The opinion Dutch people have on Ebola on Twitter versus what message about Ebola is spread in Dutch newspapers: a comparison
## Contents

1. Introduction .......................... 1
2. Related work .......................... 2
   2.1. Research on word collocations ...... 2
   2.2. Research on corpus comparison ...... 4
   2.3. Social media and other online resources for disease detection ...... 4
   2.4. Research on the statistics of word collocations ...... 5
   2.5. Wordsmith ......................... 5
3. Data .................................. 8
   3.1. Twitter database .................... 8
   3.2. Retrieving our Twitter data ........ 8
   3.3. Newspaper database ............... 11
   3.4. Retrieving our newspaper data ..... 11
4. Method ................................ 12
   4.1. Lexical richness .................... 12
   4.2. Top 100 of word collocations .... 13
   4.3. Trends of clusters per media ...... 14
5. Results ................................ 16
   5.1. Lexical richness .................... 16
   5.2. Comparing the top 100 word collocations per media ........ 16
   5.3. Trends of clusters per media ...... 21
   5.4. Comparing each cluster per media .... 22
   5.5. Pearson product-moment correlation coefficient .... 32
6. Conclusion ............................ 33
1. Introduction

Around February 2014, the most widespread outbreak of Ebola started in Guinea. The World Health Organisation (WHO) declared the Ebola outbreak an international public health emergency on 8 August 2014. Since the beginning of the outbreak, approximately 8,600 people have died of the virus disease. The course of the disease attracted a lot of media attention. WHO released an interview with Melissa Leach, the director of STEPS (Social, Technological and Environmental Pathways to Sustainability) Centre, who stated that the media has given Ebola a disproportional amount of attention. She furthermore stated that “Ebola fever is now seen as a deadly local disease requiring a universal kind of “rapid response,””. Other research has shown that news media have a great (negative) influence on peoples opinion, like the view people have on Islam and Muslims. An interesting topic is whether news media alter the opinion people have on certain subjects, and one can ask himself whether the role of the media is to spread subjective messages about those subjects, opposite to spreading objective news.

The goal of this research is to show how Dutch people on Twitter think about Ebola versus the message that is spread about Ebola in Dutch written media. Research shows that Twitter is more of a news media than a social network, with over 85% of the Tweets being headlines or persistent news. But what if we peel down the layers of retweets and headlines, and see what Twitter is in its core. What do the individuals itself write about Ebola on Twitter. Do they follow the news, or form a whole different opinion on their own?

In order to see how people form their opinion about Ebola, and whether they get their information about Ebola from news sites, we would need to proof that there is a relation between what people say about Ebola on Twitter and the news that is spread about Ebola in newspapers. If there is a strong connection, it is more likely that people form their opinion based on the information they acquired from the newspaper. Or the complete opposite, we could find that people on Twitter doubt the news that is spread in the media, by using words like ‘illuminati’. The Illuminati theory is a popular modern conspiracy theory. Conspiracy theorists propose that there exists a secret society called the Illuminati, that is trying to control the world. This theory became so popular that the term ‘Illuminati’ now jokingly stands for a conspiracy theory. So if the term Ebola is frequently mentioned with Illuminati, it indicates that these Twitterers jokingly dismiss Ebola as some kind of conspiracy theory. In order to research this, we examined a range of topics related to Ebola, such as Ebola contamination and fear of Ebola and focused on the course of these topics over time, as well as the comparison between Twitter and written media. Using varying methods including word collocations and frequency counts, the opinions on Twitter and in the news articles could be visualised. There is so
much data that it is impossible to read it all manually, so automatic methods are needed. Word collocations and frequency counts are ways to provide insight in trends that you would have missed when reading all the information by yourself. These methods are also objective, verifiable and unbiased, and therefore offer a more scientific approach to the research question.

The outcomes will show whether the Ebola outbreak is perceived by the population in the same way as how the outbreak has been covered in the written media, and whether the media play a role in forming views and opinions about Ebola. The outcomes can be used when studying discrepancies between Twitter and the media and for raising awareness about the use of the media when portraying a disease outbreak like Ebola.

The following section describes the existing research using word collocations and the use of online media for disease related topics. In addition, the section describes different statistical methods for the calculation of word collocations. We then describe the way in which we collected our data, and which methods are used for the analysis of this data. This is followed by a result section, in which we show the outcomes of our variety of analyses. The thesis ends with a conclusion, in which we reflect on the results we have seen and look forward to possible future research on this topic.

2. Related work

2.1. Research on word collocations

The main method that we will use in the research will be collocations. Collocations are words that occur more frequently together than one might expect. These words are interesting for our research, because they contain relevant information about how people discuss the subjects. For instance, if you do research on two main politicians in the Netherlands, let’s say Geert Wilder and Mark Rutte, a word collocation will give the following results (please note: this is purely hypothetical):

<table>
<thead>
<tr>
<th>Geert Wilders</th>
<th>Mark Rutte</th>
</tr>
</thead>
<tbody>
<tr>
<td>Muslims</td>
<td>Softy</td>
</tr>
<tr>
<td>Moroccans</td>
<td>Profiteer</td>
</tr>
<tr>
<td>Stupid</td>
<td>Refugees</td>
</tr>
<tr>
<td>Hatred</td>
<td>#bedbathbread</td>
</tr>
</tbody>
</table>
This is a clear indication for how these politicians are viewed.

One of the researches that also uses collocation in order to investigate the imaging of a certain topic in the media, is the research on the portrayal of Islam in the British press.[1] The researchers divided the collocation into several categories including: ethnic/national identity, characterizing/differentiating attributes, conflict, culture, religion, and group/organizations. Each collocated word was categorized as a positive (e.g. ‘social’) or a negative word (e.g. ‘intolerant’). The two most frequent collocate pairs that the researchers found were ‘Muslim world’ and ‘Muslim community’. According to the researchers, this showed that British press is used to collectivize Muslims. Other frequent collocate pairs showed that Muslims were also represented as easily offended, alienated, and in conflict with non-Muslims.

Another research using collocations is a paper on analysis on the representation of feminism in the media.[2] This researchers used English and German papers for finding collocations on the word feminism. The collocated words are categorized into several categories which is shown in table 2.

**Table 2: The most frequent collocates for the term ‘feminism’ divided into six categories.**

<table>
<thead>
<tr>
<th>Category</th>
<th>Collocates</th>
</tr>
</thead>
<tbody>
<tr>
<td>political movements</td>
<td>radical, political, anti, socialist, politics, movement, Marxism, liberal, minority, equality, liberation, black</td>
</tr>
<tr>
<td>sex, gender</td>
<td>lesbian, gay, lesbianism, gender</td>
</tr>
<tr>
<td>time</td>
<td>1970s, post, contemporary, wave, second, generation</td>
</tr>
<tr>
<td>place</td>
<td>Western</td>
</tr>
<tr>
<td>evaluative</td>
<td>new, modern</td>
</tr>
<tr>
<td>academic</td>
<td>theory</td>
</tr>
</tbody>
</table>

Furthermore, the complements are labelled as positive or negative as shown in table 3.

**Table 3: complements of 'feminism' labeled as positive or negative.**

<table>
<thead>
<tr>
<th>Complements</th>
<th>Negative meanings</th>
<th>Positive meanings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feminism is</td>
<td>dead (24)</td>
<td>news (6)</td>
</tr>
<tr>
<td></td>
<td>no + longer (5)</td>
<td>action (5)</td>
</tr>
<tr>
<td></td>
<td>bad (5)</td>
<td>alive (4)</td>
</tr>
<tr>
<td></td>
<td>seen + as negative (4)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>boring (4)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>finished (3)</td>
<td></td>
</tr>
</tbody>
</table>

The research hereby provides evidence for the subjective portrayal of feminism in the media in different cultural contexts.
These two researches only take the portrayal of their topic in the media into account. The only thing the research shows is the image of their subject in the media. They do not look at the influence the media have on the view people have on this topic. Although it is likely that the media has influence on the view that people have on these subjects, it is not proven that the media actually has any influence at all with their portrayal of Islam and feminism. This is where our research goes a step further, by also investigating the influence that the media has on the view of Ebola on Twitter.

2.2. **Research on corpus comparison**
Word collocations are one of the methods that calculate how strongly a word is related to Ebola. Another method for calculating the relation of a keyword to Ebola is keyword analysis. This method uses relative frequency as a statistical approach on which keywords are specifically related to the topic (in this case Ebola). Jurafsky et al. used this method for a corpus comparison, in order to investigate which words are more used in positive restaurant reviews, and which words a more used in negative restaurant reviews [16]. He found for instance that negative reviews are more likely to contain words related to trauma, and an increased use of the words ‘we’ and ‘us’ in negative reviews. Whereas positive reviewers use words related to ‘addiction’ more when it comes to less expensive food, the reviewers use long words and words focusing on sensory pleasure when it comes to expensive restaurants.

In our case, we used keyword analysis as a corpus comparison method in order to see which words are typically related to Ebola in the news, and on Twitter.

2.3. **Social media and other online resources for disease detection**
Our research focuses on the influence of the media and the general view that people have of Ebola on Twitter. Twitter and other online resources are very powerful databases for health related information, as has been proven by countless research on monitoring health related issues online. Google is able to not only detect an existing flu epidemic, but also predict a flu epidemic that will occur in the nearby future[3]. By analyzing Google search terms, they found that certain search terms are a good predicator of a flu epidemic. The same methods apply to Google dengue trends, where Google is able to detect Dengue activity.[4]

Twitter has also proved itself as a valuable resource for disease detection in various research, as all these studies show that Twitter can be successfully used for detecting and predicting a flu epidemic. [5][6][7][8]
2.4. **Research on the statistics of word collocations**
There has been much research done on which method is the best for using word collocations. Varying from a relatively simple frequency analysis to a very mathematically complex approach. Examples of these methods are:

- association measures, that compute an association score for two words.
- n-grams, where there can be more than two words that are linked
- variable-length sequence, where the number of words is not fixed in advance
- distributional methods, that gives the frequency distribution of a given word
- higher-order statistics, which compares an clusters similar frequency distributions

In 2004, Evert et al. gives a good overview on association measures in order to compute word collocations.[9] The book first describes frequency counts as a method for locating word collocations. After that, association scores are assigned to each collocation, the higher the association score, the stronger the association between pairs of words are. The association measures are divided into four different categories: significance of association group, degree of association group, information theory group and heuristic formulas. We explain the likelihood measurement in a later section is an example of an association measure in the significance of association group. Other association measures that fall into these categories are dice coefficient (degree of association group), mutual information and MI3 (information theory group) and t-score (heuristic formulas), which are all explained in the next section.

All these measures have different advantages and disadvantages, which I shall not all explain. The main lesson we learn here is that there is a vast amount of methods that can be used within the field of association measures alone. The main advantages of association methods is that it is a relatively simple method, that can be applied to a large amount of word pairs.

There are online software tools available, that offer a ready-to-use package for identification and analysis of n-grams, such as the Ngram Statistic Package(NSP). NSP has been proven to be successful in identifying collocations and word associations. [10]

2.5. **Wordsmith**
Wordsmith is a software tool that allows the user to search for patterns in text corpora.

One of is key features is to calculate correlations and collocations. So in fact, this is a tool that conveniently performs the same analysis that Evert discusses.

Wordsmith reads a text corpus and searches for the given word and makes a frequency list for the words surrounding the keywords within a given horizon. The default setting of that horizon (or ‘window’) is 5, which means Wordsmith looks up to five place right, and five place left of the
keyword. Wordsmith has a function for calculating the relationship score. The relationship score is a numerical value assigned to a word collocation and can be seen as a score for how strongly a word is related to the keyword (in this case Ebola). This relationship score can be calculated by six different statistical methods:

- **Z-score**
  
  \[ \frac{J - E}{\sqrt{E(1 - P)}} \]

  where
  
  \( J \) = joint frequency
  
  \( S \) = collocational span
  
  \( F_1 \) = frequency of word 1
  
  \( F_2 \) = frequency of word 2
  
  \( P \) = \( \frac{F_2}{(\text{total tokens} - F_1)} \)
  
  \( E \) = \( P \times F_1 \times S \)

- **Mutual information**
  
  \[ \log_2\left(\frac{A}{BC}\right) \]

  where
  
  \( A \) = joint frequency divided by total tokens
  
  \( B \) = frequency of word 1 divided by total tokens
  
  \( C \) = frequency of word 2 divided by total tokens

- **Log likelihood [12]**
  
  \[ 2(a\ln(a) + b\ln(b) + c\ln(c) + d\ln(d) - ((a + b)\ln(a + b)) - ((a + c)\ln(a + c)) - ((b + d)\ln(b + d)) - ((c + d)\ln(c + d)) + ((a + b + c + d)\ln(a + b + c + d))) \]

  where
  
  \( a \) = joint frequency
  
  \( b \) = frequency of word 1
  
  \( c \) = frequency of word 2
  
  \( d \) := frequency of pairs involving neither \( w_1 \) nor \( w_2 \)

  and "\( \ln \)" means Natural Logarithm

- **Dice coefficient**
  
  \[ \frac{2J}{F_1 + F_2} \]

  \( J \) = joint frequency

  \( F_1 \) = frequency of word 1 or corpus 1 word count

  \( F_2 \) = frequency of word 2 or corpus 2 word count

  Ranges between 0 and 1.

- **T-score**
  
  \[ \frac{x}{\text{total tokens}} - \frac{x}{\sqrt{J}} \]

  where
  
  \( J \) = joint frequency
  
  \( F_1 \) = frequency of word 1
  
  \( F_2 \) = frequency of word 2
  
  \( X \) = \( F_1 \times F_2 \)
For this research we used the mutual information approach. Evert already stated that it is not clear which method works best, and further research did not fully clarify this. It is clear however that each method has its own ‘preferences’, with mutual information focusing more on low frequency words, and other statistics focusing more on words that more frequently appear.[17] Other researchers used Sketch Engine for their study, and did not explain which statistical method is behind this tool [1], or used the t-score method, but did not substantiated this.[2] The reason we eventually chose the mutual information method is due to the fact that it is a solid method for finding low frequency collocation.[17] The disadvantage could be that if it focuses too much on low frequency words, but we solved this by using a frequency cut off.

The main function for this research is the ‘collocation’ function. This function finds all the words near the search word (in this case Ebola), within a certain horizon. The window in this experiment is set on 5. Which mean Wordsmith searches 5 words to the right and 5 words to the left when Ebola occurs. After the window search, and the counting of words, it assigns a relationship score to the collocates. For instance, ‘the’ could be found quite a number of times near ‘Ebola’, but ‘the’ also appears near to many other words, which indicates that the words ‘the’ and ‘Ebola’ are not strongly related at all. That is also why we added the random tweets and newspaper articles, because Wordsmith then has data of other occurrences of the words and can compute whether a word is really strongly related to Ebola. Wordsmith is able to compute this relationship with several different statistical methods. This experiment uses the statistical method ‘mutual information’. The result of this Wordsmith analysis is a table with the results: words that are related to Ebola with their relationship score.
3. Data

3.1. Twitter database
The Information Science department of the University of Groningen has its own Dutch Twitter database. This database contains tweets since December 2010, and is still collecting Dutch tweets every day. Billions of tweets are collected via an API (dev.twitter.com/docs/api/1.1/post/statuses/filter), and relevant (i.e. Dutch tweets) are retrieved based on 229 Dutch words that are unique for the Dutch language (figure 1).

![List of the 229 keywords and hashtags used by the keyword filter for collecting Dutch tweets.](image)

Although these words are unique for the dutch language, there were still many non-Dutch words found in the results. For this reason, the language software libTextCat was added as an extra language checker, and the language that the user specified in Twitter was linked to its tweets. These methods were combined in such a way that when a tweet was labeled as Dutch by either of the two systems (not by both systems), it was positively selected as a Dutch tweet. This method results in a precision score of 97.6%, a recall score of 91.2% and an f1 score of 94.3.

Twitter has restrictions about the number of tweets that can be downloaded per time period, and that is why the database only covers about 40% of all Dutch tweets. But because the stream of tweets is evenly spread throughout the day, this 40% of all tweets is believed to be a representative sample of all Dutch tweets. A possible disadvantage of a 40% sample is that it can be that there is not enough data in certain time stamps available. Fortunately, there was enough data for the months we were interested in for this research, so the 40% Twitter sample did not influence this research.

3.2. Retrieving our Twitter data
The database is stored as gzip files, each containing an hour of data. Each month about 27 million tweets are stored, which means that for the time period of 1st July 2014 – 30th April 2015 there were millions and millions of tweets available in our database. Within those millions of tweets we searched for tweets containing Ebola. Duplicates had to be removed in order to prevent a skewed dataset and
because we want to make a comparison with newspapers, all tweets and retweets from news agencies had to be removed.

The Twitter database can be searched and results can be exported by using the command-line interface terminal. The ‘grep’ commands searches texts or files for the given words or strings. An example of the grep commands that we used:

```
zgrep -i 'ebola' tweets/2014/12/*.gz > EbolaDecember2014.txt
```

Zgrep in particular allows the user to search in compressed or zipped files. The search is case sensitive, that is why we have to include both ‘Ebola’ and ‘ebola’. The –i flag prevents the search to be case sensitive, so in this case, it will search for ebola as well as Ebola. This example searches through all files in the December 2014 directory and writes the results to the file ‘EbolaDecember2014.txt’.

In figure 2 we can see the number of tweets collected for each month with the command used above.

![Ebola frequency](image)

**Figure 2: Number of tweets containing 'Ebola' per month**

Although we do have an ‘Ebola-dataset’, this dataset is not suited for this research. It contains a lot of retweets, that would skew the results. We only wanted the original Tweets, as those are the only ones that are in fact written by an individual. Besides the retweets, a lot of Tweets are posted by accounts managed by news web sites. As we wanted to study the differences between Twitter and the media, we did not want the Twitter data to be “contaminated” by news from the media.

So first, we filtered out the retweets. Each retweet starts with the combination ‘RT’, so excluding all tweets with the case sensitive ‘RT’, was an effective way of excluding retweets.

The first zgrep command showed us that there were not only many tweets in the results of news
accounts, but that even the majority of the tweets were from news accounts or related organizations. It was not possible to filter all the accounts of news agencies, because there were far too many, even filtering on words such as ‘nieuws’ did not work well enough, because many tweets and usernames about Ebola news did not contain this word.

Fortunately, all these unwanted results had one similarity: they all contained a link to the original news article. Filtering on these links led to the elimination of almost all news related tweets. The disadvantage of this method is that the tweets containing a link from regular users were also eliminated. This only involved a very small number of tweets, making this method by far the best solution to the problem.

An example of one of the final commands:

```
zgrep -e -i 'ebola' tweets/2014/12/*.gz | grep ! 'http' | grep ! 'RT' > EbolaDecember2014.txt
```

The ‘-e’ is added because the command searches for multiple search patterns. As mentioned before, we also wanted to filter certain tweets containing specific words from our results, that is why the added the ‘-v’ commands with the words ‘http’ and ‘RT’. The command writes the results to the file ‘EbolaDecember2014.txt’.

This search resulted in the following number of tweets for each month:

![Number of tweets final dataset](image)

**Figure 3:** Number of tweets containing 'Ebola' per month, without retweets and news tweets
Comparing this dataset with the original ‘Ebola-dataset’ shows that between 12-25% of the original dataset remains after excluding the retweets and news tweets.

In order to get a higher relationship score for the word collocation of Ebola, a random sample of tweets had to be added. In this way, you also have examples of tweets that are not specific for Ebola, resulting in a higher mutual relationship score for words that are specifically mentioned with Ebola. Adding 27 million random tweets was not feasible due to practical reasons such as file size and downloading time. The sample was set on 40,000 random tweets, adding a different sample of this size each month.

3.3. Newspaper database
As for the (online and offline) media, the LexisNexis database was used. This database contains articles of the vast majority of Dutch newspapers (list included in Appendix A). The oldest archives reach back to 8th January 1990, the newspaper ‘NRC handelsblad’ has their archives digitalized since that date. The other major Dutch newspapers have their archives all digitalized since the 90s. A limitation of LexisNexis is that the important Dutch news website ‘nu.nl’ is not included in their database. This is because nu.nl is not allowed to store their articles for more than three months¹. However, the ANP is the largest news agency in the Netherlands, and is included in the LexisNexis database. ANP is also supplier of news for nu.nl, the biggest news website in the Netherlands. Therefore, the news from nu.nl is also covered in the LexisNexis database.

3.4. Retrieving our newspaper data
All Dutch articles containing ‘Ebola’ were collected for each month. LexisNexis has its own option to filter out duplicate articles. With its setting on ‘moderate similarity’, it removes all the articles that are moderately and highly similar.

This resulted in the following number of news articles for each month:

¹ http://webwereld.nl/e-commerce/9396-archief-nu-nl-weg-onder-druk-anp
In order to get a higher relationship score, a random sample of news articles had to be added. In this way, you also have examples of news articles that are not specific for Ebola, resulting in a higher mutual relationship score for words that are specifically mentioned with Ebola.

Adding thousands of random news articles was not feasible, because LexisNexis only allows you to download 200 articles at a time, and each download takes a few minutes. The sample was set on 1000 random news articles, adding a different sample of this same size each month. Although 40,000 tweets and 1000 newspaper articles seems like a big difference in sample size, the actual sample size is comparable, as the average of Ebola newspaper articles per month was 564 and the average of Ebola tweets per month was 27,819, each corpus has somewhat less than twice their own size added with sample data.

4. Method

4.1. Lexical richness
The lexical richness of a dataset is the variety in vocabulary used in that dataset. This is of interest for this research, because the lexical richness has influence on the mutual information score. If one of the datasets has a higher lexical richness, that means that there are more unique words in that dataset and it is more difficult to catch all the relevant keywords. We attempted to correct this problem by including all conjugations of keywords (for example the past and present tense of a verb) and by
including the fact that Twitter and newspaper use different words for the same message. For example ‘aanstekelijk’ (‘contagious’) is used more on Twitter, while ‘besmettelijk’ (in English also ‘contagious’) is used more in newspapers, while they mean the same.

The lexical richness is calculated by the type/token ratio. The type/token ratio is the percentage of unique words in the dataset and is calculated by the following formula:

$$R = \frac{U}{T} \times 100$$

Where R is the type/token ratio in percentage, U is the number of unique words and T is total number of words.

Not all words in a text are unique, some words are used multiple times in a text. If the Twitter corpus contains many more unique words, then it can be assumed that words are less correlated with Ebola and the relation scores will be lower.

For instance, a text contains a 1000 words, but uses 560 unique words. This text has 1000 ‘tokens’ and 560 ‘types’. It’s type/token ratio (TTR) is therefore \(\frac{560}{1000} \times 100 = 56\%\).

The TTR is very dependent on the length of the text. A very short text can reach a TTR of 80\%, while a text with a million of words can have a TTR of only 2\%.

### 4.2. Top 100 of word collocations

We wanted to compare the overall message spread by the media with the opinion that Dutch people have about Ebola on Twitter. One way to do so, is by adding all the data we gathered in each of the different months into one big dataset, and then calculate the mutual information score for Ebola with other words with Wordsmith. We added a random sample to the original dataset because we want to know how often a word appears in random text. This is necessary information in order to investigate whether a word is indeed specific for Ebola. For instance, if a word appears frequently with Ebola, but also appears frequently in random text, the word is not specific for Ebola, and the mutual information score should be lowered.

This random sample was the exact same size as the original dataset. This means that we added exactly 278.189 random tweets to the Twitter dataset and 5638 random newspaper articles to the newspaper dataset. Because we added the exact same portion of random data to each dataset, we were now able to really compare the mutual information score of the different media.

A well known limitation of mutual information is that it is very sensitive to low frequency. If for instance a keyword appears only twice together with Ebola, but it does not appear in the random sample, than the keyword gets a high mutual information score assigned, while the occurrence of the
keyword could be a coincidence. That is why we applied a frequency cut-off for the keywords. There were much more word collocations on Twitter than in the newspapers, the frequency cut-off is therefore higher for the Twitter results than for the newspaper results. A keyword in the Twitter results had to occur at least nine times before we would take it into consideration and a keyword in the newspapers had to occur at least five times before we would consider it as a valid word collocation.

4.3. **Trends of clusters per media**

A shortcoming of previous studies is that they investigate a long time period, but that they only have a single conclusion, for instance, that the use of the word Muslim in the UK press is linked to armed issues, social conflict and terrorism.[1] We wanted to go a step further, and discover trends about Ebola over time. The data showed us that it is best to look at the data per month. In this way, there is enough data to compare the months June 2014 – April 2015.

In order to see which words are strongly correlated to the word Ebola, the tool ‘Wordsmith’ is used. In order to successfully compare the words related to Ebola, words have to be split into categories. These categories, or clusters are formed on the basis of the mutual information score. For each month, the mutual information score is calculated with Wordsmith. The words are ranked from high to low mutual information score and the highest 10% is considered as a possible keyword. Not all words are interesting for our results, for instance ‘binnenland’ (in this case probably best translated as ‘national’) says nothing about a view or opinion about Ebola. That is why a selection from the top 10% is made.

The trend of these clusters can then be plotted in order to see whether there is a lot of attention on these words or not. The relationship score cannot be used for this purpose, as not every month has the same amount of data. For instance, the Ebola tweets from July contain a total of 17,413 words, while the Ebola tweets from October contain a total of 343,611 words, but every month has the same amount of random tweets added (namely 40,000). This resulted in a much higher relationship score for the keywords related to Ebola in July than the related keywords in October. In this case, we used relative frequency [16]. We compared the relative frequency of the words appearing with Ebola within the search horizon. If for instance ‘disease’ appears within the search horizon 20 times in July, and 50 times in October, the relative frequency in July is 20/17,413*100=0,1149 while the relative frequency in October is 50/343,611*100=0,0146. Plotting the relative frequency of the clusters each month give clear understanding of the amount of attention that is given in that media to that subject.
There were twelve clusters which we took into account:

- ‘bestrijding’ (fighting the disease)
- ‘behandeling’ (treatment)
- ‘medicijnen/vaccins (medicine/vaccines)
- ‘maatregelen’ (measures)
- ‘doden/overleden/etc.’ (deaths, deceased, etc.)
- ‘besmettelijk/versprijding/ etc.’ (contagious/spreading/etc.)
- ‘gevaar’ (danger)
- ‘angst/paniek’ (fear/panic)
- ‘grap’ (joke)
- ‘oplossing’ (solution)
- ‘problem’ (problem)
- ‘patiënten’ (patients)

Each cluster contains all related words and all singulars, plurals, tenses and synonyms are included. These clusters can be compared between the sources, that is how we can see the similarities and differences between the sources, and see if one source possibly influences the other source.

The scores are compared by relative frequency. The words in each cluster are summed up to a total and this total is divided by the number of Ebola mentions in the source that month. The total is not divided by the total amount of words in the source that month, because Twitter uses a lot less words than newspaper articles, and this would make a poor comparison. Because the window of each Ebola mention is the same (namely 5), dividing the total of a cluster by the total number of Ebola mentions is the best comparison. This is done automatically through a Python programming code.

The pseudo code to achieve this is:

Read the excel file

If the string in column B = Ebola

Then Number of Ebola mentions = column F

If the string in columns B contains ‘strijd’ or ‘aanpak’ or ‘tegengaan’

Then add the value of column F to the total cluster number

Divide the total cluster number by the number of Ebola mentions and multiply by 100
5. **Results**

5.1. **Lexical richness**

The newspapers dataset has a total 2,242,405 words and the Twitter dataset has a total 1,001,853 words, so it would not be fair to compare the two corpora purely based on their type token ratio as the type token ratio is highly influenced by the length of the text. That is why we used the tool Wordsmith for the calculation of a standardized TTR. A standardized TTR takes large chunks of texts, and calculates the TTR for this piece of text, then moves on to the next chunk of text and calculates the TTR, and so on until all chunks have a TTR, the standardized TTR is the average of these TTRs.

In this case, the TTR of the newspaper corpus is \( \frac{75.132}{2.242.405} \times 100 = 3.35\% \) and the TTR of the Twitter corpus is \( \frac{67.095}{1.001.853} \times 100 = 6.70\% \). The standardized TTR however is 24.92\% for the newspapers and 22.00\% for Twitter, using chunks of 10,000 words. This implies that the newspapers are actually lexically more enriched than the Twitter corpus.

5.2. **Comparing the top 100 word collocations per media**

We calculated the mutual information score and the top 100 word collocations are shown in table 4. Keep in mind that the cut-off is nine for Twitter and five for the newspapers. The words that are highlighted in yellow are the words that we believed are indicative for this research. They are not entirely objective, but show a certain point of view towards the subject, in this case Ebola. The words that are highlighted in pink are diseases.

**Table 4: Overview of the top 100 word collocations per media.**

<table>
<thead>
<tr>
<th>Twitter</th>
<th>Relationship score</th>
<th>Frequency</th>
<th>Newspapers</th>
<th>Relationship score</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>MANÃ</td>
<td>6,05</td>
<td>11</td>
<td>UGANTEES</td>
<td>5,32</td>
<td>5</td>
</tr>
<tr>
<td>EBOLAÁ</td>
<td>4,05</td>
<td>14</td>
<td>TASK</td>
<td>5,32</td>
<td>5</td>
</tr>
<tr>
<td>STEPHANIEMLW</td>
<td>3,88</td>
<td>9</td>
<td>FRONTLINE</td>
<td>5,32</td>
<td>5</td>
</tr>
<tr>
<td>THUISOPÃ</td>
<td>3,88</td>
<td>9</td>
<td>PUBLICATIETIJDEN</td>
<td>5,06</td>
<td>6</td>
</tr>
<tr>
<td>JOLIENJOOSSEN</td>
<td>3,88</td>
<td>9</td>
<td>UITWIJST</td>
<td>5,06</td>
<td>5</td>
</tr>
<tr>
<td>BESTRIJDERS</td>
<td>3,72</td>
<td>9</td>
<td>VERERGERING</td>
<td>5,06</td>
<td>5</td>
</tr>
<tr>
<td>OPGEDOEN</td>
<td>3,72</td>
<td>9</td>
<td>HERMSE</td>
<td>4,84</td>
<td>7</td>
</tr>
<tr>
<td>ENOLA</td>
<td>3,59</td>
<td>11</td>
<td>BESTRIJDER</td>
<td>4,84</td>
<td>7</td>
</tr>
<tr>
<td>WEETIK</td>
<td>3,59</td>
<td>10</td>
<td>WÁS</td>
<td>4,84</td>
<td>7</td>
</tr>
<tr>
<td>STERFTECLJFER</td>
<td>3,59</td>
<td>10</td>
<td>BEHANDELMETHODEN</td>
<td>4,84</td>
<td>6</td>
</tr>
<tr>
<td>Term</td>
<td>Score</td>
<td>Position</td>
<td>Definition</td>
<td>Score</td>
<td>Position</td>
</tr>
<tr>
<td>----------------------</td>
<td>-------</td>
<td>----------</td>
<td>---------------------------</td>
<td>-------</td>
<td>----------</td>
</tr>
<tr>
<td>GETEISTERDE</td>
<td>3.59</td>
<td>10</td>
<td>TEKORTGESCHOTEN</td>
<td>4.84</td>
<td>6</td>
</tr>
<tr>
<td>ILLUMINATIE</td>
<td>3.59</td>
<td>9</td>
<td>REEUWIJKSE</td>
<td>4.84</td>
<td>6</td>
</tr>
<tr>
<td>GEINFECTEERD</td>
<td>3.59</td>
<td>9</td>
<td>TWAALFDUIZEND</td>
<td>4.84</td>
<td>5</td>
</tr>
<tr>
<td>BESTRIJDT</td>
<td>3.59</td>
<td>9</td>
<td>VOORLICHTEN</td>
<td>4.84</td>
<td>5</td>
</tr>
<tr>
<td>NIGELTELLES</td>
<td>3.59</td>
<td>9</td>
<td>DANSMARATHON</td>
<td>4.84</td>
<td>5</td>
</tr>
<tr>
<td>VUURWERKGELD</td>
<td>3.46</td>
<td>12</td>
<td>GRAPT</td>
<td>4.84</td>
<td>5</td>
</tr>
<tr>
<td>GEPATENTEERD</td>
<td>3.46</td>
<td>11</td>
<td>AGAINST</td>
<td>4.84</td>
<td>5</td>
</tr>
<tr>
<td>PANIEKZAAIERIJ</td>
<td>3.46</td>
<td>9</td>
<td>TESTVACCIN</td>
<td>4.65</td>
<td>6</td>
</tr>
<tr>
<td>MUTEERT</td>
<td>3.35</td>
<td>12</td>
<td>RISICOPASSAGIERS</td>
<td>4.48</td>
<td>7</td>
</tr>
<tr>
<td>DIEJEN</td>
<td>3.35</td>
<td>12</td>
<td>VERERGERD</td>
<td>4.48</td>
<td>5</td>
</tr>
<tr>
<td>CHIKUNGUNYA</td>
<td>3.35</td>
<td>11</td>
<td>PARLEMENTARIËR</td>
<td>4.48</td>
<td>5</td>
</tr>
<tr>
<td>LIKEMERIAM</td>
<td>3.35</td>
<td>10</td>
<td>TEGENGEGAAN</td>
<td>4.48</td>
<td>5</td>
</tr>
<tr>
<td>SIGUIRI</td>
<td>3.35</td>
<td>9</td>
<td>BEGRAFENISTEAM</td>
<td>4.48</td>
<td>5</td>
</tr>
<tr>
<td>INGEZAMELD</td>
<td>3.35</td>
<td>9</td>
<td>BEWAAR</td>
<td>4.48</td>
<td>5</td>
</tr>
<tr>
<td>VIRUSJE</td>
<td>3.35</td>
<td>9</td>
<td>VERSPILLEN</td>
<td>4.32</td>
<td>6</td>
</tr>
<tr>
<td>PANORAMATV</td>
<td>3.24</td>
<td>14</td>
<td>BETEUGELD</td>
<td>4.32</td>
<td>6</td>
</tr>
<tr>
<td>HULPACTIE</td>
<td>3.24</td>
<td>11</td>
<td>RESPONSE</td>
<td>4.32</td>
<td>6</td>
</tr>
<tr>
<td>XCURLYBALA</td>
<td>3.24</td>
<td>11</td>
<td>MALIKA</td>
<td>4.32</td>
<td>5</td>
</tr>
<tr>
<td>PASIA</td>
<td>3.24</td>
<td>10</td>
<td>GEDIAGNOSTICEERD</td>
<td>4.32</td>
<td>5</td>
</tr>
<tr>
<td>HOERE</td>
<td>3.24</td>
<td>10</td>
<td>CURE</td>
<td>4.19</td>
<td>7</td>
</tr>
<tr>
<td>OVERHEERST</td>
<td>3.24</td>
<td>10</td>
<td>SYMPTOOM</td>
<td>4.19</td>
<td>7</td>
</tr>
<tr>
<td>AMELIEWEN</td>
<td>3.24</td>
<td>10</td>
<td>ONTDEKKER</td>
<td>4.19</td>
<td>6</td>
</tr>
<tr>
<td>NOODFONDS</td>
<td>3.24</td>
<td>10</td>
<td>PATIENTEN</td>
<td>4.19</td>
<td>5</td>
</tr>
<tr>
<td>HETISESMEEX</td>
<td>3.14</td>
<td>14</td>
<td>VOORGELICHT</td>
<td>4.19</td>
<td>5</td>
</tr>
<tr>
<td>VERSPREI</td>
<td>3.14</td>
<td>12</td>
<td>HOMEOPATHIE</td>
<td>4.06</td>
<td>8</td>
</tr>
<tr>
<td>ERGERS</td>
<td>3.14</td>
<td>11</td>
<td>PROOF</td>
<td>4.06</td>
<td>8</td>
</tr>
<tr>
<td>GONORROE</td>
<td>3.14</td>
<td>10</td>
<td>BINNENBLIJVEN</td>
<td>4.06</td>
<td>7</td>
</tr>
<tr>
<td>AGYTEVOL</td>
<td>3.14</td>
<td>9</td>
<td>HOMEOPATEN</td>
<td>4.06</td>
<td>6</td>
</tr>
<tr>
<td>NIALLFTBETUL</td>
<td>3.14</td>
<td>9</td>
<td>ELIBOL</td>
<td>4.06</td>
<td>6</td>
</tr>
<tr>
<td>MONDIALE</td>
<td>3.14</td>
<td>9</td>
<td>RASIT</td>
<td>4.06</td>
<td>6</td>
</tr>
<tr>
<td>LIJDERS</td>
<td>3.05</td>
<td>13</td>
<td>PIERIK</td>
<td>4.06</td>
<td>5</td>
</tr>
<tr>
<td>GEVALLE</td>
<td>3.05</td>
<td>13</td>
<td>AANRICHT</td>
<td>4.06</td>
<td>5</td>
</tr>
<tr>
<td>BIOWAPEN</td>
<td>3.05</td>
<td>12</td>
<td>INWEST</td>
<td>3.95</td>
<td>6</td>
</tr>
<tr>
<td>OPDUIKT</td>
<td>3.05</td>
<td>10</td>
<td>JUICHPAK</td>
<td>3.95</td>
<td>6</td>
</tr>
<tr>
<td>OPGEDOKEN</td>
<td>3.05</td>
<td>10</td>
<td>BEMOEDIGEND</td>
<td>3.95</td>
<td>6</td>
</tr>
<tr>
<td>Word</td>
<td>Score</td>
<td>Type</td>
<td>Word</td>
<td>Score</td>
<td>Type</td>
</tr>
<tr>
<td>------------------</td>
<td>-------</td>
<td>---------------</td>
<td>------------------</td>
<td>-------</td>
<td>---------------</td>
</tr>
<tr>
<td>WOIII</td>
<td>3.05</td>
<td>10</td>
<td>BUITENLANDREDACTIE</td>
<td>3.95</td>
<td>6</td>
</tr>
<tr>
<td>SVDNESSJPG</td>
<td>3.05</td>
<td>9</td>
<td>SOMETHING</td>
<td>3.95</td>
<td>5</td>
</tr>
<tr>
<td>LEpra</td>
<td>2.96</td>
<td>14</td>
<td>VERTOONDEN</td>
<td>3.95</td>
<td>5</td>
</tr>
<tr>
<td>REXTREST</td>
<td>2.96</td>
<td>11</td>
<td>AFHERSERS</td>
<td>3.95</td>
<td>5</td>
</tr>
<tr>
<td>BUILENPEST</td>
<td>2.96</td>
<td>10</td>
<td>BESMETTINGSRISICO</td>
<td>3.95</td>
<td>5</td>
</tr>
<tr>
<td>ONTEKTE</td>
<td>2.96</td>
<td>9</td>
<td>BESTRIJVERS</td>
<td>3.84</td>
<td>9</td>
</tr>
<tr>
<td>GEZONDHEIDSWE</td>
<td>2.88</td>
<td>11</td>
<td>HEMORRAGISCHE</td>
<td>3.84</td>
<td>8</td>
</tr>
<tr>
<td>RKERS</td>
<td>2.88</td>
<td>10</td>
<td>DOODDEGAAN</td>
<td>3.84</td>
<td>7</td>
</tr>
<tr>
<td>UITBRAKEN</td>
<td>2.88</td>
<td>9</td>
<td>TILBURGERS</td>
<td>3.84</td>
<td>5</td>
</tr>
<tr>
<td>KUTLEVEN</td>
<td>2.88</td>
<td>9</td>
<td>JIHAGANGN</td>
<td>3.84</td>
<td>5</td>
</tr>
<tr>
<td>OBERHAUSEN</td>
<td>2.88</td>
<td>9</td>
<td>VOORLICHTINGSCAMPAGNES</td>
<td>3.84</td>
<td>5</td>
</tr>
<tr>
<td>MATTHYSSANDER</td>
<td>2.88</td>
<td>9</td>
<td>ENZ</td>
<td>3.84</td>
<td>5</td>
</tr>
<tr>
<td>LANGEFRANS</td>
<td>2.88</td>
<td>9</td>
<td>BESCHIKBAARHEID</td>
<td>3.84</td>
<td>5</td>
</tr>
<tr>
<td>STUITEND</td>
<td>2.80</td>
<td>19</td>
<td>FORCE</td>
<td>3.74</td>
<td>7</td>
</tr>
<tr>
<td>DENCILEK</td>
<td>2.80</td>
<td>17</td>
<td>TSAAR</td>
<td>3.74</td>
<td>7</td>
</tr>
<tr>
<td>DOCTER</td>
<td>2.80</td>
<td>13</td>
<td>OVERLEVERS</td>
<td>3.74</td>
<td>7</td>
</tr>
<tr>
<td>VERDBELT</td>
<td>2.80</td>
<td>13</td>
<td>SCHROEVEN</td>
<td>3.74</td>
<td>6</td>
</tr>
<tr>
<td>STIERVEN</td>
<td>2.80</td>
<td>11</td>
<td>VEGAS</td>
<td>3.74</td>
<td>5</td>
</tr>
<tr>
<td>JANDROSEFB</td>
<td>2.80</td>
<td>10</td>
<td>SEARCH</td>
<td>3.65</td>
<td>7</td>
</tr>
<tr>
<td>MUTEREN</td>
<td>2.80</td>
<td>10</td>
<td>OVERDRAAGBAAR</td>
<td>3.65</td>
<td>6</td>
</tr>
<tr>
<td>MERS</td>
<td>2.80</td>
<td>10</td>
<td>LOTTE</td>
<td>3.65</td>
<td>5</td>
</tr>
<tr>
<td>ROSALINFENTY</td>
<td>2.80</td>
<td>9</td>
<td>GENOOTSCHAP</td>
<td>3.65</td>
<td>5</td>
</tr>
<tr>
<td>HETISLUUK</td>
<td>2.80</td>
<td>9</td>
<td>LEMGERMEDEWERKERK</td>
<td>3.56</td>
<td>22</td>
</tr>
<tr>
<td>EQUATEUR</td>
<td>2.72</td>
<td>20</td>
<td>FOUCHIER</td>
<td>3.56</td>
<td>8</td>
</tr>
<tr>
<td>EPICENTRUM</td>
<td>2.72</td>
<td>17</td>
<td>ONBEOEDLDE</td>
<td>3.56</td>
<td>7</td>
</tr>
<tr>
<td>ESSIEEN</td>
<td>2.72</td>
<td>17</td>
<td>BINNENKWAM</td>
<td>3.56</td>
<td>7</td>
</tr>
<tr>
<td>KRIJGE</td>
<td>2.72</td>
<td>14</td>
<td>DUSVERRE</td>
<td>3.48</td>
<td>12</td>
</tr>
<tr>
<td>HERPES</td>
<td>2.72</td>
<td>12</td>
<td>GELYNCHT</td>
<td>3.48</td>
<td>7</td>
</tr>
<tr>
<td>LUCKYTV</td>
<td>2.72</td>
<td>12</td>
<td>HERKEN</td>
<td>3.48</td>
<td>7</td>
</tr>
<tr>
<td>VERPLEGEN</td>
<td>2.72</td>
<td>11</td>
<td>BESTRIEED</td>
<td>3.48</td>
<td>6</td>
</tr>
<tr>
<td>TBMIXMASTER</td>
<td>2.72</td>
<td>11</td>
<td>KORRELS</td>
<td>3.48</td>
<td>6</td>
</tr>
<tr>
<td>REISVERBOD</td>
<td>2.72</td>
<td>11</td>
<td>BLOEDENS</td>
<td>3.48</td>
<td>5</td>
</tr>
<tr>
<td>AGV</td>
<td>2.72</td>
<td>11</td>
<td>FABELS</td>
<td>3.40</td>
<td>18</td>
</tr>
<tr>
<td>KIEK</td>
<td>2.72</td>
<td>10</td>
<td>CRISICOORDINATOR</td>
<td>3.40</td>
<td>13</td>
</tr>
</tbody>
</table>
The Dutch are known to curse a lot with diseases, with examples as kolere (Eng: Cholera), tyfus (Eng: typhus) and kanker (Eng: cancer). This is also reflected in the top 100 word collocations, as Ebola is frequently mentioned with other diseases on Twitter (highlighted in pink). Another thing the Dutch are famous for is their straightforwardness and irony. The Dutch don’t beat around the bush and often they mean the opposite of what they write or say. This effect is illustrated by the following sample of Tweets, with their literal translation:

```
Tweets/2014/10/20141001:17.out.gz:GWNjeffrey2312     Hoop dat al me docenten ebola krijgen ofs👏👏
```

“I hope all my professors will get Ebola.”
“Jesus, I’m following cancer emo ebola people. Just nicely die all of you.”

“Get the ebola.”

“I wouldn’t be surprised if I got ebola #ill.”

“I just see it happening that I get ebola. Never been ill, never broke anything. Pay attention, karma’s gonna get me.”

“Blegh, I’m also ill at home… I’m guessing Ebola #ill #Ebola.”

“So what, ebola and free Palestina. Lefties, those are having a really difficult time!”

Interesting word collocations are highlighted in yellow. Other words are noise from the dataset, in the case of Twitter, many words are usernames that happen to mention Ebola a few times. The diseases mentioned earlier are not considered as interesting word collocations as they only appear on Twitter and not in the newspapers, they cannot be used for comparison. Other keywords are not noise, but also do not say something about the message that is spread about Ebola, or the opinion that people have on Ebola.


There are a number of very interesting words that could not find a place in a cluster, but definitely say something about how Ebola is viewed. Specific words for a single months can say something interesting, like a shift in the opinion about Ebola in that particular month. These words are shown in table 5.
Table 5: words that are not clustered, but contain relevant information about Ebola.

<table>
<thead>
<tr>
<th>Twitter</th>
<th>English translation</th>
<th>Newspapers</th>
<th>English translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illuminatie</td>
<td>Illuminati</td>
<td>Bemoedigend</td>
<td>Encouraging</td>
</tr>
<tr>
<td>Paniekzaaierij</td>
<td>Alarmism</td>
<td>Hysterie</td>
<td>Hysteria</td>
</tr>
<tr>
<td>Virusje</td>
<td>‘Little virus’</td>
<td>Grapt</td>
<td>Joking</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tekortgeschoten</td>
<td>Failed/inadequately</td>
</tr>
</tbody>
</table>

For all words we examined whether they fit into a cluster and whether they occurred in several months. If they did not occur in several months, we did not include them in a cluster.

5.3. **Trends of clusters per media**

The following two graphs show how much attention is given to the clusters of word on a medium (Twitter or newspaper) in the examined months. This is calculated with the relative frequency.

![Twitter Graph](image_url)

*Figure 5: Amount of attention given to clusters of words on Twitter calculated by relative frequency*
The first thing to notice are the overall scores on Twitter and the newspapers. The highest scoring word clusters in the newspapers score 0.6 to about 1.5, while the highest score on Twitter is only 0.6.

Twitter focusses on the deaths and deceased. The spreading of Ebola was highly discussed in September, but rapidly decreased after September. The fight against Ebola gets quite some attention, with peaks in September and November. Medicines are discussed mostly in December, and less in other months.

5.4. **Comparing each cluster per media**
The following graphs compare a cluster of words between the two media (Twitter and newspaper) during the examined months.
The newspaper score significantly higher than Twitter when it comes to words associated with the fight against Ebola. Twitter shows minor peaks in September and November, while the newspapers scores overall higher with peaks in November and March. There was a national Day of Action on the 28th of November, on which money was raised for the fight against Ebola. This explains why there was a high peak for this cluster in November, and it is interesting to see that this gets quit a lot more attention in the newspapers than on Twitter. In March there is news about emergency measure taken in fight against Ebola in the West and Southwest of Guinea.
Overall, the newspapers score higher for words associated with death and death toll. There is a sharp drop measured in February, and a high peak after that in March. The peak in March can be explained by two events:
- the discovery that the Ebola outbreak began with the death of a toddler in Guinee
- the headlines stating ‘thousands of Ebola patients died because of inertia at WHO’

Apparently, this news does not affect the opinion of the people on Twitter.

Figure 9: comparison of the cluster 'joke' between Twitter and newspapers

This is one of the few graphs where Twitter have higher relative frequency than the newspapers, indicating that people on Twitter associate words like ‘joke’ and ‘funny’ a lot more with Ebola than in the newspapers. There are quite a lot of jokes about Ebola on the Dutch Twitter, some examples:

Tweets/2014/08/20140804:21.out.gz:boosvogel ebola klinkt als een online versie van bola, al heb ik geen idee wat bola is.
“ebola sounds like the online version of bola, although I have no idea what bola is.”

Tweets/2014/09/20140917:16.out.gz:ChageNijn Waarom doen die mensen met Ebola eigenlijk geen #IceBucketChallenge? Willen ze niet beter worden ofzo?
“Why are the people with Ebola not doing the #IceBucketChallenge? Don’t they want to get better?”
Figure 10: comparison of the cluster 'treatment' between Twitter and newspapers

Words associated with treatment score higher in the newspapers than on Twitter, but the overall trend is the same, except for a large peak in March. This peak in March in the newspapers is about the fact that there are hardly any Ebola patients left, and that many treatment centres are abandoned. The absence of a peak on Twitter in March suggests that the Twitterers do not talk about such positive news.

Figure 11: comparison of the cluster 'spreading of Ebola' between Twitter and newspapers
The trends in the cluster ‘spreading’ are similar in the newspapers and on Twitter. The major difference is the peak in April in the newspapers while there is a decrease in April on Twitter. The peak in September can be explained by the fact that the Ebola outbreak really took off in September, causing a lot more media attention. Although one might expect that the peak for this cluster would be higher in November, when the number of infections per day was the highest, as can be seen in figure 12.

![Figure 12: Cases and deaths of Ebola per day](image.png)

The peak in April in the newspapers stems from two news items:

- Dutchman Peter Jan Graaff is the new leader of the Ebola mission of the United Nations, a mission that fights the contagious disease.

- More than 26,000 people are infected with Ebola since the outbreak.
Figure 13: comparison of the cluster 'solution' between Twitter and newspapers

This graph stands out from other graphs. There were zero words associated with a solution in the newspapers, while there are on Twitter. This indicates that newspaper do not focus on a solution, while the Dutch Twitterers do talk about a solution. Although this statement must be qualified by the fact that this cluster contains a very limited amount of words. The peak in August is partly caused by a Dutch woman stating that “homeopathy could be the solution to Ebola”, this results in a discussion that mostly mocks her. This is however a small portion of the mentions of ‘solution’. Other mentions are variegated, there is not a specific topic that is discussed in the other tweets. Most of them are about the fact that there is (still) no solution for Ebola and some of them are saying that Ebola might be the solution for overpopulation.

Figure 14: comparison of the cluster 'medicine' between Twitter and newspapers
The trends in the cluster ‘medicine’ are very similar in the newspapers and on Twitter. Although the overall relative frequency is higher in the newspapers than on Twitter, the trends are almost the same. The peak in August in the newspapers is about a new experimental medicine called ‘ZMapp’, that can cure Ebola in monkeys. The tweets about medicines and Ebola on Twitter are very different in August, there is not a specific source of the tweets, but most of them are concerned that there is no medicine or cure for Ebola.

![Figure 15: comparison of the cluster 'danger' between Twitter and newspapers](image)

The cluster ‘danger’ follows a very different trend in the newspapers than on Twitter. Twitter only peaks in August, while there is a decrease in frequency in the same month in the newspapers. The newspapers have minor peaks in September and December and a major peak in April. In September, the newspapers write about the misconception how contagious Ebola really is, and that the danger of contamination is not that high. This is an interesting development, the newspapers put the danger of Ebola into perspective and there is a decrease of the mentioning of danger on Twitter. Most of the news mentioning danger in relation to Ebola speak about medical workers risking their own life in the fight against Ebola.

The peak in April are from newspapers warning that the danger about Ebola is not over, although there is a smaller change that the disease will spread again. Some of the newspapers are even stating that ‘the most dangerous period has arrived’, because the attention to Ebola is now weakening.
Figure 16: comparison of the cluster 'problem' between Twitter and newspapers

The trends of the cluster ‘problems’ are almost the opposite on Twitter and in the newspapers. There is a large amount of attention given to this subject in July on Twitter, while there is zero attention to
the cluster ‘problems’ in the newspapers the same month. There is also a large peak in March on newspapers, while there is no peak on Twitter.

The peak on Twitter in July on Twitter cannot be explained by studying the tweets. Each tweet uses the word ‘problem’ differently, there cannot be a source pinpointed.

The peak in March in the newspapers is due to the fact that the Ebola outbreak started exactly one year ago and ‘Ebola is still a problem and a stigma’.

![Figure 17: comparison of the cluster 'measurements' between Twitter and newspapers](image)

The attention on Twitter about the measurements against Ebola is very low, there are however high peaks for this cluster in the newspapers.

In March there were newspapers reporting that new Ebola measures were taken in Guinea.
Figure 18: comparison of the cluster ‘fear’ between Twitter and newspapers

The trend of the cluster ‘fear’ is very stable for Twitter. In the newspapers there is a major peak in October. The newspapers that month write about the fear that the local population has for the medical staff. Disbelief and superstition are still part of the view that the local population has on Ebola.

Figure 19: comparison of the cluster 'patient' between Twitter and newspapers

The overall trend for the cluster ‘Patients’ is approximately the same on Twitter and in the newspapers. There is a peak in December. There was talk of the possibility that an Ebola patient would come to the Netherlands. This clearly got a lot of attention, mostly on Twitter.
5.5. **Pearson product-moment correlation coefficient**

In order to prove whether there is a link between the Twitter and newspaper results, the Pearson product-moment correlation coefficient is calculated for each cluster of words between the two media. The Pearson product-moment correlation coefficient is a measure of the strength of a linear association between two sets of numerical values. The correlation coefficient is always a value between -1 and +1. A negative value means that the data sets are negatively correlated, in our results this would mean that the trends are opposite on Twitter and in the newspapers. We therefore expect a negative correlation coefficient with the cluster ‘problem’.

The Pearson product-moment correlation coefficient for two sets of values, $x$ and $y$, is given by the formula:

$$ r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} $$

where $\bar{x}$ and $\bar{y}$ are the sample means of the two arrays of values.

The software package SPSS automatically calculated the Pearson product-moment correlation coefficient with the given statistics. Table 6 gives an overview of the results of this calculation for each cluster.

**Table 6: Pearson correlation coefficient for each cluster of words between Twitter and newspapers.**

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>fight</td>
<td>0.583741</td>
</tr>
<tr>
<td>treatment</td>
<td>0.618976</td>
</tr>
<tr>
<td>medicine</td>
<td>0.712593</td>
</tr>
<tr>
<td>measurement</td>
<td>0.293057</td>
</tr>
<tr>
<td>death</td>
<td>0.69257</td>
</tr>
<tr>
<td>spreading</td>
<td>0.60573</td>
</tr>
<tr>
<td>danger</td>
<td>-0.15922</td>
</tr>
<tr>
<td>fear</td>
<td>0.388675</td>
</tr>
<tr>
<td>joke</td>
<td>0.469253</td>
</tr>
<tr>
<td>solution</td>
<td>!</td>
</tr>
<tr>
<td>problem</td>
<td>-0.39823</td>
</tr>
<tr>
<td>patients</td>
<td>0.485526</td>
</tr>
</tbody>
</table>
Calculating the Pearson correlation coefficient for the word cluster ‘solution’ was not possible, as this cluster has no results in the newspapers, and the value therefore had to be divided by zero. Clusters with a high correlation score are: fight, treatment, medicine, death and spreading. This means that Twitter and the newspaper are mutually consistent of the attention they give to a certain cluster. The clusters measurement, danger, fear, joke, solution, problem and patients do not have a high correlation score. This means that the amount of attention given to these clusters on Twitter and the newspapers differ from each other. The clusters ‘danger’ and ‘problem’ even have a negative correlation coefficient, which means that Twitter and the newspapers rather show an opposite trend for these clusters.

6. Conclusion

In order to draw a conclusion, we have to get back to our original research question. The goal of this research was to get to the core of Twitter show how Dutch people on Twitter think about Ebola versus the message that is spread about Ebola in Dutch written media. Is Twitter still a news media if you strip away the retweets and headlines?

When we look at the results, there is a clear dichotomy. We can clearly make a separation into two categories. For the clusters with more objective concepts like treatment, medicine, death and spreading, the media and Twitter are closely linked, with correlation coefficients above 0.5. When it comes to subjective concepts like fear, danger, joke and problem it is clear that the media and Twitter are not linked to one another, and sometimes even follow an opposite trend, with correlation coefficients below 0. This means that for some topics, people on Twitter do follow the news media, but on some topics they form a completely different view or opinion. Newspapers are a far more objective medium while Twitter is clearly a subjective medium. The results show that especially the objective news in the newspapers are shared at Twitter, while the Dutch people on Twitter form an opinion completely on their own when it comes to the subjective topics. This opinion is in most of the cases that you don’t have to take Ebola seriously, indicated by the term ‘illuminati’, the frequent mentions of other diseases that appear in the word collocations and the jokes made about Ebola on Twitter.
Literature


Appendix A

Dutch sources of LexisNexis

A

A&C Media
Accountant
AD/Algemeen Dagblad
AD/Amersfoortse Courant
AD/De Dordtenaar
AD/Groene Hart
AD/Haagse Courant
AD/Rivierenland
AD/Rotterdams Dagblad
AD/Sportwereld Pro*
AD/Utrechts Nieuwsblad
AFX - NL*
Algemeen Dagblad*
Algemeen Nederlands Persbureau - ANP
AllAboutFeed (English)
Almere Vandaag*
Alphen.cc*
Amersfoortse Courant*
ANP English News Bulletin*
AP Dutch Worldstream
ASAPII Database
Avicultura Prof (Spanish)*

B

Beleggers Belangen
BiZZ*
BN/DeStem
BNR.nl
Boerderij
Boerderij Melkvee 100Plus (Dutch Language)*
Boerderij Regionaal (Dutch language)*
Boerderij Vandaag
Boerderij Verdieping (Dutch Language)*
Brabants Dagblad
Business Biographical
Business Wire Nederlands (BW)

C

CASH (Dutch)*
CASH (French)**
CASH (French) FR*
Cobouw
Company Briefs - Gale Group (Intl)
CompanynewGroupe - Dutch
Corporate Immigration Law
Curatoren.nl
**D**

Dag*
Dagblad De Limburger
Dagblad De Limburger (PL)
Dagblad De Pers*
Dagblad Rivierenland*
Dagblad van het Noorden
de Volkskrant 16:00*
de Gelderlander
de Gooi- en Eemlander
de Groene Amsterdammer
de Stentor
De Stentor/De Apeldoornse Courant
De Stentor/Dagblad Flevoland
De Stentor/Deventer Dagblad
De Stentor/Gelders Dagblad
De Stentor/Nieuw Kamper Dagblad
De Stentor/Overijssels Dagblad*
De Stentor/Sallands Dagblad
De Stentor/Veluws Dagblad
De Stentor/Zutphens Dagblad
De Stentor/Zwolse Courant
de Telegraaf
de Twentsche Courant Tubantia
de Volkskrant
de Weekkrant
Distrifood
Duns Market Identifiers - Netherlands
Dutch Company Information
Dutch Company Information, Managers and Directors
Dutch Company Information Mergers & Acquisitions*
Dutch Company Information Ownership Structures
Dutch Company News Bites - Market Report*
Dutch Company News Bites - Results*
Dutch Company News Bites - Sector Report*
The Dutch Economy*

**E**

Economist Intelligence Unit (EIU) Country Commerce
Economist Intelligence Unit (EIU) Country Forecasts
Economist Intelligence Unit (EIU) Country Reports
Economist Intelligence Unit (EIU) Country Risk Service
Economist Intelligence Unit (EIU) Executive Briefings
Economist Intelligence Unit (EIU) Financial Services Report
Economist Intelligence Unit (EIU) Risk Briefing
Economist Intelligence Unit (EIU) ViewsWire
Eindhovens Dagblad
Elsevier
Elsevier Juist
Elsevier Thema (Dutch Language)*
F
Factiva Nieuwsoverzicht (NL)*
FD.nl
Federal News Service
Feed Mix (English)*
Feed Tech (English)*
Feiten & Cijfers
FEM Business & Finance*
Het Financieele Dagblad
Het Financieele Dagblad (English)*
FlowerTech (English)
Forum
Fruit & Veg Tech (English)*

G
GlobalAdSource (Netherlands)
Goudsche Courant*
Groenten en Fruit

H
Haagsche Courant*
Haarlems Dagblad
Hoover's Company Records - Basic Record
Hoover's Company Records - In-depth Records

I
IAC (SM) Company Intelligence (R) - International
IJmuider Courant
Instant ID International Netherlands
International Institutional Database
I&PN Investment and Pensions Nederland*

J
J/M*

K, L
Leeuwarder Courant
Leidsch Dagblad
LexisNexis® Corporate Affiliations(TM)
Limburgs Dagblad
Limburgs Dagblad (PL)
Logistiek

M
Market Guide Company Profiles - Netherlands
MD Business News*
Meat International (English)*
Media Digitaal Info (Abstracts)
Metro (NL)
Middle East Newsfile (Moneyclips)*
Mining Journal*
Misset Café (Dutch Language)*
Misset Catering
Misset Horeca
Misset Hotel (Dutch Language)
Misset Restaurant (Dutch Language)*
Money(Dutch)*
MTI Econews

N
National Journal Daily Extra PM
Nederlands Dagblad
News Bites - Benelux
News Bites - Benelux: Netherlands
NeXT!
Nieuwsblad Transport
Noordhollands Dagblad
NOVUM
NRC Handelsblad
NRC.NEXT

O
Offshore
Oil & Gas Journal
Opzij*

P
Pakblad*
Het Parool
Pensions and Investments
Pig Progress (English)
Platts NuclearFuel
Platts Nucleonics Week
PLATTS NORTH SEA LETTER*
Pluimveehouderij
Poultry Processing Magazine (English)*
Property Casualty 360 - National Underwriter
Provinciale Zeeuwse Courant
The PRS Group International Country Risk Guide
The PRS Group Political Risk Service
Psychologie Magazine

Q
Quote
Quotenet
Quotenet.nl (Dutch Language)

R
RDS Business & Industry Selected Documents
RDS Business and Management Practices - Selected Documents
Rechtspraak.nl
Reed Business Feeds
Reformatorisch Dagblad
Resolution of International Boundary Disputes Involving Quasi-States

S
SeeNews Netherlands
South China Morning Post
Spits*

T
Tenders Electronic Daily
Tijdschrift voor de Politie
Transport & Opslag (Archive)*
Trekker
Trouw

U
Utrechts Nieuwsblad*

V
Vakblad AGF (Dutch Language)*
Volkskrant Banen*
Vrij Nederland

W
WebNews - Dutch
World Poultry (English)
Worldscope-International Company Profiles

X - Z
Zorg + Welzijn
Zorgvisie