1. Introduction

Web news are often criticised for being poor quality, superficial, sensational. Some claim that online news are getting worse every year, but are they really? This paper aims to look at how online news have changed during the last decade. In particular, we are going to investigate changes in the both subject and complexity of online news. The analysis will be done using methods from the field of Social Data Science, such as web-scraping, cleaning, visualization and machine-learning.

2. Data Collection Process

In order to explore how the news have developed within the past decade, we have scraped article data from DR's website. DR's website is excellent for exploring the news development for multiple reasons: (i) DR has its articles publically available for everybody, unlike most newspaper, (ii) they have articles going back a long time (more than ten years) and (iii) their website is fairly easy to scrape, as the structure of the articles pages are almost identical with some exceptions however.

Our data collection process can broadly be split up as follows:

i. Generate strings of the dates we want to scrape.
ii. Scrape an archive page for links to all articles at a given date. We use the strings of dates in the url for the archive page
iii. Scrape articles from these links.
iv. Save a dataset for the date containing data from step (iii) and create a few variables from the scraped variables.
v. Repeat step (ii)-(iv) for all dates from (i).
vi. Collect all the datasets from step (iv) in one big dataset we can use for analysis.

Regarding (i), we have chosen to take wednesday every fourth week for from the 9th of august 2017 and back to the start of 2007. This is close to one date every month and by taking the same weekday every time, we hope to have more homogenous data.

In step (iii), we scrape seven variables: *article title, published date and time, summary, author 1, author 2, sub-titles* and *article text.* In addition to this we create two variables, *category* and *subcategory,* from the link string as this string includes this information. Some articles do not have a subcategory and will instead have a link string for this variables. We fix this in next section, *Cleaning of Data.*

In step (iv), we save one dataset for each day, which allows us to be able to resume the process at a given date, if it is interrupted for some reason. In step (iv) we also create a few new variables from the scraped variables in previous step: *number of periods, commas, exclamation marks, question marks, words, long words* (longer than six character) and *number of words in title*. Finally, we also split *published date and time* into two variables: *published date* and *published time.* We have opted for creating a lot of variables and scraping many variables, even though all of them might not be usable. We much rather be on the safe side and not be limited in our analysis because we were missing variables we could have scraped. In the following section, we will, among other things, access the quality of the dataset we end up with.

3. Cleaning of data

Before analysing the data we inspect it, looking for any mistakes there might be. It is worth noting that the scraper could be modified to take the following data-cleanings into account. This would be a more robust method and if more time was at hand, we would have done this. Before cleaning the data we have 22.685 articles.

Duplicate articles

One thing we quickly realize is that some articles are duplicated. Upon exploring DR's website, we realize that this is not a mistake in our scraper, but rather DR sometimes listing an article twice. A related problem is that multiple links found on DR's news archive pages are dead, i.e. an error-page comes up when trying to load them. DR seems to have converted their old links to new ones and in the process, in some cases, listed both old and new link. In some cases the old link redirects to the new link, in other cases the old link does not work.

We find 3.900 duplicate title of which four articles have two duplicates. This results in 1.954 duplicates having to be deleted. We have prioritized deleting observations with the old link format. We go from 22.685 articles in total to 20.731.

Missing values

Next we inspect all the articles that are missing most values, to see whether our scraper has failed to include them for a reason (could be a different page setup) or if something else is the case. We have 9.030 cases without title, timestamp or main text. That is a huge number and if these are real articles, our scraped should be remade to better include these. We see that of these 9.030 cases, 3.594 has the subdirectory /P4/ and upon trying many of these, we see that these redirect either to a page indicating the page does not exist anymore or to a site of a local radio channel. In a situation with more time on hand, all the links could be looped over to make absolutely sure that this is the case for all of subdirectory. Alternatively, it could also implemented in the scraper. However, in this situation with less time on hand, we will assume that all links with /P4/ as a subdirectory are articles that are gone, why we drop these observations. It is possible that these were changed in the same way as the duplicated articles, but with no redirection.

Furthermore, we see that 5.133 cases have /Regioner/ as a subdirectory. Like with the /P4/ subdirectory, these do not seem to redirect to articles, but to overview papers for the present day. We, therefore, treat these alike the /P4/ subdirectory.

A few other minor cases more can quickly be discarded: Links with sub-subdirectory /Regionale/, for which the the case is the same as the subdirectory /Regioner/; observations with “live” as sub-subdirectory or their link containing either “live” or “minut-minut”, as these are not articles, but live-feeds from previous events; links containing “taet-paa” or “tidslinje” as these are not articles.

We are now left with 259 cases. Many of these links lead to theme pages, which are just overview of other articles, that do show up on the archive-pages. Thus we can safely remove these. Some links lead to an old article page design, specifically articles with the subdirectory “/sundhed/”. We will have to re-scrape these. Other links lead to error-pages, theme-pages, Tour de France-pages and a few other minor cases. We will discard these as well. We note again, that with more time, the original scraper could be made to take these different cases

into account. Creating a scraper is an iterative process, where the scraper is modified gradually as new cases are found. We end up discarding 140 error pages, 30 Tour de France-pages, 8 theme pages and 4 pages from the few other minor cases. We rescrape 77 articles, all with subdirectory /sundhed/. Thus all 259 cases are taken care of.

Additional Cleaning

After having discarded the above described articles a few minor modifications have to be made. We make the category and sub-category variables lower-case; we calculate lix-number for the articles as $LIX = \frac{words}{periods} + \frac{(long\ words * 100)}{words}$, which is a measure of a text's readability, the lower the easier the text; we delete subcategories, that are only present once and replace them with the category (this is necessary because of the way we created the category and subcategory, cf. *Collection of Data*).

The Final Dataset

After all the cleaning is done we end up with 11.778 unique articles with twenty variables: *link, author 1, author 2, category, published date and time, subcategory, sub-titles, summary, article text, title, periods , commas, exclamation marks, question marks, words, long words (*longer than six character), *words in title, published date, published time* and *lix number* for the article. Not all variables are relevant for all articles, e.g. not all articles have a second author. Furthermore, we note again that we have opted for more variables rather than fewer, such that we do not restrict ourselves in the later analysis, because we are missing variables.

4. Descriptive statistics and visualization

In order to see how news have changed over the past decade, we have made time series graphs that plots descriptive statistics from our dataset. The plots are based on both the number of articles on DR, and their lix number. Before plotting anything an essential ethical consideration is deciding what information to disclose and how to disclose it, without the risk of harm to the individual. Our dataset has information regarding individual authors at DR, but revealing personal information for each author is not relevant for our research question. Therefore, due to both ethical and privacy concerns, only content such as graphs and tables that do not disclose any individual information are included in this paper.

Notice that the first plot is based on articles which are divided into subcategories: *Indland, kultur, nyheder, penge, politik, regionale, sports, udland, vejret og viden*. These are the same subcategories as DR uses, except the sports-subcategory. This is a joint category made from all the smaller sports categories.

Our first figure shows the average daily number of articles posted on DR in each subcategory for each year from 2007 to 2017. The second graph shows yearly average of the lix-number for all articles on DR. As shown in figure 1 below, the number of posted articles in each subcategory are quite stable over time. There are yearly fluctuations, but no clear upward or downward trends for any of the subcategories. The sports-subcategory is the only exception, since it's been a steady decline in the number of sports articles on DR from 2012 to 2016. DR posted half as many sports articles in 2016 as in 2012, which is a significant decline. Another thing to notice, is that there are only sports articles in 2007. This may have had an effect on the lix number in this year. Furthermore we see that there are no articles in the subcategory "nyheder" in 2014, which may be because it is used as a residual category for DR. The category "regional" was first introduced in 2013, where there are a lot of articles, after which is has been at a much lower level.

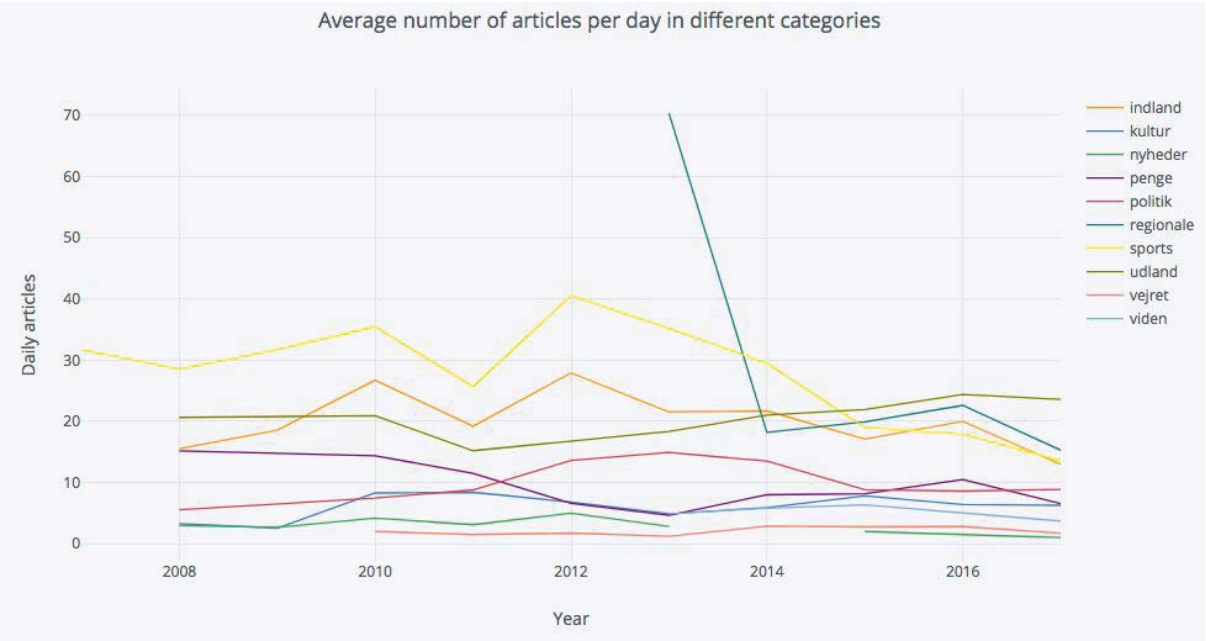


Figure 1: Number of articles in each subcategory

We are using the lix-number as an indicator of whether articles are getting more complex or not. Below is a figure which shows average lix numbers for all articles on a monthly basis from 2008 to 2016. There are big differences from month to month, but it's still clearly a slight downward trend towards lower lix-number for each year. This indicates that articles are getting less complex.

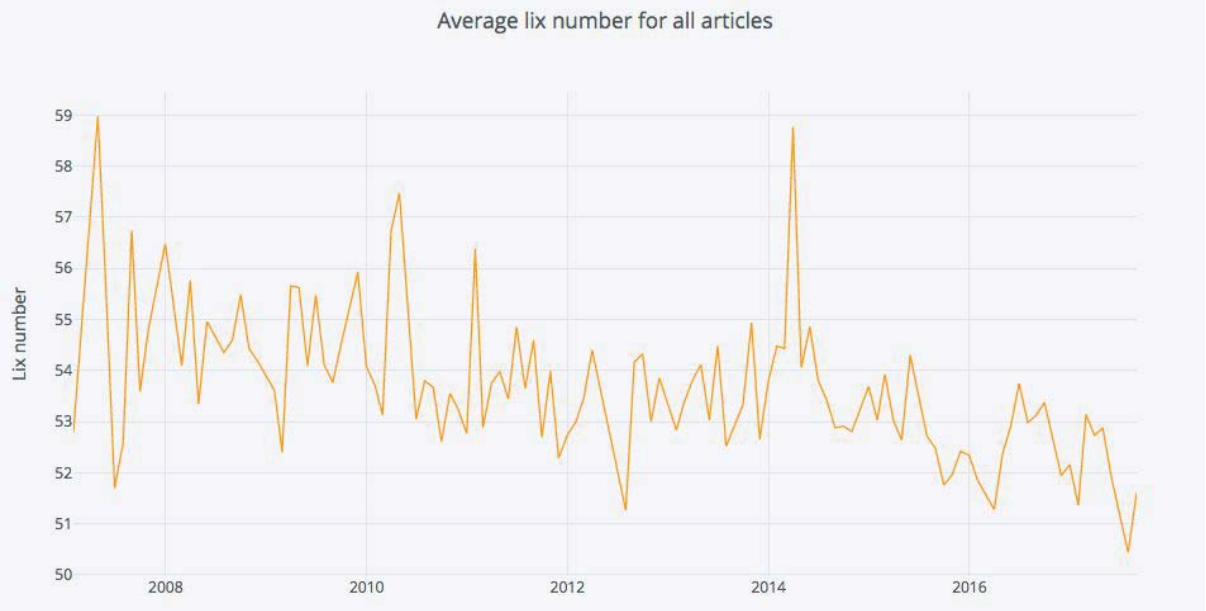


Figure 2: Average lix for all categories

Finally, we have also looked at how the article titles' length has changed over time. This is presented in Figure 3. The figure shows the average number of words in the title over time. We could have also looked at the number of characters.

As can been seen from the figure the average length of the title have gone drastically up since 2012, before which the average number of words in the title were stable at around 5 words. In 2016-2017 this average number of words in the title has doubled to around 10 words on average.

One could speculate that this change might have occurred because people are increasingly reading titles only and not the actual articles, why DR has to put as much information as possible in the titles.

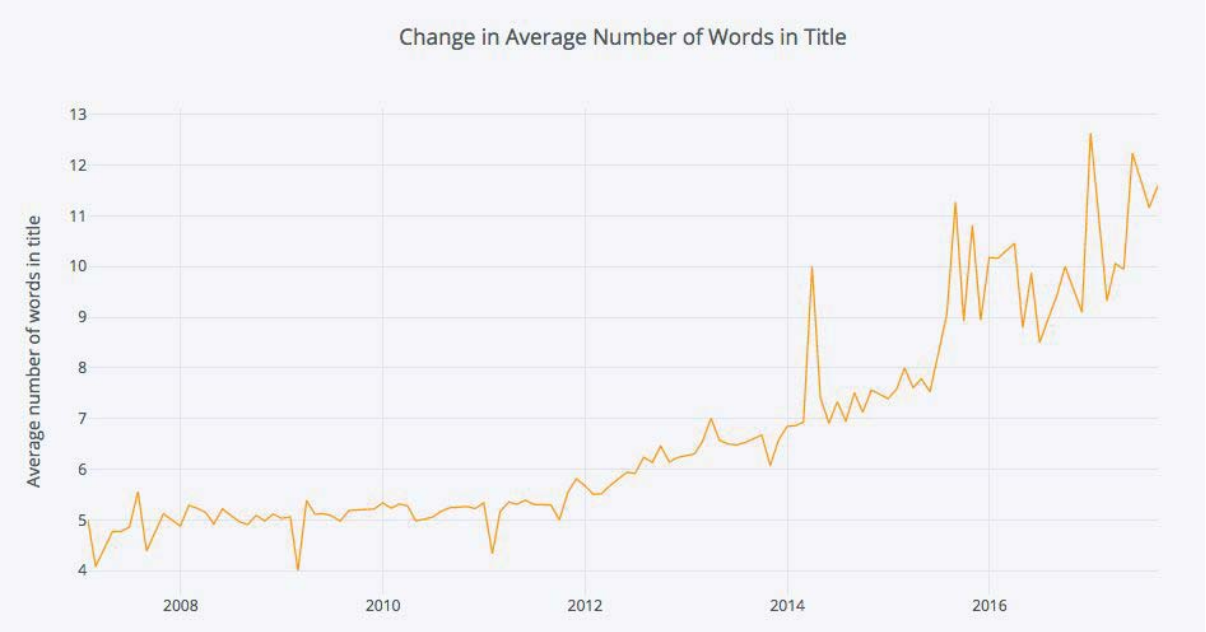


Figure 3: Change in average article title length

5. Predictive Analysis

To see what have affected the lix number in the articles throughout the last decade we thought that it would be interesting to see if it is possible to predict the lix number of a current article. We will use number of commas, category, length of title and date published for this prediction. These variables will be used to fit the best possible model in order to predict the lix number of a current article. We wanted to use sex as a variable as well, however the creation of this variable could not be finished in time (we had made a code, which ran through all allowed names in Denmark and using these to assign sex to author, it took too long time unfortunately). We do not use number of words, number of long words or periods as these are all used in the construction of the *lix* variable. We have omitted subcategory to avoid overfitting as category and subcategory are very similar, but subcategory takes too many values.

We will include the following models in our analysis: OLS, Lasso and Random forest.The accuracy measure we will be using is the RSME (Root Mean Square Error). RSME is used as a measure of the difference between predicted population values and the real population values. By using the RMSE we can compare different models' ability to predict.

RMSE is calculated as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{n}}$$

I.e. RMSE is the square root of the average of squared errors.

By choosing the model with lowest RMSE as our final model we get the model with the best ability to predict.

As we work with a single dataset, it is best avoid models with high bias. However, there is a tradeoff between bias and variance. Having a little bias often allows for a better predictive model than otherwise. However having too much bias can result in missing relevant patterns in the data. We can remove some bias by random assignment. In practice it is done by splitting the data frame into a training and a test set and performing cross validation (CV). The split is chosen manually, but the observation and the training and test set are both assigned randomly. For all of our models we have chosen to split 60/40. This means that 40% of the observations are assigned to the test set and 60% to the training set. We then run the model on the training set only and perform prediction on the test set only. By performing cross validation we average the result of multiple runs with different test and train sets. Doing this we have less bias in our models. In our cross validation we average the RMSE from 10 folds (runs).

To make *publish_date* a useable variable, we have converted the oldest date to 0, and added 1 for each month passed. This means that we can use it as 'time passed', which is a quantitative variable.

OLS

OLS is a short term for ordinary linear regression. The OLS is the simplest approach, which means that it weight insignificant variable, as is has no way to perform variable selection. Furthermore, correlation between predictors, variance inflation are common problems of the model. For a 10 fold CV average OSL gives a RMSE of 7.50, which is considerably higher than allowed for a good model. The results from OLS will be discussed below.

LASSO

The LASSO model uses OLS as a base model. In addition to the OLS it performs variable selection and regularization to enhance to prediction. This means the LASSO extension of the OLS will remove to most insignificant variable in order to improve the overall accuracy.
The LASSO Model performs variable selection by minimizing the loss function:

$$Minimize \sum_{i=1}^{n}(Y_i - \sum_{j=1}^{n} X_{ij} * B_j)^2 + \lambda \sum_{j=1}^{n} |b_j|$$

λ is the tuning parameter (shrinkage parameter) , which means that λ controls the size of the coefficients and the amount of regularization if λ goes towards 0 we have to least squares solution. If lambda goes towards ∞, logically betas will shrink to zero and we are left with the intercept only model. Therefore as we minimize the above function some of the betas will possible be set to zero depending on the value of lambda[1]. For a 10 fold CV average LASSO gives a RMSE of 7.45

If we compare the coefficients of the OLS and LASSO, we see that indeed some of the betas have been set to zero, when using LASSO. Below is a table with the name of the coefficients on top, and name of model on the left side. In our model, we use category-sundhed as the reference category:

	intercept	commas	titl_words	publish_date	category-levnu	category-nyheder	category-sporten
OLS	55.305930	-0.065358	-0.207199	-0.004887	-3.062427	1.441196	-0.123313
LASSO	55.804003	-0.060800	-0.167653	-0.004470	-0.000000	0.000000	-0.000000

Here LASSO has removed 'category', i.e. β_4 to β_6, as dummies for category. We can see that title words generally have a negative effect on the lix number of the article. Furthermore there is a minor trend in publish_date, as the newer the article is the lower lix number. Lastly we can see that 'levnu' articles have a lower lix number than the other categories.

[1] http://statweb.stanford.edu/~tibs/sta305files/Rudyregularization.pdf

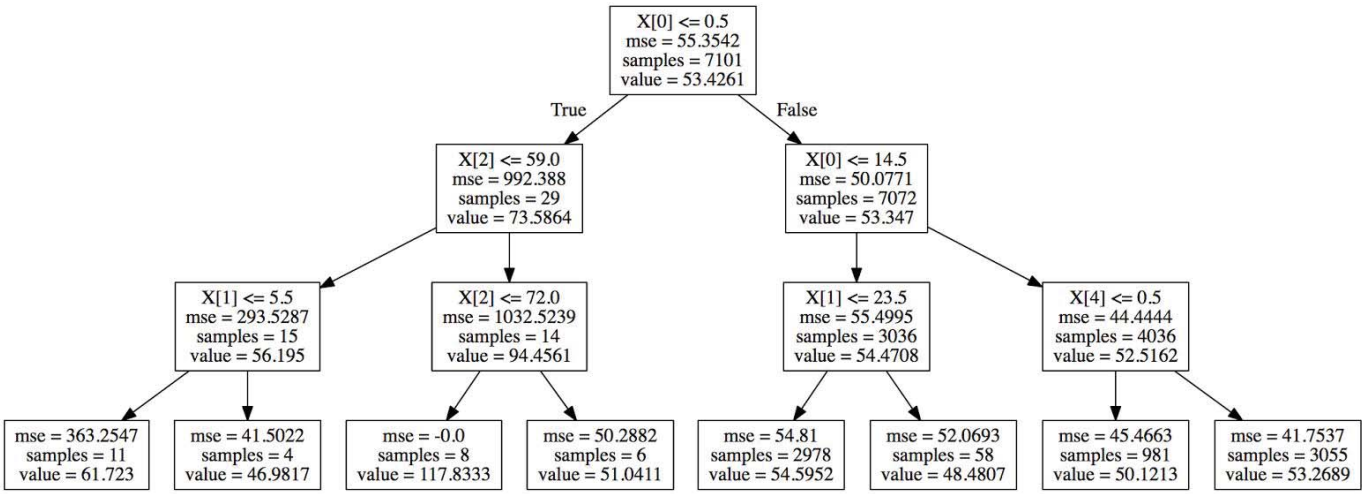
Random Forest (Regression Tree)

One of the more common machine learning algorithm is random forest. Random forest is based on the decision/regression tree. The regression tree is based on recursive partitioning, which means that you partition the space of the sample in smaller groups again and again until it's possible to describe the data in the sub-partitions with a simple model.[2] In the regression tree the trees represents recursive partitioning, and a node represents the cell of the partition. We move through these nodes by comparing our data to the statement in each note. When we reach a root node we have to value which to assign the our data.

Random forest algorithm works as a large collection of uncorrelated regression trees. It is used improve the regression tree model as random forest averages the predictions of multiple regressions trees trained on different patch of the same training set. This is called bagging, which is used to reduce the the overall variance. In our model we have chosen to set the max number nodes to 3, even though a larger number of nodes increases prediction accuracy, however having too many nodes can result in overfitting[3].

An example of a regression tree from one of our forests is shown below. The split values are values conditions with is either True or False, depending on the value of a variable. The split values are shown in the top of each box. In each box MSE, sample and value is shown too. the value is the current value which to assign the observation to, where sample is the number of observation which belongs to the node:

[2] http://www.stat.cmu.edu/~cshalizi/350-2006/lecture-10.pdf
[3] http://www.math.usu.edu/adele/RandomForests/UofU2013.pdf



(Tree 25 visualized by graphviz, x0 = commas, x_1 = titl_words, x_2 = publish_date, x_3 = category-levnu, x_4 = category-nyheder, x_5 = category-sporten, x_6 = category-sundhed)

We see that x_4 (subcategory) is not used in this particular tree, and that a few nodes have very high split values ranging between 94 and 118, which give the impression that the model in certain cases overestimates the lix number a lot. For a 10 fold CV average RF gives a RMSE of 7.48.

RMSE Results

We have compiled the RMSE of each model to compare them:

```
_____________________________________________
CrossValidation - Root Mean Squared Error
_____________________________________________
rmseLASSO:
[7.4967855041067626]

rmseOLS:
[7.4534839406935003]

rmseRF:
[7.4754319192420002]
```

All the models perform very similar. The RMSE is quite high as a lix difference of 7 almost could be the difference between a tough scientific publication (say a lix number of 51) and a magazine (say a lix number of 44). To sum up, it is not very easy to predict the lix number with our current models and data.

6. Discussion

There are multiple limitations in our analysis of the developement of online news. First of all, DR is not representative for all online news. One of the biggest reasons for this is that DR is public service. Thus DR does not have to compete for readers the same way other online news providers do. This might have a significant effect on the content DR creates. Therefore, we cannot not draw any general conclusions about the online news, but are limited to talk about how DR's news have developed, which may or may not be the same way for other news providers.

Another limitation is that lix is just one of many measurements of a text's complexity. Other measurements of a text's complexity may be better and could give other conclusions. A somewhat similar limitation regards the categories/sub-categories, which are also just DR's own division of articles, for which there might be a much more precise division. What DR might consider politics, could just be gossip. Thus, changes or lack thereof, cannot not be definitive proof that DR has or has not made their articles more sensational.

In the cleaning of the data we deleted a lot of articles. We had good reason to do this, but in the process some articles, which should not have been deleted, might have been deleted. If this is the case it could have skewed our estimation.

On a final note, the way online news are today is way different from how they were in 2007. Today we have way more video-articles, live feed commentary, interactive arcticles and more. Because of this structural change in the online news it is very difficult to directly compare how news articles were in 2007 and how they are today. The change may just have been caused by the structural change. If one were to continue work on this project, it could be interesting to work more with this structural change and see how the number non-classic articles types (i.e. video-articles, live feed commentary and so on) has increased over time.

7. Conclusion

So, how did news change during the last decade? There have been a change in the number of articles in each subcategory, but mostly small yearly fluctuations. The exception is *sports,* where the number of written articles have fallen. If news articles where getting more sensational, it should have been less articles regarded as serious. Like articles about *penge* and *politik*. However, this is not the case. This could indicate that news articles are isn't getting more sensational, but rather that much of the news content have not changed much the last 10 years.

The lix number going down indicates that articles are becoming less complex with time. This result does not have any clear indications for the quality of articles: The articles may deal with less complex subjects, they may be written in a simpler/better language or something else could have affected this change.

Looking at number of words in the article's titles, we see a drastic increase from 2012 until today. The titles have doubled. This could be because DR adapts to readers mostly reading titles and not articles, thus needing to put as much information in titles as possible.

In the predictive analysis we found that there is a very minor trend in lix number as it seems that older article have lix number a bit higher can newer ones. This is in line with what we saw in descriptive analysis.

Overall, we cannot say much about the general development in the news, because DR is non-representative for the general news. We can, however, say that there seems to be a tendency for the complexity of the news articles on DR to have gone down. We cannot say what is the cause of this with the data at hand. Furthermore, the number of articles on different subjects on DR seems to be almost unchanged except sports, in which the number of articles has gone down.