What makes news social media news

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 research 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?

Recent research has shown that the media scene in general are more focused on domestic news than previously and that the news on social media are to an even greater extent when considering the American situation (Rewire 2013). The vast majority if not all social media monitor online behavior and use various algorithms to determine what specific stories to target the individual. 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 on the same time helps people find what they want to find by using this information, it seems to be a self-enforcing effect making people get indoctrinated in their own beliefs.

The fact that beliefs in this sense are amplified within the closed system is referred to in the literature simply as "echo chambers" (Sunstein (2001)). The existence of echo chambers potentially narrows peoples' view of the world. An example of this, taken from "The Filter Bubble": two similar friends search for the same term on Google but get different results and also the number of results vary (Pariser, Eli. The Filter Bubble. 1st edition, VIKING: Penguin Press, 2011, pages 1-3). This happens because of Google's search algorithm, exposing persons to different results, pending on previous search history. Though, not everyone agrees that this is in fact the case. 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's, thus making the 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 implying that selection occurs. But how is this selection made? Is it a random draw or is at a specific type of news qualifying for the Facebook page? In this specific setting, it could seem reasonable to assume news media simply 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. 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 creating 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 if a specific type of articles gets most attention on Facebook and what characterizes these. We use the number of likes, comments and shares as a proxy for how many people getting exposed to the article(XX DEN HER LINJE SKAL TILPASSES TIL, HVAD VI VÆLGER AT FORTÆLLE OM).

We find that there indeed is a bias in what type of articles are being posted on Facebook, both for DR and Politiken. A higher share of domestic than international articles are being posted on Facebook, and likewise domestic articles are more popular than the international ones. This indicates echo chambers also appear 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. In the last section we will 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 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.
The Danish alphabet has some special letters (æ, ø, å) which turned out to have some implications. The letters were not displayed properly in Rstudio and this meant that we could not perform our supervised learning models on the data. This was fixed by substituting in the correct letters.

Scraping the Facebook pages

We Scraped the medias' Facebook pages 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

Now that all the data has been put together in a single data frame, we check if it need to be revised. 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 we scraped. If we did not drop sections with a separate Facebook page our results would be bias. 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 medias.

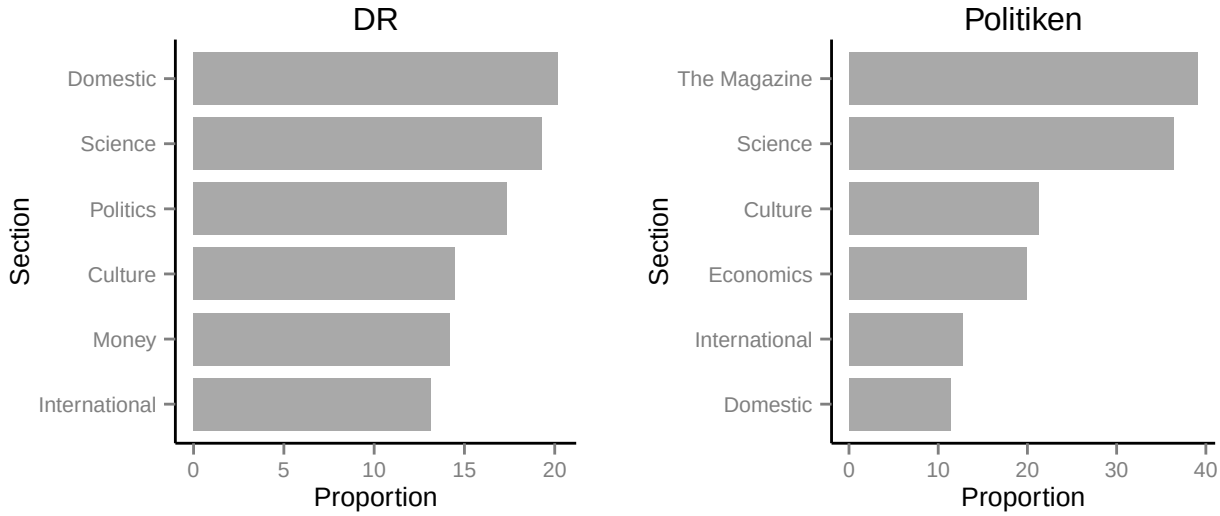
From website to Facebook

The first thing to notice when considering the data is that it a fairly small fraction of the articles getting posted on Facebook. 4,499 (11.9 %) of DR's articles and 5,188 (13.9 %) of Politiken's articles are posted on Facebook. As mentioned in the introduction we want to find out whether this is a random draw or there is a certain type of selection. And more precise we want to find out whether this shift in news platform creates some kind of echochambers - 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 xx below. It shows the different interactions between media-consumers, media-producers and finaly social media. Our starting point is to investigate the selection proces of which articles from the websites that also gets shared on Facebook. We start by looking at descriptive figures and a binary-choice model and later on we build a model to test the overall hypothesis, that the content of an article is a good predicter for whether an article is shard 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 Magasine. It is af part og Politikens 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 is getting shared on Facebook. It's 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 shared on Facebook and 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

	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

Table 1 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 International.

With this 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 inland 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. If we succeed in using the content of an article as a predicter 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 to big for the following algortihms 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 will 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 descripted 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 2232 features.

We now split the subsample into a training- and a test-dataset. We let the training set consist of 4/5 of the total observation 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 to heavy due to the huge amount of data. Since we do not want to limit the size of the sample or the number of features, 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
##
##        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 (Lidt fishy). The F-score is a weighted average of the above mentioned numbers.

It is clear that no single algorithm can make solid predicting that larges 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. In this case though, 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 definitely is a system in the share-rate on Facebook depending on the section and the actual 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 general 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 at the internet, 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). As we see below the articles regarding foreign stuff is heavily overrepresented in this subsample compared to 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.

```
## Source: local data frame [7 x 2]
##
##    section amount
##     (fctr)  (int)
## 1  indland     94
## 2   kultur     20
## 3 oekonomi      7
## 4    penge     13
```

```
## 5  politik     20
## 6   udland     97
## 7    viden      4
```

```
## Source: local data frame [8 x 2]
##
##          section amount
##            (chr)  (int)
## 1        Culture   6156
## 2       Domestic  13541
## 3      Economics   3400
## 4 International  12069
## 5          Money   1590
## 6       Politics   3876
## 7        Science   1929
## 8  The Magazine    683
```

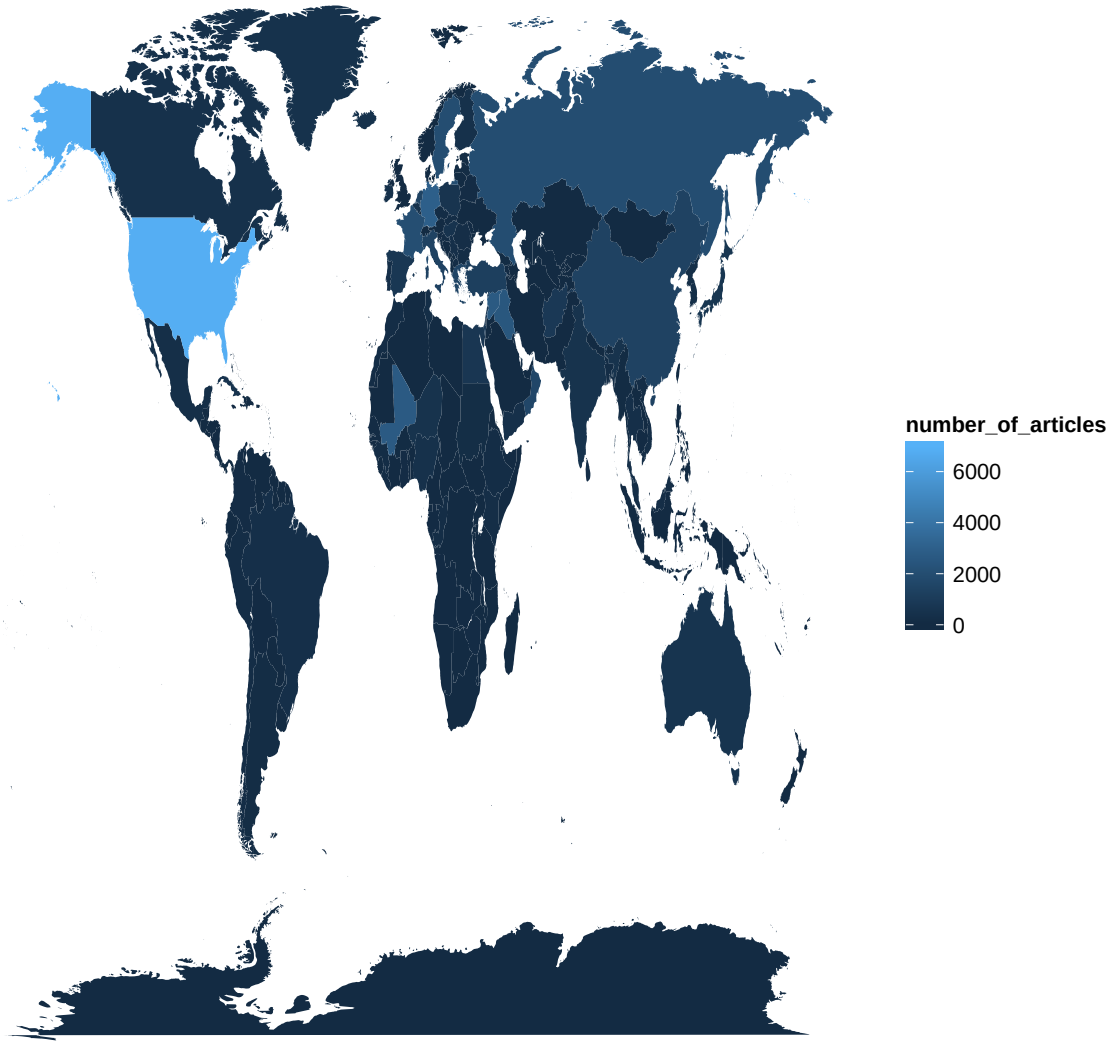
Is it international news getting left out

As stated in the introduction, one of the conclusion from a similar study in USA 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 (andel af sektioner p? facebook) show 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 getting 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 containing stories from different countries worldwide and see how they differ between the official website and Facebook. To do this we search true all articles to see how often each country is mentioned in the articles on

Facebook and on the official website. The numbers are illustrated in figure xx. If we start 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. There is however a few surprises in the top ten - as 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 this is creating a lot of refugees to Europe and 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 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. But 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 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. 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 countries' 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.

Figure x - Article intensity by country

All articles

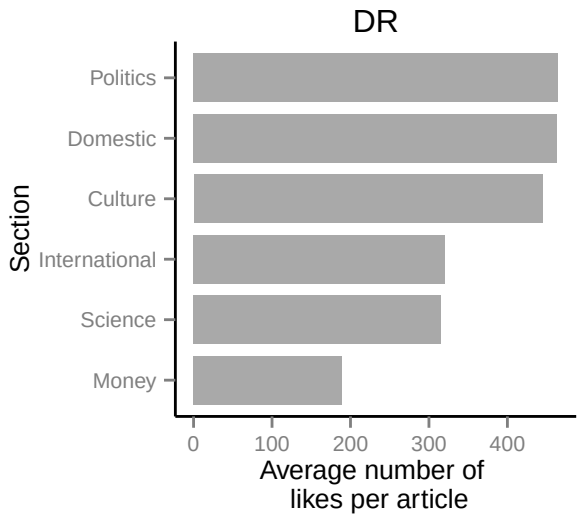


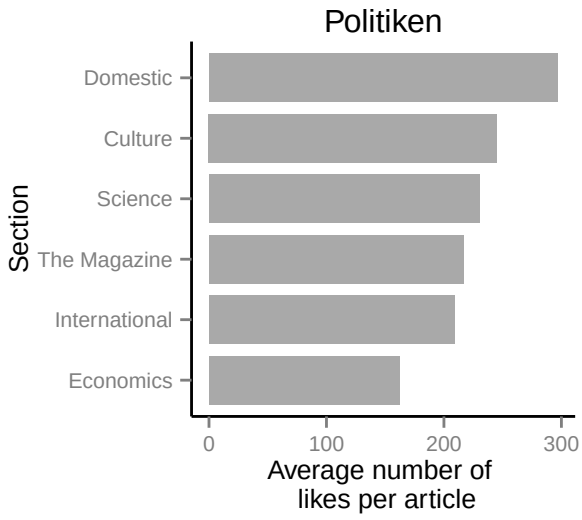
Article on Facebook

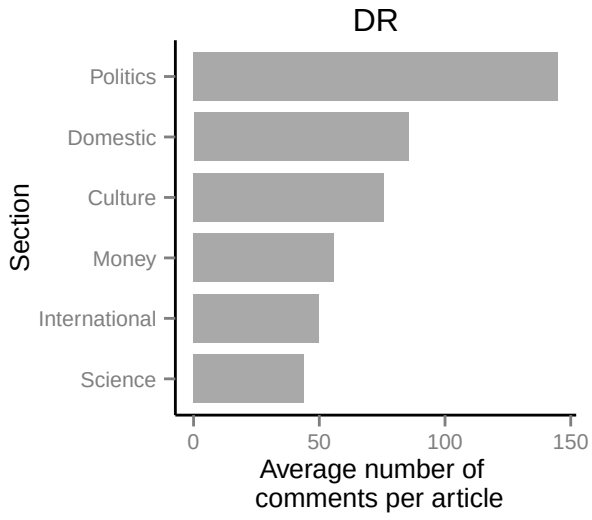
What gets viewed

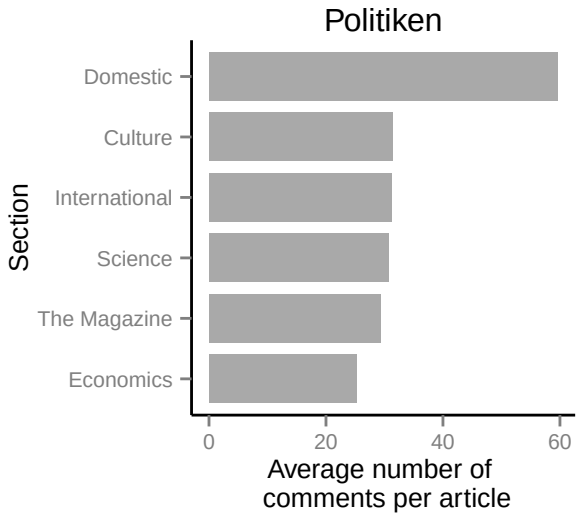
In this section we now consider the other part of figure xx in explaining the potential forming of an echo chamber namely the interaction between the news consumers and the social media to see if there seem to be a selection bias here in regards to what type of news people on the social media gets exposed to, once the articles are posted 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 we 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 share's articles in each section on average receives (figure x.) there seem to be quite large differences. The results are more in line with what the theory of echo chambers suggests. 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 share, which in our context indicates a larger number of views. In all cases international news seem to be liked and commented little compared to 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, news people can relay 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 medias primarily choosing domestic news to put on Facebook.

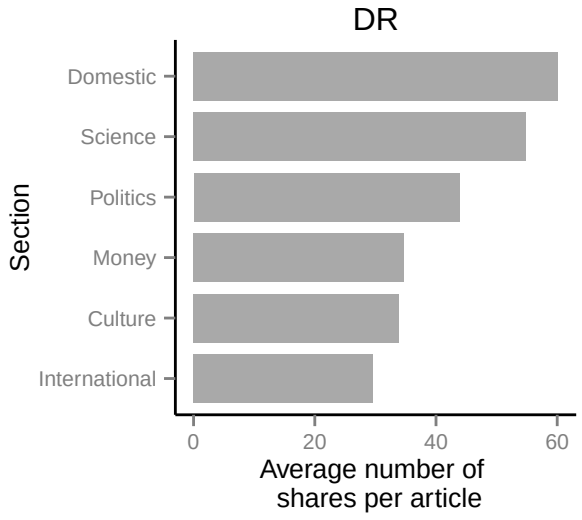
Figure xxx - Amount of average likes, comments and shares on Facebook

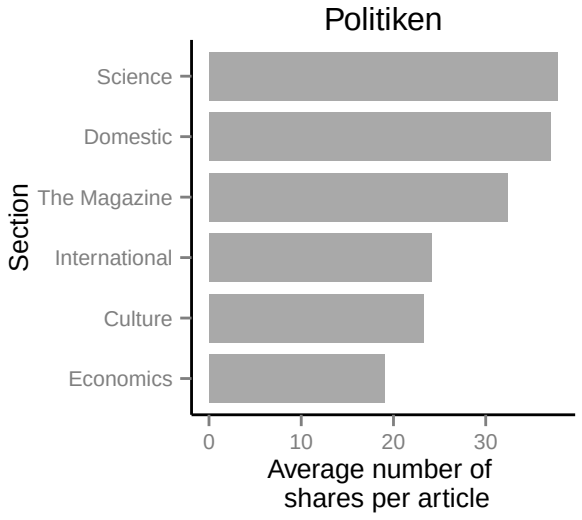

DR
Politics
Domestic
Culture
International
Science
Money
Section
0
100
200
300
400
Average number of
likes per article


Politiken
Domestic
Culture
Science
The Magazine
International
Economics
Section
0
100
200
300
Average number of
likes per article


DR
Politics
Domestic
Culture
Money
International
Science
Section
0
50
100
150
Average number of
comments per article


Politiken
Domestic
Culture
International
Science
The Magazine
Economics
Section
0
20
40
60
Average number of
comments per article

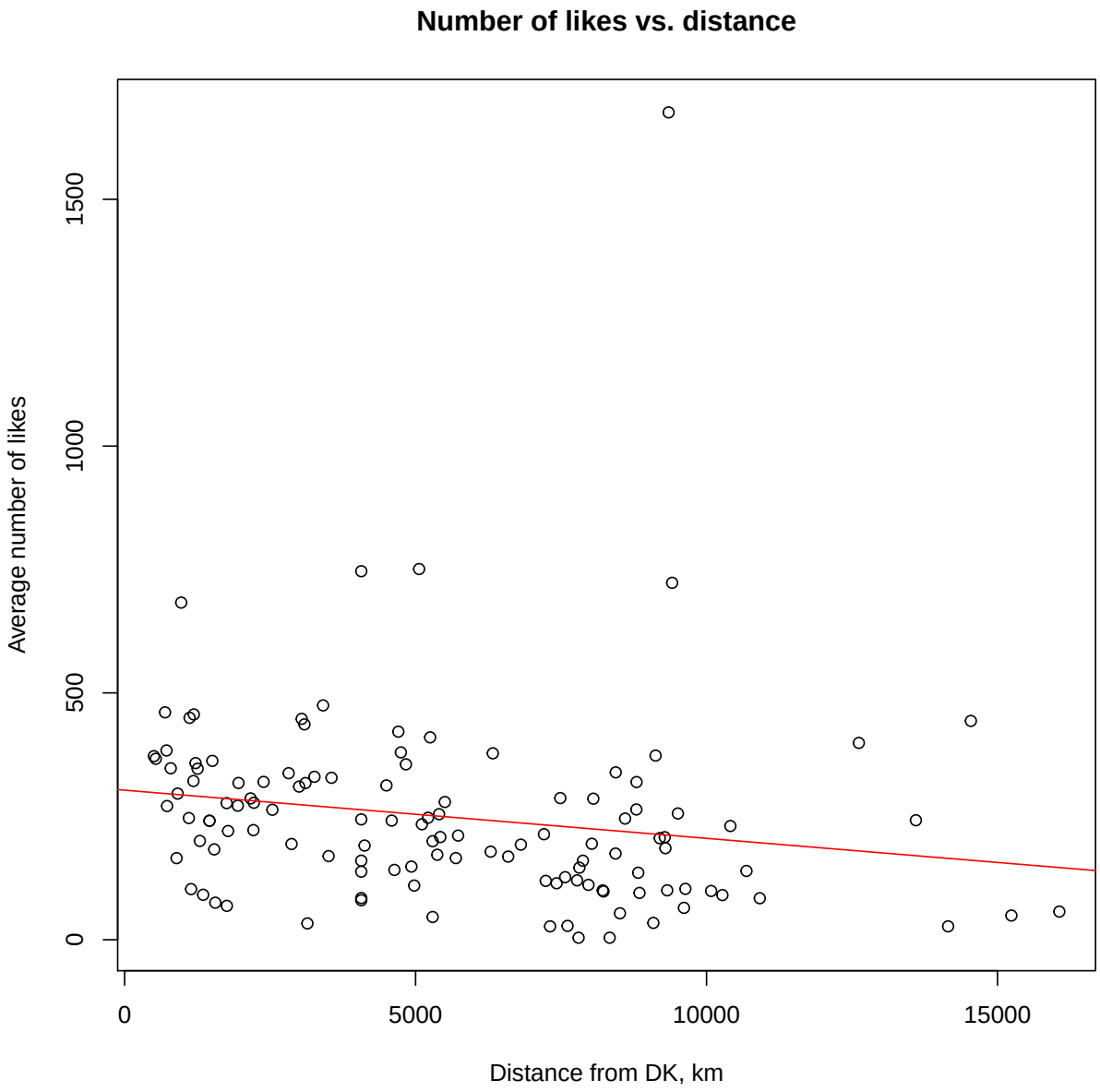

DR
Domestic
Science
Politics
Money
Culture
International
Section
0
20
40
60
Average number of
shares per article


Politiken
Science
Domestic
The Magazine
International
Culture
Economics
Section
0
10
20
30
Average number of
shares per article

Distance and exposure

As seen in the previous section there seem to be a larger exposure to domestic news on Facebook, judging by the average number of likes, comments and sharing's. 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 x show 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 show there do 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.

Figure xx - Number of likes vs. distance



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