Examination number

9, 45, 59 - Group 16

Course title: Social Data Science

Winter Exam 2015

Submission date: 15th December 2015

Number of pages including front page:

24

What makes news on social media

Benjamin Wedel Mathiasen, Bjarke Dahl Mogensen, Mikkel Mertz

Contents

Introduction

Social media and in particular Facebook has become a part of the daily life during the last decade. Recent studies has shown that a growing part of the younger generations get the majority of their news from social media with Facebook as a popular choice (Reuters Institute Digital News Report 2012, p. 13). The increasing use of social media and the fact that a growing part of the younger generations get their news from social media might explain why an increasing proportion of Danes read news every day (Danske unges museums- og mediebrug, p. 55).

But is the news on social media representative?

Zuckerman (2013) finds that the media scene in general has become more focused on domestic news than previously and that news on social media to an even greater extent when considering the American news scene. The vast majority if not all social media monitor online behavior and use various algorithms to determine what specific stories to target the consumer. On Facebook, this means they control what news should pop up in the top of the newsfeed and thus gets most attention. This may ensure that people get to know what they want to know but not necessarily, what they need to know. Since there is a tendency for people to follow news closer to them more intensive and media tools at the same time expose people to specific stories, using this information, it seems to be a self-enforcing effect potentially indoctrinateing people in their own beliefs. In other words people only get exposed to stories supporting their beliefs.

The fact that beliefs in a sense are amplified within the closed system is referred to in the literature as "echo chambers" (Sunstein (2001)). The existence of echo chambers potentially narrows peoples' view of the world. Because of various algorithms two almost identical persons potentially get very different results using a search engine as for example Google. This happens because of Google's search algorithm, exposing persons to different results, pending on previous search history. While this makes searching online more efficient, vital information might get left out. Not everyone agrees that this is in fact the case and research made by The Media Insight Project found that in America 86% of people in the age group 18 to 34 years meet views different than their own on social media. This would make the

potential problem smaller or insignificant (How Millennials Get News: Inside the Habits of America's First Digital Generation. AP, University of Chicago and American Press Institute, march 2015).

The majority of Danish news media are represented both on an official website and a corresponding Facebook page - the Facebook page posting links to selected stories. However, far from all articles are posted on Facebook stating that selection occurs. But how is this selection made? Is it a random draw or is it a specific type of news qualifying for the Facebook page? In this specific setting, it might seem reasonable to assume that news media select the articles expected to be most popular thus in its very sense, subjecting people getting their news solely from Facebook, to a very selected subset of news - namely the news simular to previous stories getting a large amount of attention. It could be the case that a higher share of domestic than international articles are posted on their Facebook page since domestic articles may be more relatable than international articles, thus getting more views. If this is true it might be a significant factor in existens and creation of echo chambers. The aim of this paper is twofold:

1) We wish to investigate what type of selection process, if any, is going on between the official website and Facebook page for a subsection of Danish media

2) We wish to investigate if a specific type of articles ending up on Facebook catches most attention

We investigate the first question by collecting articles published on DR's and Politiken's official website and look into what characterizes the articles making their way to Facebook. We do this, using a probit model and supervised learning models. We chose these specific media since they are two of the most online read media, thus hopefully catching a large fraction of the overall picture. For the second question, we consider the articles selected for Facebook and investigate whether it is a specific type of articles getting most attention on Facebook. We use the number of likes, comments and shares as a proxy for how many people getting exposed to the article.

We find that there is a bias in what type of articles being posted on Facebook, both for DR and Politiken. A higher share of domestic than international articles are being posted on

Facebook, compared to their official website. Likewise domestic articles are more popular than the international ones. This indicates echo chambers potentially exist in Denmark.

The following section, describes how the data is collected. Hereafter, we will proceed to examine what characterizes the news entering Facebook, followed up by analyzing what makes news on Facebook popular. Hereafter, we look into the distribution of articles in the International section. In the last section we give some concluding remarks.

Data

The data used in this paper consists of articles scraped from DR and Politiken's website and their corresponding Facebook pages, DR Nyheder and Politiken. We have scraped all articles from 18th of November 2014 and one year forward. We have chosen to scrape articles for one year to put an upper limit on the number of articles scraped, but still have a serious amount of articles to base our analysis on. A total of 74,966 articles were scrapped but after some revising of the data we ended up with a data frame containing 43,244 articles – 21,938 articles From DR, and 21,306 articles from Politiken.

Scraping the websites

In order to scrape DR and Politiken's website we used Google Chrome's CSS selector extension -SelectorGadget. We scraped the media's news archive for the article's href link (article link) and date of the article release. When this information was obtained we used the article link to scrape the title and text of all articles. At last this information was merged together in one data frame by the articles link. The news section in which the articles were posted was obtained from the articles' link. Some article links were directing to an error-page. These links were left out. Also, some of the articles' text contained links to other articles and the belonging title. The links had the same CSS path as the article text, meaning the links could not be deselected while the text was selected. These links/titles were removed by substitution since they potentially could bias our results of the supervised learning models.
For practical reasons the articles were scraped in several rounds. For DR each news section

was scraped separately whereas for Politiken the articles were scraped by time periods. All the articles were merged together in a data frame at the end. Some articles appeared in more than one section. With a little inspection it was clear that these articles fitted into more than one section so we randomly removed duplicates.

Scraping the Facebook pages

We scraped the medias' Facebook pages "DR Nyheder" and "Politiken" using Facebook's API, which is convenient. We scraped information of the number of likes, comments and shares for the last 10,000 posts to make sure we had posts going back to 18th of November 2014.

Websites and Facebook pages merged

The data from the medias' websites and Facebook pages are merged together by the article link. Everything posted on the Facebook pages that is not posted on the websites, is discarded during the merge. If what systematically is left out, is a specific type of stories it can potentially bias our results. However, the removed stories were mainly videos from different sources and would not have made a difference since we only consider articles in the current setting. The new data frame contains all articles scraped from the websites and Facebook information about the articles posted on Facebook.

Revising the data

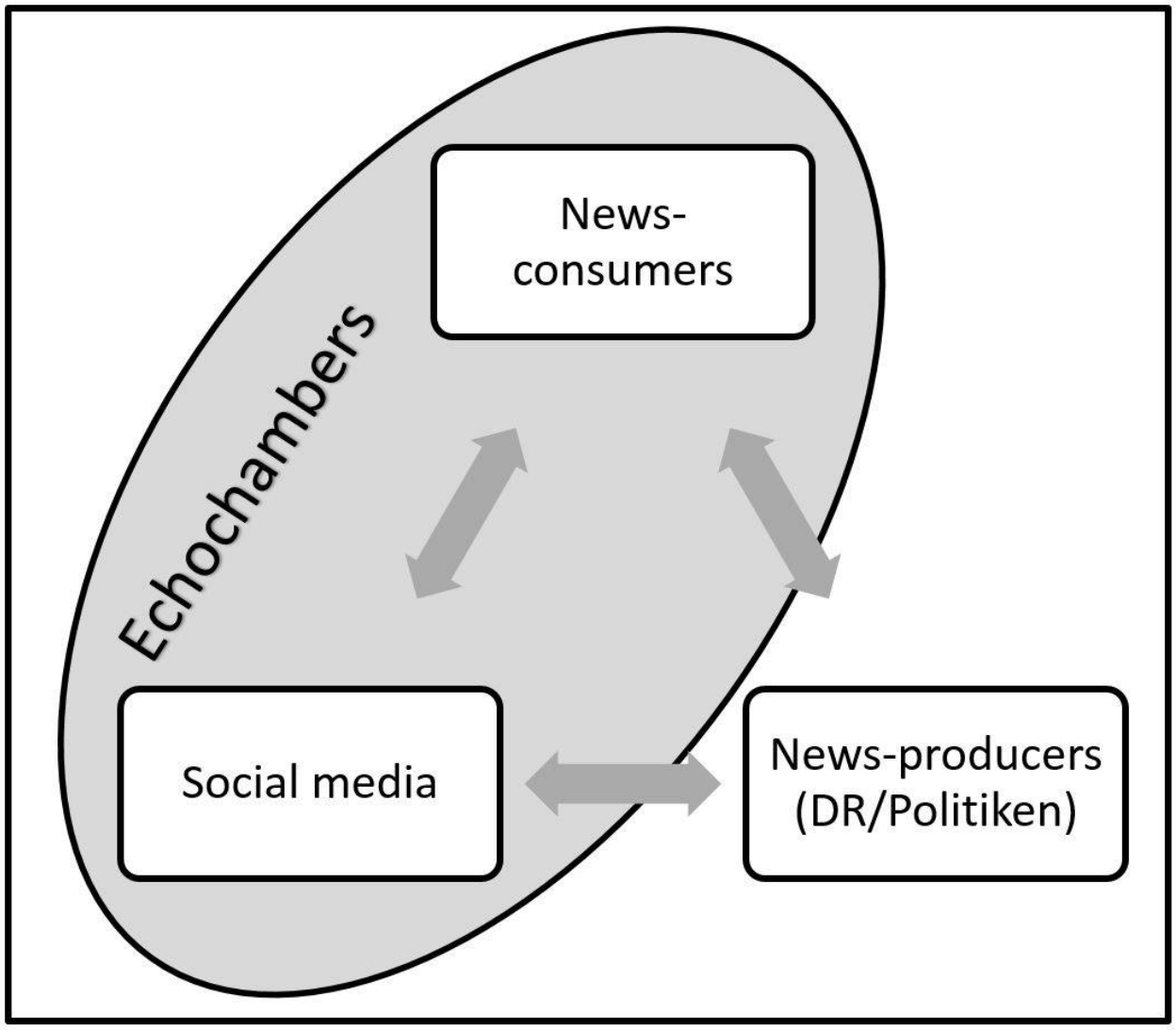
The main purpose of this paper is to consider news articles, so from a subjective view, all sections that do not contain news articles are dropped. Also, sections which has its own Facebook page are dropped since articles from these sections mainly are posted on other Facebook pages than considered here. If we did not drop sections with a separate Facebook page our results would be bias, since these would see, to be left out of Facebook more frequently than in reality. The data has been trimmed from containing 36 sections to only contain 8 sections: Domestic, International, Politics, Money, Economics, Culture, Science,

and The Magazine. The sections money and politics are only present at DR whilst economics and the magazine only are present at Politiken. The remainder sections are included in both media.

From website to Facebook

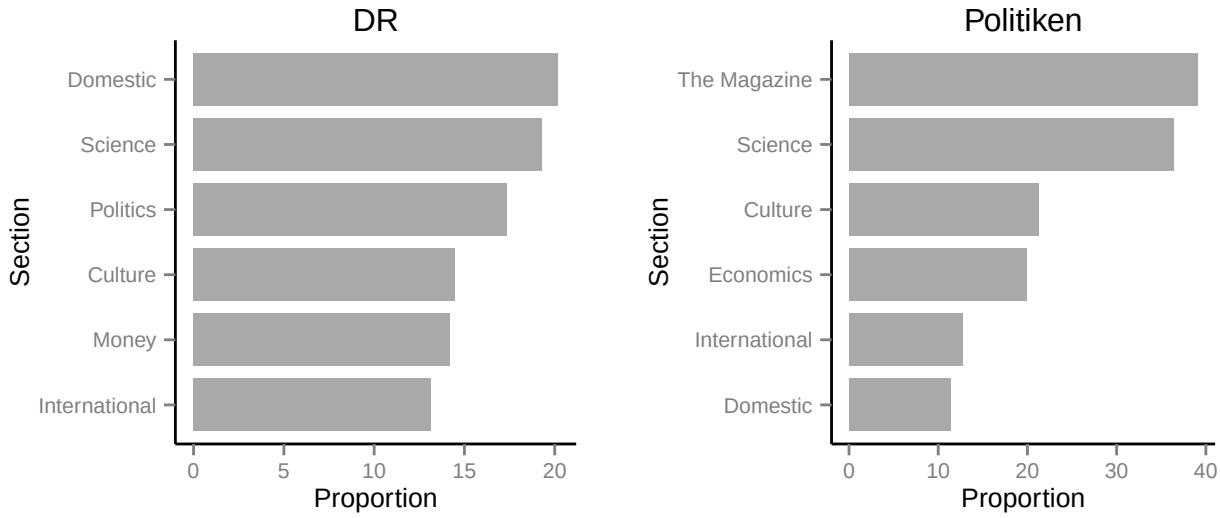
The first thing to notice when considering the data is that a fairly small fraction of the articles getting posted on Facebook. 3,602 (16.4%) of DR's articles and 3,479 (16.3%) of Politiken's articles. As mentioned in the introduction we want to find out whether this selection seems to be random draw or if there is a certain type of selection. More precise, we want to find out whether this shift in news platform creates some kind of echo chambers - a situation in which information, ideas, or beliefs are amplified or reinforced by transmission and repetition inside an "enclosed" system. Our framework of analyzing and discussing this issue, is Figure 1 below. It shows the different interactions between media-consumers, media-producers and finally social media. Our starting point is to investigate the selection process of which articles from the websites that also gets shared on Facebook. We start by looking at descriptive figures and a binary-choice model. Later on we build a model to test the overall hypothesis, that the content of an article is a good predictor for whether an article is shared on Facebook or not.

Figure 1 - Interrelationship between media consumers, media producers and social media



If we look at how these shares are distributed across sections as shown in Figure 2 we find that they differ quite a lot, also across the two media sources. At Politiken the most shared section is The Magazine. It is a part of Politiken's premium content, so one could imagine that the high sharing-rate is because they have a special interest in promoting this particular content. On DR the section with lowest share-rate is international - only 13 % of the articles in this section are getting shared on Facebook. It is about the same rate on Politiken (12.7 %).

Figure 2 - Share of articles from website getting posted on Facebook



As Figure 2 shows there seems to be differences across sections in regards to what gets on Facebook. A simple binary choice model may shed light on whether there is significant correlation between the section and whether the article is shared on Facebook. The dependent (binary) variable in our binary choice model is whether or not the article is posted on Facebook. The explanatory variable is what section the article comes from. The model is constructed using the glm-function. The output from the model is reported in Table 1.

Table 1 - Binary choice model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.13	0.01	-77.87	0.00
section2Domestic	0.10	0.02	5.13	0.00
section2Science	0.36	0.03	10.34	0.00
section2Politics	0.19	0.03	6.70	0.00
section2Culture	0.26	0.02	10.97	0.00
section2Money	0.06	0.04	1.36	0.17
section2Economics	0.28	0.03	9.95	0.00
section2The Magazine	0.85	0.05	16.76	0.00

The reference-section is International news and as we see, a lot of the section-categories have a significantly different effect on the sharing-probability than the International section. With these estimates we cannot say anything about the marginal effect from one section or another, but the sign of the coefficients is directly interpretable. We see that all sections, except Money, have a significant, positive influence on the sharing-probability. There is no surprise here since it is interpretable as the averages. However it does underline our hypothesis that social media, in this context Facebook, has a larger focus than the official website on domestic news compared to foreign news.

Supervised learning: Can we predict whether an article is shared on Facebook or not?

From the very simple binary choice model we now move to a more sophisticated method in our attempt to understand the underlying reasons for an article to be shared or not shared on Facebook. If we succeed in using the content of an article as a predictor for the sharing-probability, we are one step closer to finding out how a bias may look like.

We start out by creating a sub-sample of the total scrape. The amount of data is simply too big for the following algorithms to run on a normal laptop. We make a random draw of 4000 obs., which we will continue to work with below.

With our sub-sample we now build an algorithm using RTextTool. We do this to investigate whether or not there is a clear selection process from the official website to Facebook. The following algorithm-design builds upon the nine steps described in the article RTextTools: A Supervised Learning Package for Text Classification by Jurka et al. (2013). However this process involves a series of choices that we will shortly walk through here. First of all, we need to prepare our data for the analysis. We do that by creating a document-featuring matrix. In that process we also change all words to lowercase letters, remove punctuations and separators, stemming the words and ignoring stop words. This gives us a total of 93,002 features or words.

We choose to reduce the number of features, and we do this for two reasons: first - the amount of memory required performing training and classification of a model containing

nearly 100,000 features and 4,000 unique observations exceed an ordinary laptops capacity. Second, we do not want rare words to deliver the majority of the leverage to the model. We remove all words that show op in less than 80 articles (2% of total), which gives us a total of 2,232 features.

We now split the sub-sample into a training- and a test-dataset. We let the training set consist of 4/5 of the total observations and let R do the random split. After the split, we create a so-called container, which is a matrix that can be used for training, and classifying different model types.

Now we are ready for training our models. We have chosen to use all available algorithms except Bagging, for later comparison of performance and are creating a so-called ensemble agreement for enhancing the labeling accuracy. It is a shame that we have to drop the Bagging-model, since averaging over bootstrap-samples can reduce errors from variance but the process is simply too heavy due to the large amount of data. Since we do not want to limit the size of the sample or the number of features further, we have to drop this model. However, it is our belief that with a total of eight algorithms and the possibility of creating an ensemble agreement, the precision of our classifier will be acceptable.

After training the model, we are ready to classify the models. We do that by using the classify model-statement from RTextTool for each of the eight models. We create the analytics of the models using create_analytics. The summary of the models is showed below:

Table 2 - Ensemble summary

```
## ENSEMBLE SUMMARY
##
##        n-ENSEMBLE COVERAGE n-ENSEMBLE RECALL
## n >= 1                1.00              0.82
## n >= 2                1.00              0.82
## n >= 3                1.00              0.82
## n >= 4                1.00              0.82
## n >= 5                1.00              0.82
## n >= 6                0.96              0.83
## n >= 7                0.75              0.85
##
##
## ALGORITHM PERFORMANCE
##
##        SVM_PRECISION           SVM_RECALL           SVM_FSCORE
##                0.410                0.500                0.450
##       SLDA_PRECISION          SLDA_RECALL          SLDA_FSCORE
##                0.615                0.550                0.555
## LOGITBOOST_PRECISION    LOGITBOOST_RECALL    LOGITBOOST_FSCORE
##                0.410                0.500                0.450
##    FORESTS_PRECISION       FORESTS_RECALL       FORESTS_FSCORE
##                0.410                0.500                0.450
##     GLMNET_PRECISION        GLMNET_RECALL        GLMNET_FSCORE
##                0.620                0.510                0.480
##       TREE_PRECISION          TREE_RECALL          TREE_FSCORE
##                0.410                0.500                0.450
## MAXENTROPY_PRECISION    MAXENTROPY_RECALL    MAXENTROPY_FSCORE
##                0.555                0.560                0.555
```

In the table above Precision refers to how often the particular algorithm predicts correct. So in this context how often an article, that the algorithm predicts to be shared on Facebook, actually was shared on Facebook? On the other hand, Recall refers to the percentage of the articles shared on Facebook the algorithm correctly predicts to be shared on Facebook. The F-score is a weighted average of the above mentioned numbers.

It is clear that no single algorithm can make solid predictions alone. The largest F-scores is for SLDA and maximum entropy. Therefore, in line with the recommendation from Jurka et al., we now create an ensemble agreement to enhance labeling accuracy. The purpose of this exercise is to maximize the accuracy of our predictions. RTextTools include a function for this called create_ensembleSummary, but as we see above the result is also been printed when you use the summary(analytics)-function. To choose which ensemble to use is basically a trade-off between accuracy and coverage, the greater Recall-accuracy the lower coverage. Though, in this case there is 100 % coverage up until the 6th algorithm. For the first five algorithms, the Recall accuracy is 82 %, which is pretty good for such a high coverage.

Both the probit-model and the algorithms indicate that there is a system in the share-rate on Facebook depending on the section and the content of the article. If we found very low or no fit for both the probit-model and the algorithm, there would be no reason to think that there should be any difference in the kind of stories you will find at DR and Politiken's website and the articles that they share on Facebook. But as mentioned, that is not the case.

One of the overall hypothesis in the Filter Bubble is, that our lives online, including social media, is narrower in the sense of confirming views and not making a representative representation of the world. This assignment sets out to investigate this hypothesis in a Danish context. We want to do that by comparing the foreign/domestic ratio of articles shared on Facebook and looking further on the characteristics on the articles, that have the greatest impact in terms of leveraged in our prediction model.

Now we take a closer look at the articles not shared on Facebook and with the highest ensemble probability (above the median). The articles regarding foreign stuff are heavily overrepresented in this sub-sample compared to the total dataset. 38 % of the articles in the sub-sample comes from the international section whereas this share only is 27.9 % in the total

dataset. That is a clear sign, that our ensemble system finds a clear connection between the content of the articles and the sharing-probability.

Are international news being left out

As stated in the introduction, one of the conclusions from a similar study in America were that the media in general has become more focused on domestic news, and that social media even more so. Since we only consider a year, our data does not allow us to consider if there has been an evolution towards larger focus on domestic news. However, figure 2 suggest that a larger fraction of domestic news than international news end up on Facebook. But what about the distribution of international news - is it specific countries being left out of Facebook?

To see whether this is the case in Denmark, we compare the article intensity by countries on Facebook compared to the official website, leaving domestic news out. In other words, we want to see how many articles contain stories from different countries and see how they differ between the official website and Facebook. To do this, we search through all articles to see how often each country is mentioned in the articles on Facebook and the official website. The numbers are illustrated in figure 3.

If we start by looking at the figure displaying the distribution from the official website, the picture is partly as expected. The majority of foreign news comes from western countries (USA and Europe) with USA and Germany as the countries with the majority of the articles (the exact numbers are included in appendix A). Russia also has a fairly large proportion of the news with a total of 1,946 articles. There is however a few surprises in the top ten - Mali with 2,663 articles, Syria with 2,639 articles and Oman with 1,577 articles. This might not be as surprising after all, since the civil war in Syria has gotten a lot of media attention, because it is creating a lot of refugees to Europe. Syria can thus be related to a lot of articles describing refugees in Denmark, making this a closer history than it seems at first sight.

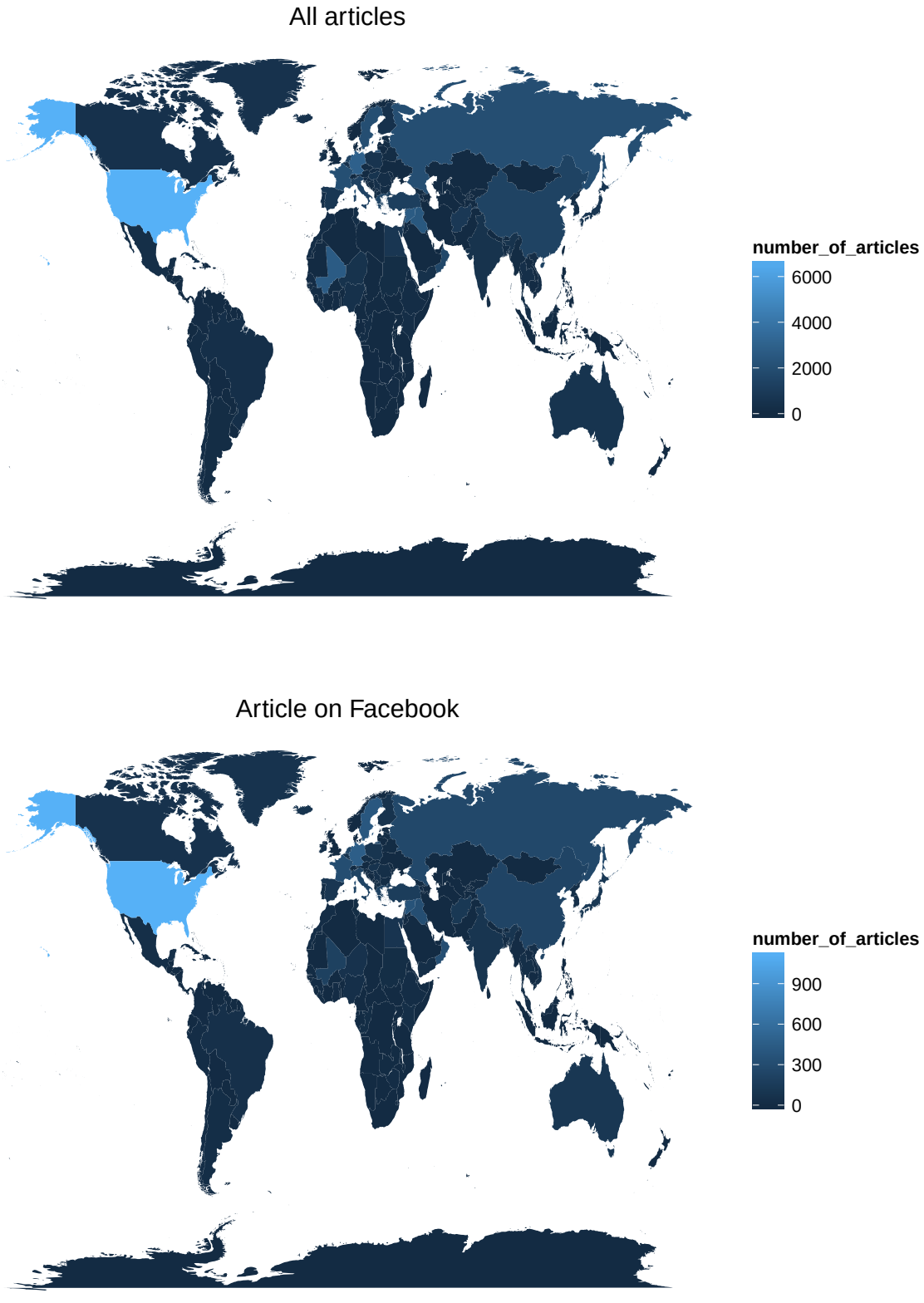
The amount of news concerning Mali is due to a civil war taken place during this period and the possibility of posting Danish soldiers, thus also making this a more relatable story. The large amount of news from Oman cannot be explained that easily. But, even with these cases

explaining news from different places than we might expect, there seems to be a large amount of places with few or no news. This is the case for the majority of Africa, large parts of Asia, South America, Canada and Australia, thus there seem to be some selection bias towards news from specific countries.

However, if we compare it to the figure next to, displaying the same for the articles on Facebook, the difference does not seem to be that big. There is logically fewer news from all countries, however there does not seem to be a smaller fraction left out from the countries close to Denmark in neither distance nor behavior. There are a few changes in the top of the table, like Sweden which is moving from having the seventh to third most appearances when considering Facebook. At the same time Mali moves out of the top ten while Oman and China moves up the latter, thus not creating a one sided picture, towards for example European news.

While it does not seem as if the media we are considering in this paper seem to make a very clear cut selection about what country news they are going to post on Facebook, there might still be a selection internally on Facebook invoking people to be more aware of some stories than others. In the next section we are going to consider what makes news on social media popular, and in this context what type of news people in general are exposed to.

Figure 3 - Article intensity by country



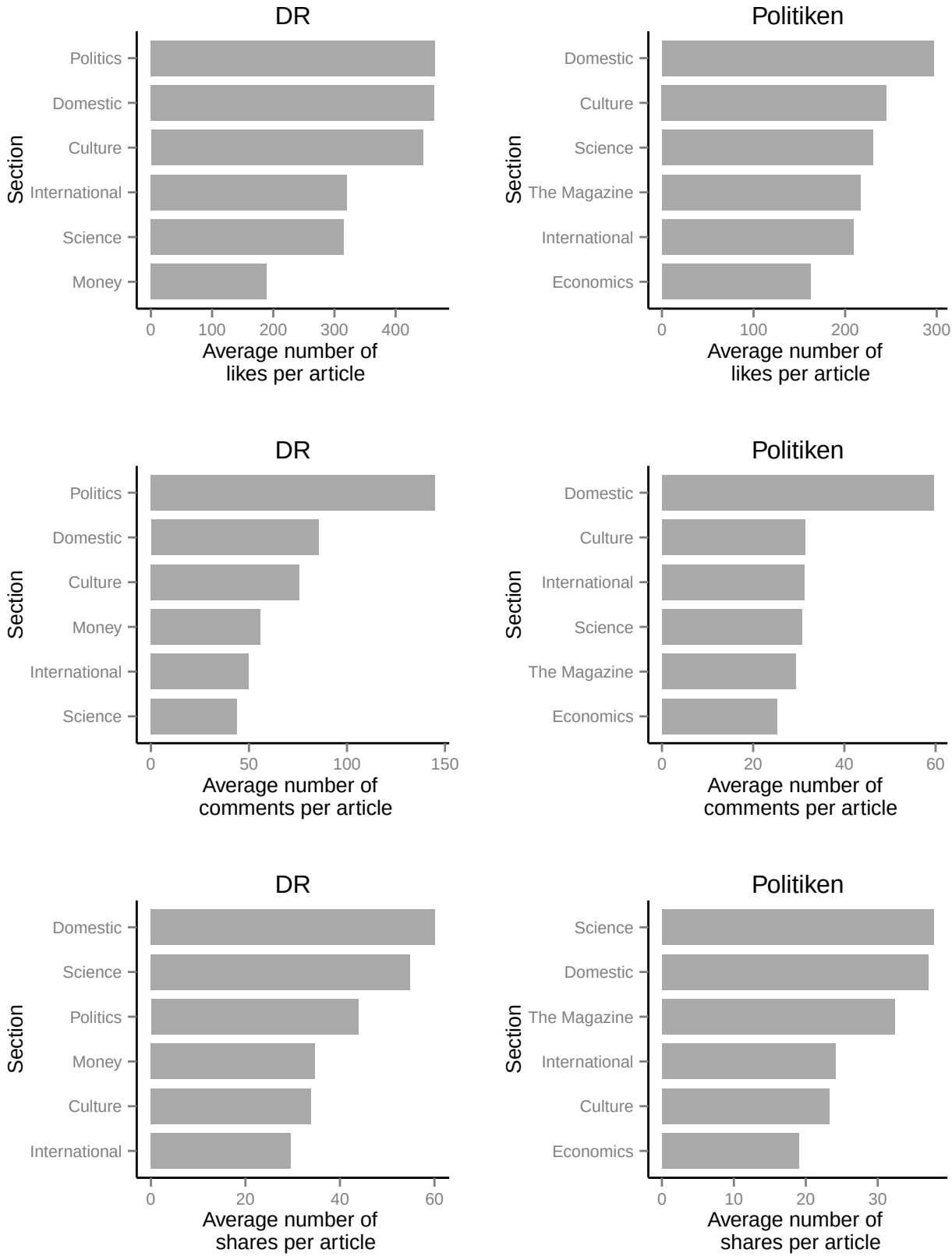
What is viewed

In this section we consider the other part of figure 1 in explaining the potential forming of an echo chamber, namely the interaction between the news consumers and the social media. We do this to see if there seems to be a selection bias here in regards to what type of news people on the social media gets exposed to, once on Facebook.

As mentioned previously we use the number of likes, comments and sharing's as a proxy for how many people getting exposed to the specific news story. We do this since we cannot observe the number of views a specific post has on Facebook, but expect there to be a positive correlation with this and the before mentioned. If we consider the different sections of our media and mainly the average number of like's, comments and shares articles in each section on average receives (figure 4), there seem to be quite large differences.

The results are in line with what the theory suggests (closer news equals interesting news). The average number of likes, comments and shares are in general highest for domestic news, politics (which in this case primarily is domestic politics) and culture for DR. The highest rated on Politiken show the same picture with domestic and culture articles as the most liked, commented and shared, which we interpret as a larger number of views. In all cases international news seem to be viewed less than domestic news, politics and culture. This suggest that we see parts of the same pattern in the Danish media scene as they do in the States, namely that there seems to be a higher focus on news close to people corresponding to news people can relate to. If it is in fact true that the number of likes etc. can be used as a valid proxy for people getting exposed to the news, it also seems to be the case that news on the social media is less focused on the international news and that this is not driven by the media primarily choosing domestic news to put on Facebook.

Figure 4 - Amount of average likes, comments and shares on Facebook

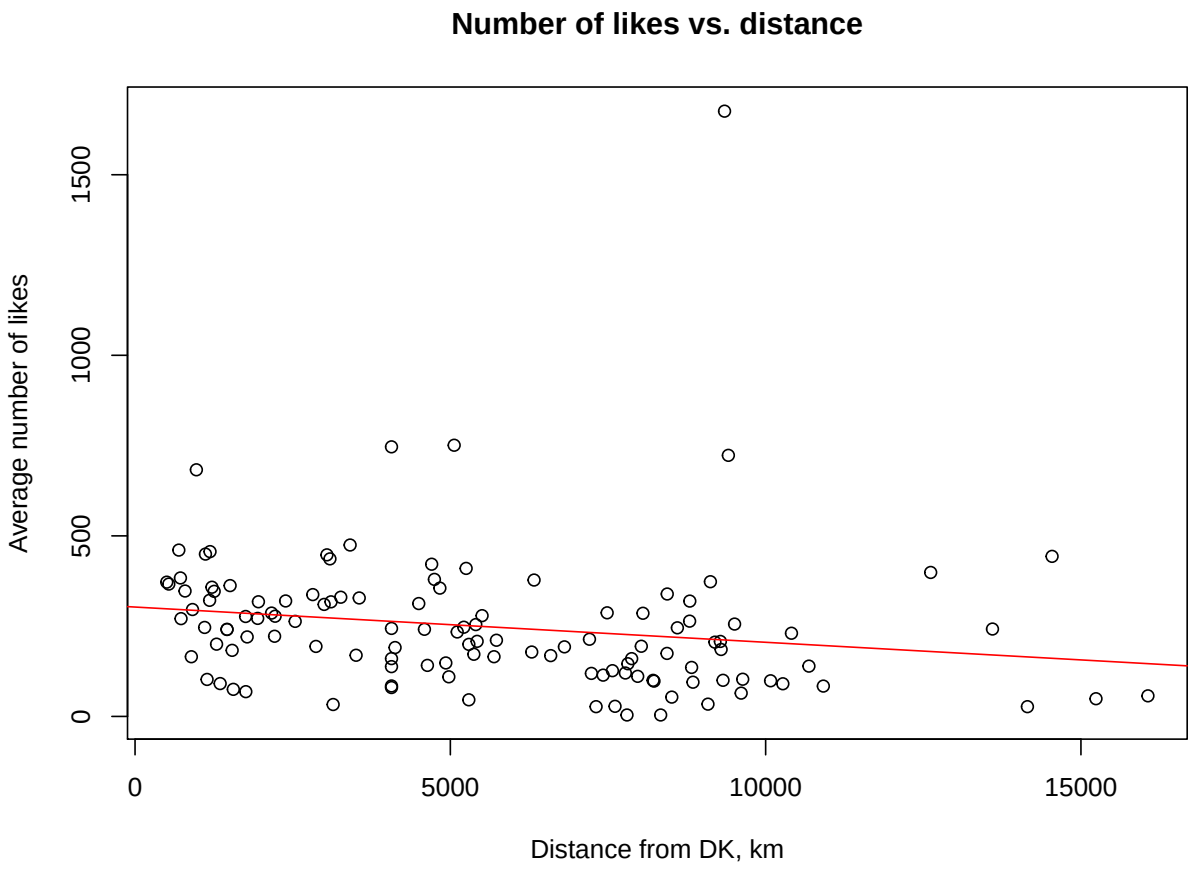


Distance and exposure

As seen in the previous section there seems to be a larger exposure to domestic news on Facebook, judging by the average number of likes, comments and shares. But what if we once again leave out domestic news, is there a correlation between distance from Denmark and the exposure of the news?

Figure 5 shows a scatterplot with the average number of likes on the secondary axis and the distance from Denmark on the first axis, thus every country is represented once, with one distance and the average number of likes. As figure xx shows there does seem to be a negative slope, implying that there is a negative relationship between distance to Denmark and the average exposure. The downward slope is however not that large and we should not put to much in to this. However it is interesting to see that even when just considering the raw numbers, the closeness of a news story in its purest form seems to affect the exposure.

Figure 5 - Number of likes vs. distance



Conclusion

"A squirrel dying in front of your house may be more relevant to your interests right now than people dying in Africa" - Mark Zuckerberg, Facebook founder. The quote is from The Filter Bubble by Eli Pariser, and it sums up the essence of this project: Which news are we exposed to on social media and how do we react on these exposures? A study from the US finds that the media scene in general are more focused on domestic news than previously and that the news on social media to an even greater extent when considering the America (Rewire 2013). The purpose of this project was to investigate whether we see the same patterns in Denmark. The motivation occurred from the fact, that these echo chambers are potentially a threat to our democracy, since the public debate and the general level of knowledge, could be the losers in such a scenario. Therefore, the overall conclusion, that two of the biggest media-producers in Denmark in general are biased regarding which articles they share on Facebook, deserves general attention. Especially the public funded and public-service-obliged DR is biased against sharing articles in the domestic section - these articles are shared nearly 50 % more than articles in the international section.

But also when we consider the actual content of the articles we see a clear tendency of the bias. Our supervised learning experiment, where we considered a subsample of 4,000 articles, showed that with a pretty high accuracy (82 %) we could predict whether an article will be shared on Facebook or not - alone depending on the content.

The question is why this bias occurs? The paper does not give answer to this question, but some of the possible explanations have been revealed. For example, it has been shown that there in general are less feedback (in form of liking, commenting and sharing) on articles in the international section. That could imply that the media producers to a lesser extent will tend to share this type of articles. As mentioned the papers does not offer a causal explanation, but Figure 1 gave a framework in which some of the dynamics can be analyzed and interpreted. For example, it is a highly possible hypothesis that the media chooses what to share on Facebook given the same knowledge we reported in Figure 4 - In other words: the identified bias may not be a deliberately bias, but solely occurring from the purpose of getting as many people to interact as possibly.

If that is the case, we are in a situation where the more knowledge the media gets about what people like - for example via Facebook Insight-service - the more they are contributing to creating an "enclosed" system. The findings in this papers are not groundbreaking, since we do not find causal structures, new trends nor discoveries. Nevertheless, we find evidence that the Danish media have a biased nature on Social media and in the worst case contributes to the creation of echo hambers. This conclusion deserves general attention as it may have great influence on the wellbeing of our democracy.

References

Reuters Institute. (2012). *Reuters Institute Digital News Report 2012: Tracking the Future of News*

Dream. (2015). *Danske unges museums- og mediebrug*

Zuckerman, Ethan. (2013). *Rewire, digital cosmopolitans in the age of connection.* New York: W. W. Norton & Company

Pariser, Eli. (2011). *The Filter Bubble.* 1st edition, VIKING: Penguin Press

Sunstein. (2001). - *Echo Chambers.* Princeton University Press

Jurka, Timothy P; Collingwood, Loren; Boydstun, Amber E.; Grossman, Emiliano; Atteveldt, Wouter van. (2013). *RTextTools: A Supervised Learning Package for Text Classification.* The R Journal Vol. 5/1, June

Appendix

Number of articles from official website

```
##          region Articles
## 1           USA     6874
## 2       Germany     2943
## 3          Mali     2663
## 4         Syria     2639
## 5          Iraq     2471
## 6        France     1998
## 7        Sweden     1953
## 8        Russia     1946
## 9          Oman     1577
## 10       Greece     1472
## 11 Netherlands     1367
## 12        China     1344
## 13        Italy     1331
## 14       Turkey     1153
## 15       Israel      908
## 16 Afghanistan      769
## 17        Spain      692
## 18      Lebanon      662
## 19        Egypt      640
## 20        Japan      631
```

Number of articles from Facebook page

```
##          region Articles
## 1           USA     1119
## 2       Germany      458
## 3        Sweden      401
## 4         Syria      368
## 5          Oman      321
## 6          Iraq      320
## 7        France      314
## 8        Russia      265
## 9   Netherlands      237
## 10        China      234
## 11        Italy      220
## 12       Greece      203
## 13         Mali      202
## 14       Turkey      165
## 15        Japan      127
## 16 Afghanistan      119
## 17   Australia      111
## 18        Spain      105
## 19       Israel       99
## 20        India       94
```