Data sampling

- All the links from the PDF newsletters from 02 September 2016 to 02 February 2017 have been downloaded (the column used was “Link to the Disinformation”). There were 16 newsletters with 803 links in total.
- The links with the following domains have been ignored because considered reliable source of information (or contain videos which are not easy to analyse). Not being an expert I suppose more link should be filtered out in order to keep the sources of propaganda only.

  ```
  'youtube.com', 'un.org', 'europa.eu', 'nato.int',
  'state.gov', 'securitycouncilreport.org', 'defense.gov',
  'stopfake.org', 'theguardian.com', 'amnesty.org',
  'justice.gov', 'telegraph.co.uk', 'euobserver.com',
  'martenscentre.eu'
  ```
- Each news (skipping the above domains) was processed with the following machine learning steps:
  - The text of the article was extracted (removing other parts of the page like menu, footer, headers, etc)
  - The text was translated to English
  - Semantic e sentiments analysis was applied to the text once translated to English
  - The results where stored in a database and analysed with Python code

The resulting dataset was:

Number of links to articles	803
Links ignored because from reliable source of information list or containing videos	210
Links not processed due to errors (web site down, text too big or other errors)	92
Links successfully processed	501

This means that of all the 803 links, 501 where analyse to produce the following results.

Results

Most common websites

On the left the web site, on the right the frequency in %, filtered by 1% or above.

```
www.parlamentnilisty.cz           5.172414
miaistok.su                       3.879310
protiproud.parlamentnilisty.cz    3.663793
cz.sputniknews.com                2.586207
ria.ru                            2.370690
sputniknews.com                   2.370690
baltnews.lv                       2.370690
aeronet.cz                        2.370690
ac24.cz                           1.939655
prvnizpravy.parlamentnilisty.cz   1.293103
www.rt.com                        1.293103
www.hlavnespravy.sk               1.293103
tvzvezda.ru                       1.293103
tass.ru                           1.077586
ukraina.ru                        1.077586
www.facebook.com                  1.077586
www.nwoo.org                      1.077586
www.bezpolitickekorektnosti.cz    1.077586
lenta.ru                          1.077586
russia-insider.com                1.077586
www.osce.org                      1.077586
```

Most common terms

On the left the term, on the right the frequency in %, first 30 most frequent. Terms can refer to people, events, locations, organisations.

Some terms could be merged, for example Ukraine with Ukrainian.

Russian	2.795699
Ukrainian	2.322581
people	1.913978
United States	1.053763
Ukraine	0.967742
Vladimir Putin	0.774194
Russian Federation	0.709677
Moscow	0.709677
European Union	0.688172
country	0.666667
countries	0.602151
Germany	0.602151
Donald Trump	0.580645
Syrian	0.580645
Kiev	0.559140
war	0.537634
citizens	0.473118
territory	0.473118
European	0.451613
children	0.451613
Donbass	0.430108
authorities	0.430108
Russia	0.408602
President	0.408602
elections	0.408602
Hillary Clinton	0.387097
NATO	0.387097
Aleppo	0.387097
Czech Republic	0.365591
Europe	0.344086

Most mentioned locations

On the left the location type or name, on the right the frequency in %.

Russian	8.795670
Ukrainian	7.307172
United States	3.315291
Ukraine	3.044655
Russian Federation	2.232747
country	2.097429
Germany	1.894452
Syrian	1.826793
Kiev	1.759134

countries	1.623816
Moscow	1.488498
territory	1.488498
European	1.420839
Donbass	1.353180
Aleppo	1.217862
Czech Republic	1.150203
Russia	1.150203
Europe	1.082544
region	1.082544
world	1.014885
American	1.014885
city	0.947226
Donetsk	0.879567
Crimean	0.879567
state	0.811908
Finland	0.676590
Crimea	0.676590
village	0.676590
US	0.676590
states	0.676590

Most mentioned organisations

On the left the organisation type or name, on the right the frequency in %.

European Union	3.686636
NATO	2.073733
government	1.612903
army	1.497696
RIA Novosti	1.497696
group	1.382488
EU	1.382488
Moscow	1.267281
LNR	1.267281
schools	1.152074
Government	1.036866
party	1.036866
military	1.036866
Kremlin	0.921659
CIA	0.921659
forces	0.921659
United Nations	0.921659
team	0.806452
ATO	0.806452
DNI	0.806452
NGOs	0.806452

organization	0.806452
OSCE	0.691244
APU	0.691244
organizations	0.691244
parties	0.691244
TASS	0.691244
Militia	0.576037
newspaper	0.576037
HSP	0.576037

Most mentioned events

On the left the event, on the right the frequency in %.

war	6.393862
elections	4.859335
conflict	3.836317
meeting	3.324808
attack	3.324808
election	2.301790
interview	2.046036
opening	1.790281
introduction	1.790281
incident	1.790281
press conference	1.534527
coup	1.534527
visit	1.278772
crisis	1.278772
attacks	1.278772
speech	1.023018
meetings	1.023018
ceasefire	1.023018
collapse	1.023018
bombings	0.767263
crash	0.767263
death	0.767263
fight	0.767263
terrorist attacks	0.767263
bombing	0.767263
migration	0.767263
conference	0.767263
assault	0.767263
Convention	0.767263
exercise	0.767263

Most mentioned other terms

On the left the term, on the right the frequency in %.

situation	0.953429
information	0.843418
media	0.770077
part	0.770077
number	0.660066
relations	0.623396
fact	0.550055
case	0.513385
regime	0.476714
side	0.440044
policy	0.440044
one	0.440044
work	0.403374
fire	0.403374
right	0.403374
way	0.403374
sanctions	0.403374
question	0.330033
law	0.330033
threat	0.330033
interest	0.330033
ships	0.330033
propaganda	0.293363
problem	0.293363
data	0.293363
head	0.293363
convoy	0.293363
power	0.293363
democracy	0.293363

Authors

Authors on the left, number of articles post on the right. Most of the articles do not have clearly specified authors.

Vedoucí kolotoče	10
Jana Petrova	3
Világfigyelő	3
admin	2

Alena Novotná	2
Caroline B. Glick	2
Baxter Dmitry	2
Eliot Higgins	2
Олесь Бузина	2
Svět kolem nás	2
Tyler Durden	2
Virginia Hale	2
Redakce	2
Juraj Pokorný	2
Václav Dvořák	2
Orosz Hírek	1
ČTK/Pavlíček Luboš	1
Martin Walsh	1
Justin Seitz	1
Игорь Скрыпач	1
Volodymyr Petrov	1
Christopher Black	1
Алексей Громов	1
Артем Филипенок	1
Adam Garrie	1
Investigative Bureau	1
Nick Waters	1
Paul Goble	1
The Saker	1
CBC News	1
	..
Bethania Palma	1
Edmondo Burr	1
Gilbert Doctorow	1
Emil Kalabus	1
RISI TV	1
Ricky Twisdale	1
Nicu Goncear	1
Jiřina Holotová	1
BGŻ BNP Paribas	1
Андрей Шаврей	1
Děnis Kločkov	1
Graham Lanktree	1
redaktor editor	1
Martin Kohout	1
Rudy Panko	1
Boris Reitschuster	1
Jiří Baťa	1
Adam Mosseri	1
Paul Joseph Watson	1
Юрий Канцанс	1
Василий Волга	1

```
Interpreter Staff                          1
Фото: Дмитрий Духанин / «Коммерсантъ»      1
F. William Engdahl                         1
Georg Diez                                 1
Фото: Максим Блинов / РИА Новости          1
John Helmer                                1
Oscar Platt                                1
Roger Annis                                1
Александр ЛУКЬЯНОВ                         1
```

Most common languages

On the left the language, on the right the frequency in %.

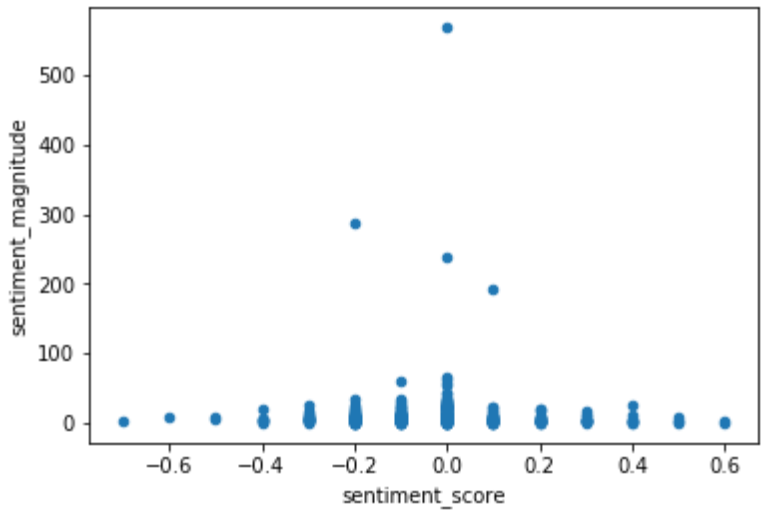
About 90 documents were not translated due to service costs quota limit, this table refers to the percentage of languages of the article which were successfully analysed.

```
en    32.543103
ru    30.818966
cs    25.431034
de     2.801724
sk     2.586207
hu     1.508621
ro     0.862069
bg     0.862069
uk     0.646552
pl     0.431034
nl     0.431034
lt     0.431034
sv     0.215517
lv     0.215517
be     0.215517
```

Sentiment

Sentiment score is a value from -1 to 1, where -1 means very negative, 0 neutral and 1 very positive. Sentiment magnitude is how strong is the sentiment in the news. Each bullet represents a news.

The chart shows that the sentiment of the news is well distribuited between negative and positive around the neutral, which means that the sentiment analysis is probably not very useful to spot fake news.



Conclusion

Some of the articles were not processed to keep the costs under a certain quota. In addition more domains should be probably filtered out because are a reliable source of information, while this study should focus on propaganda articles only. I am not an expert on this field and I do not know which additional domain should be filtered out.

Even with these easy-to-fix limits, 500 articles represent a good starting point to show how articles can be automatically processed with scalable machine learning techniques to extract some valuable information.

Spotting fake news is a very hard task, giants like Facebook and Google are cooperating with reliable source of information to spot misinformation, which means that the human part is still crucial to identify the reliability of news.

Text analysis however can be useful to search for potentially misleading information all over the Internet and alert the staff for potential misinformation before the news spreads the Internet and traditional media.

Possible applications of automated text analysis

- Spot new trends, for example new targeted countries, people, organisations, publishers of disinformation
- Scan main disinformation websites and alert when an article contains sensitive terms
- Scan social networks (eg Twitter) to find accounts posting disinformation before it reaches bigger audience