Environmental topic classification and parliamentarian recognition

Elías Bjartur Einarsson
University of Iceland
ebe@hi.is

Karl James Pestka
University of Iceland
kjp3@hi.is

Abstract

In this project we sought applications of natural language processing to simplify the task of classifying news and political texts on the topic of sustainability and the environment. We also tracked mentions of Icelandic parliamentarians and political parties associated with news on these topics.

We used a corpus of news articles in Icelandic from the past 10 years as our data set. Articles were tokenized and lemmatized in Icelandic, then grouped using unsupervised topic discovery with a Latent Dirichlet Allocation (LDA) classifier (Blei et al., 2003) in order to find clusters which were more likely to predict environmental subject matter. Latent Semantic Indexing (LSI) provided even more targeted filtering for environmental topics in the news.

We then counted mentions of politicians and their positions in the text using a fine-tuned XLMR-enis model for the task of named entity recognition (NER). Finally, we gathered statistics on the tracked entities.

Topic classification did not show expected promise, since any use of either LDA or LSI would either exclude articles with environmental focus or include many others unrelated to the topic. However, gathering statistics on parliamentarian mentions in the news provided much more interesting results and the bulk of our research went into this task.

Further research in tracking political mentions offers much promise, and several directions are suggested for expanding upon our efforts.

1 Introduction

In early September, 2021, The Icelandic Youth Environmentalist Association (Ungir Umhverfissinnar) published Sólin 2021, a graded assessment of political parties' environmental platforms ahead of parliamentary elections (Umhverfissinnar, 2021). The goal was to aid the democratic process, allowing "all Icelanders, young and old, to make an informed decision on which way to vote in September." The report involved significant text processing and manual sorting, and we were interested to see if natural language processing (NLP) could be employed to simplify future reports, or even provide ongoing assessments of members of parliament through monitoring environmental news and parliamentary reports.

1.1 The Sólin 2021 report

The Youth Environmentalist Association (UU) used a multi-stage process to achieve an impressively detailed and specific report on each party's platform (see Figure 1).

UU began by hiring an interdisciplinary team of graduate students: a biologist, a political scientist, and a psychologist, all of whom were completing further studies in environmental science and sustainability.

The grading criteria used for the assessment was determined through a consensus of the UU board, the interdisciplinary team, as well as outside experts. Suggestions for the criteria were gathered from the 1200 members of UU.

Next, UU sent the grading schematic for Sólin to all political parties in May 2021, inviting representatives to an introductory meeting. Party platforms would be graded according to how well they addressed concerns around three main topics: climate change, nature conservation, and circular society (Table 1).

Statements and platforms were solicited from each party on these topics, and UU assembled a large corpus of text in pdf and Microsoft Word format. Next, references to politicians, parties, and any identifying information were removed in an anonymization stage to assist the judges in grading impartially. Finally, the interdisciplinary team

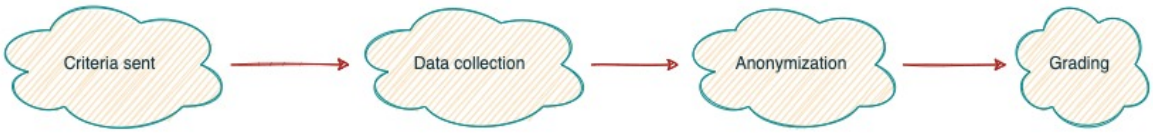

Figure 1: The multi-stage assessment process employed by the Sólin team.

Category	Percent of final grade
Climate Change	40%
Nature Conservation	30%
Circular Society	30%

Table 1: Main categories used in the Sólin assessment.

and multiple experts submitted scores for each anonymized policy. Throughout this process, the determination of grading criteria and all judging excluded anyone with political ties.

The average scores on each policy were then calculated and assigned to back to the originating parties. The final scores for each party's platform were published in early September 2021.

1.2 Natural language processing tasks

At which stage(s) could NLP simplify this process? We began by asking the Sólin team what their main bottlenecks were.

Data collection proved to be a significant project, as the Sólin team had to chase down answers from parties in order to assemble the text in a timely manner. We intuited that an NLP classifier might be able to provide some assistance in filtering environmental policies from unrelated platform topics at this stage (see Figure 2).

Additionally, named entity recognition (NER) could automate the process of anonymization. However, we determined this task to be outside the scope of this project as we were primarily focused on classification and statistics. This relatively straightforward application of NER offers a promising direction for future research.

Another difficulty the Sólin team encountered was the task of interpreting whether an environmental policy that *appeared* to specifically address a Sólin goal was *actually in favor of that goal*, or simply pivoting by rewording the goal in a clever way. For example, one party might have received points based on a literal interpretation of their stated goal of "discussing" the restoration of wetlands. A more careful reading of this policy, however, indicated that the party's intention was to downplay wetland renewal by suggesting that the topic required more discussion (before any action could be taken). In other words, the statement sought to subtly cast doubt upon the scientific consensus on the value of the wetlands. Such a nuanced interpretation could only be made in light of rest of the party's platform, a broad context which might elude all but the most finely-tuned machine interpretations.

The Sólin team thus made a convincing point of the necessity of human interpretation at the final grading stage, in order to catch such edge cases. NLP might then be better suited to provide an intermediate sentiment analysis of environmentally topical statements, such as indicating whether they appeared to be positive or negative; however any such simple binary would be a "naive" baseline necessarily to be improved upon by human interpretation in the final stage.

An additional task well-suited to NLP would be using NER to tally mentions of politicians and parties, and to track their appearance in environment-related texts. It was this final task which eventually became our focus, being better suited to the particular data set with which we ended up working.

2 Data Collection

In order to train a classifier that could identify text relevant to the environment and sustainability, we needed labeled training data. Such a classifier could be trained, for example, on Sólin's labeled political texts; it could then be applied to predict the topicality of other input, such as the news, parliamentary laws, election platforms, and future assessment projects. However, we did not receive the Sólin data in time and instead commenced training with a corpus of 640,136 news articles in Icelandic from 2011 to November 2020 (see Table 2).

By beginning with the news data set, we could get started on training a general model, which would be ready for fine-tuning to Sólin's data at a later stage. Furthermore, a larger training and testing set seemed to suggest a more accurate model.

Using a news corpus had the additional benefit of allowing us to explore NER and gather statistics

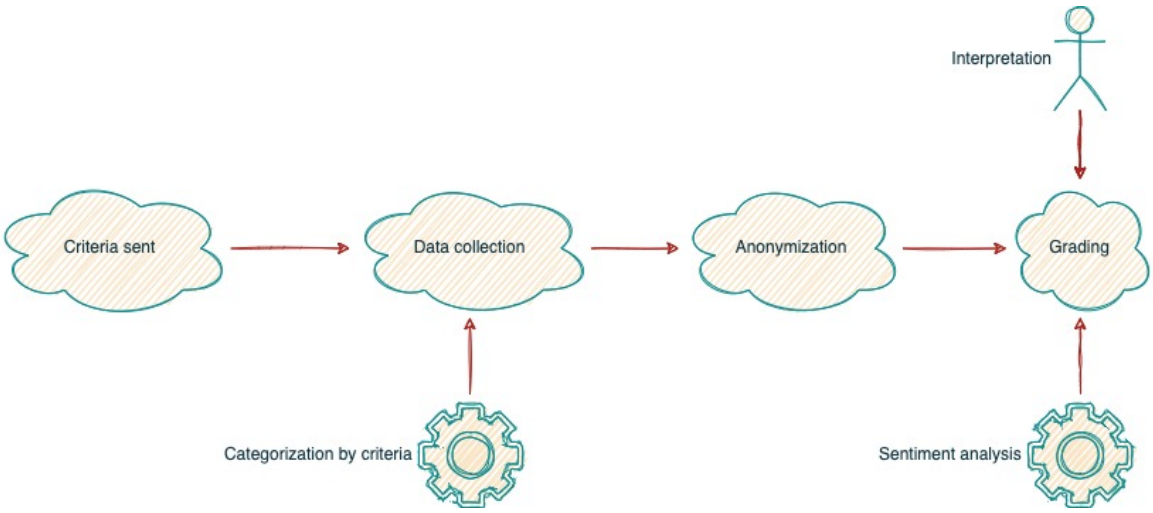

Figure 2: Tasks that could be delegated to NLP.

on mentions of politicians and parties.

2.1 Limitations

One issue with the news data set was that articles relevant to the environment or political party platforms were relatively rare. For whatever the reason that may be, it certainly makes sense to specifically discuss the platforms of non-governing parties in the rare pre-election season.

Another issue related to the NER task Icelandic news tends to interview the governing coalition and its politicians most, leaving out minority and non-parliamentary parties. Thus any statistics on parliamentarians from this data set would be less useful for informing future elections and more useful for assessing the past.

We randomly sorted the data into a training and test set of 85% to 15%. Since none of the data was labeled by topic, we hand-labeled a subset of each of the training and test sets as shown in Table 2.

2.2 Data pre-processing

We tokenized our articles by splitting sentences into words, then removing punctuation and the most common Icelandic words. Then we performed a naive lemmatization where we only retained nouns that had a single possible lemma using the Bin package from Miðeind (Þorsteinsson et al., 2021). After pre-processing we had 320,706 unique tokens.

Data set	Articles	Percent
Total data set	640,136	100%
Training set	544,115	85%
Test set	96,021	15%
Total labeled	155	
Training set labeled	117	
Test set labeled	38	
Total unique tokens:	320,706	

Table 2: Split of data used for training and testing.

3 Methodology

3.1 Labeling with semi-supervised learning and LDA

To obtain the large sample of labeled data needed for training a classifier, we utilized semi-supervised learning in order to discover the clusters of news articles that contained the most environment-related news. To do this, we had to first manually label some data as environment-related and unrelated news. We could then use unsupervised topic discovery and observe which topic cluster the positive and negative examples end up in. If all positive examples ended up in a single cluster, the the rest of the articles in that cluster could potentially be labeled as environmental as well. If a mix of related and unrelated articles appeared in the cluster, one could at least use eliminate clusters containing none of the hand-labeled examples, filtering out all but a subset of "likely" data for further classification. This technique is often useful when dealing with large amounts of unlabeled data (Lin, 2020).

To obtain the first articles for manual labeling, we wrote scripts that searched for environmental news articles (positive examples) by keyword, and then labeled the output. The first keywords were simple: "náttúruvernd", "loftslag" and "umhverfi". The search included all words that contain these words or their Icelandic inflections, which is important as Icelandic words are often compound. This filter returned about 7% of articles, and with manual labeling only about 1/6 of those actually proved positive. Thus about one percent of articles seemed to be positive examples.

After having manually labeled over 20 positive examples we created a Latent Dirichlet Allocation classifier (Blei et al., 2003). In simple terms LDA works by placing both the bag-of-word representations of the articles in a Dirichlet distribution among the number of topics and the topics in a Dirichlet distribution among the number of tokens. The optimal arrangement, found with machine learning optimization, "discovers" the topics, while the number of topics to be found is a tuneable hyperparameter. We hoped of catching all labeled examples under one topic and began training for 10 topics, using the Gensim package to create the model (Řehůřek and Sojka, 2010). Further classification on that single topic could then narrow in on our focused subject.

3.2 Finding similar articles using Latent Semantic Indexing

After mixed results from the LDA model we moved to a Latent Semantic Indexing (LSI) model. Here used the same pre-processed corpus that we used in our LDA model but first implemented Term Frequency - Inverse Document Frequency (TF-IDF). TF-IDF assigns weights to each token proportional to its rarity in the corpus. That is, a term that appears often in a document but seldom in others is assigned a higher weight in defining the topic of that document. On the other hand, common words that often appear in documents aren't weighted as much, such that they contribute less towards the classification of documents in which they appear.

The LSI model makes use of TF-IDF weights when identifying topic areas. It does so by factoring a matrix listing included tokens for each document, as weighed by TF-IDF. The resulting matrix factor determines the topics. We were able to train a model with 200 topics and another with 500 topics.

The resulting model was then used to assess the distance between different articles and a bespoke topic vector containing hand-selected keywords relating to the environment and sustainability. The included keywords are listed in Table 3.

Topic vector		
'andrúmsloft'	'loftslagsaðgerðir'	'sjávarborð'
'auðlindaráðuneyti'	'loftslagsbreyting'	'sjór'
'aðgerðasinni'	'loftslagslíkan'	'skref'
'bráðna'	'loftslagsmál'	'stóriðja'
'friðland'	'loftslagsráðstefna'	'sóun'
'friðun'	'loftslagsstefna'	'súrnun'
'gráðurhúsaáhrif'	'loftslagsvá'	'umhverfi'
'gróðurhúsalofttegund'	'loftslagsvísindi'	'umhverfisdómstóll'
'hamfarahlýnun'	'loftslagsáhrif'	'umhverfisfræði'
'heildarlosun'	'losun'	'umhverfissjónarmið'
'hlýnun'	'mengun'	'umhverfisskattur'
'hringrásarhagkerfi'	'náttúra'	'umhverfisvernd'
'jarðefnaeldsneyti'	'náttúruauðlind'	'umhverfisverndarsinni'
'jökull'	'náttúruhamfarir'	'umhverfisverndarstofa'
'jörð'	'náttúruvernd'	'umhverfisáhrif'
'kol'	'náttúruverndarlög'	'umhverfisþing'
'kolefnishlutleysi'	'náttúruverðmæti'	'vistkerfi'
'landnotkun'	'orkugjafi'	'ísöld'
'lifnaðarháttur'	'orkunotkun'	'úrgangur'
'loftslag'	'parísarsamkomulag'	'útblástur'
'loftslags-breyting'	'sjálfbærni'	'útlosun'
		'þjóðgarður'

Table 3: Keywords used to define environmentalism in LSI.

This provided some additional accuracy in finding related articles over LDA, although optimizing accuracy through hyperparameter tuning (in particular, adjusting the threshold for the similarity score) proved challenging.

3.3 Counting parliamentarian mentions with NER

Our final task was to perform Named Entity Recognition (NER) on our data in order to count mentions of parliamentarians in news articles. For that we used an XLMR-enis model (Ingólfsdóttir et al., 2020) which we had previously fine-tuned on Icelandic text. We obtained data from the Icelandic parliament website on current members of parliament, as well as all previous parliamentarians. We used both the aforementioned model as well as regular expression methods to create a lookup table of parliamentarians to parties. We then searched different articles for the appearance of these names and kept count of their appearances and positions in the texts. Lastly, we took note of the most interesting statistics gathered from a set of random articles, as well as a set of topical articles that had been selected by tuning the LSI model's similarity threshold of 0.15.

Data	Sum of score	Top score
Positive	6, 8, 3	6, 3, 8
Filtered but negative	6, 8, 0	6, 3, 8
Random negative	3, 8, 6	3, 8, 0

Table 4: LDA classification results showing top 3 categories for the sum of scores as well as the most frequent top score category

Figure 3: Wordcloud for topic 6, the most frequent category for positive examples as well as others

4 Results and Evaluation

4.1 LDA topic discovery

Each of the ten topics "discovered" by the LDA classifier returned a set of uniquely weighted keywords found among articles within the topic. Some topics were fairly straight forward to interpret from viewing the keywords, such as sports and finance; others were more vague and open to interpretation. After training the model, our goal was to observe which categories the positive examples appeared in the most. We measured this in two ways, both by summing up all the scores from the different categories for all positive labeled training examples, and then by discreetly counting which category most often had the top score (the top category would effectively be chosen for that example). The top three categories from both orderings turned out to be the ones indexed by 6, 8 and 3. Unfortunately the same categories proved top contenders for the articles that had passed the keyword filter but were hand-labeled as negative, as well as for a random sample of negative examples, though to a lesser degree. The top scores and top sums can be seen in table 4.

These results meant that this classifier could not give us a reliable way to filter for positive results. It could be used to filter out negative articles by excluding categories other than 6, 8 or 3 but since these seemed to be the most popular categories for the whole data set such filtering would not be of much use.

The only possibility of using LDA seemed to be that category 6 was seldom the top category for the random negative data set, suggesting that articles that had their highest score for category 6 would be more likely to relate to matters of the environment and sustainability. In reality, however, many non-topical articles would clearly pass that filter and many truly environmental articles would fail to pass it.

4.2 LSI topic filtering

For these reasons we moved on to the LSI model. We compared two models, one which had been trained with 200 topics and another with 500, using the same pre-processed corpus as the LDA model but adding TF-IDF weighting as previously mentioned. The 200 topic model proved more useful and the discussion of results will refer only to this.

The resulting model was used in conjunction with a hand made topic vector for environmentalism. The similarity distance (the cosine of the angle between the vectors) between the topic vector and different news articles was measured and an arbitrary threshold of 0.15 was used. By using this threshold, we were able to guarantee that most articles passing that threshold were related to environmentalism, although quite a few were unrelated. Bafflingly, only 4 of the 21 training articles initially hand-labeled as environmentally topical passed that threshold. The situation was such that a higher threshold would exclude many positive examples but a low threshold would include too many unrelated articles. Thus 0.15 provided the needed functionality to label more training data, but was not an optimal predictor for the larger task of classification.

4.3 NER mention counting

Due to our limited success, we started work from the other end of our task process. The original task was to classify texts into environmental and non-environmental, then process the positive cases in order to find political references. Now we started looking for parliamentarians in articles in general, and observing how they were distributed among parties and times.

For this we used a fine-tuned XLMR-enis model as well as regular expressions. Unfortunately do to time- and computational limitations we did not

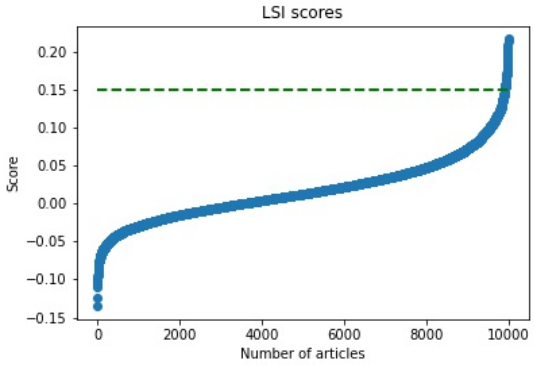

Figure 4: LSI similarity scores for the topic vector and 10,000 random articles. The threshold, 0.15, was determined arbitrarily.

7000 random	LSI filtered
Katrín Jak.	Guðmundur I. G.
Bjarni Ben.	Katrín Jak.
Sigmundur D. G.	Bjarni Ben.

Table 5: Top three current parliamentarians found in random articles, as well as in LSI-filtered environment-related articles.

manage to process a significant portion of the training data. We did however process the articles that passed the 0.15 LSI threshold as well as a random sample for comparison, larger in size.

At this point it is important to note again the limitations of the corpus, as they will influence any statistics about parliamentarians. The corpus of news articles over-represents parliamentarians who have been in parliament for a long time. Furthermore, our implementation of name recognition only recognizes full names. It is quite customary to include the full name of a person at least once in a news article but this might still over represent those that don't have a middle name, for instance, since they are referred to in full more often to for disambiguation purposes. Moreover, it is likely that the ruling parties for each time period are more represented since appointed ministers appear more often in the news than parliamentarians in general.

We focused mainly on statistics for *current* parliamentarians but compared that to statistics for all previous parliamentarians. The top represented current parliamentarians for all time periods are shown in table 5.

The main difference between the random sample and LSI-filtered sample seems to be that Guðmundur Ingi moves to the top in the filtered articles.

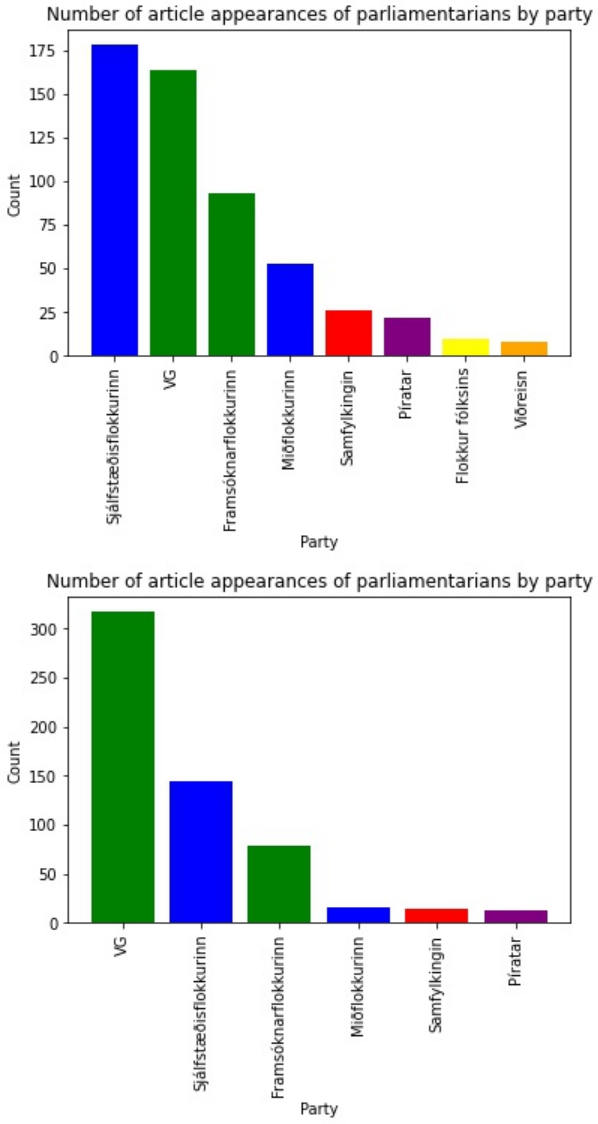

Figure 5: Party representation in random news (above) and LSI-filtered news (below).

This is not surprising since he was the minister of the environment and natural resources (up until the writing of this article). For a more detailed view of the statistics see the the Appendix. It is revealing to see to which parties these parliamentarians belong and whether they have equal representation. That can be seen in Figure 5.

In order to negate the effects of time spent in parliament we also looked at similar statistics for different years. Suffice it say that for the last year of the data set (2020) the results were similar with Katrín Jakobsdóttir and Bjarni Benediktsson most prominent in random news and Guðmundur Ingi Guðbrandsson surpassing them in the filtered set. Previous years of course had different figures in the foreground but generally the main ministers were

heavily represented.

The analysis of the same data but not only for current parliamentarians but every current and previous member did not change the results for the top number of mentions but mainly added many more to the list of included.

5 Conclusion and Next Steps

We experienced limited success in classifying articles with environmental topics, as we were unable to accurately process our unlabeled data set with unsupervised models. Due to the lack of labeled articles, there seemed little promise in to the task of sentiment analysis within environmentally topical data.

The most viable application of NLP we discovered turned out to be examining statistics on mentioned parties and parliamentarians. This task was derived less from the initial needs of the Sólin assessment and more from the nature of our news dataset. However, such statistics could potentially provide the public with constantly updated track records on each current parliamentarian and their level of involvement with environmental policies and related news.

In retrospect, more work might have gone into obtaining a topical corpus for training, possibly by drawing news from environmentalist publications and political platforms from party websites. Topical news articles on political platforms could also be targeted by filtering for articles during pre-election season. While it was unfortunate not to have been able to access Sólin's data, gathering a corpus of party platforms by hand would not have been far off the mark towards classifying policy.

5.1 Further Research

The "mention counter" invites many avenues of further investigation. For example, we did track where in each article the parliamentarian or entity appears and that could be used to perform topic classification or sentiment analysis in their vicinity. Furthermore it would be possible to track the most likely topics of each party and parliamentarian.

Another interesting research question would be to investigate whether parliamentarians favored by the "mention counter" are actually representative of the Icelandic parliament. Once the bias towards ruling parties is factored in, are there any other factors that predict over-representation in the news? In particular, is there a tendency to mention a particular gender, age, or race? It would also be compelling to examine outliers who are over- and undermentioned.

Acknowledgments

We would like to thank Vésteinn Snæbjarnarson for his enthusiastic help and support as well as Radim Řehůřek, the creator of Gensim, for making what would otherwise have been a tedious job close to effortless.

Bibliography

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Svanhvít Lilja Ingólfsdóttir, Ásmundur Alma Guðjónsson, and Hrafn Loftsson. 2020. MIM-GOLD-NER – named entity recognition corpus. CLARIN-IS.

Jianan Lin. 2020. Semi-supervised learning with k-means clustering. *Towards Data Science*.

Vilhjálmur Þorsteinsson, Hulda Óladóttir, and Sveinbjörn Þórðarson. 2021. BinPackage. CLARIN-IS.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Ungir Umhverfissinnar. 2021. Sólin: Einkunnagjöf ungra umhverfissinna fyrir alþingiskosningar.

A Appendix

The following plots display the statistics for parliamentarian mentions in news articles.

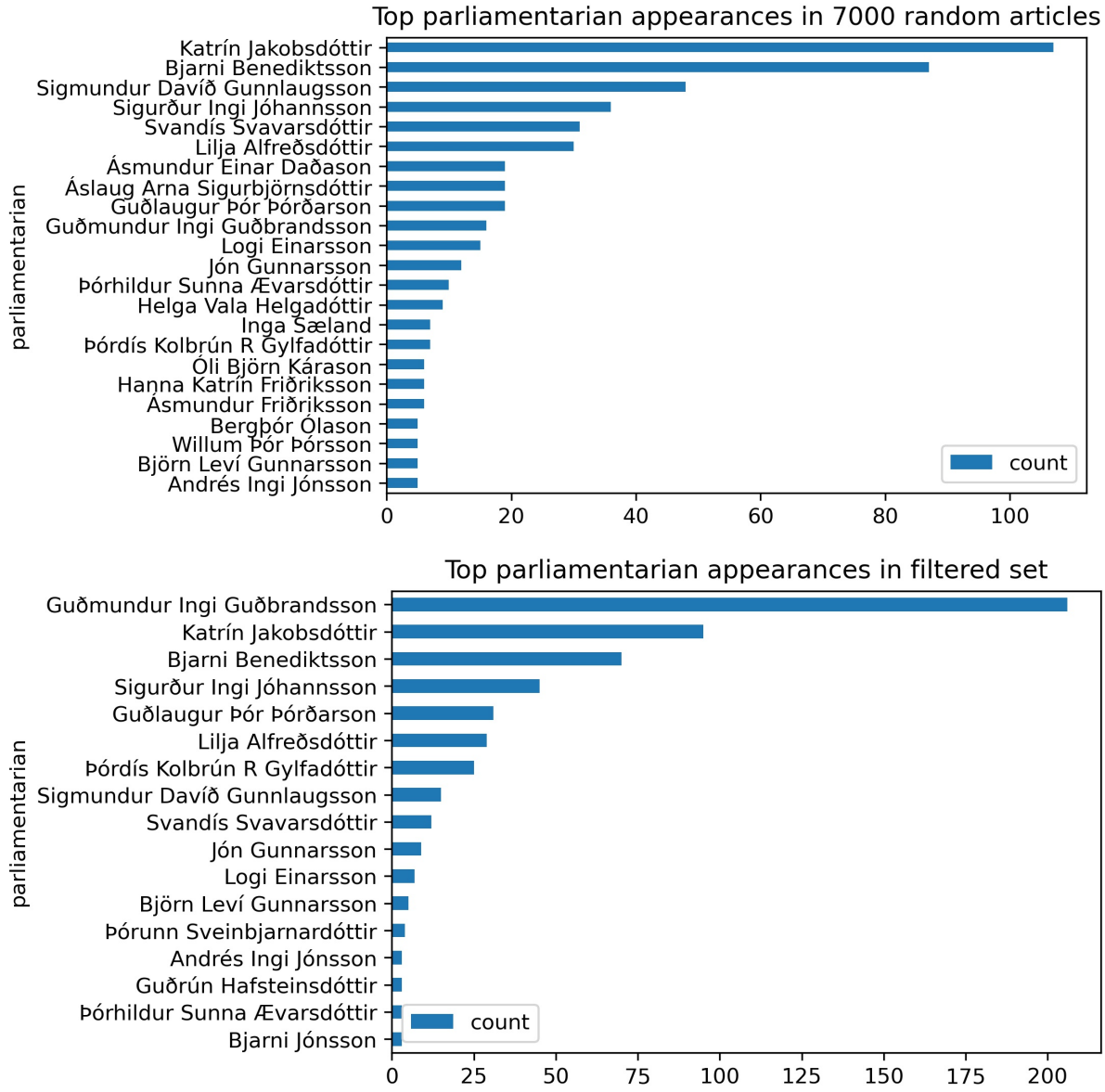

Figure 6: Top current parliamentarians for random articles (above) and LSI filtered articles (below)

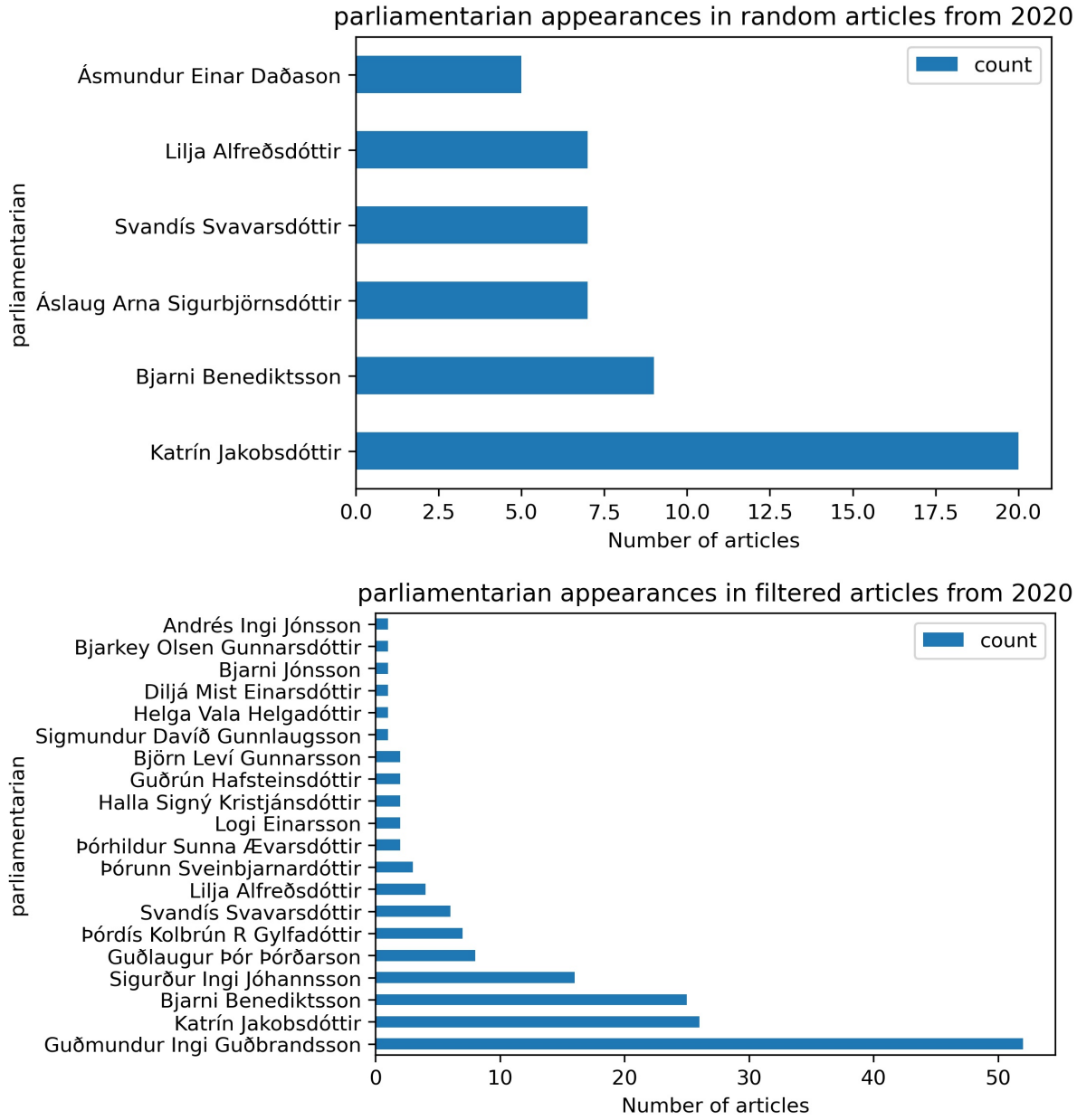

Figure 7: Top current parliamentarians in the year 2020 (until November) for random articles (above) and LSI filtered articles (below)

The following plots show the distribution of scores in each LDA discovered topic.

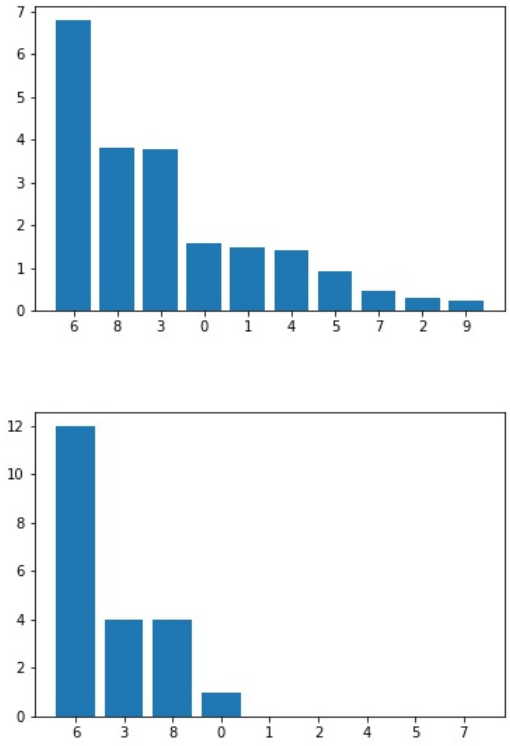

Figure 8: Positive examples, sum of category scores (above) and top category (below)

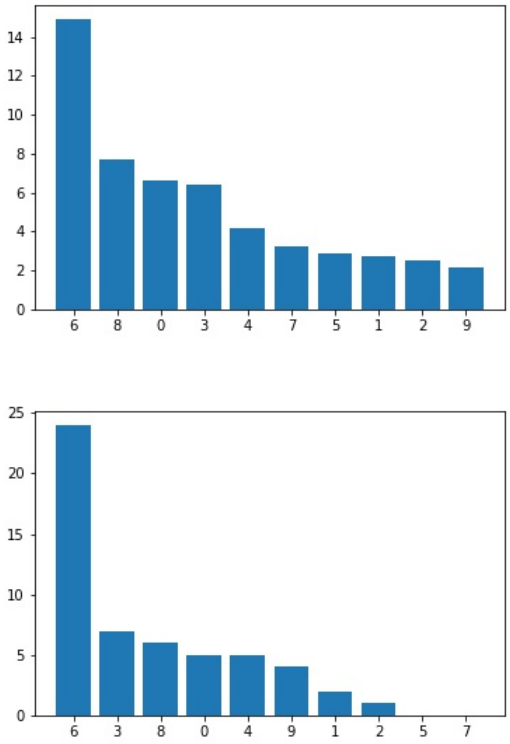

Figure 9: Negative examples that passed the keyword filter, sum of category scores (above) and top category (below)

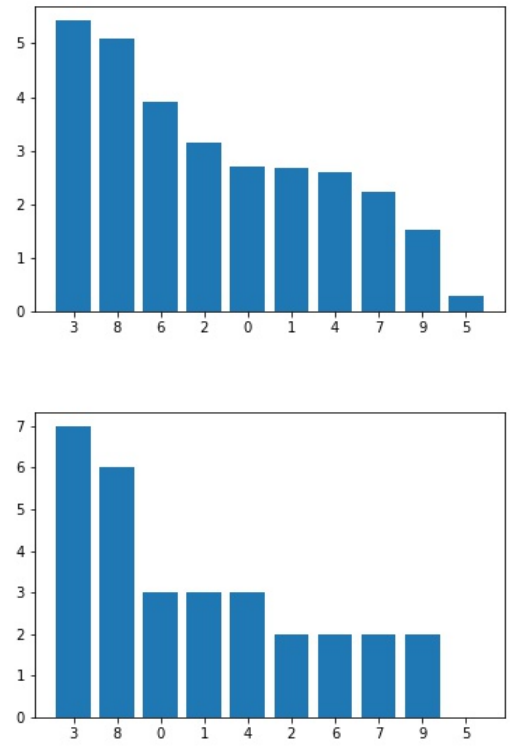

Figure 10: Random negative examples, sum of category scores (above) and top category (below)