Visualizing fakenews as reported by euvsdisinfo.eu

Jesper Henrichsen
Supervised by: Pedro Ferreira

December 13, 2017

Contents

1 Introduction

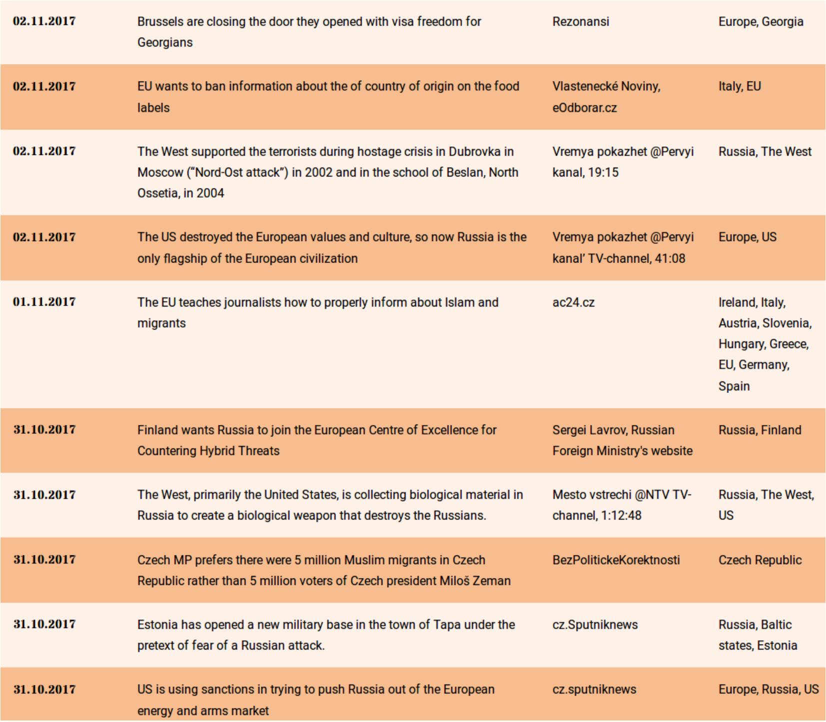
Since the introduction of the internet, information has been able to reach further and faster than ever before. Recent developments of omnipresent social media has created the perfect platform for the spread of this information within the internet. For the sake of advertisement social media platforms has, since their introduction, only become better at targeting their audience with information that will capture their attention. As put by the team behind the newsfeed at Facebook: "*The goal of News Feed is to deliver the right content to the right people at the right time (...)*". However, a relevant question would be to whom it is implied to be *right* for. That it is not necessarily the user, has recently emerged as a likely answer.

The nature of social media, has proved to be an effective way for the spread of information, genuine as well as misinformation. This report will describe the process of scraping and visualizing information from euvsdisinfo.eu. One objective was to investigate whether these cases could be used to view fakenews as attacks. And if so, could there be identified a perpetrator either an outlet or country, and a victim as in the country being mentioned in the information.

2 Background

Since 2015 the campaign euvsdisinfo.eu, has been run by the European External Action Service East Stratcom Task Force. The primary focus of the campaign is to identify and debunk pro-Kremlin disinformation. The campaign includes cases of debunked information from sources spanning from official news sites to twitter accounts, to non-online content such as interviews. The website euvsdisinfo.eu includes, at the time of writing, more than 3400 such cases, these are the cases that this report is based on. The cases are all reported either by the campaign staff themselves, or one of the 400 collaborating organizations and individuals. In Figure 1 is seen how cases are listed on the website, the order of which is chronological with respect to the date it was reported.

Figure 1: Example listing of disinformation cases on euvsdisinfo.eu



02.11.2017	Brussels are closing the door they opened with visa freedom for Georgians	Rezonansi	Europe, Georgia
02.11.2017	EU wants to ban information about the of country of origin on the food labels	Vlastenecké Noviny, eOdborar.cz	Italy, EU
02.11.2017	The West supported the terrorists during hostage crisis in Dubrovka in Moscow ("Nord-Ost attack") in 2002 and in the school of Beslan, North Ossetia, in 2004	Vremya pokazhet @Pervyi kanal, 19:15	Russia, The West
02.11.2017	The US destroyed the European values and culture, so now Russia is the only flagship of the European civilization	Vremya pokazhet @Pervyi kanal' TV-channel, 41:08	Europe, US
01.11.2017	The EU teaches journalists how to properly inform about Islam and migrants	ac24.cz	Ireland, Italy, Austria, Slovenia, Hungary, Greece, EU, Germany, Spain
31.10.2017	Finland wants Russia to join the European Centre of Excellence for Countering Hybrid Threats	Sergei Lavrov, Russian Foreign Ministry's website	Russia, Finland
31.10.2017	The West, primarily the United States, is collecting biological material in Russia to create a biological weapon that destroys the Russians.	Mesto vstrechi @NTV TV-channel, 1:12:48	Russia, The West, US
31.10.2017	Czech MP prefers there were 5 million Muslim migrants in Czech Republic rather than 5 million voters of Czech president Miloš Zeman	BezPolitickeKorektnosti	Czech Republic
31.10.2017	Estonia has opened a new military base in the town of Tapa under the pretext of fear of a Russian attack.	cz.Sputniknews	Russia, Baltic states, Estonia
31.10.2017	US is using sanctions in trying to push Russia out of the European energy and arms market	cz.sputniknews	Europe, Russia, US

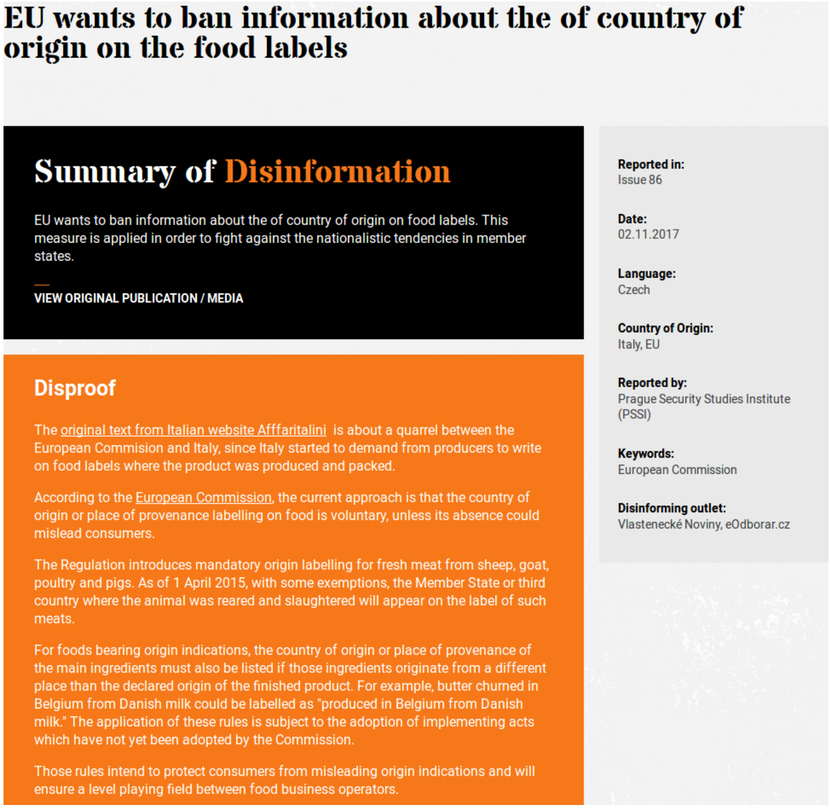
3 Acquiring a dataset

In this section, the approach to acquiring and extracting information for the visualization will be described. The initial dataset was scraped from euvsdisinfo.eu, however, this dataset was gradually extended in a number of steps. These steps will be described in this section.

3.1 Scraping the cases

The cases listed at euvsdisinfo.eu consists of information regarding the case, such as who reported it, in what country or countries did it originate in, as well as the disproof that debunks the information as invalid. Other information for each case is meta information about the information and its source. In Figure 2 is seen an example of the information listed for a specific case.

Figure 2: An example of the information for each case



EU wants to ban information about the of country of origin on the food labels

Summary of Disinformation

EU wants to ban information about the of country of origin on food labels. This measure is applied in order to fight against the nationalistic tendencies in member states.

VIEW ORIGINAL PUBLICATION / MEDIA

Disproof

The original text from Italian website Affaritalini is about a quarrel between the European Commision and Italy, since Italy started to demand from producers to write on food labels where the product was produced and packed.

According to the European Commission, the current approach is that the country of origin or place of provenance labelling on food is voluntary, unless its absence could mislead consumers.

The Regulation introduces mandatory origin labelling for fresh meat from sheep, goat, poultry and pigs. As of 1 April 2015, with some exemptions, the Member State or third country where the animal was reared and slaughtered will appear on the label of such meats.

For foods bearing origin indications, the country of origin or place of provenance of the main ingredients must also be listed if those ingredients originate from a different place than the declared origin of the finished product. For example, butter churned in Belgium from Danish milk could be labelled as "produced in Belgium from Danish milk." The application of these rules is subject to the adoption of implementing acts which have not yet been adopted by the Commission.

Those rules intend to protect consumers from misleading origin indications and will ensure a level playing field between food business operators.

Reported in:
Issue 86

Date:
02.11.2017

Language:
Czech

Country of Origin:
Italy, EU

Reported by:
Prague Security Studies Institute (PSSI)

Keywords:
European Commission

Disinforming outlet:
Vlastenecké Noviny, eOdborar.cz

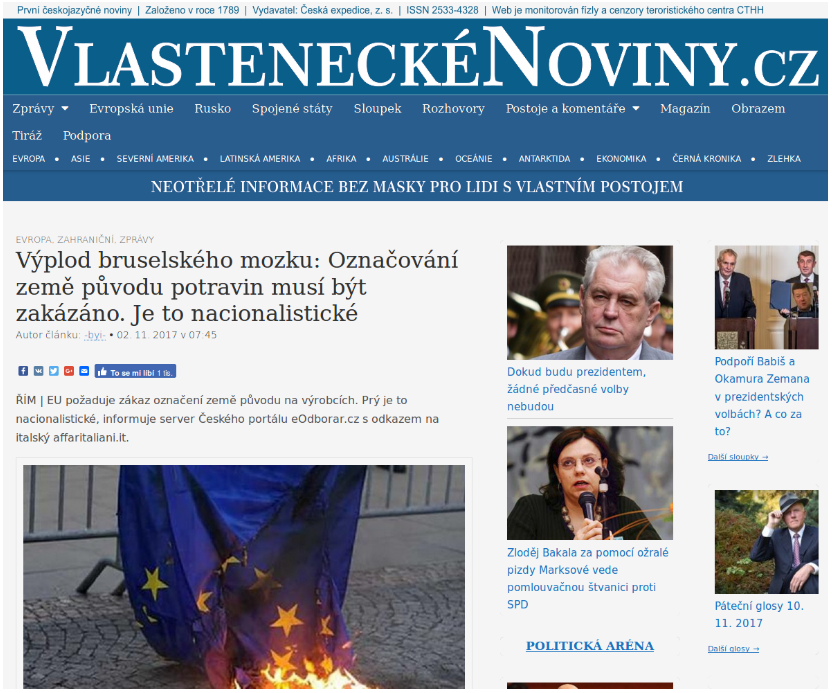
There is no API, or otherwise download button on the campaign website to get the whole dataset as is. However, there is no mention in their robots.txt file that indicates that scraping is not permitted. Therefore, the intial dataset was achieved after writing a scraper in Python[1]. The list of cases is an overview list with a pagination of 10 cases per page, an example of this list is shown in Figure 1. The offset is given by the URL, and so the crawling to acquire links to each specific page can be done in a well defined way, without having to worry about the specific structure of the website. However, parsing of the structure of the website is necessary in order to scrape the information of each specific case. For this, the python library BeautifulSoup[2] was heavily used.
In the end, the information was saved into a csv file containing one disinformation case per row, and in total 3071 data rows.

[1] https://github.itu.dk/jeshe/fakenews-scraper/blob/master/news_scraper.py
[2] https://www.crummy.com/software/BeautifulSoup/

3.2 Article scraping and content extraction

In order to get more information that was already available for each case at euvsdisinfo, information from the original source was also included and added to the dataset. This section will describe the approach that was used in order to reliably get information across the many differently structured websites.

Figure 3: An example of a debunked news article from a Czech website


První českojazyčné noviny | Založeno v roce 1789 | Vydavatel: Česká expedice, z. s. | ISSN 2533-4328 | Web je monitorován fízly a cenzory teroristického centra CTHH
VLASTENECKÉNOVINY.CZ
Zprávy
Evropská unie
Rusko
Spojené státy
Sloupek
Rozhovory
Postoje a komentáře
Magazín
Obrazem
Tiráž
Podpora
EVROPA • ASIE • SEVERNÍ AMERIKA • LATINSKÁ AMERIKA • AFRIKA • AUSTRÁLIE • OCEÁNIE • ANTARKTIDA • EKONOMIKA • ČERNÁ KRONIKA • ZLEHKA
NEOTŘELÉ INFORMACE BEZ MASKY PRO LIDI S VLASTNÍM POSTOJEM
EVROPA, ZAHRANIČNÍ, ZPRÁVY
To se mi líbí 1 tis.
POLITICKÁ ARÉNA
Další sloupky →
Další glosy →

Výplod bruselského mozku: Označování země původu potravin musí být zakázáno. Je to nacionalistické

Autor článku: -byi- • 02. 11. 2017 v 07:45

ŘÍM | EU požaduje zákaz označení země původu na výrobcích. Prý je to nacionalistické, informuje server Českého portálu eOdborar.cz s odkazem na italský affaritaliani.it.

Dokud budu prezidentem, žádné předčasné volby nebudou

Zloděj Bakala za pomocí ožralé pizdy Marksové vede pomlouvačnou štvanici proti SPD

Podpoří Babiš a Okamura Zemana v prezidentských volbách? A co za to?

Páteční glosy 10. 11. 2017

A general approach to extracting content is difficult because source of information can be any sort of media, whether digital or physical, in writing or a video. And, because each website is different, then even building a scraper to extract the content of the subset of sources that are online articles, will be difficult. In Figure 3 is seen an example of a source article.

One important property for the purpose of this project is that in order to consider the names of locations that are being mentioned in the articles, the entire content is not necessary. Therefore, I chose to use the meta tags to extract information about the content, summary, titles, author, and descriptions of each online resource. I would also filter on html tags such as the header tags, h1, h2, h3, h4, h5, h6 as well as the title tag used for setting the window title. I found through experiment that only considering the first found header tag worked best, the reason is that other header tags than the first one would often be titles of other news articles that the news site wants the user to click on.

Figure 4: Meta tags defined in python scraper

```
metatag_list = [
    ("name", "description", "content"),
    ("property" "og:title", "content"),
    ("property", "og:description", "content"),
    ("name", "twitter:title", "content"),
    ("name", "twitter:description", "content"),
    ("name", "language", "content"),
    ("name", "keywords", "content"),
    ("name", "subject", "content"),
    ("name", "topic", "content"),
    ("name", "summary", "content"),
    ("name", "subtitle", "content"),
    ("itemprop", "name", "content"),
    ("itemprop", "description", "content"),
]
tag_list = [ ("h%s" % i, None) for i in range(1, 7) ] + [("title", None)]
```

Extracting content by meta tags proved to be a very reliable approach, since most news sites are interested in their content being shared. So, in order to optimize sharability almost all online resources had optimized for search engines, often reffered as just: *SEO*. In Figure 4 is seen the defined metatags that the scraper was looking for, defined in the variable: metatag_list, as well as the list of header and title tags defined in tag_list. Each metatag is defined as a tripplet. The first value defines what attribute to look for when finding a HTML tag of type meta, the second what value such attribute should have. When a match is found for the first two, then the third string in the tripple is used to know from what attribute to extract text from. In the metatags listed in Figure 4 all content was extracted from attributes of the same name regardless of the match on the first two. However, defining triples makes this method generalize to any other metatags that might be relevant to add in the future. The reasoning behind the list of other HTML tags being defined as tuples, is similar. Although, for my purposes, the second string was set to None, in which case the script would scrape the inner HTML of the element instead of the content of a named attribute.

In order to avoid duplicating content as much as possible the content of a meta tag or html tag was compared to the content that was already found in previous tags. This approach resulted most often in a short paragraph of information about the article, including keywords, title and summary. Before applying the approach to the articles scraped from euvsdisinfo.eu, it was tested on arbitrarily chosen articles, such as articles from danish news outlets. It was also tried on

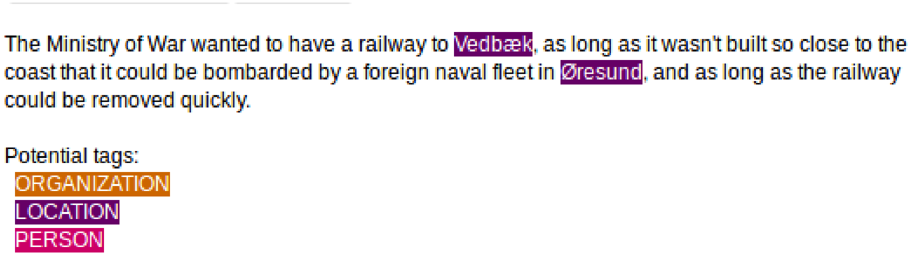
a sample of articles taken from the subreddits: r/politics, r/news, and r/worldnews[3], because these were collections of news articles with a good variety of news outlets. The results were used to evaluate qualitatively on the smaller sample. None of which turned out to have no content, and only one resulted in an extract only consisting of keywords. Interestingly, among the keywords for that particular result were also the name of the country that the news story revolved around.
One thing that was clear from this, was that locations were not always mentioned if the news concerned well known state leaders such as Putin, Merkel or Trump. In such cases, often only the names of the state leaders were present as indication of what countries were mentioned in the articles.

3.3 Named entity recognition

One initially desired outcome was the possibility to extract location information about what countries were being mentioned in the articles. This section will describe the approach to using stanford NER tagger to quantitatively extract location names from the meta information acquired for each article.

Named entity recognition is a research field within natural language processing concerned with recognizing names of things such as people, company names, or, as relevant for this project, locations. The study of recognizing location names within texts is one of the most studied areas of named entity recognition, however, still an ongoing research. As such the NER tagger used in this project is released open source as part of the Stanford nlp library[1] which also includes many other models and methods relevant to the field of natural language processing. A more in depth explanation of named entity recognition or even natural language processing, is however, beyond the scope of this project. In Figure 5 is seen an example of tagging a sentence using the graphical interface of the NER tagger. The sentence in Figure 5 is arbitrarily chosen from Wikipedia, and is not part of the content of the dataset.

Figure 5: NER tagging on a sample sentence



The Ministry of War wanted to have a railway to Vedbæk, as long as it wasn't built so close to the coast that it could be bombarded by a foreign naval fleet in Øresund, and as long as the railway could be removed quickly.

Potential tags:
ORGANIZATION
LOCATION
PERSON

The input that was given to the NER tagger was the concatenation of title,

[3] https://reddit.com/r/politics+news+worldnews

summary (as provided by euvsdisinfo), keywords, and the meta tags content scraped from each individual site.

3.4 Facebook likes

The initial dataset, provides only a uniformly weighted list of debunked news cases. In order to provide different weights to the articles in any future visualization, each source URL in the dataset was looked up via facebooks open graph API. The result of the API lookup was a mapping for each source article to a number of likes on facebook, in the case that the information was ever shared on facebook. In Figure 6 is seen the essential part of the python code used for fetching the number of likes of each url. The function get_shares takes as argument a url and returns dictionary of the response from Facebook's API.

Figure 6: Fetching facebook likes for each source URL

```python
def get_shares(url):
    enc_url = encode(url)
    access_token = get_access_token()
    fields = "og_object{engagement}"
    api_endpoint = "https://graph.facebook.com/v2.11/%s?fields=%s&access_token=%s"
    return requests.get(api_endpoint % (enc_url, fields, access_token)).json()
```

4 Results

The resulting dataset from the previous section includes 17 columns as can be seen in Table 1.

Table 1: Resulting columns in the dataset

Column name	**Description**
issue	The issue number as given by euvsdisinfo.eu.
date	The date the case was reported.
outlet	The outlet that published the information.
language	The language the information is provided in.
origin	Origin of the story, this is sometimes a satirical piece from a different country.
reported by	The person or organisation who reported the case.
keywords	Keywords describing the published piece of information.
source	The URL to the original story, this is not always present i.e. if the source is an interview.
title	The title of the debunked information.
summary	The summary of the information that was debunked.
disproof	The reasoning for euvsdisinfo to flag the information as dishonest.
metatags	The types of metatags that included information from the original source during scraping.
metatags_content	The content of the metatags that was scraped from the original source.
locations	Any locations found by the NER tagger.
misc	Miscellaneously tagged words or phrases tagged by the Stanford NER model.
people	Recognized names of people by the NER tagger.
likes	Number of facebook likes registered to the source URL, if any.

The last 6 columns of Table 1 are additions to the initial dataset acquired from euvsdisinfo.eu, namely: metatags, metatags_content, locations, misc, people, and likes.

The most mentioned location is, expectedly, Ukraine. Equally unsurprising is the most frequent language, russian, given the focus of the task force behind the campaign on pro-Kremlin news. The top 5 mentioned locations and the 5 most frequent languages is seen in Figure 7.

Figure 7: 5 most frequent languages and top 5 mentioned locations

language	**% of articles**
Russian	55
Czech	16
English	9
Slovak	2
Georgian	2

location	**% of articles**
Ukraine	27
Russia	23
USA	21
Europe	8
Syria	5

In general the NER tagger performed well, only 426 of the 3071 cases did not have any locations tagged. A sample of the articles tagged is shown in Figure 8. The articles shown were chosen based on having the shortest titles in the dataset, simply because they would be easilier presented. As can be seen from Figure 8, most sources have recognizable locations which seems intuitively correct given the title. However, at least one in these samples, have falsely tagged locations such as PravdaReport, and the word Having. Since measuring such false positives and equally false negatives, cannot be done quantitatively, this inspection has only been done manually on similar samples. However, it is safe to say that the tagged locations cannot be taken for granted as there will be noise in terms of falsely tagged locations. With that in mind, it can serve as a way of depicting the general pattern of locations named in the disinformation sources.

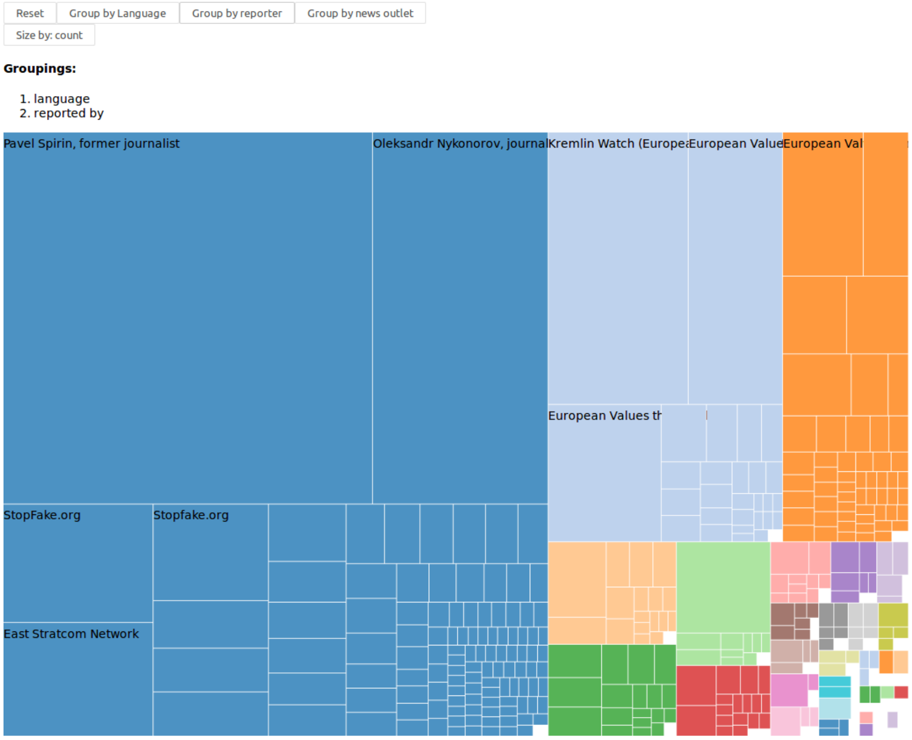
Figure 8: Sample of cases shown by their title, language and locations

title	**language**	**locations**
Crimea decided its own fate	Russian	Crimea,Russia,China,India
Ukraine is a neo-Nazi state	Russian	Ukraine,West
NATO is encircling Russia.	Czech	Russia
Ukraine is governed by Nazis.	Russian	Ukraine
Sweden wants to leave the EU.	Spain	Sweden
Georgia is a US colony.	Georgian	Georgia,US
Ukraine has a Nazi identity.	Russian	Ukraine
Ukraine is a part of Russia.	Russian	Ukraine,Russia
Savchenko is a US spy.	Russian	US
Nazis control Ukraine.	Russian	Ukraine
NATO kills Serbian children.	English	PravdaReport,West,Russia,Having, Syria,russia,Serbia,Yugoslavia
Ukraine is governed by nazis.	Russian	Ukraine
All Ukraine is Russia.	Russian	Ukraine,Russia

Another interesting aspect appears when looking at the values in the reported by column. More than $\frac{1}{3}$ of the cases has been reported by either of two journalists: Pavel Spirin (780 cases) and Oleksandr Nykonorov (379 cases). In third is the European think-tank European Values. Disregarding possible duplicates in the naming of reporting providers, only 33 entities have reported more than 10 cases out of 193 unique entity names. This is an important aspect to consider, besides the focus of the campaign, when looking at any skewness in the dataset. Oleksandr Nykonorov appears to be a Ukrainian journalist, while the European Values think-tank is an NGO based in the Czech Republic.

For the purpose of doing a more visual inspection of the dataset, an online visualization was created using d3. The visualization uses either of 3 predefined columns from the dataset: Language, reported by, and outlet to create a treemap visualization. While the treemap may not be optimal for all purposes, it goes a long way of showing the above mentioned proportionality within the dataset. In Figure 9 is seen a screenshot of this visualization, the screenshot shows the treemap of the number of articles reported by each person or organization within the different languages. The labels are hard to capture, since, for visual clarity, only entities having reported more than 50 articles have a label. The other will show a label upon mouse hovering. The colors represents different languages, the dark blue, which is most present, represents russian. Each square represents how many articles an entity has reported. As an example it can be seen that for russian, Pavel Spirin has reported the most cases.

Figure 9: Screenshot of the treemap visualization using grouped data by `language` and then `reported by` column



Groupings:

1. language
2. reported by

Another possible option for the user is to set the size of the squares not by number of cases, but after how many facebook likes each source url accumulated. For information that has not been shared on facebook, the number of likes defaults to zero. The visualization is available as a proof-of-concept at http://k4lk.dk:3000.

5 Discussion

Visualizations are difficult as they easily become misrepresentative of the dataset, this is especially true for the domain of fakenews. One example is to consider the cases as attacks by a perpetrator on another vicim, where the roles could be countries or news outlets. However, even as the news outlets, the perpetrator, is provided as part of the data from euvsdisinfo.eu, then it is hard to automize the extraction of what country is the victim. Often it will be seemingly obvious who the victim is, based on the location and the title or summary of the information. An example of such is russian articles that mentions Ukraine while the NER tagger has recognized the two people mentioned: Adolf Hitler and Petro Poroshenko.

However, more often it will be a russian article mentioning two locations: Russia and Ukraine. In such a case it is much more difficult to automize a ruleset of what mentioned country is the actual victim. The reason is that no context is given as to how the locations are being mentioned, so even if an article mentions Russia in a very different light than how it mentions Ukraine, then that context is lost when only the location is extracted, and through a visualization based on this information, both will look like the victim.

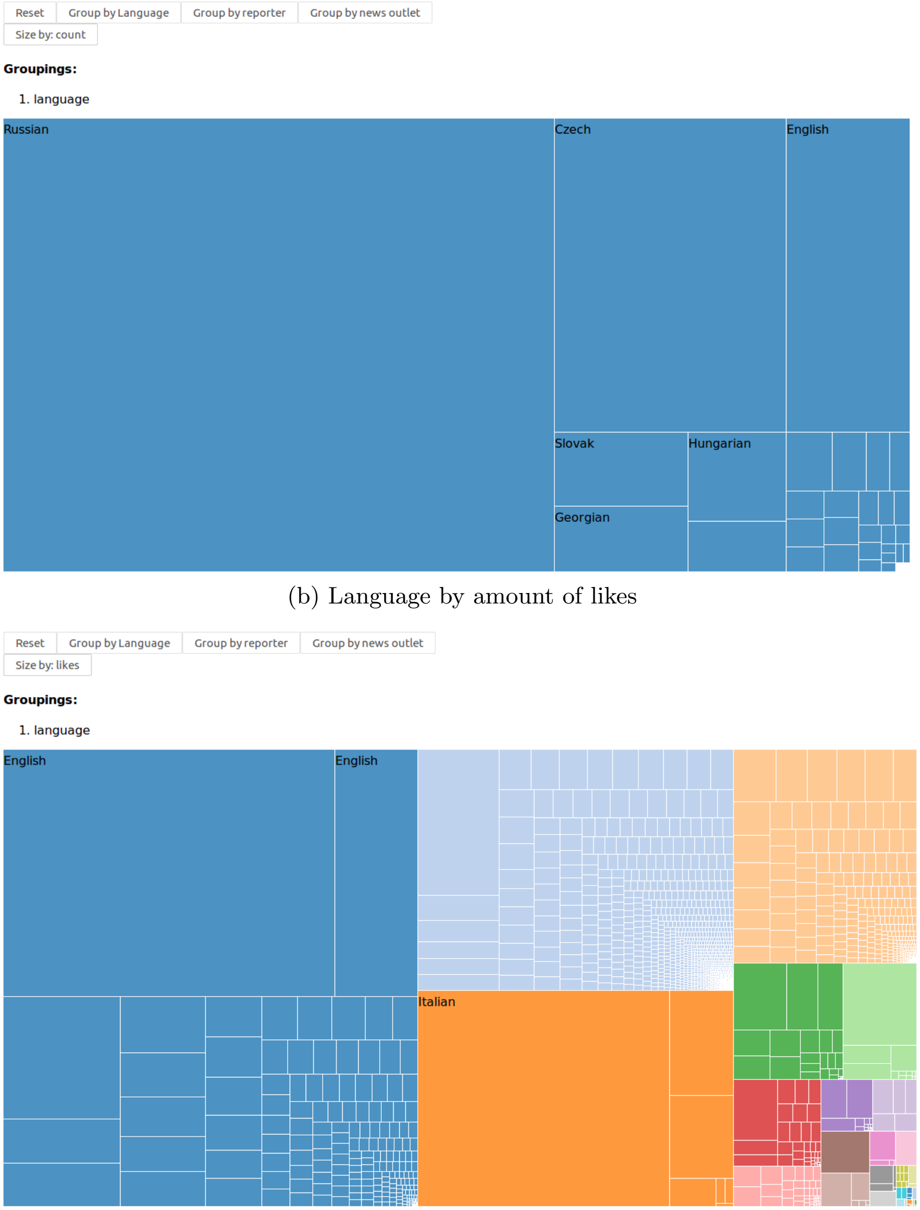
Furthermore, the dataset itself introduces a lot of biases that if not handled carefully might itself promote a false view of the world to the reader. The fact that the euvsdisinfo.eu was started in 2015 which correlates with the war in Crimea, and the fact that the campaign already focuses on pro Kremlin news stories, will risk overrepresenting russian news articles. This is especially important if as to not present these data as representative of the state of fake news in the world. The same is true for the campaign itself. When debunking news stories that was never shared on social media in the first place, there is a risk of helping the misinformation more than countering it, simply by giving it publicity that it would never have received otherwise. Even if that publicity is negative, it might be more in the interest of the original outlet than simply not having any attention.

That attention is important is stressed by the fact that the majority of news outlets that were online, had optimized for search engines. Their objective to be as shareable as possible, since they will want their information to spread in order to be effective, by definition makes these websites easily available to crawl and scrape in a structured, reliable way. At least as reliable as webscraping goes.

In this project the attempt to extract information about people and locations that were mentioned in the online articles, worked well using the metatags provided for search engines. The lack of context the words appeared in, made it hard to automize any interpreting visualization. However, it is still interesting without a layer of interpretation to be able to look at the distribution of locations as it provides a perspective on how the focus on Russia manifests itself in the dataset.

The aspect of disproportionality is also interesting in regards to how these cases are presented on the campaign website. All cases as illustrated in Figure 1 is shown with a uniform importance. However, as seen when requesting the likes via the facebook API for each URL, it becomes clear, that the distribution of likes is very different from this uniform presentation. Whether or not likes is a good measure of impact, the perspective is arguably different when sizing the treemap using the likes of each case, instead of just counting each case. The difference is visually inspectable through the online visualization, http://k4lk.dk:3000. In Figure 10 is seen a screenshot of each of these two cases. The figures shows the cases grouped by the language they were published in. However, as can be seen in Figure 10a, the by far most present language in terms of information cases, is russian, but if the same view is measured in number of likes, seen in Figure 10b, then english accumulates more than double the amount of likes as the second most liked russian languaged sources (light blue). In Figure 10b each square represents an article, thereby it can be seen that a few specific cases dominates in terms of accumulating likes. A few specific cases being exponentially more liked than others, can be seen from the visualization as a pattern across all languages.

Figure 10: Cases grouped by language and sized either by number of cases, or accumulated likes of the cases



(a) Language by number of articles

Groupings:

1. language

(b) Language by amount of likes

Groupings:

1. language

6 Conclusion

In order to be able to better counter the impact of fakenews, we need a better understanding of the nature of fakenews. In this project a very small part of the entire body of fakenews were looked at, namely the cases available from euvsdisinfo.eu.

Even though the nature of fake news makes it nearly impossible to automate the visualization or analysis of fake news, it is still possible through semi automatic methods, as illustrated in this report. This is only achievable because of resources such as euvsdisinfo.eu, that makes the metainformation available as collections in a structured manner. In the same way that the various ways information can be debunked, is not achievable in an automatic way, the same can be said for the deeper interpretations of the eventually debunked cases. However, creating a visual perspective goes a long way to better our understanding of underlying patterns. As is true for many complex phenomenons, then no single perspective can provide the whole truth. Similarly, to visualize the underlying patterns in the dataset used for this project, many different views is needed. For this reason, the explorable visualization is a good approach, however the concrete treemap that was the outcome of this project, can only be seen as a limited part of the perspectives needed to understand the phenomenon of fake news.

References

[1] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.