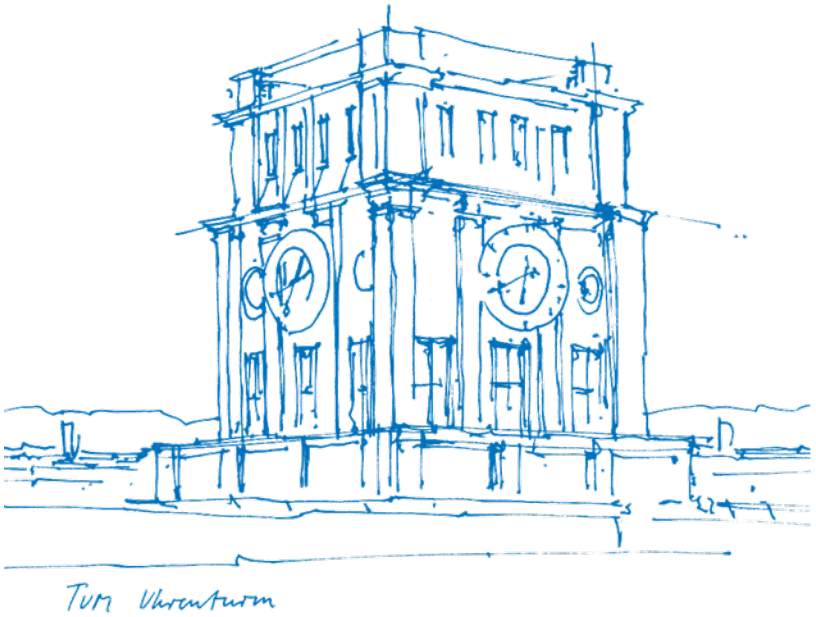
Clustering-Based Sentiment Analysis for Media Agenda Setting

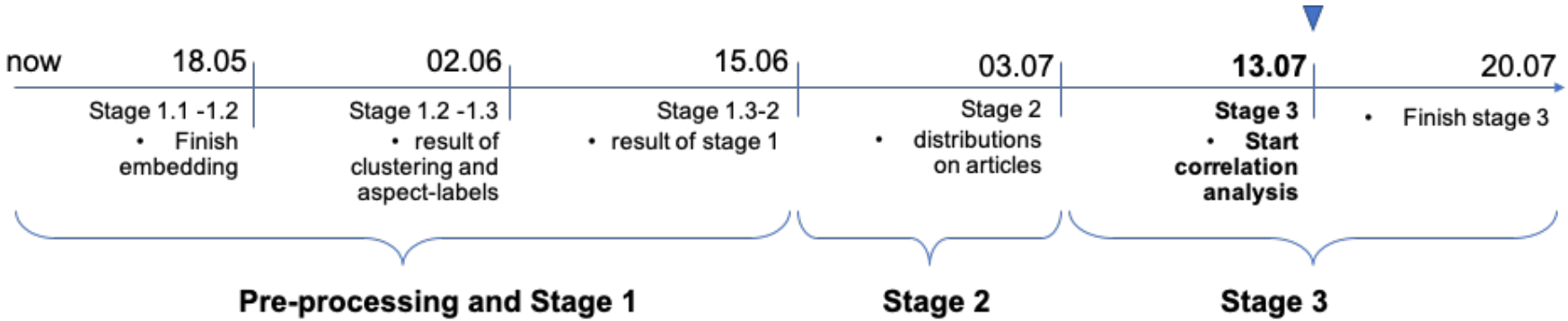
Opinion Lab Group 2.3, presentation 6

Wing Sheung Leung, Qiaoxi Liu

July 13, 2020



Milestones


now
18.05
Stage 1.1 -1.2
• Finish embedding
02.06
Stage 1.2 -1.3
• result of clustering and aspect-labels
15.06
Stage 1.3-2
• result of stage 1
03.07
Stage 2
• distributions on articles
13.07
Stage 3
• Start correlation analysis
20.07
• Finish stage 3
Pre-processing and Stage 1
Stage 2
Stage 3

Overview

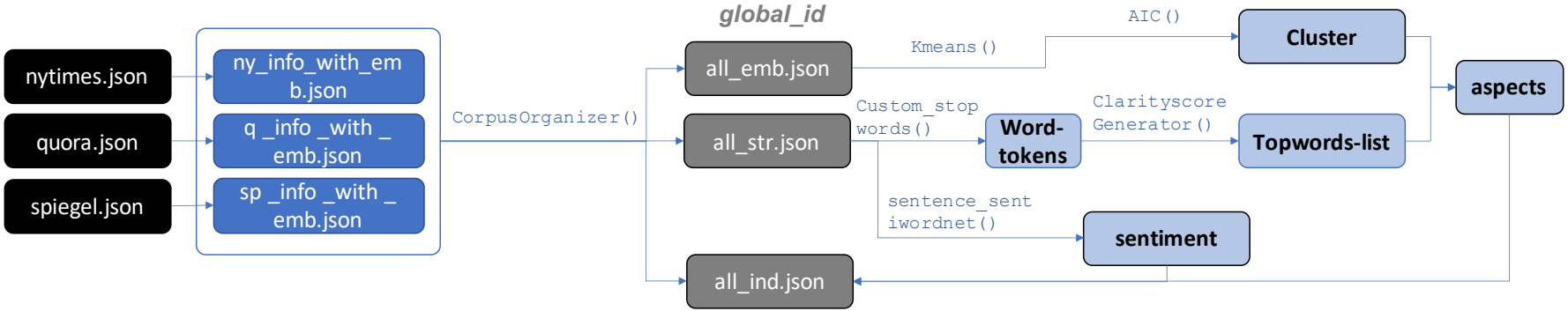
Stage 1: Generate sentence embeddings with our corpus

Stage 2: Distribution

Stage 3 Correlation Analysis

Top words

Output from stage 1



	global_id	corpus_name	doc_id	com_id	date	cluster	sentiment
0	0	nytimes	0	NaN	2005-11-01	7	0.12500
1	1	nytimes	0	NaN	2005-11-01	7	-0.12500
2	2	nytimes	0	NaN	2005-11-01	9	0.07500
3	3	nytimes	0	NaN	2005-11-01	7	0.50000
4	4	nytimes	0	NaN	2005-11-01	7	-0.03125

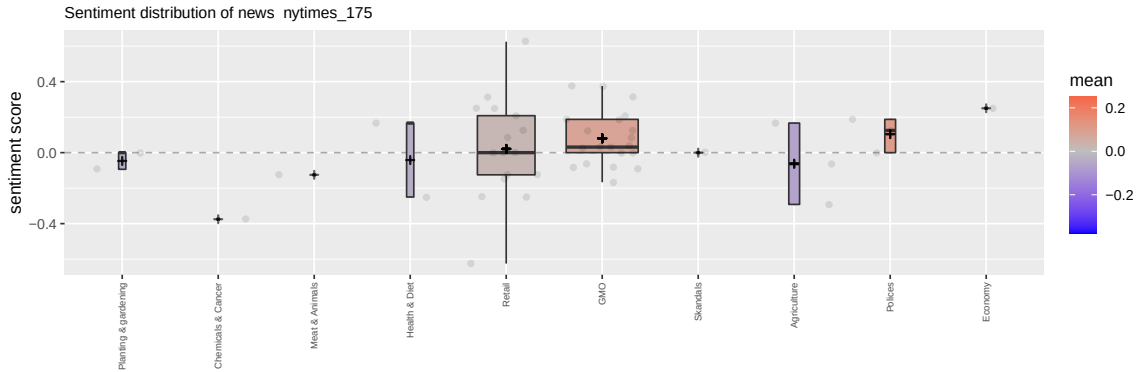
Stage 1: Generate sentence embeddings with our corpus

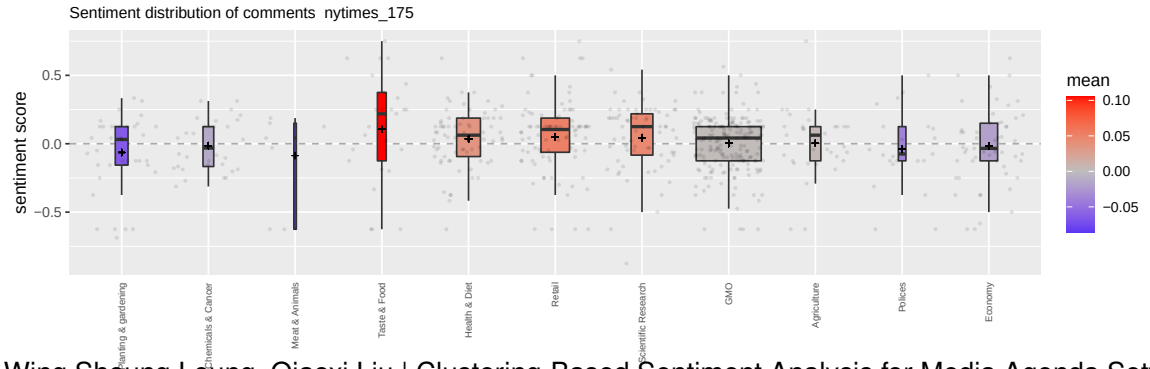
Stage 2: Distribution

Stage 3 Correlation Analysis

Top words

NYTimes: Labeling GMO





NYTimes: Labeling GMO

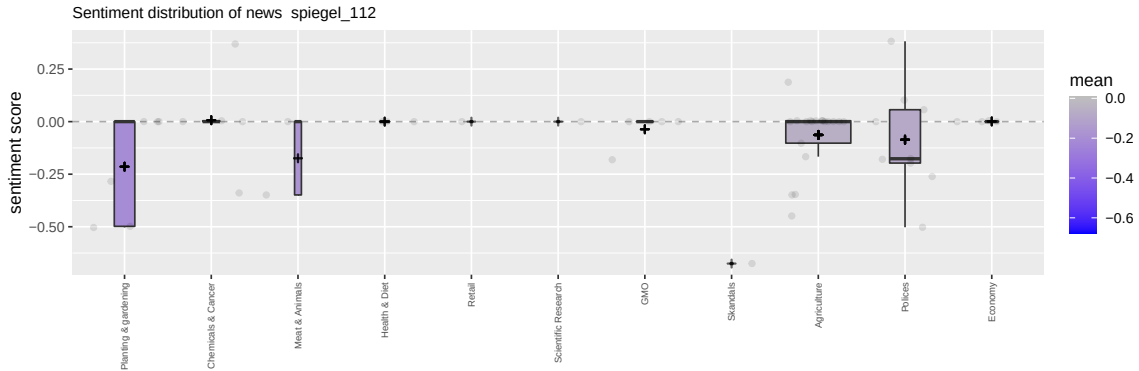
article:

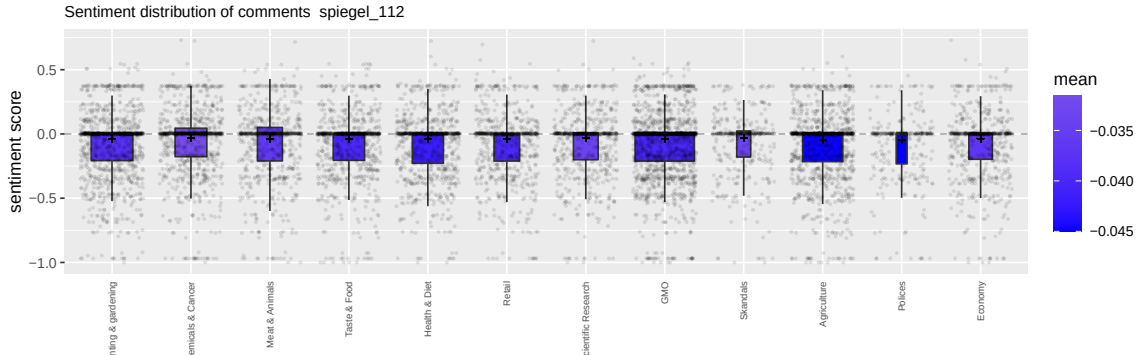
Whole Foods Market, the grocery chain, on Friday became the first retailer in the United States to require labeling of all genetically modified foods sold in its stores, a move that some experts said could radically alter the food industry. ... Whole Foods, which specializes in organic products, tends to be favored by those types of consumers, and it enjoys strong sales of its private-label products, whose composition it controls. ... He said Whole Foods looked forward to working with suppliers on the labeling.

comments:

- Sometimes genetically modifying food has unintended consequences like bad tasting tomatoes (or worse)
- I have a cousin who works at Whole Foods. He is a **happy** employee and loves it. Thinks it is a great company.
- I am not sure if non-GMO foods are **healthier** to eat but they are certainly better for the environment.
- I **applaud** Whole Foods for at least taking a stand.
- consuming red meat is an emotionally charged issue for many people.
- The conclusion: "Red meat consumption is associated with an **increased risk** of total, heart, and cancer mortality"
- Since apples are apparently the most pesticide-ridden fruit, I have gotten to like the more expensive but sweeter ones.

Spiegel: Der Skandal um Dioxin





Spiegel: Der Skandal um Dioxin

article:

Es ist einer der größten Giftskandale der vergangenen Jahre: Bis zu 3000 Tonnen dioxinverseuchtes Fett wurden laut Bundeslandwirtschaftsministerium an 25 Futtermittelhersteller in mindestens vier Bundesländern geliefert. Wo das Gift von dort aus hingelangte und welche Mengen an Nahrungsmitteln belastet sind, ist weitgehend unklar. Verbraucher reagieren zunehmend verunsichert: Der Verkauf von Hühnereiern ist "spürbar" gesunken, teilte die landwirtschaftliche Marktberichterstattungsstelle MEG mit. Welche Gefahren drohen durch die Einnahme von Dioxin? Welche Vorsichtsmaßnahmen können getroffen werden? SPIEGEL ONLINE gibt Antworten auf sieben Fragen.

comments:

- Der Dioxin-Skandal war **hoffentlich nicht der letzte**. Es sollten so viele wie möglich vorkommen. **Am besten aber wäre, wenn ein paar Konsumenten nachweislich an solchen oder anderen Giftstoffen in Lebensmitteln sterben.**
- 3000 Tonnen verseuchtes Tierfutter - das ist ein Terroranschlag.
- Ich kann und will niemandem verbieten Fleisch von deutschen Rindern zu essen.
- den gesetzlichen Vorgaben ist so eine Sache. **Fahren Sie mal mit einem Fiat 500 mit 80 km/h frontal gegen eine S-Klasse!**. Ihr Vergleich hinkt doch wohl. **Wer sich mit Bio-Artikeln überfrisst, stirbt auch. Also sind Bio-Lebensmittel auch lebensgefährlich, wenn man falsch damit umgeht. Guten Appetit.**

Observed Features

- Spiegel
  - news: Kritical (suspection)
  - comments: **Sarkasmus, irony**, indirect, spreading among more topics
- NYtimes
  - news: Tend to be neutral or positive (like an advertisment for stakeholder)
  - comments: **straightforward**, clear statement (against or for), benefit or not

Stage 1: Generate sentence embeddings with our corpus

Stage 2: Distribution

Stage 3 Correlation Analysis

- 3.1 Global Sentiment distribution on NYTimes and Spiegel
- 3.2 Visualization of Correlation of sentiment on NYTimes and Spiegel

Top words

Stage 3: Distribution of articles and comments Overview

Comments from NYtimes and Spiegel

NYtimes
99 out of 327 has comments.
Sentences of comments from each news:

- minmax = (3, 2535)
- mean = 481.7
- median= 216.0

Spiegel
61 out of 152 has comments.
Sentences of comments from each news:

- minmax=(8, 23255)
- mean = 2309.0
- median= 697.0

Our distribution is based on...

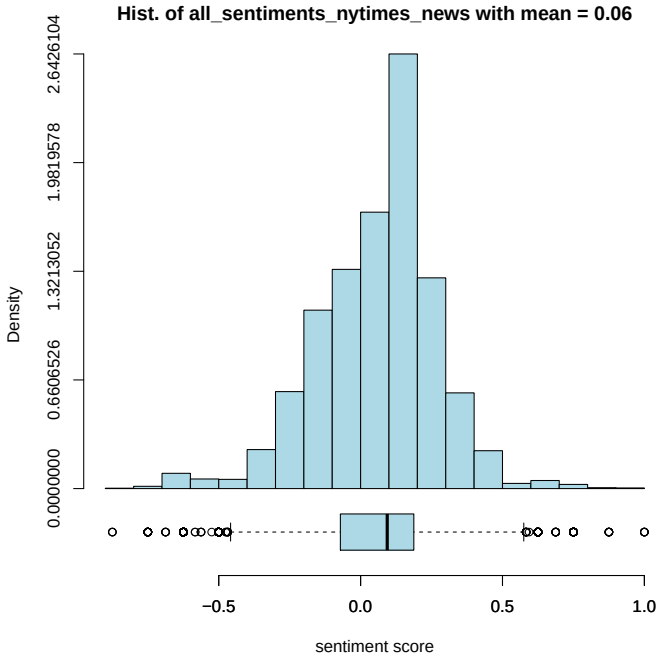
To analyse the correlations between news and its comments (sentences-wise), we select the news which has comments.

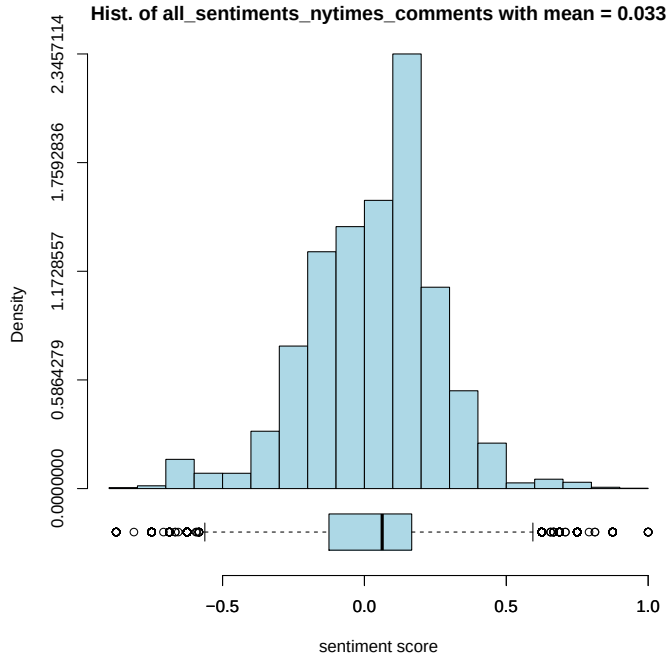
We filter out the sentences belonging to garbage-aspects.

- NYtimes: We pick up all 99 news (14660 sentences) with its comments (47691 sentences)
- Spiegel: We pick up all 61 news (5689 sentences) with its comments (140855 sentences)

3.1.1 Global Sentiment Score Distribution on NYTimes

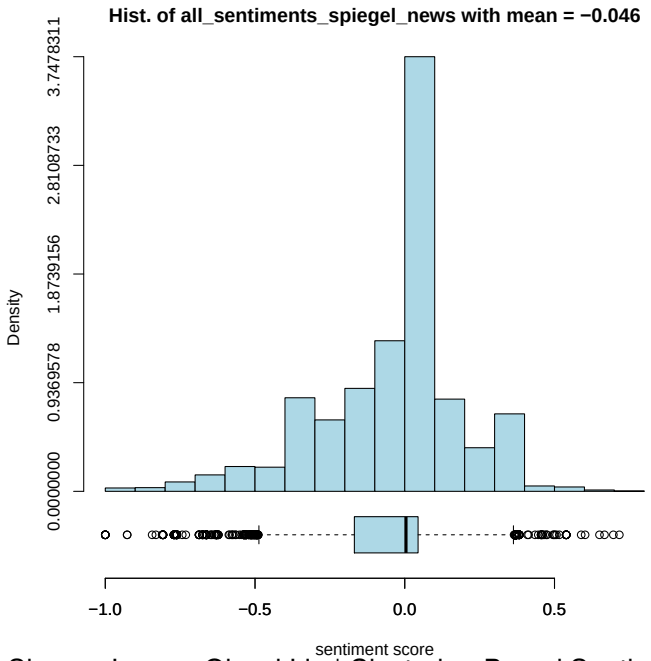
- mean(news) = 0.06 > mean(comments) = 0.033

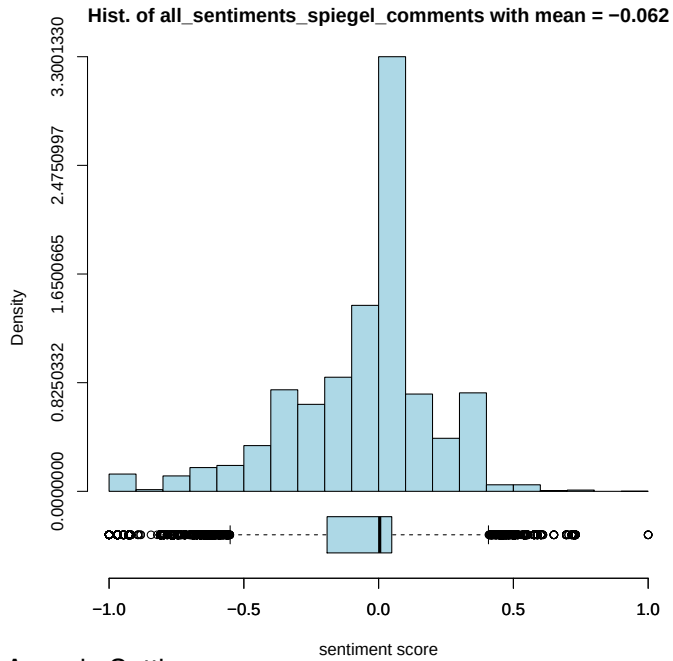




3.1.1 Global Sentiment Score Distribution on Spiegel

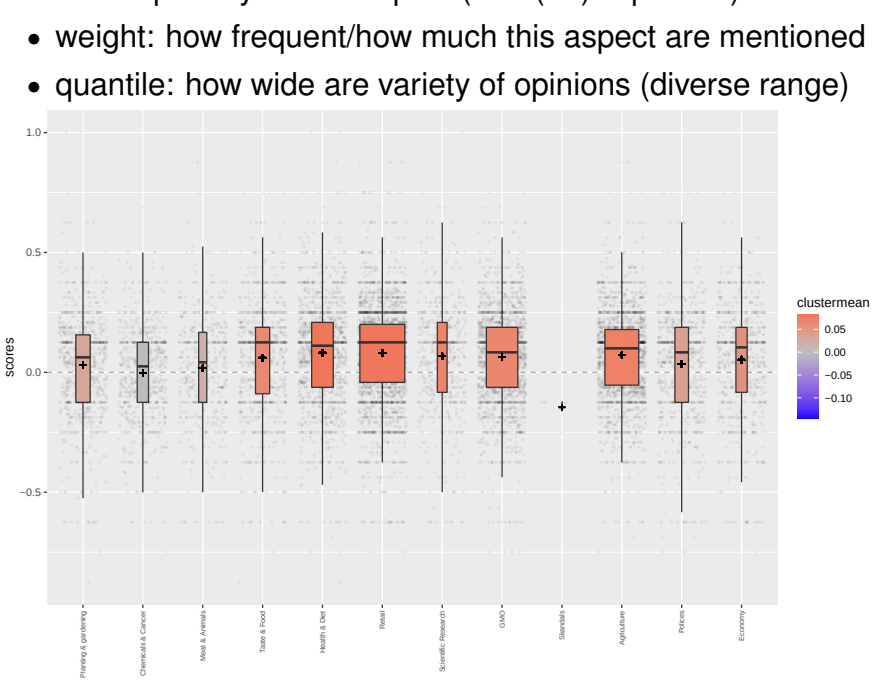
- mean(news) = -0.046 > mean(comments) = -0.062

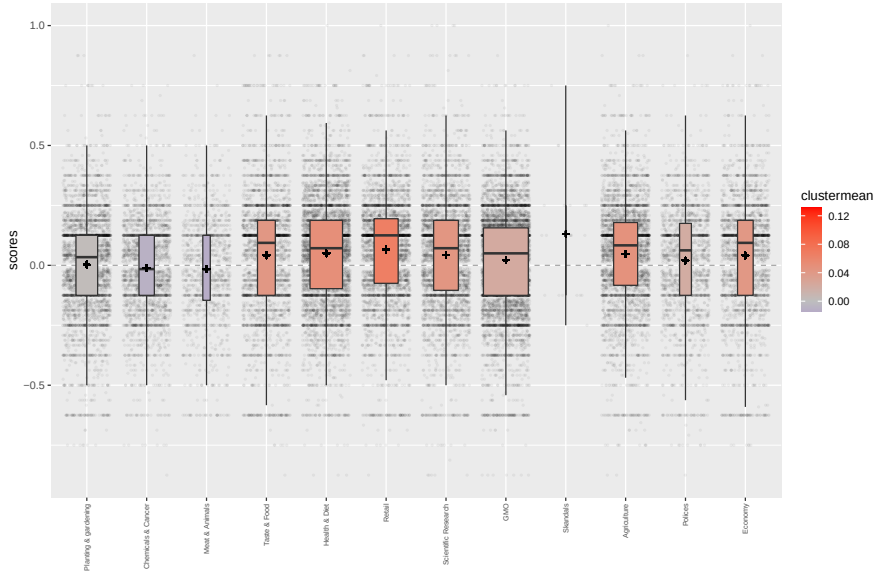




3.1.2 Global Sentiment distribution per topic on NYTimes

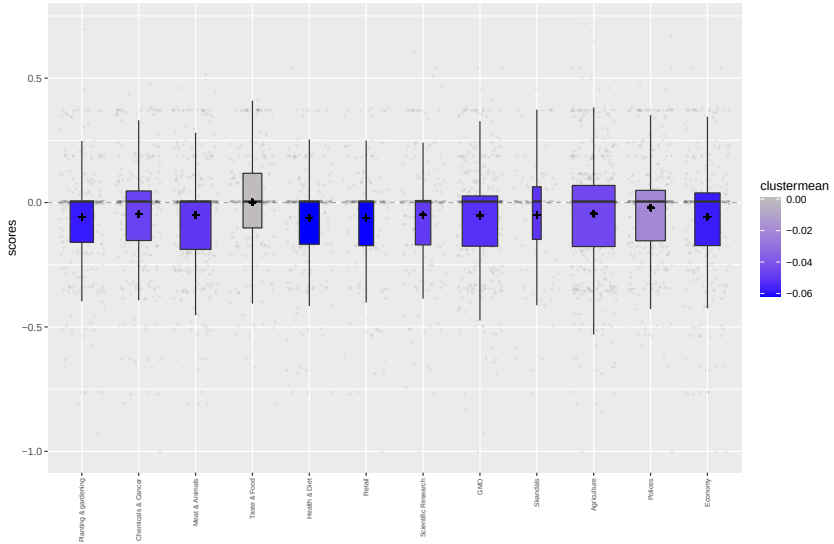
- mean: polarity of one aspect (red (>0) is postive)
- weight: how frequent/how much this aspect are mentioned (all weights sum up to 1 in one article)
- quantile: how wide are variety of opinions (diverse range)

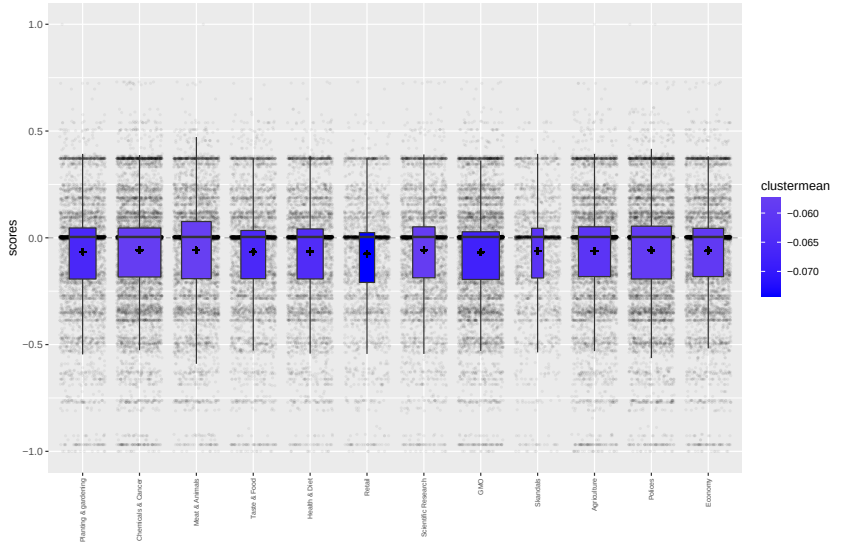




3.1.2 Global Sentiment distribution per topic on Spiegel

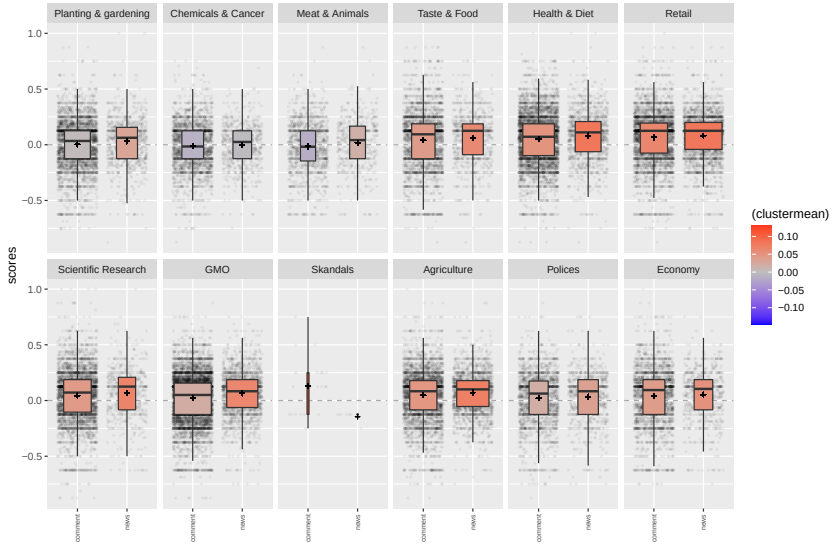
- news: left, comments: right

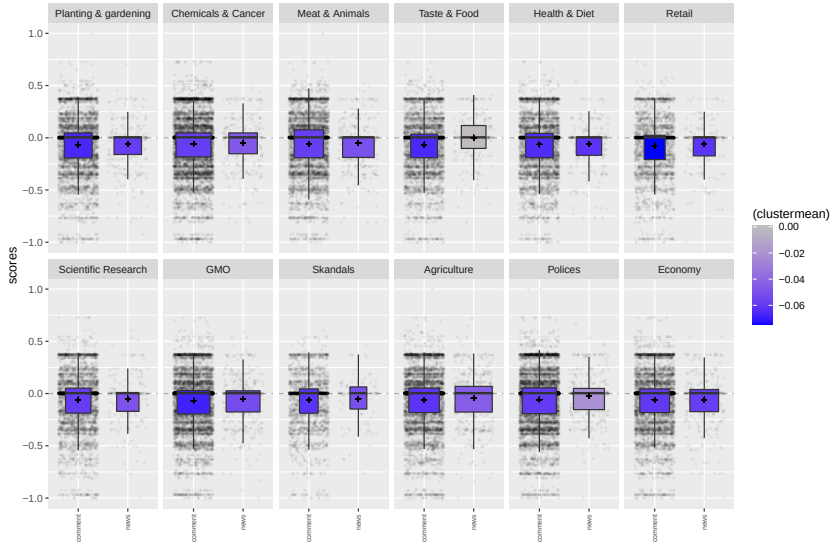




3.1.3 Relation between articles and comments sentiment

- nytimes: left, spiegel: right





Stage 1: Generate sentence embeddings with our corpus

Stage 2: Distribution

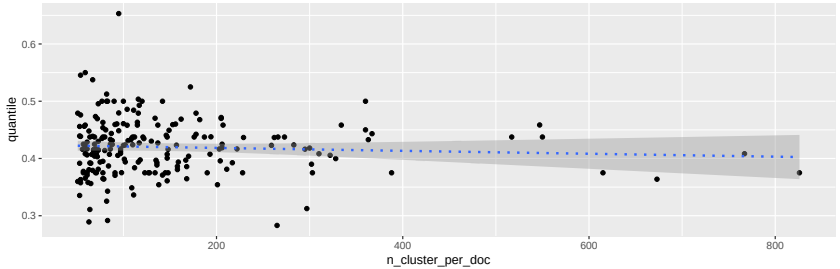
Stage 3 Correlation Analysis

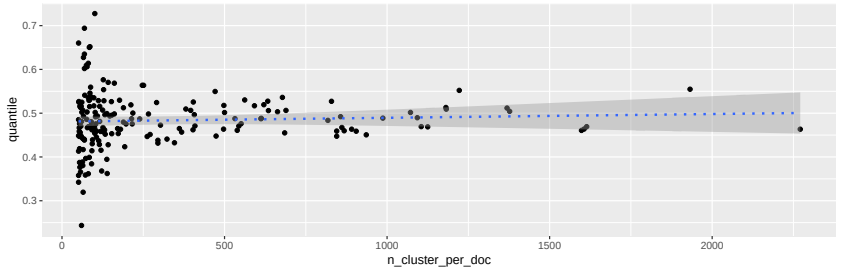
- 3.1 Global Sentiment distribution on NYTimes and Spiegel
- 3.2 Visualization of Correlation of sentiment on NYTimes and Spiegel

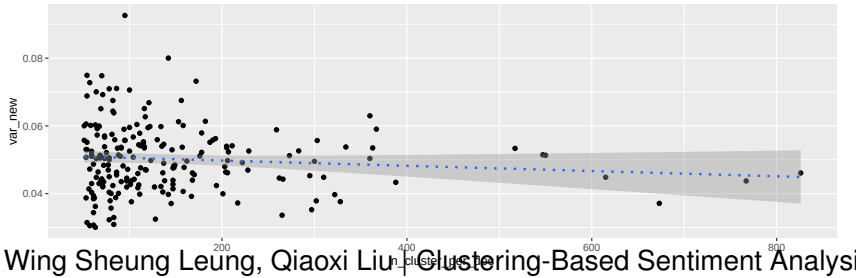
Top words

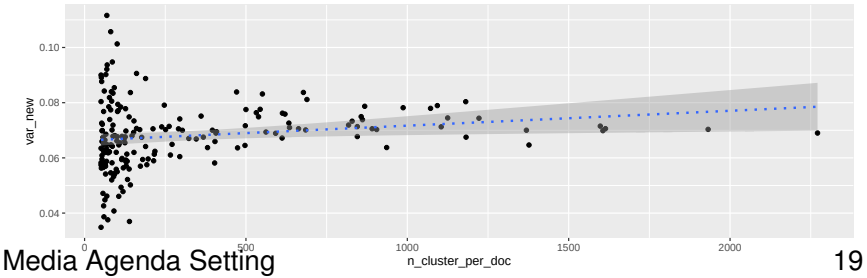
3.2.1 Global Relation between number of sentences with sentiment (15%-85%)quantile/variance

- obersavation: quantile/var varies for fewer comments, as the num of sentences of comments increase, it converges.
- nytimes (left) : there is no significant relation between the num of sentences and its quantile/var
- spiegel (right) : slightly getting larger when num of sentences grows



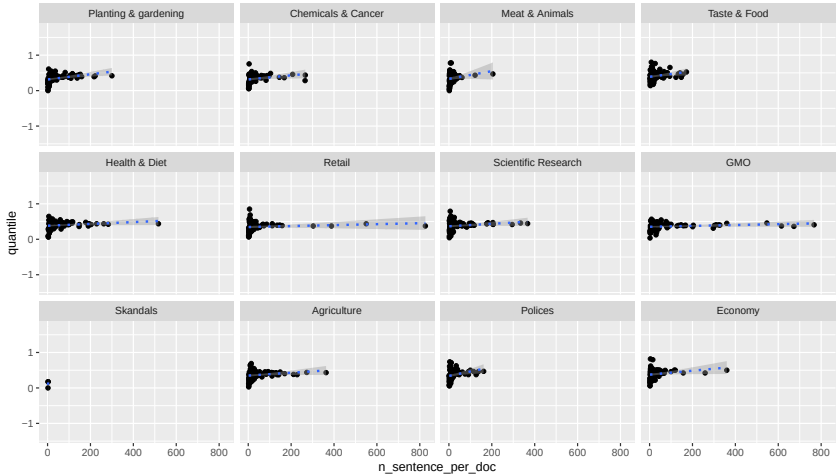


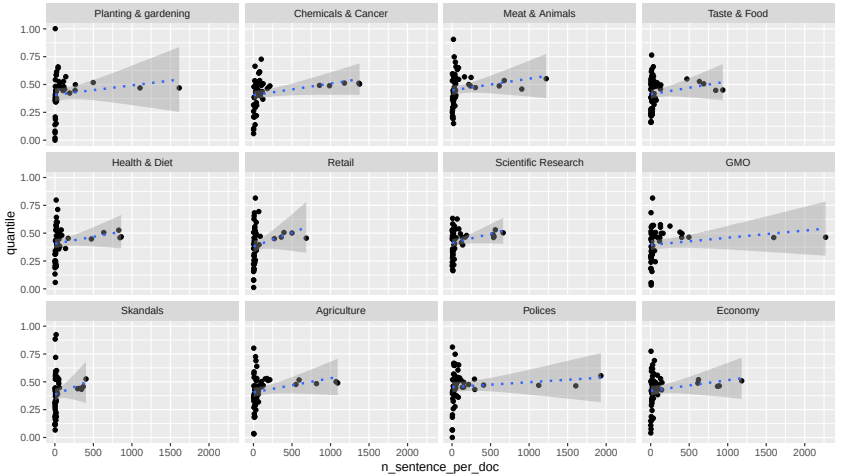




3.2.1 Relation between number of sentences with sentiment quantile per topic

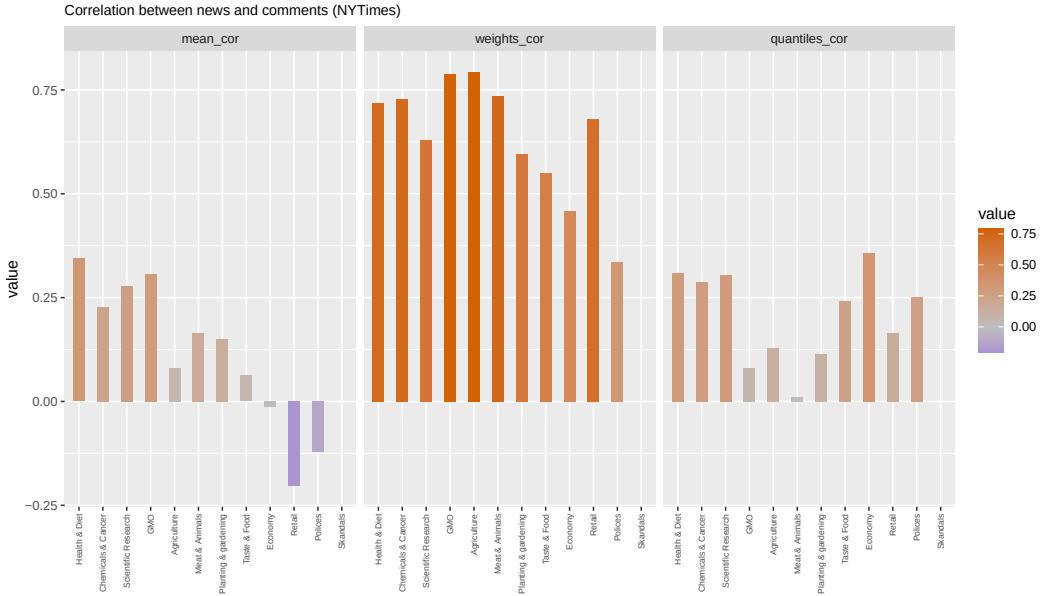
- obersavation: some topics appear only in shorter comments
- nytimes (left) per topic: very weak pos-correlation between the num of sentences and its quantile
- spiegel (right) per topic: slightly getting larger when num of sentences grows





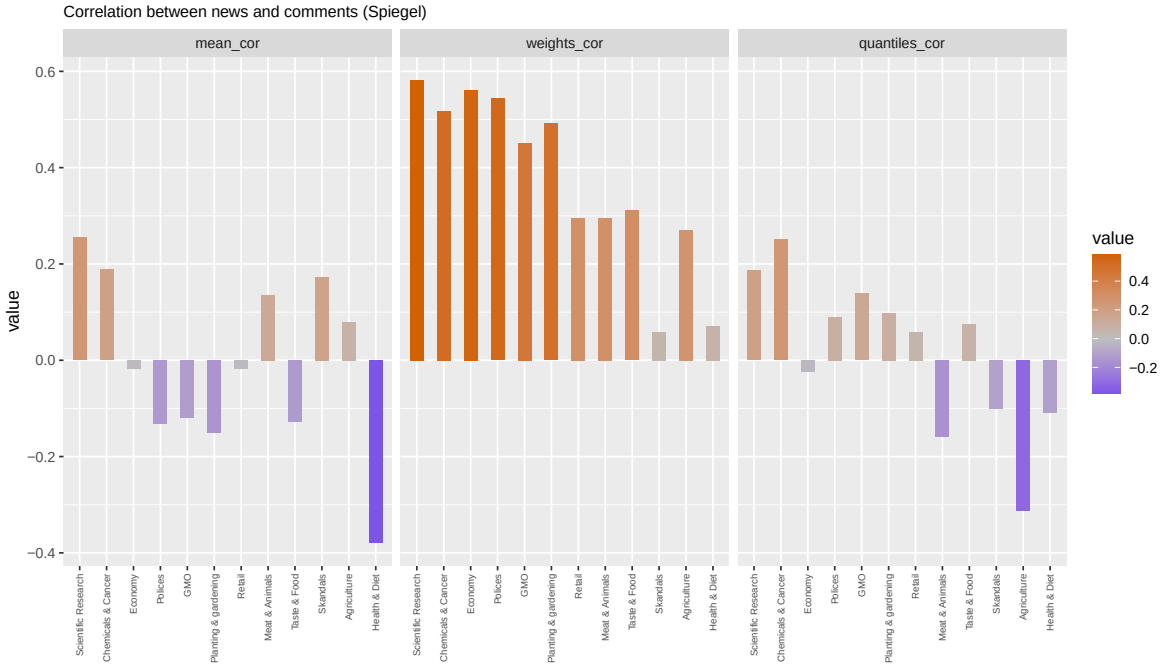
3.2.2 Correlation between NYtimes news and its comments

Variables are means(normally distributed)/weights/quantile .
Apply cor.test(var_new,var_comments). Value is its correlation coefficient(pearson,spearman)



- **means** pos correlated: ☺ → ☺ otherwise ☺ → ☹
- **weights** also (highly) pos correlated: the topics in news remain discussed mainly
- **quantile** pos correlated : greater/smaller in news → greater/smaller in comments

3.2.2 Correlation between Spiegel news and comments



- **means** more topics neg correlated ☺ → ☹ than NYtimes
- **weights** (highly) pos correlated: the topics in news remain discussed mainly. Same as NYtimes
- **quantile** more topics neg correlated : small range variety in news → greater (fierce discussion) in comments

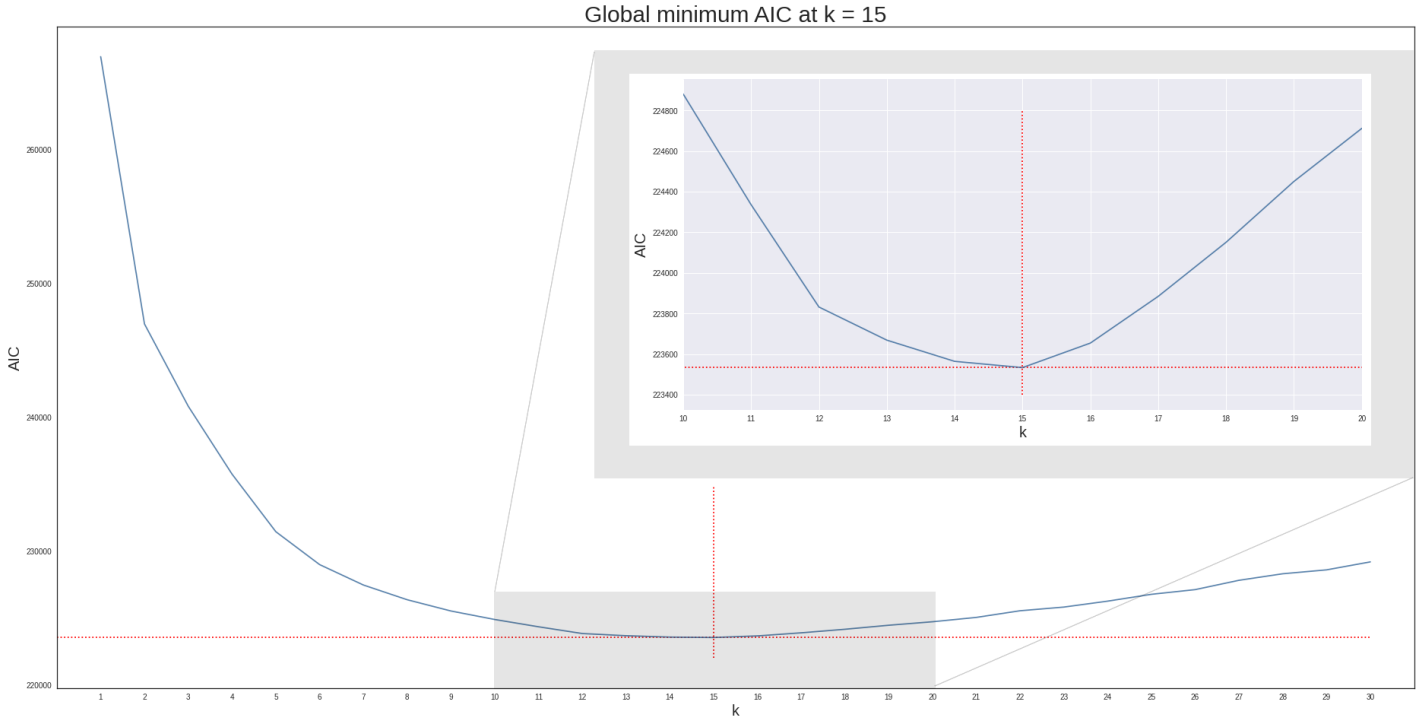
Tackled issue on top words selection

- Problems related to clarity score:
  1. Imbalanced existence of English and German words from original top words list
  2. Incoherent top words among English and German corpus
- Solutions:
  1. Pre-processed tokenized sentences
     - Replaced URLs with string 'url'
     - Ignored sentences with length smaller than 15
  2. Re-generated sentence embeddings for pre-processed sentences with XLING
  3. Re-run kmean clustering with the new sentences embeddings
  4. Revised clarity score calculation

```
vectorizer = TfidfVectorizer(stop_words = stopwords, norm = 'l1')
corpus_bag_of_word = vectorizer.vocabulary_
corpus_idf = vectorizer.fit(corpus_sentences)
corpus_tfidf = vectorizer.transform([' '.join(corpus_sentences)]) # i.e. t(w)

# for each cluster
cluster_vectorizer = TfidfVectorizer(stop_words = stopwords, norm = 'l1',
                                     vocabulary = corpus_bag_of_word)
cluster_idf = cluster_vectorizer.fit(cluster_sentences)
cluster_tfidf = cluster_vectorizer.transform([' '.join(cluster_sentences)]) # i.e. t_a(w)
```

k = 15 is optimal from AIC plot



English and German top words are coherent when k =15

Figure: English topwords (left) from NYTimes and Quora German top words (right) from Die Spiegel