Project T16

Magnus Paal, Alvar Antson, Tõnis Hendrik Hlebnikov

Business analysis

Business goals

The idea of our project is to scrape data from Postimees Group news websites and to analyse the data and perform machine learning to find the best possible article that will bring in the most money for news websites. We consider our project to be successful, when we have found concrete patterns, that indicate, what are the best things to do for an article to get the most user engagement. We consider machine learning generated articles to be successful, when the article is almost readable and/or generally shows what a popular article looks like.

Inventory of resources

Right now we have a basic scraper program, and approximately 24000 Postimees Group news website articles from 19th of June to the 24th of November. We have the first iteration of our scraper program, which we intend on making faster and extending the types of data, that it is able to retrieve. We also have a team of eager people, who are very interested in this project.

Requirements, assumptions and constraints

We hope to retrieve Postimees Group articles from at least the last 5 years. We might be constrained by the data we can get, like comments, article view count or some other statistics. Also, while we have the possibility, we can’t retrieve too much data, because it will take too long to process.

Risks and contingencies

One risk is that the format of Postimees group articles won't be the same for example 10 years ago, and we can’t retrieve all the columns that exist on more recent articles. There is also a risk that Postimees group hasn't kept articles that are too old. We might run into unseen problems while scraping, which we don’t know how to solve.

Terminology

Web scraping - A way to extract data from websites.

Machine learning - method of data analysis, which includes a system which can learn from data and make decisions based on patterns that emerge.

API - An interface which is used to communicate between two parts of an application. In our case the database and website.

Costs and benefits

Main cost of our project is our team members time and effort. Our project mainly benefits news websites, since we find content that will be the most profitable for them, because it improves user engagement. Other benefits are experience for team members in data mining and science, web scraping, machine learning, natural language processing and many more unforeseen fields we might encounter. The benefits outweigh the costs significantly.

Data mining goals and success criteria

Find what makes an article successful, by analysing its content and title. Main things we plan to analyse are text and title length, number of sentences and paragraphs, amount of specific words, number of names and anything else we can retrieve about words, by using Python's estonian processing library estnltk. Success is measured by shares in Facebook, comment count, view count and other metrics we can retrieve. We consider this to be successful, when there is a clear correlation between an article title or content and its success.

Data Understanding

Gathering data

We ended up gathering data from Postimees Group API, because there was a very interesting column 'read_count' which is very helpful for us and it made things a little bit faster. Unfortunately, we decided to not get comments, because it would have made the process slower and generate even more data, which would have taken too long to process. We ran into many problems, and still haven't fixed all of them, so we couldn't retrieve as many articles as we had hoped.

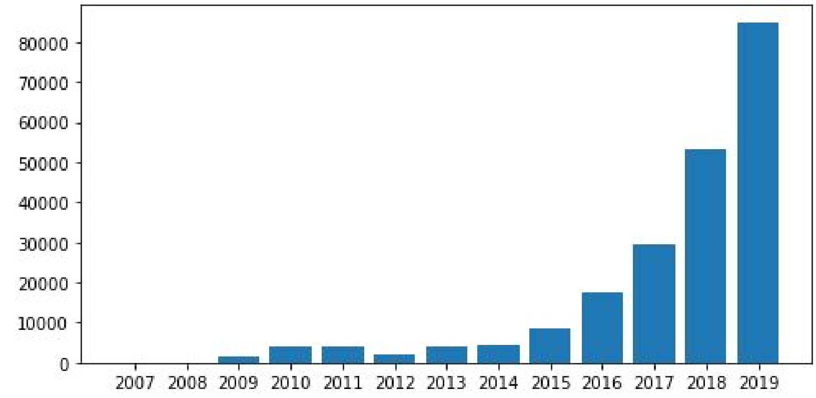
Describing data

From approximately 250k articles we got around 214000 after dropping duplicates. Our plan was to gather as many columns as we could find. In the end we got the article's newspaper, datetime, id, title, content, read count, comment count, facebook share count and author, section. Since section is not really helpful, we excluded it. Also newspaper only had one value, so we excluded that as well. Columns that might need explaining are id, which is a unique identifier of the article, read count, which is the number of people who have seen or read the article. The data collected satisfies our requirements, because we have all the necessary columns that are required to analyze the article (title, content, author) and to give a rating of its successfulness (read count, comment count, Facebook share count).

Exploring data

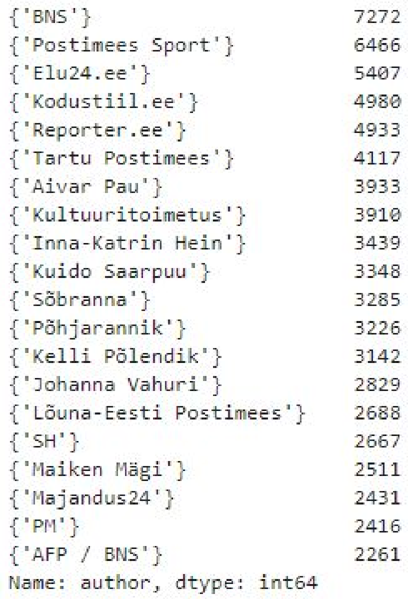
As mentioned above, we gathered 214000 unique articles.

Most articles are from 2019 and the earliest articles are from 2009. (Graph 1)



Graph 1. The yearly distribution of the articles.

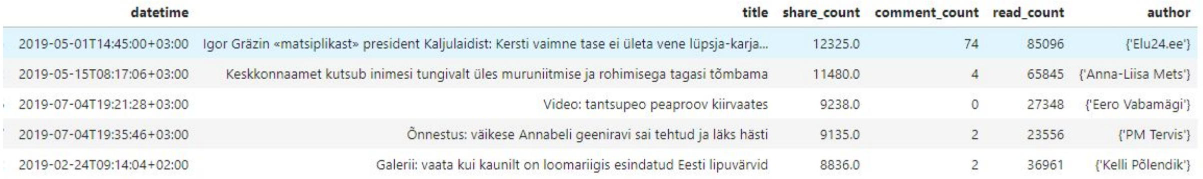
The author column is a mix of newspapers and people/groups of people. In the graph below are the 20 most frequent authors. (Graph 2)



```
{'BNS'}                     7272
{'Postimees Sport'}         6466
{'Elu24.ee'}                5407
{'Kodustiil.ee'}            4980
{'Reporter.ee'}             4933
{'Tartu Postimees'}         4117
{'Aivar Pau'}               3933
{'Kultuuritoimetus'}        3910
{'Inna-Katrin Hein'}        3439
{'Kuido Saarpuu'}           3348
{'Sõbranna'}                3285
{'Põhjarannik'}             3226
{'Kelli Põlendik'}          3142
{'Johanna Vahuri'}          2829
{'Lõuna-Eesti Postimees'}   2688
{'SH'}                      2667
{'Maiken Mägi'}             2511
{'Majandus24'}              2431
{'PM'}                      2416
{'AFP / BNS'}               2261
Name: author, dtype: int64
```

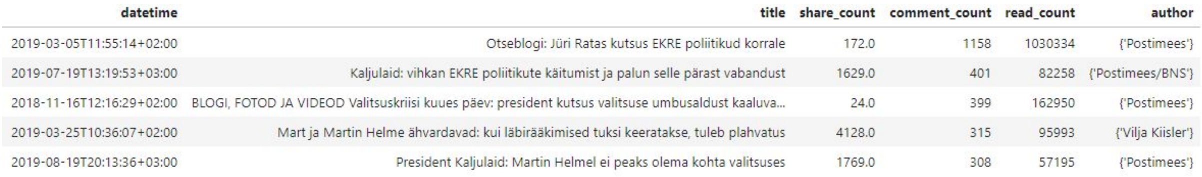
Graph 2. The yearly distribution of the articles.

We also tried to sort the articles by share, comment and read count, to see if any pattern emerge from the top 5. At first glance most shared articles seem to be uplifting news or interesting pictures, videos and events (Graph 3). Most commented articles are about politics (Graph 4) and most read are mostly sport and politics (Graph 5).



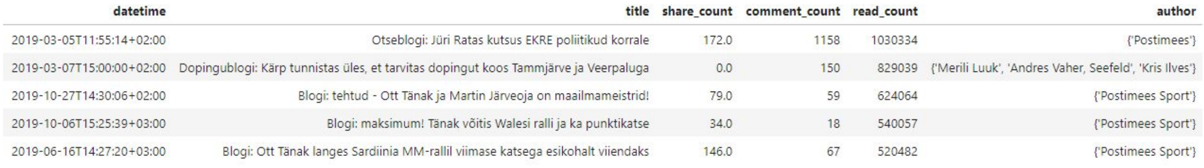
datetime	title	share_count	comment_count	read_count	author
2019-05-01T14:45:00+03:00	Igor Gräzin «matsiplikast» president Kaljulaidist: Kersti vaimne tase ei ületa vene lüpsja-karja...	12325.0	74	85096	{'Elu24.ee'}
2019-05-15T08:17:06+03:00	Keskkonnaamet kutsub inimesi tungivalt üles muruniitmise ja rohimisega tagasi tõmbama	11480.0	4	65845	{'Anna-Liisa Mets'}
2019-07-04T19:21:28+03:00	Video: tantsupeo peaproov kiirvaates	9238.0	0	27348	{'Eero Vabamägi'}
2019-07-04T19:35:46+03:00	Õnnestus: väikese Annabeli geeniravi sai tehtud ja läks hästi	9135.0	2	23556	{'PM Tervis'}
2019-02-24T09:14:04+02:00	Galerii: vaata kui kaunilt on loomariigis esindatud Eesti lipuvärvid	8836.0	2	36961	{'Kelli Põlendik'}

Graph 3. Top 5 most shared articles.



datetime	title	share_count	comment_count	read_count	author
2019-03-05T11:55:14+02:00	Otseblogi: Jüri Ratas kutsus EKRE poliitikud korrale	172.0	1158	1030334	{'Postimees'}
2019-07-19T13:19:53+03:00	Kaljulaid: vihkan EKRE poliitikute käitumist ja palun selle pärast vabandust	1629.0	401	82258	{'Postimees/BNS'}
2018-11-16T12:16:29+02:00	BLOGI, FOTOD JA VIDEOD Valitsuskriisi kuues päev: president kutsus valitsuse umbusaldust kaaluva...	24.0	399	162950	{'Postimees'}
2019-03-25T10:36:07+02:00	Mart ja Martin Helme ähvardavad: kui läbirääkimised tuksi keeratakse, tuleb plahvatus	4128.0	315	95993	{'Vilja Kiisler'}
2019-08-19T20:13:36+03:00	President Kaljulaid: Martin Helmel ei peaks olema kohta valitsuses	1769.0	308	57195	{'Postimees'}

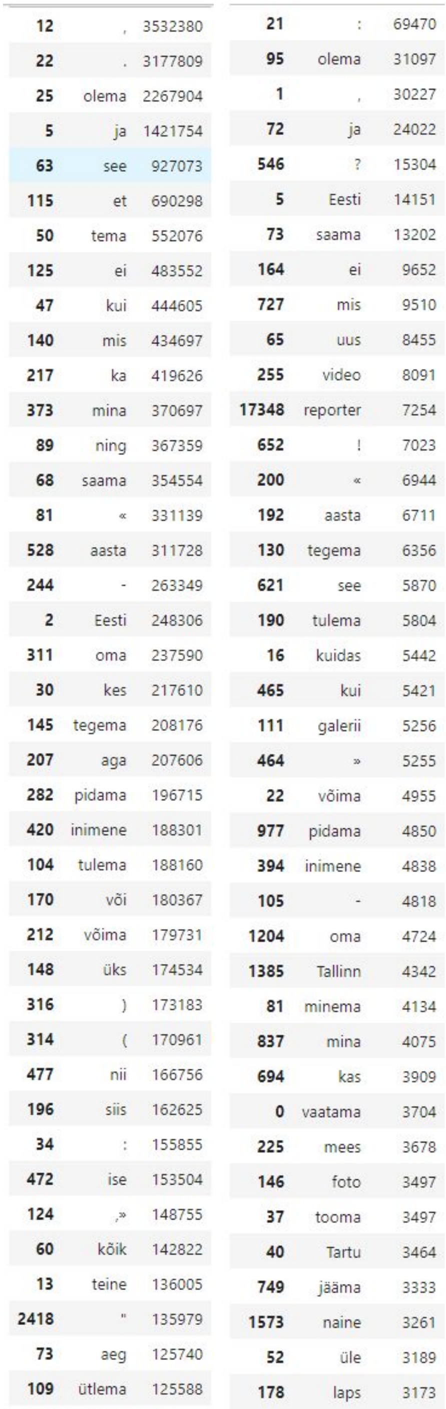
Graph 4. Top 5 articles with most comments.



datetime	title	share_count	comment_count	read_count	author
2019-03-05T11:55:14+02:00	Otseblogi: Jüri Ratas kutsus EKRE poliitikud korrale	172.0	1158	1030334	{'Postimees'}
2019-03-07T15:00:00+02:00	Dopingublogi: Kärp tunnistas üles, et tarvitas dopingut koos Tammjärve ja Veerpaluga	0.0	150	829039	{'Merili Luuk', 'Andres Vaher, Seefeld', 'Kris Ilves'}
2019-10-27T14:30:06+02:00	Blogi: tehtud - Ott Tänak ja Martin Järveoja on maailmameistrid!	79.0	59	624064	{'Postimees Sport'}
2019-10-06T15:25:39+03:00	Blogi: maksimum! Tänak võitis Walesi ralli ja ka punktikatse	34.0	18	540057	{'Postimees Sport'}
2019-06-16T14:27:20+03:00	Blogi: Ott Tänak langes Sardiinia MM-rallil viimase katsega esikohalt viiendaks	146.0	67	520482	{'Postimees Sport'}

Graph 5. Top 5 articles with most reads.

To get a quick overview of the text, we found the most frequent words in the title as well in the content of the article. The words were converted into their roots using a python language processing library estnltk. This gives us an overview of what we need to get rid of, or what we need to take into consideration, when diving deeper into the data. There are a lot of single punctuation characters and conjunctions which we don't need to take into account. (Graph 6, 7)



12	,	3532380
22	.	3177809
25	olema	2267904
5	ja	1421754
63	see	927073
115	et	690298
50	tema	552076
125	ei	483552
47	kui	444605
140	mis	434697
217	ka	419626
373	mina	370697
89	ning	367359
68	saama	354554
81	«	331139
528	aasta	311728
244	-	263349
2	Eesti	248306
311	oma	237590
30	kes	217610
145	tegema	208176
207	aga	207606
282	pidama	196715
420	inimene	188301
104	tulema	188160
170	või	180367
212	võima	179731
148	üks	174534
316	)	173183
314	(	170961
477	nii	166756
196	siis	162625
34	:	155855
472	ise	153504
124	,»	148755
60	kõik	142822
13	teine	136005
2418	"	135979
73	aeg	125740
109	ütlema	125588

21	:	69470
95	olema	31097
1	,	30227
72	ja	24022
546	?	15304
5	Eesti	14151
73	saama	13202
164	ei	9652
727	mis	9510
65	uus	8455
255	video	8091
17348	reporter	7254
652	!	7023
200	«	6944
192	aasta	6711
130	tegema	6356
621	see	5870
190	tulema	5804
16	kuidas	5442
465	kui	5421
111	galerii	5256
464	»	5255
22	võima	4955
977	pidama	4850
394	inimene	4838
105	-	4818
1204	oma	4724
1385	Tallinn	4342
81	minema	4134
837	mina	4075
694	kas	3909
0	vaatama	3704
225	mees	3678
146	foto	3497
37	tooma	3497
40	Tartu	3464
749	jääma	3333
1573	naine	3261
52	üle	3189
178	laps	3173

Graph 6, 7. 40 most frequent words in the content of the article (Graph 6) and in the title of the article (Graph 7).

Data quality

By looking over all the values in the columns, there doesn't seem to be any missing values. We also checked if older articles have facebook shares and website comments, because it might be a newer feature. Very few older articles seem to have them, but this might be because we don't have as many articles from those dates, and because fewer people used online news websites 10 years ago. The titles and article content should all be correct.

Project plan

1. Developing a scraper and scraping data from Postimees API, that can scrape from Postimees Group sites. Is written in Python and retrieves data from Postimees API. (Tõnis, Alvar) (10 hours per person)
2. General analysis of data, finding patterns on how title, content, author or anything else influences the popularity of the article which is measured by facebook shares, comments and view count. Plan to use natural language processing python library called estnltk. (Magnus) (20 hours)
3. Machine learning generated news using python. Generating fake news articles based on all articles. (Tõnis, Alvar) (20 hours per person)
4. Creating the poster for the poster event. (Magnus) (10 hours)