................................................................. 10
Figure 2: Machine Learning Process (Portilla, 2018) ................................................................. 16
Figure 3: Result of testing FakerFact as a fake news detecting website (FakerFact, 2019) 25
Figure 4: Displaying diffusional network with Hoaxy (Hoaxy, 2019) ......................................... 27
Figure 5: Likes per post and Liekes per user histograms for the dataset (Tacchini et al., 2017) ................................................................................................................................. 28
Figure 6: Hoax versus non-hoax likes per user heat-map (a) and users is common between hoax and non-hoax pages (b) (Tacchini et al., 2017) ................................................................................................................................. 29
Figure 7: Botometer showing evaluation result on the Twitter account @realDonaldTrump (Botometer, 2019) ................................................................................................................................. 31
Figure 8: Botometer showing evaluation result on a @realDonaldTrump follower’s Twitter account (Botometer, 2019) ................................................................................................................................. 32
Figure 9: Trends in Twitter Data between January 2015 and December 2019 (Fake News Tracker, 2019) ................................................................................................................................. 33
Figure 10: Word cloud in fake news versus real news content (Fake News Tracker, 2019) 35
Figure 11: Facebook’s taxonomy on questionable content (Facebook, 2019) .............................. 40