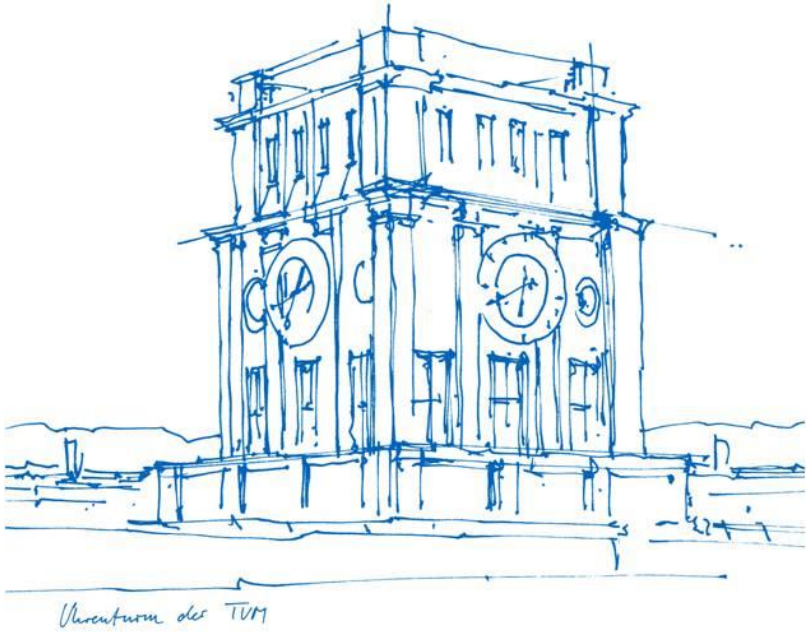
Technical University of Munich

Clustering-Based Sentiment Analysis for Media Agenda Setting

Opinion Lab Group 2.3, Final Presentation

Wing Sheung Leung, Qiaoxi Liu

July 20, 2020



Motivation

Media Agenda Setting

- Idea of influence on media towards its readers on what issues to think about

> *'Readers learn not only about a given issue, but also how much importance to attach to that issue from the amount of information in a news story and its position.' (McCombs & Shaw, 1972)*

With organic food related articles from New York Times, Quora and Der Spiegel

- Aims to investigate the relationship between these news articles and their corresponding readers' responds on *different aspects or topics* on organic food by observing

  1. occurrences of discussions or comments
  2. degree of sentiment polarity, i.e. positivity or negativity

Data

Articles and readers' comments on the domain of organic food with search terms organic food and organic farming

- **Articles** from two online newspapers,
  New York Times (English) and Der Spiegel (German)

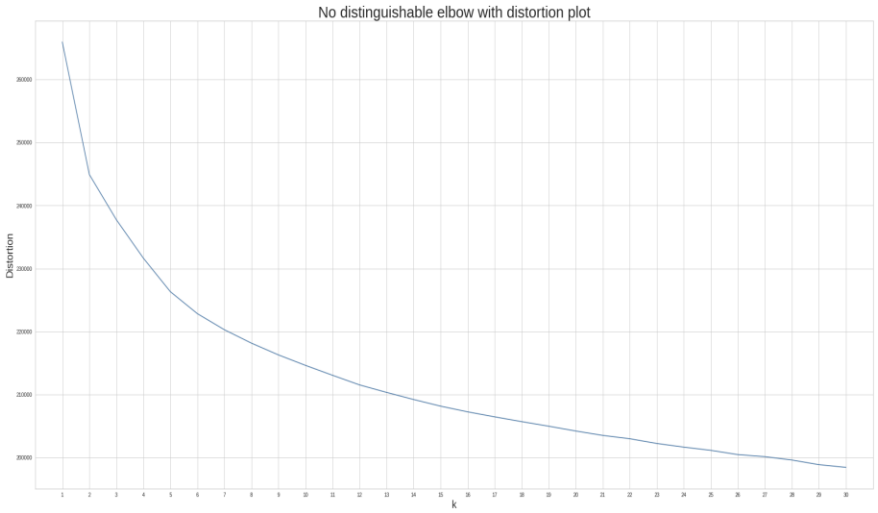
- **Direct response** (bilingual):
  comments right under those articles

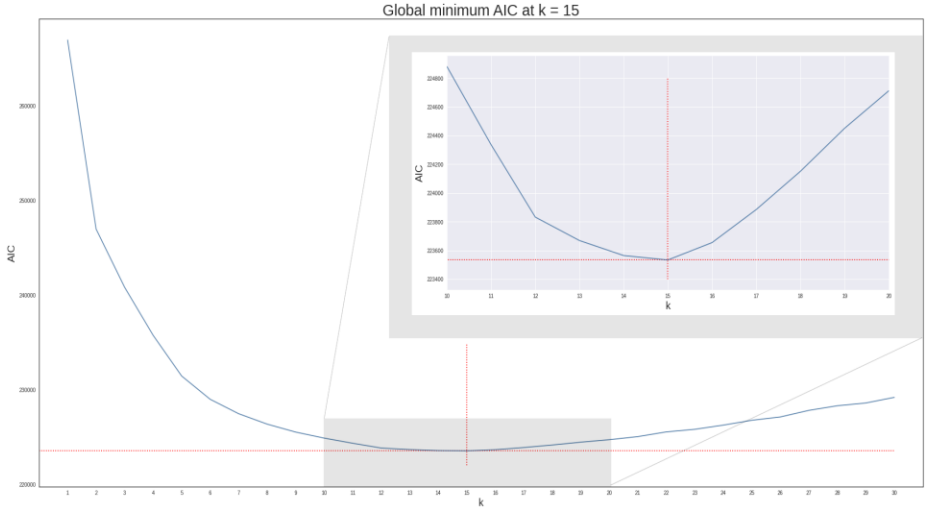
- **Indirect response**:
  posts in discussion forums, Quora

Experiment Pipeline

k-means clustering

- with sentences embedding (512-dim.) generated by XLING (Chidambaram et al., 2018)
- potential $k = \{13, 14, 15, 16\}$





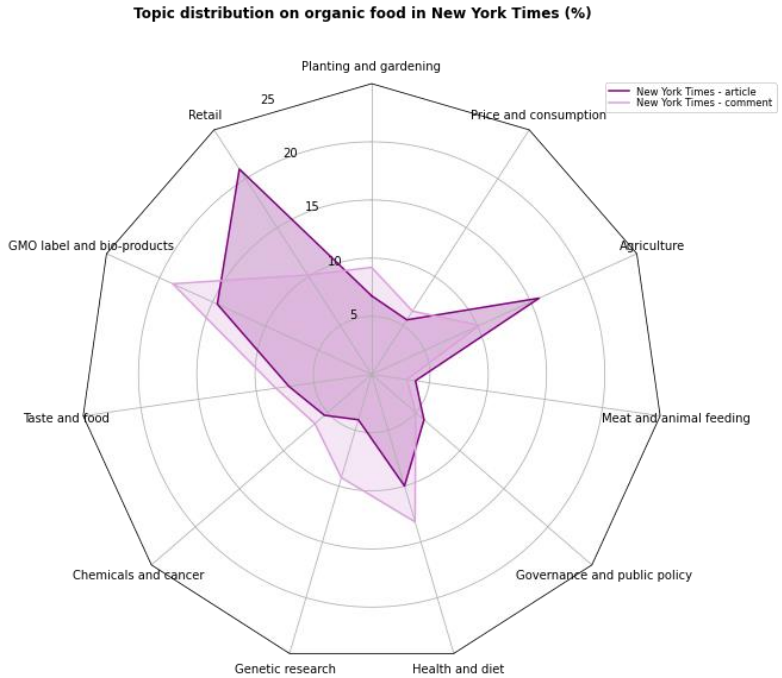
Top words

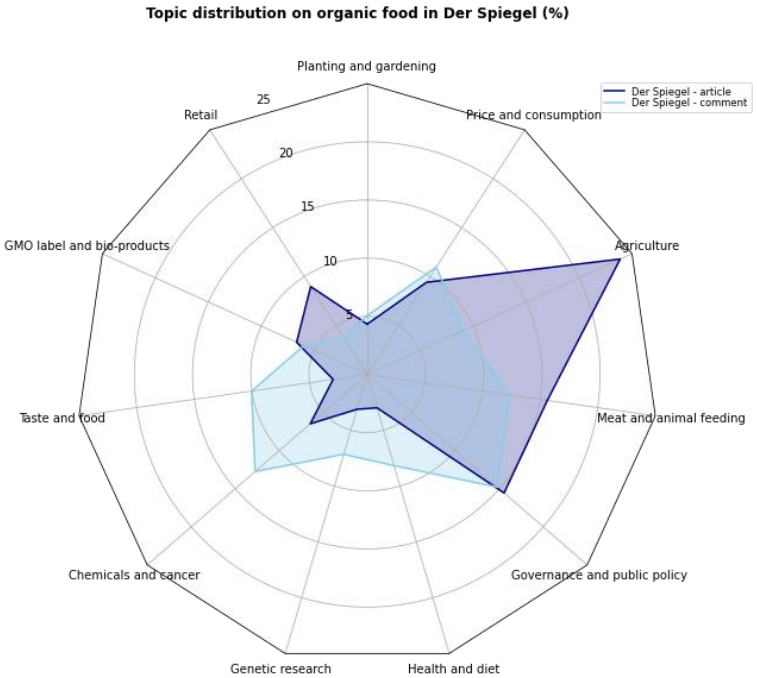
- Clarity scoring, (Angelidis & Lapata, 2018) $score_{s,a}(w) = t_{s,a}(w) log_2 \frac{t_{s,a}(w)}{t_s(w)}$

Word cloud for English corpus with stemmed words

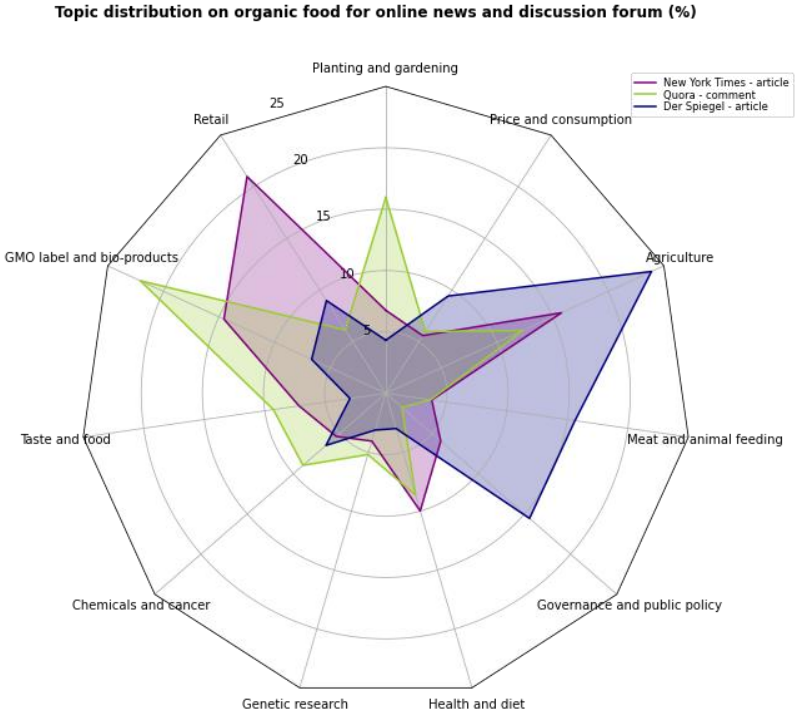
Word cloud for German corpus with stemmed words

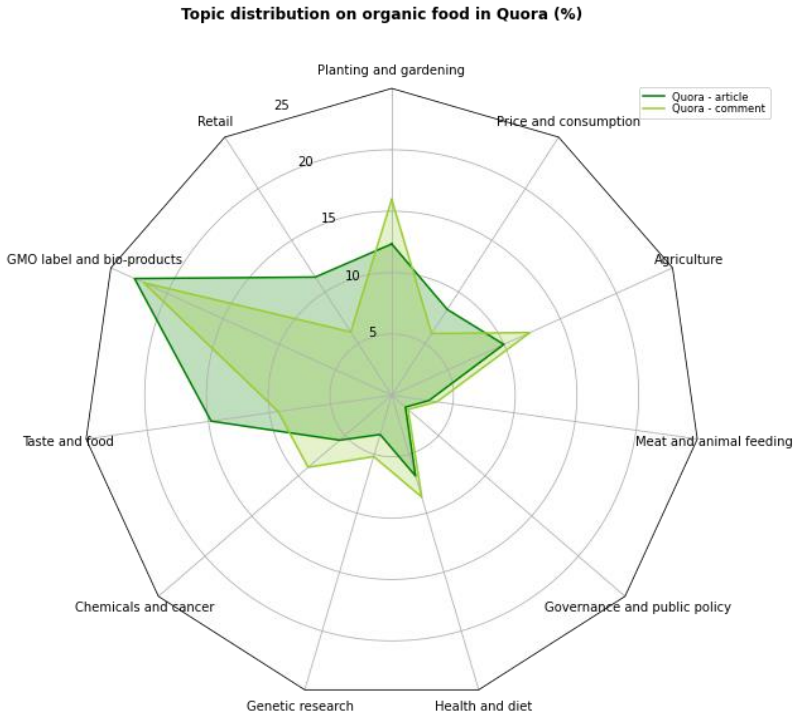
Mild evidence telling readers' discussion rate positively correlated to media's posting rate





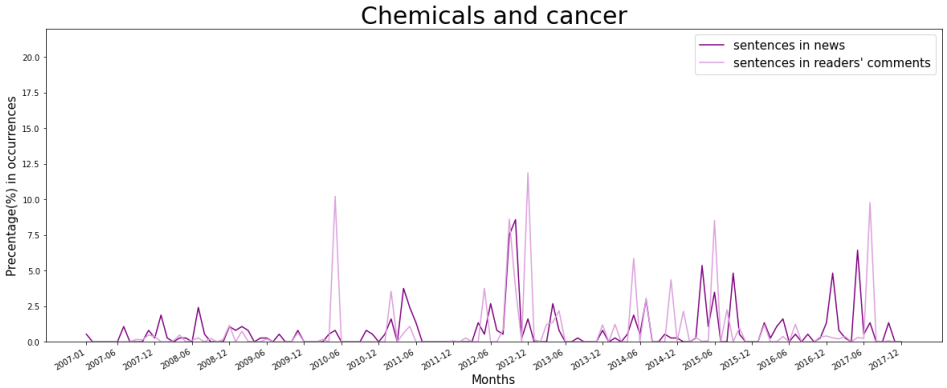
Cannot conclude on how media affect public to think

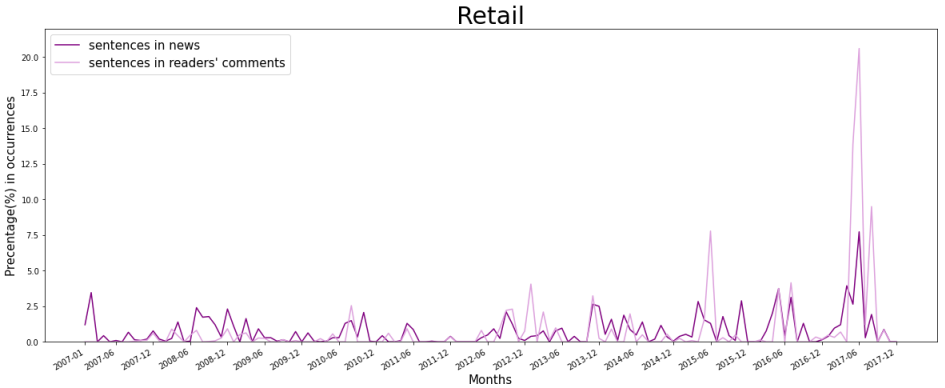




Mild evidence telling change of readers' discussion rate follows change of media's posting rate

New York Times





Sentiment Assignment on pre-trained model

Procedure

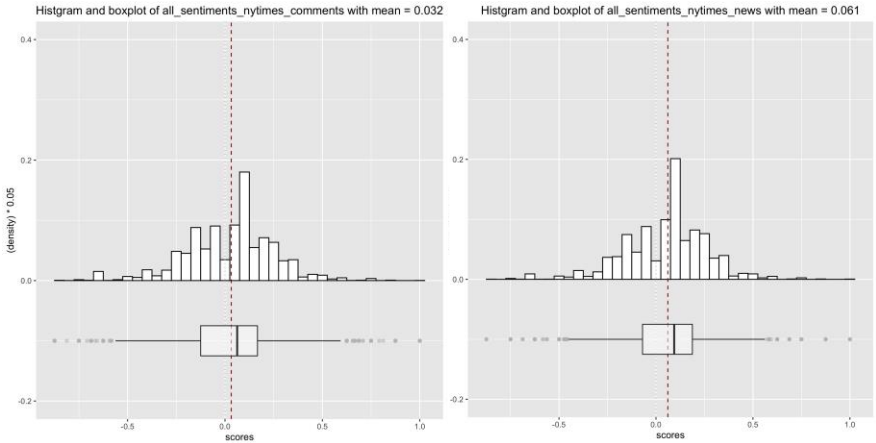
1. word-source from POS tagging: ADJ,NOUN,ADV,VERB.
2. computing score for each tagged word with Interval [-1, 1] based on:
   - English: synonym set to compute average sentiment score (SentiWordNet 3.0,Baccianella et al., 2010)
   - German: semantic orientation of pos/neg based on PMI (SentiWS, Remus et al., 2010)
3. averaging all token scores in a sentence to assign final score of this sentence

Example:

- "Like most grocery chains, Whole Foods does not release sales data on individual stores."('Like', 'IN'), ('most', 'JJS'), ('grocery', 'NN'), ('chains', 'NNS'), (',', ','), ('Whole', 'NNP'),...result: -0.625

Global View: Sentiment score distribution (density)

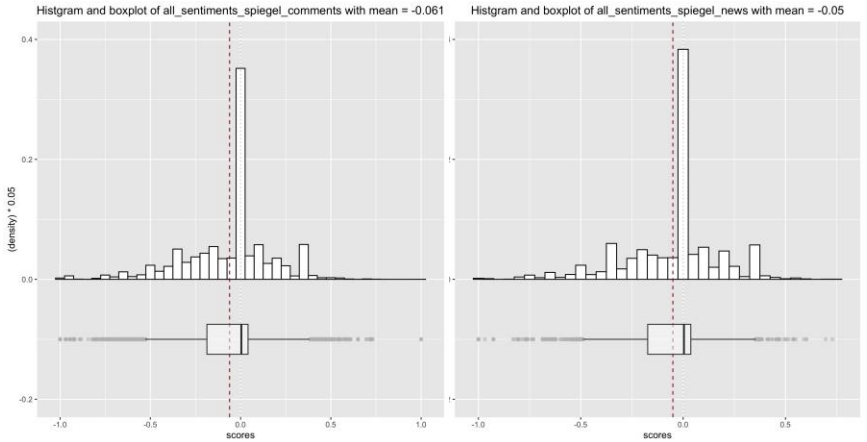
NYTimes: total 99 news



#S_comments : 45,597 Mean = 0.061

#S_news: 14,202 Mean = 0.032

Spiegel: total 61 news

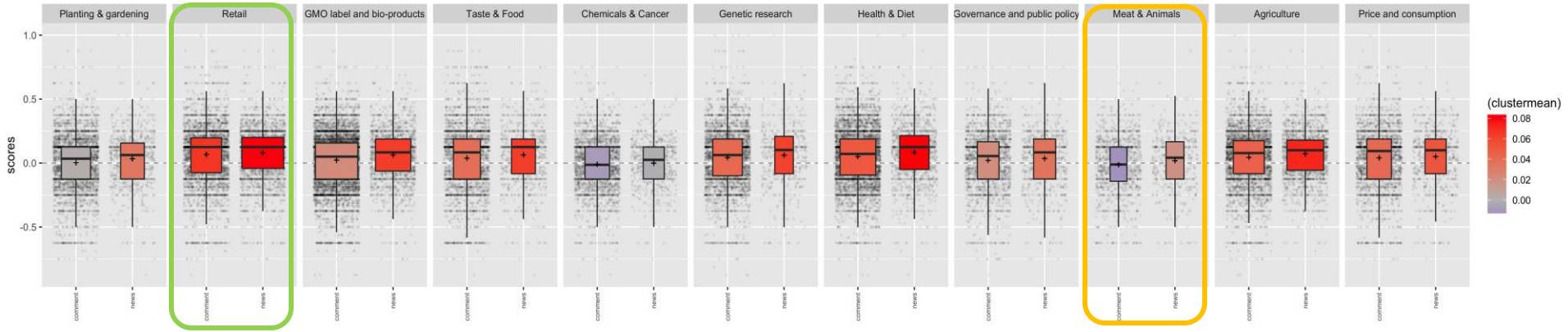


#S_comments:121,369 Mean = -0.061

#S_news: 6,795 Mean = -0.05

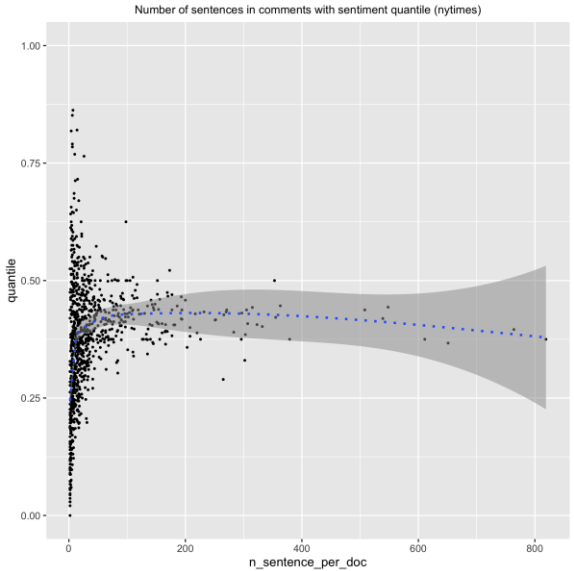
Global View: Four components to interpret (NYTimes)

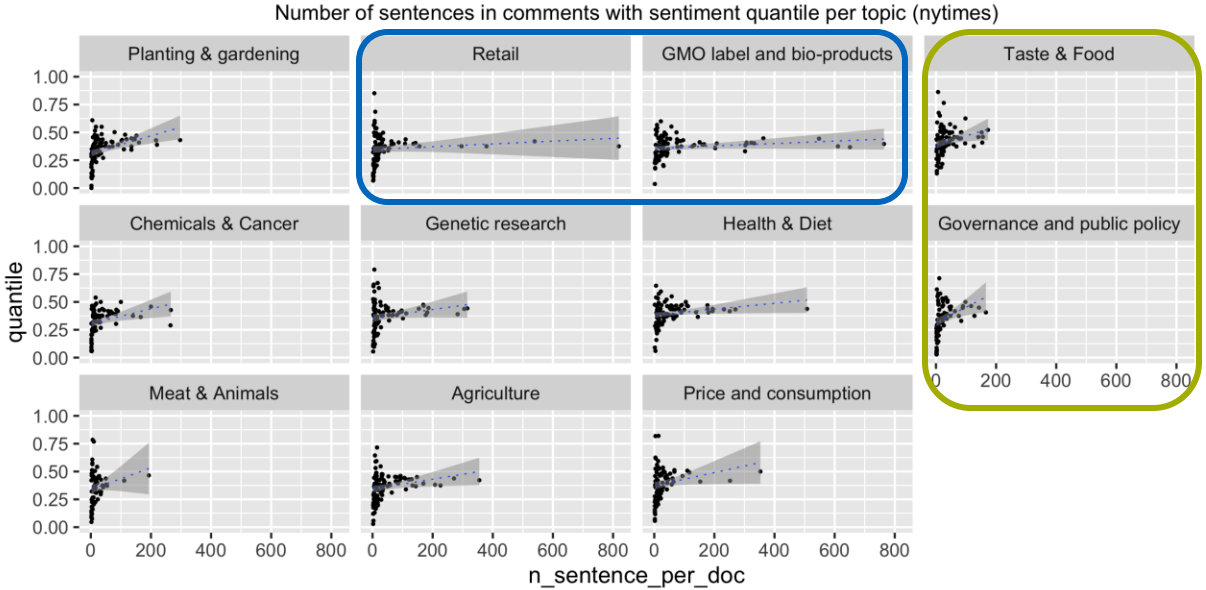
- "color" --> **polarity** of sentiments (per topic) --> **mean** of sentiment scores --> interval (-1,1)
- "width" --> **dominance** of a topic --> **weight** (num_s in one topic / total_num_s) in an article --> interval (0,1)
- "height" --> **fluctuation** of sentiments (per topic) --> **quantile** of sentiment scores --> interval (0,2)
- "area of the box" --> **hotness** of a topic --> dominance * fluctuation
- Some aspects to observe:
  - Ranking inside one corpus (I.e. nytimes news)--> the hottest is Retail / **fluctuation** is similar globally (**varies** in single article!)
  - Changes per topics --> (Meat & Animals) the **polarity** turns to negative ; (Retail) less **dominant**



Global View: Shorter comments, more fluctuate opinion?

NYtimes comments (998 points represent a mentioned topic in an new's comments. x-axis = len(comments) avg. 10 topics are mentioned in comments of one news)





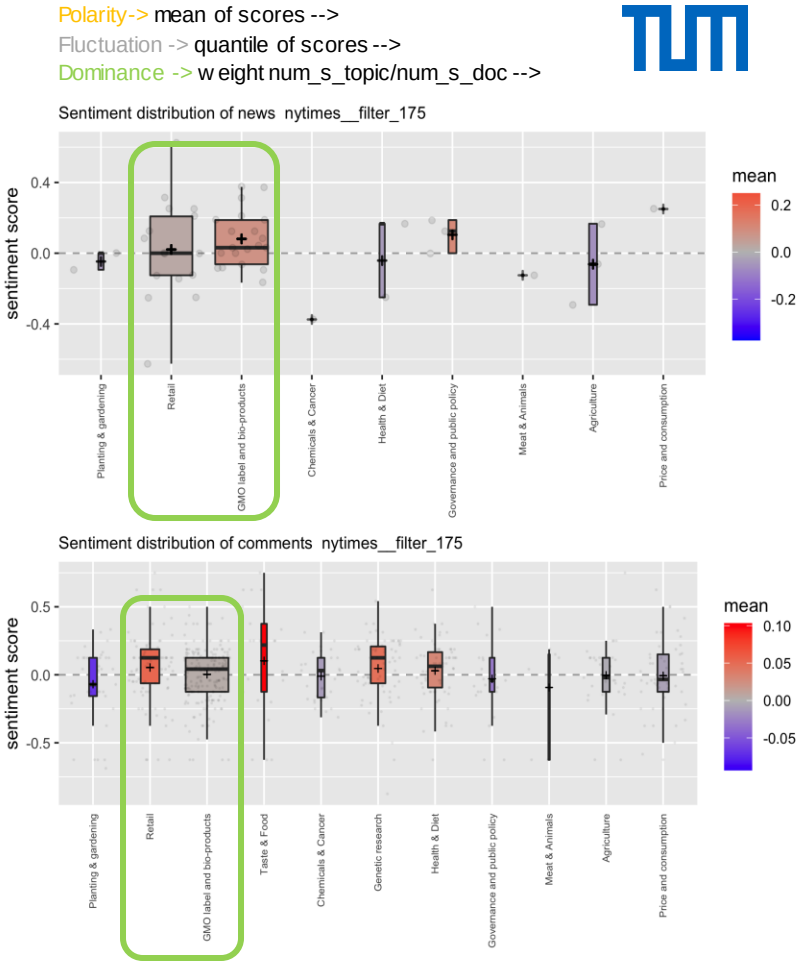
Individual View on NYTimes

article:

Whole Foods Market, 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. ... tends to be favored by those types of consumers, and it enjoys strong sales of its private-label products, ...

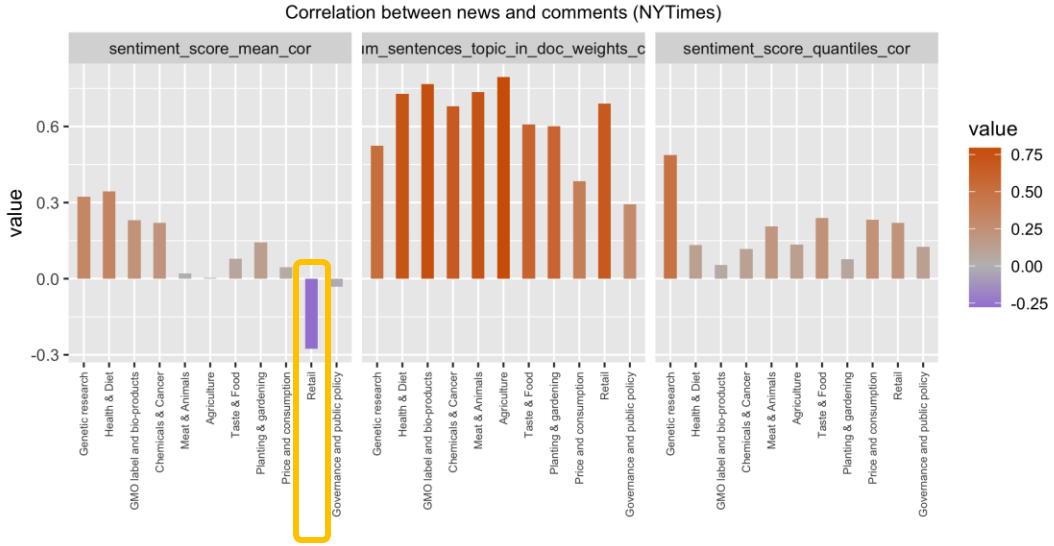
comments:

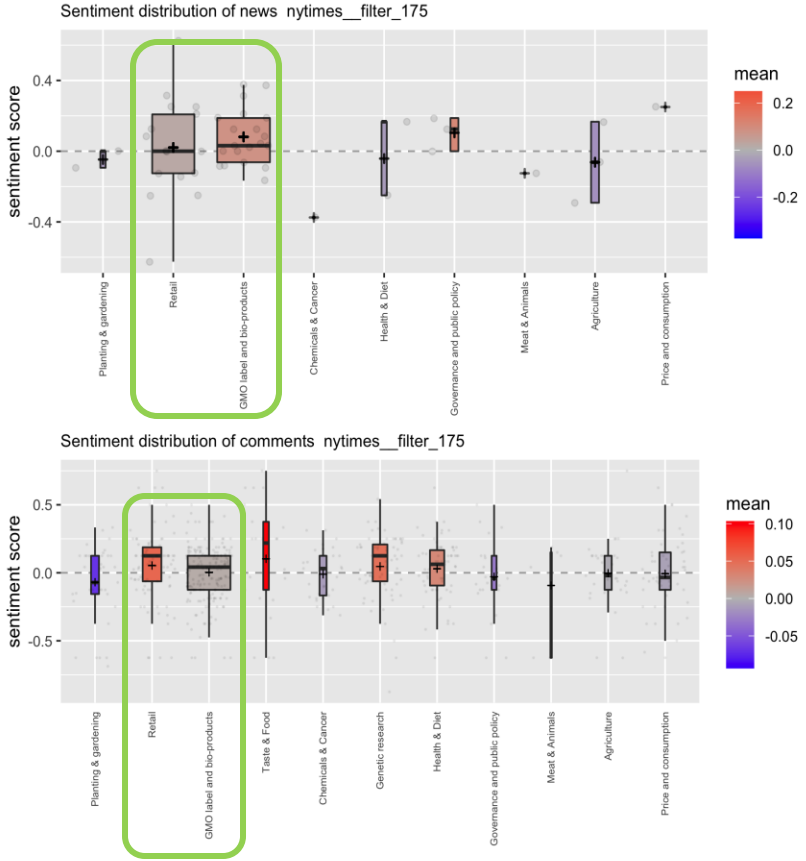
- I have a cousin who works at Whole Foods. He is a *happy* employee and *loves* it. Thinks it is a *great company.*
- I am not sure if *non-GMO* foods are *healthier* to eat but they are certainly *better for the environment.*
- I *applaud* Whole Foods for at least taking a stand.
- consuming *red meat is* an emotionally charged issue for many people.
- "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.



Individual View on NYTimes

Polarity-> mean of scores --> weak correlation/ neg
Fluctuation -> quantile of scores --> week/not significant
Dominance -> weight num_s_topic/num_s_doc --> strong





Individual View on Spiegel

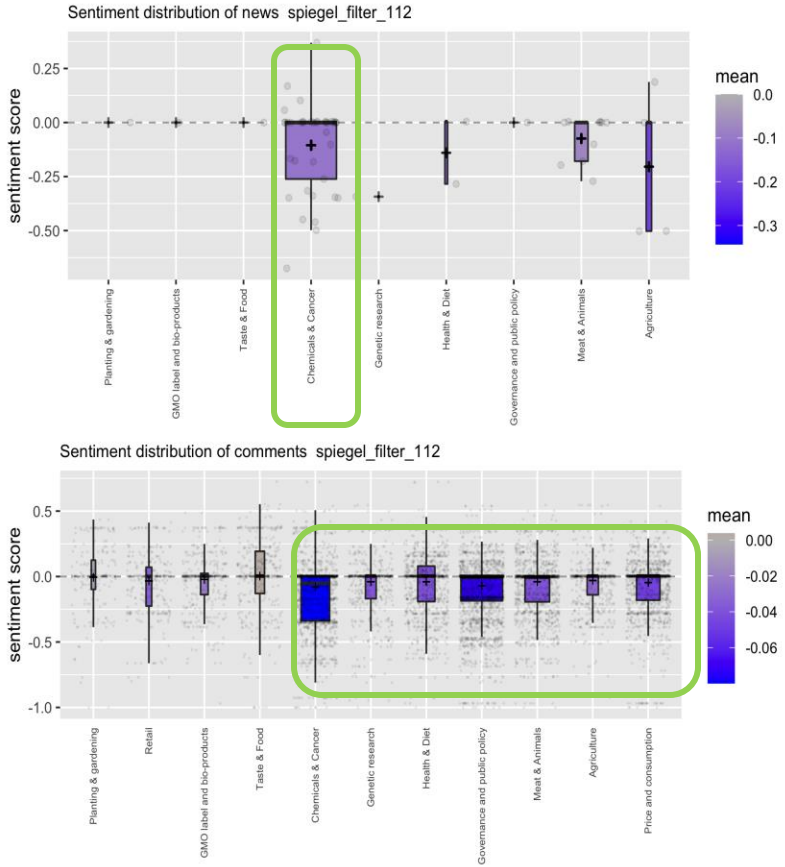
article:

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...

comments:

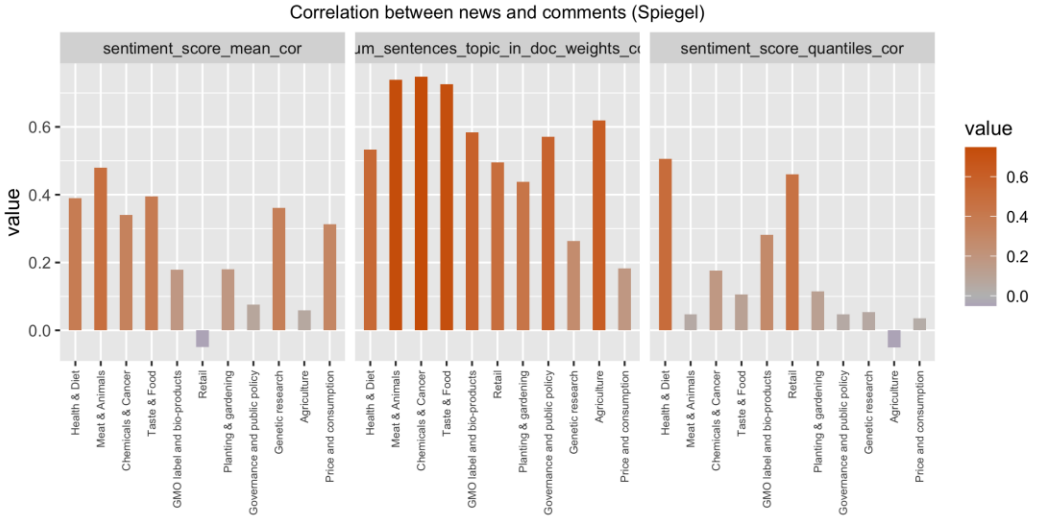
- ...Skandal war hoffentlich nicht der letzte. Es sollten so viele wie möglich vorkommen. *Am besten aber wäre,* wenn ein paar Konsumenten nachweislich an solchen oder anderen Giftstoffen in Lebensmitteln *sterben.*
- 3000 Tonnen verseuchtes Tierfutter - das ist *ein Terroranschlag.*
- *Ich will niemandem 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.

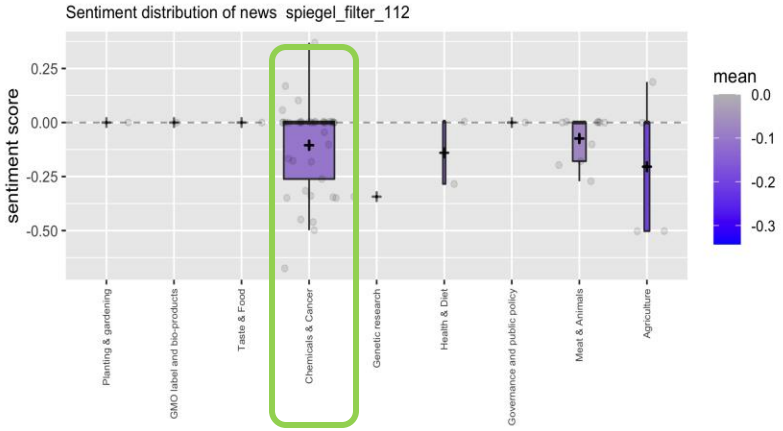
Polarity-> mean of scores
Fluctuation -> quantile of scores
Dominance -> w eight num_s_topic/num_s_doc

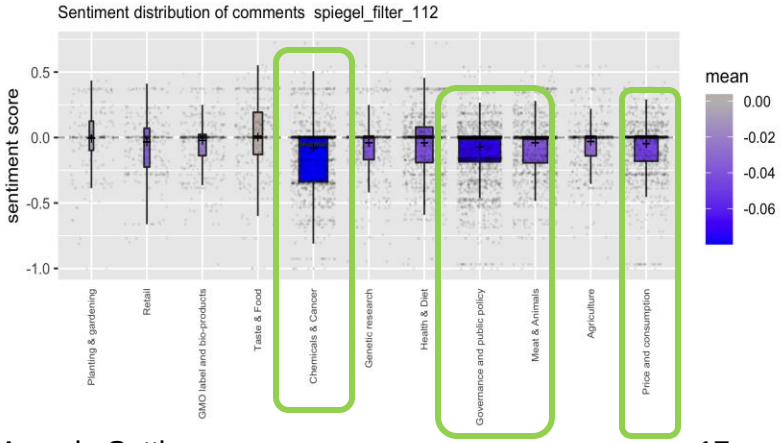


Individual View on Spiegel

Polarity-> mean of scores -> weak correlation
Fluctuation -> quantile of scores --> some are correlated
Dominance -> weight num_s_topic/num_s_doc -> strong







Conclusion

1. Coherent sentences clustering results
2. Mild evidence showing, in same source, both
   - Occurrences of sentences
   - Sentiments polarity

   in readers' comments are positively correlated to those in news respectively
3. The **dominant** topics in media has strong correlations with readers for most topics in both corpus
4. The opinion **polarity** in
   - NYTimes : shows weak/negative/no significant relation with readers per topic-respectively
   - Spiegel: shows (weak) relation with readers per topic-respectively
5. The **fluctuation** level shows
   - NYTimes : no significant in most topics
   - Spiegel: weak relation for specific topics
6. Further investigation on concrete correlation due to the size of corpus
7. Further study on culture/language influences

Questions?

References

Angelidis, S., & Lapata, M. (2018). Summarizing Opinions: Aspect Extraction Meets Sentiment Prediction and They Are Both Weakly Supervised. *ArXiv, abs/1808.08858*.

Baccianella, S., Esuli, A. & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (eds.), *LREC*, : European Language Resources Association. ISBN: 2-9517408-6-7

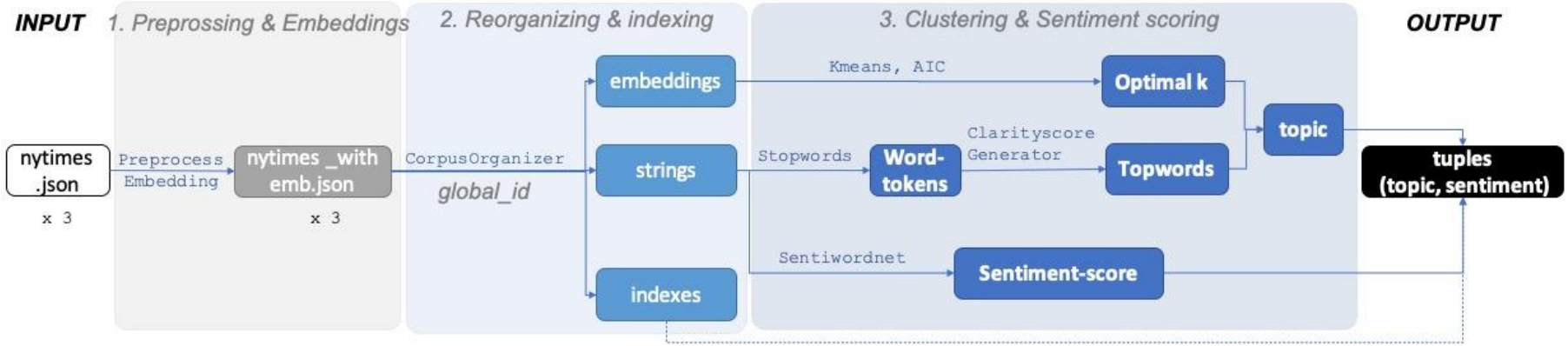
Chidambaram, M., Yang, Y., Cer, D., Yuan, S., Sung, Y. H., Strope, B., & Kurzweil, R. (2018). Learning cross-lingual sentence representations via a multi-task dual-encoder model. *arXiv preprint arXiv:1810.12836*.

McCombs, M., & Shaw, D. (1972). The Agenda-Setting Function of Mass Media. *The Public Opinion Quarterly, 36*(2), 176-187. Retrieved July 16, 2020, from www.jstor.org/stable/2747787

Remus, R., & Quasthoff, U., & Heyer, G. (2010). SentiWS - A Publicly Available German-language Resource for Sentiment Analysis. *Proceedings of the 7th International Language Resources and Evaluation (LREC'10), 168-1171.*

Saif, H., Fernández, M., He, Y., & Alani, H. (2014). On stopwords, filtering and data sparsity for sentiment analysis of twitter.

Appendix: Data flow



Preprocessing includes:

- Used NLTK sentence tokenizer
- Replaced URLs, i.e. <a> tags
- Ignored sentences with character length < 15

Appendix: Stopwords

Most frequent words (TF-High) as stopwords

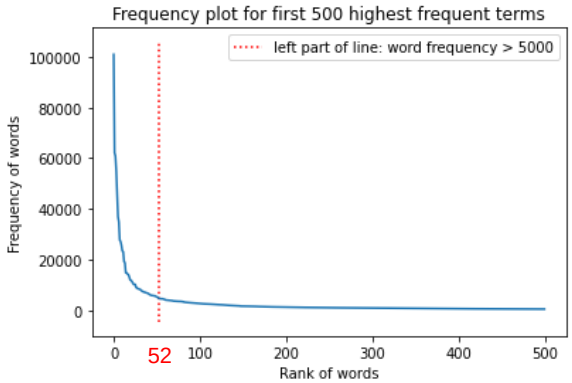
- Fit more to our own organic food domain compared to pre-defined stopword modules

English

Total number sentences: 127,464

Number of distinct words: 43,255

Number of words with frequency > 5000: 52, i.e. **0.12%**



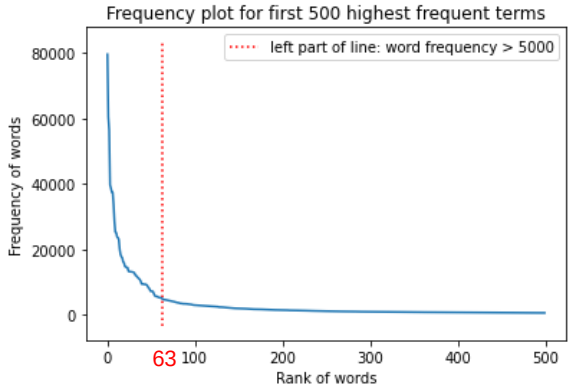
TF-High stopwords: ['the' 'and' 'to' 'of' 'is' 'in' 'that' 'organic' 'it' 'are' 'not' 'for' 'you' 'food' 'as' 'have' 'be' 'on' 'with' 'they' 'or' 'but' 'this' 'from' 'more' 'we' 'do' 'can' 'there' 'if' 'by' 'at' 'all' 'foods' 'about' 'what' 'has' 'will' 'so' 'their' 'an' 'your' 'would' 'than' 'people' 'no' 'which' 'like' 'was' 'one' 'some' 'my']

German

Total number sentences: 200,627

Number of distinct words: 73,081

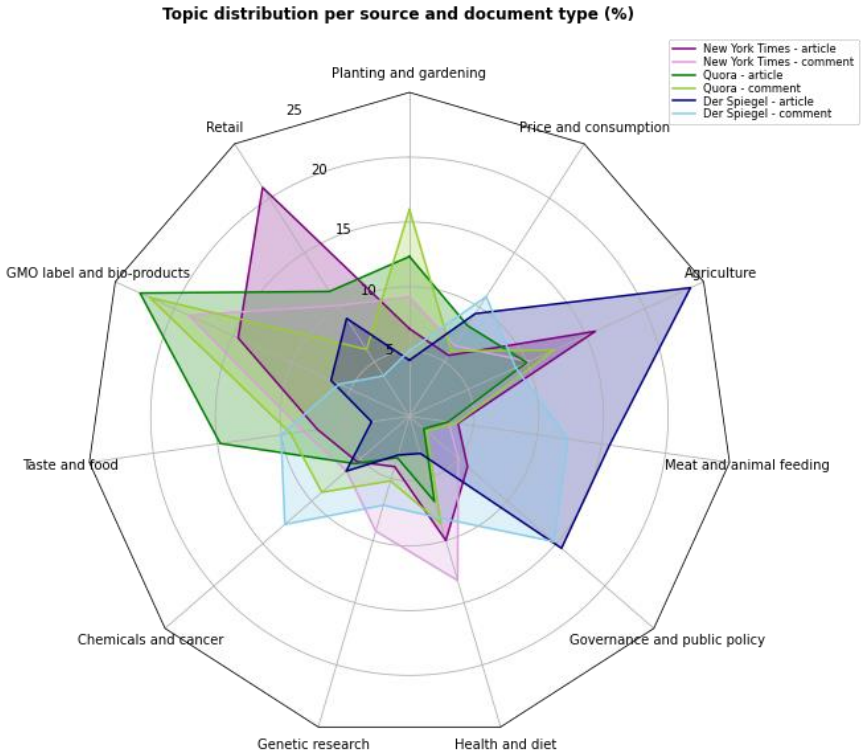
Number of words with frequency > 5000: 63, i.e. **0.09%**



TF-High stopwords: ['die' 'und' 'der' 'ist' 'das' 'nicht' 'von' 'in' 'es' 'sie' 'zu' 'den' 'ich' 'auch' 'mit' 'zitat' 'ein' 'sich' 'für' 'auf' 'man' 'sind' 'dass' 'aber' 'werden' 'wie' 'im' 'nur' 'oder' 'wenn' 'eine' 'so' 'bei' 'als' 'wird' 'aus' 'was' 'dem' 'noch' 'bio' 'an' 'dann' 'haben' 'kann' 'da' 'hat' 'mehr' 'wir' 'um' 'mal' 'doch' 'schon' 'ja' 'nach' 'sein' 'keine' 'immer' 'einen' 'des' 'gibt' 'hier' 'diese' 'durch']

Appendix: Total number of sentences and their distribution

Source	Total number sentences among Articles	Comments
English		
New York Times	14,202	45,597
Quora	1,022	43,118
German		
Der Spiegel	6,795	121,369



Appendix: Word cloud

$k = 13$

Word cloud for English corpus with stemmed words

Word cloud for German corpus with stemmed words

$k = 14$

Word cloud for English corpus with stemmed words

Word cloud for German corpus with stemmed words

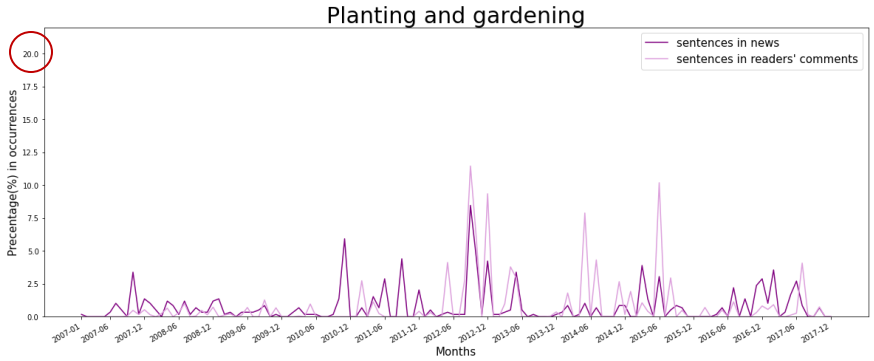
$k = 16$

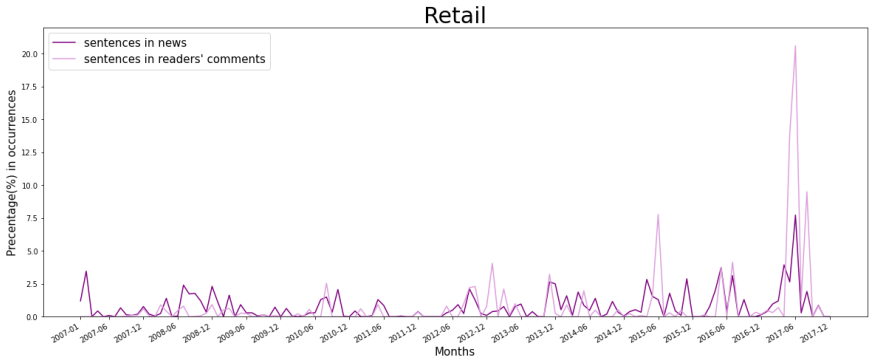
Word cloud for English corpus with stemmed words

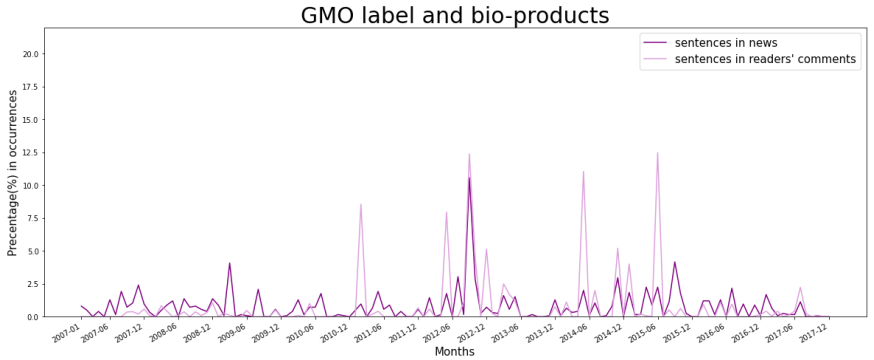
Word cloud for German corpus with stemmed words

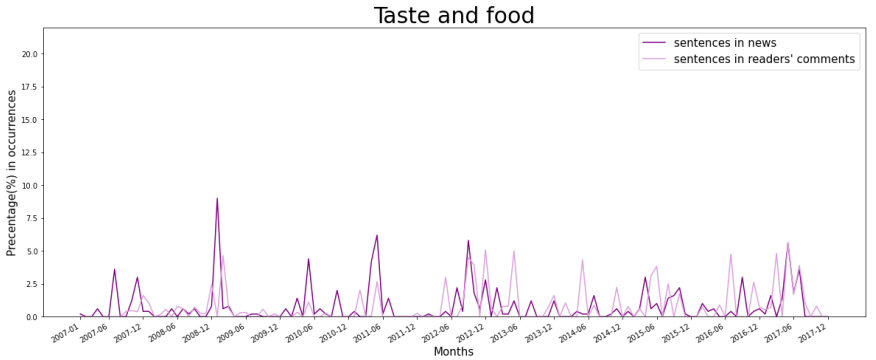
Appendix: Change of sentences occurence rate

New York Times



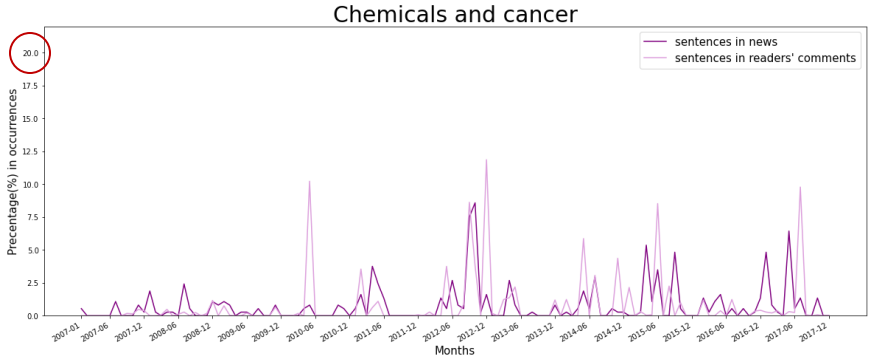


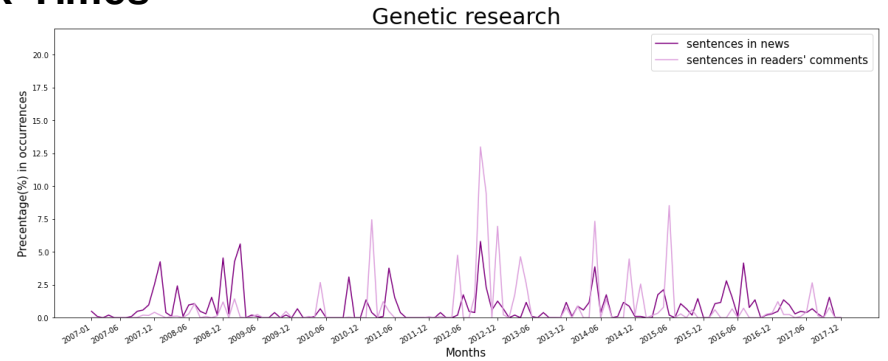


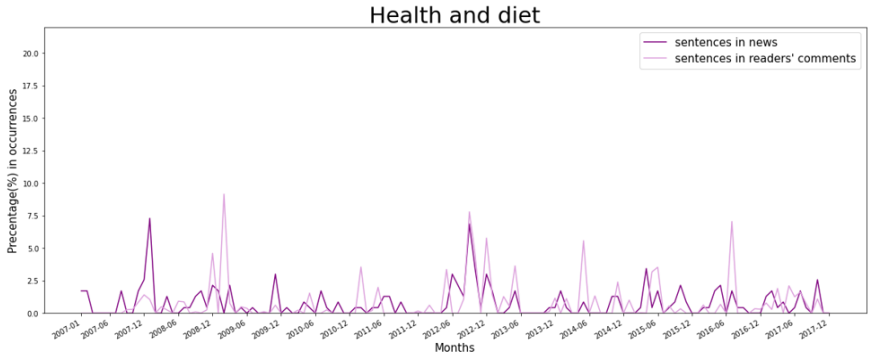


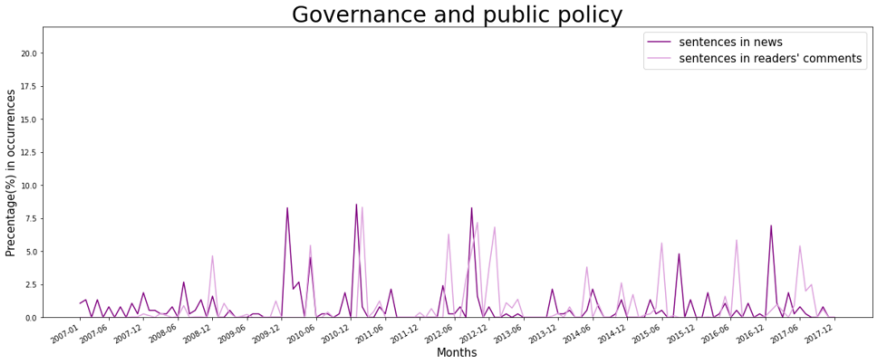
Appendix: Change of sentences occurence rate

New York Times





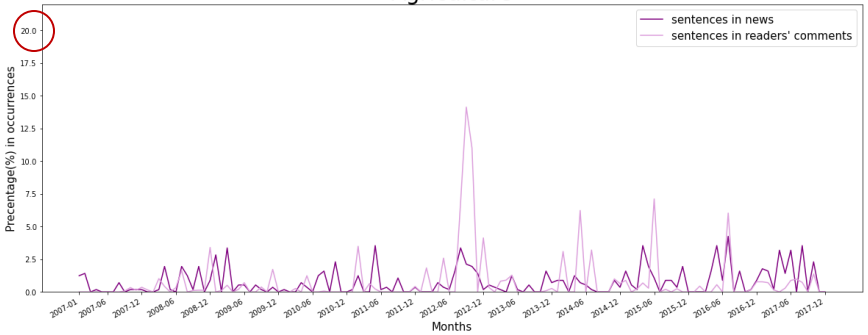




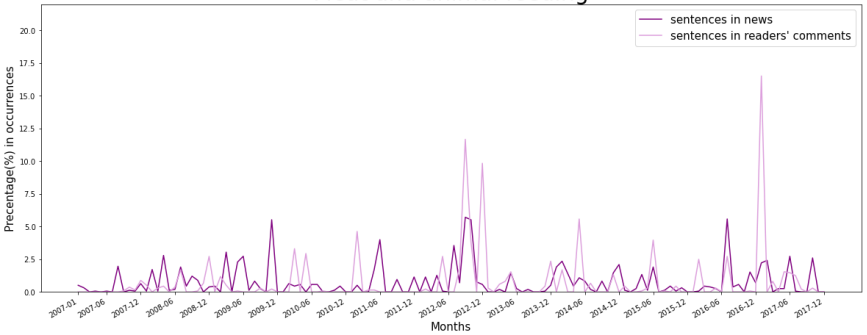
Appendix: Change of sentences occurence rate

New York Times

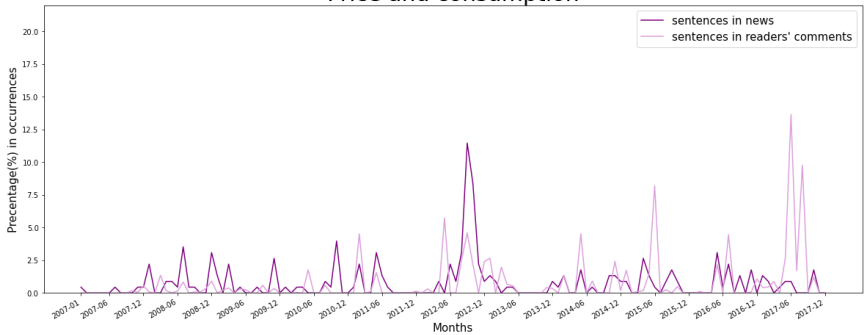
Agriculture



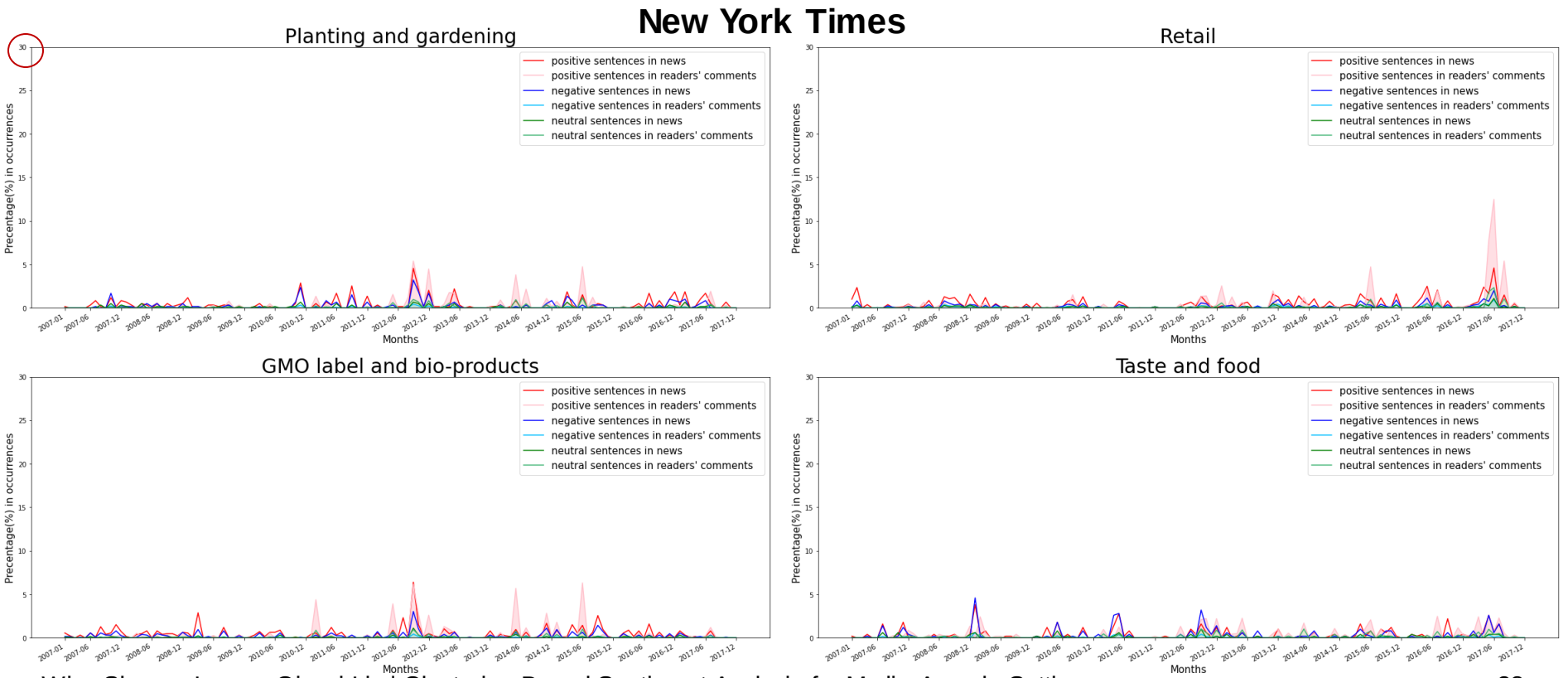
Meat and animal feeding



Price and consumption

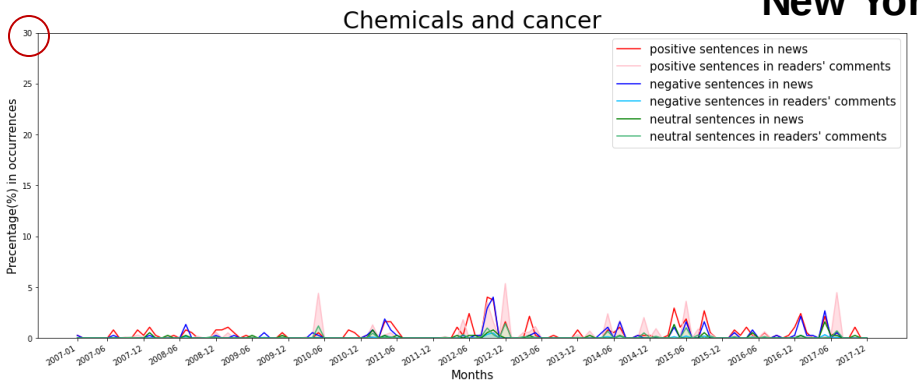


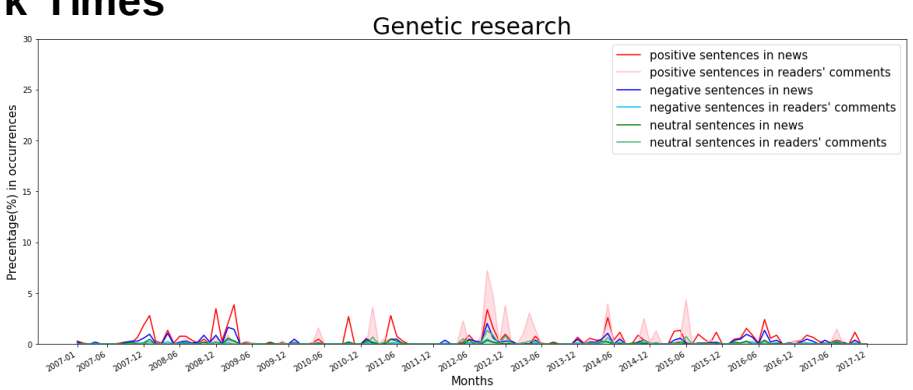
Appendix: Change of *sentiment* occurence rate

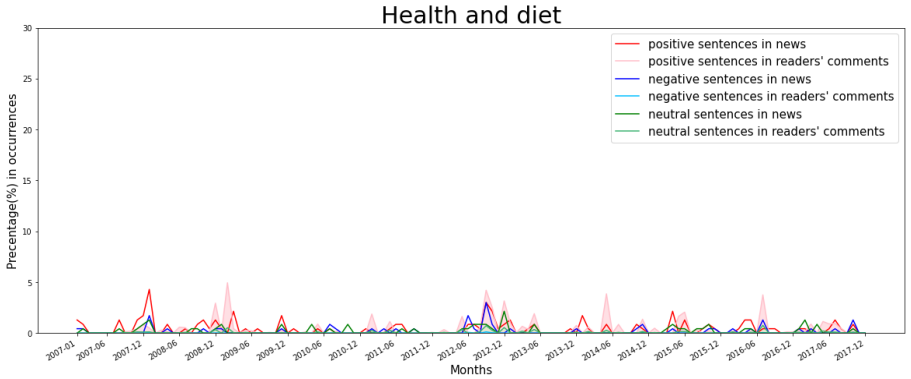


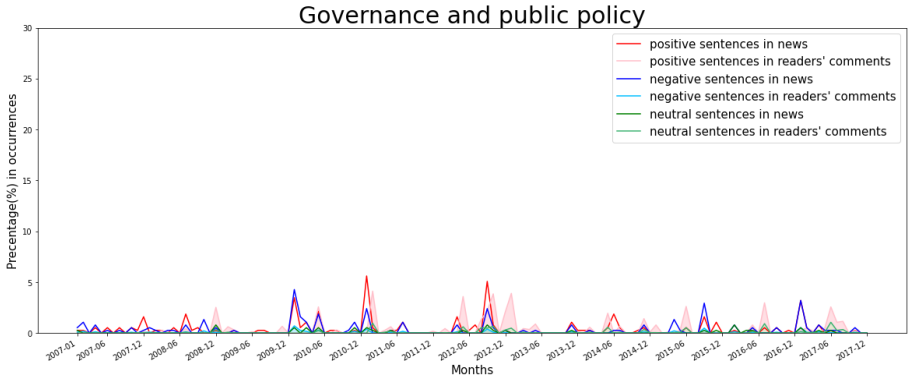
Appendix: Change of *sentiment* occurence rate

New York Times





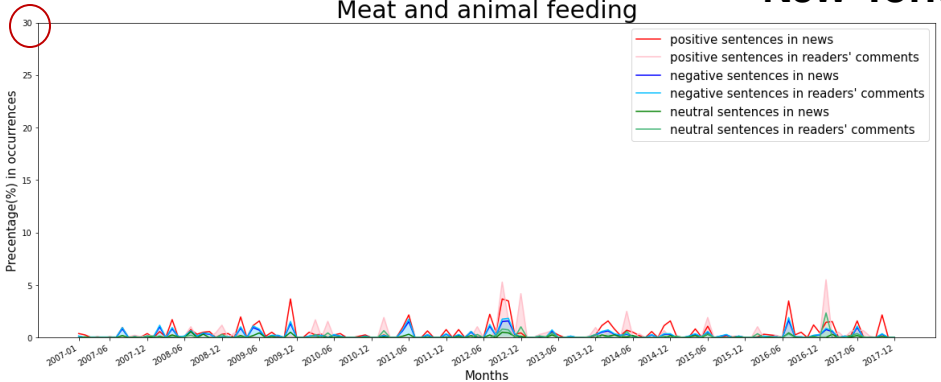




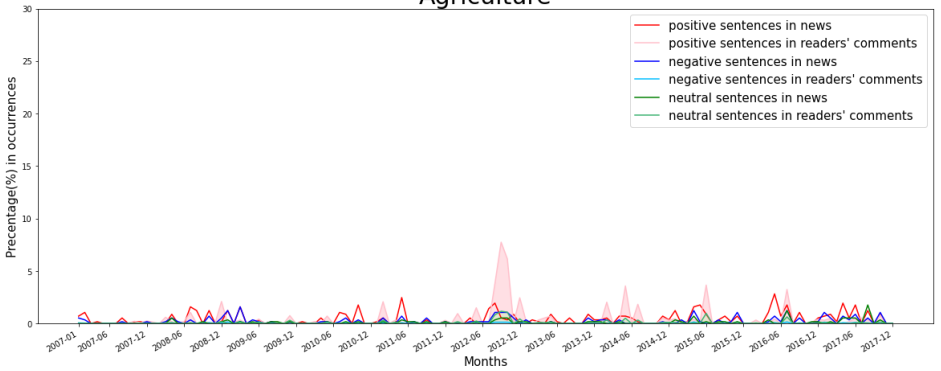
Appendix: Change of *sentiment* occurence rate

New York Times

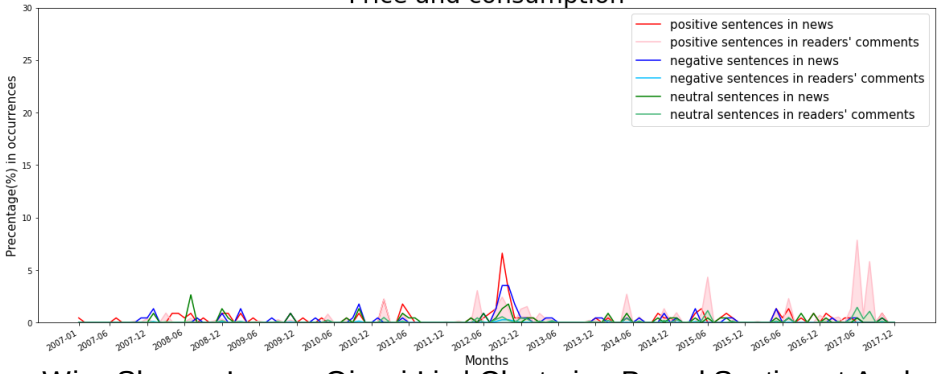
Meat and animal feeding



Agriculture

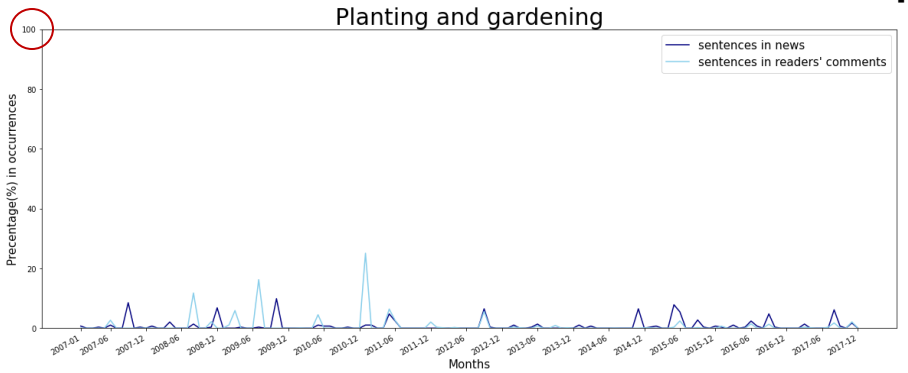


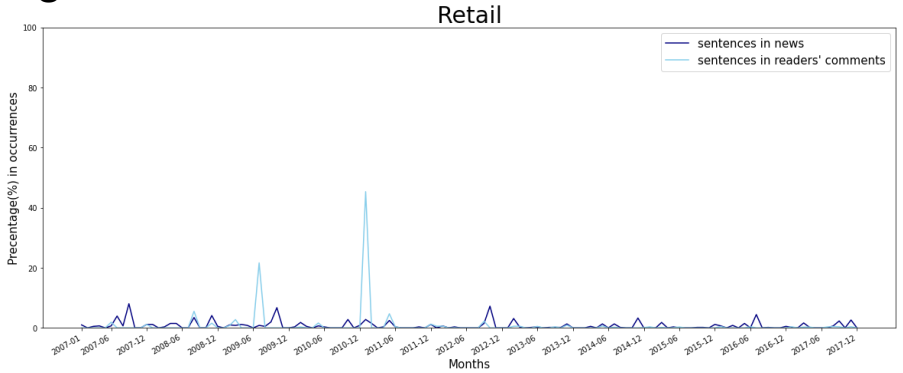
Price and consumption

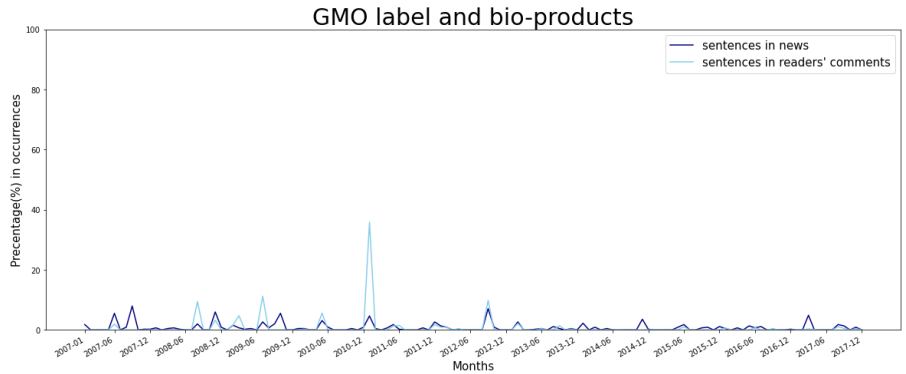


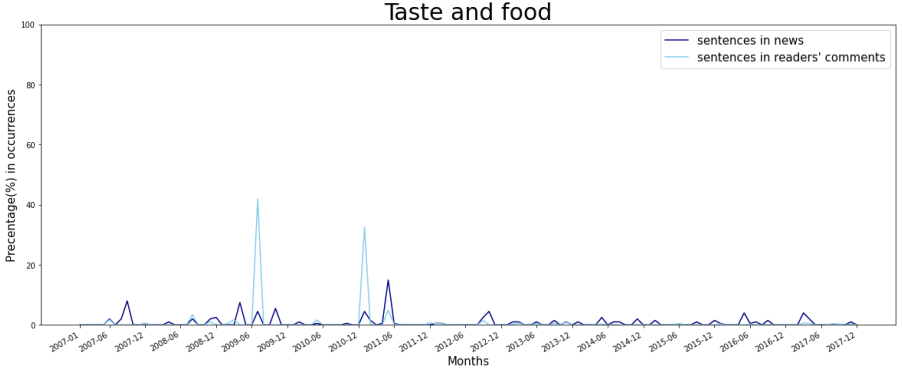
Appendix: Change of sentences occurence rate

Der Spiegel



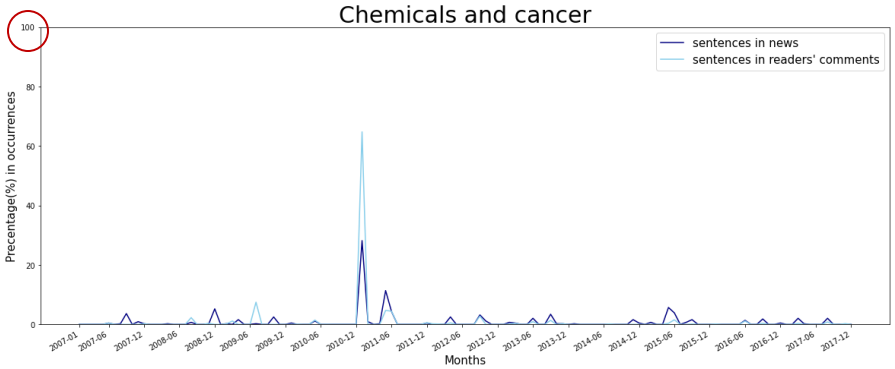


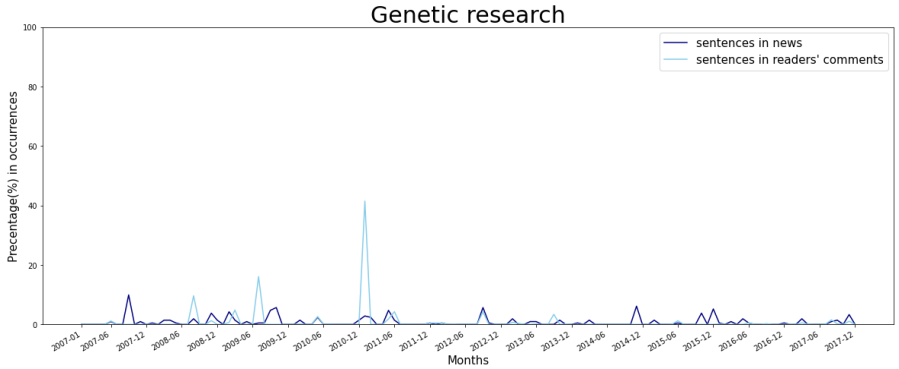


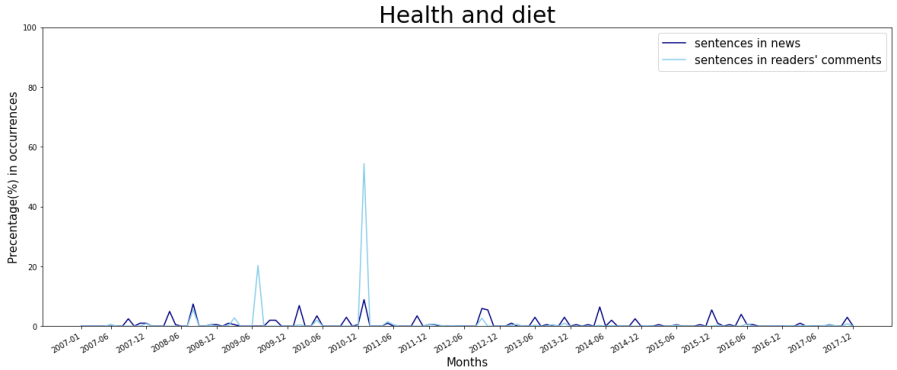


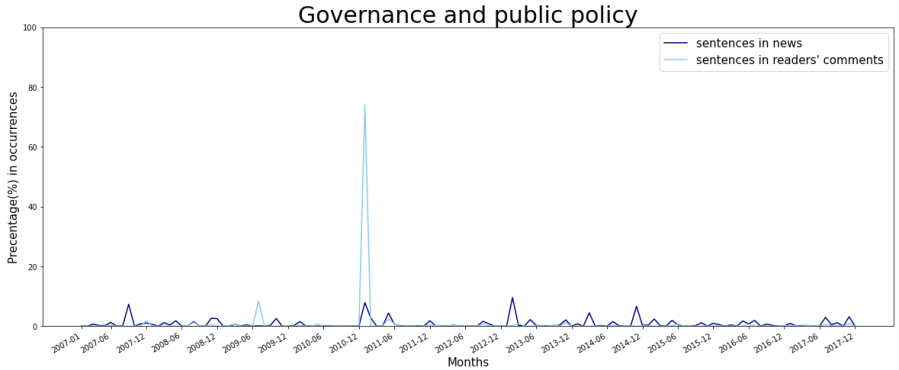
Appendix: Change of sentences occurence rate

Der Spiegel



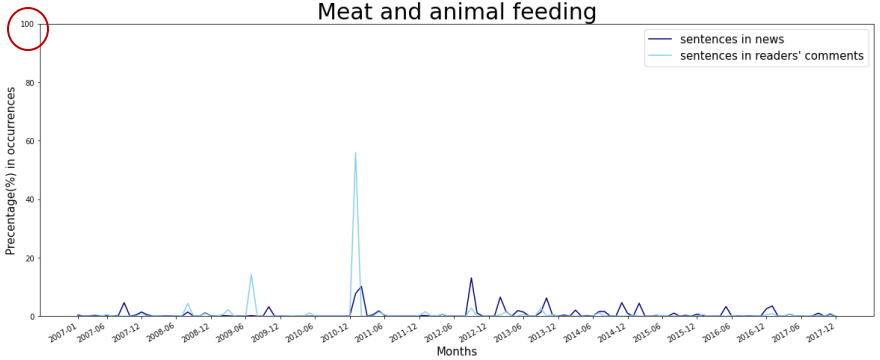


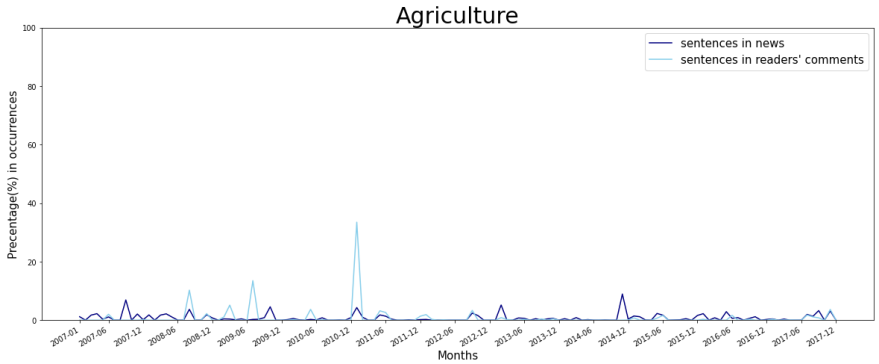


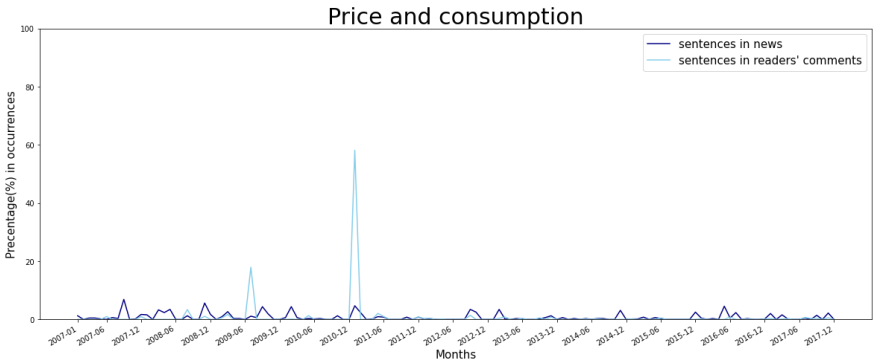


Appendix: Change of sentences occurence rate

Der Spiegel

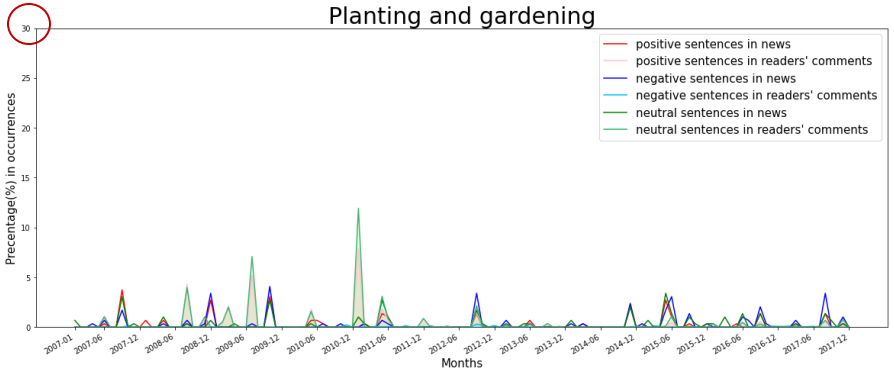


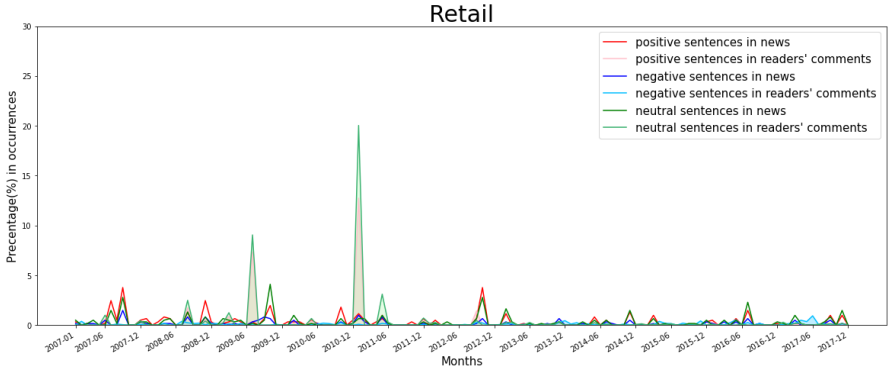


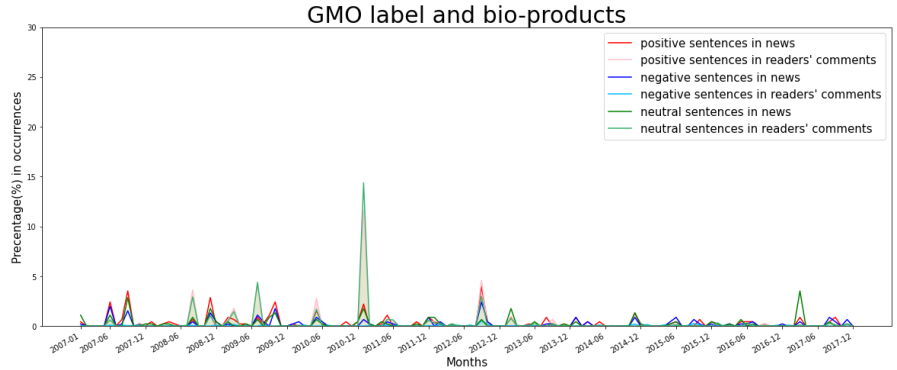


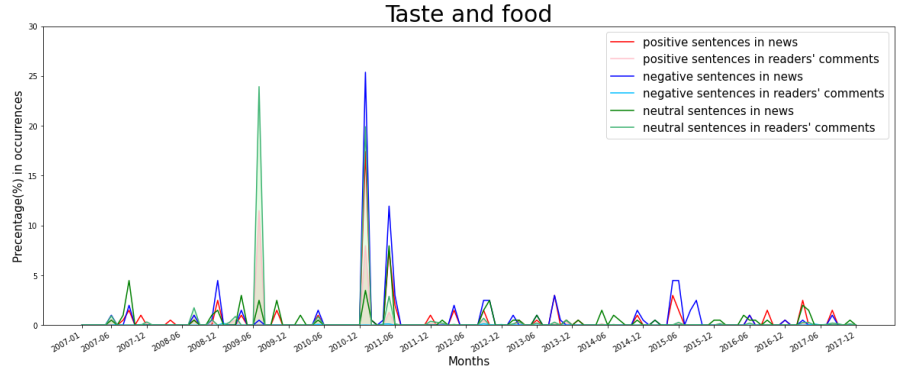
Appendix: Change of *sentiment* occurence rate

Der Spiegel



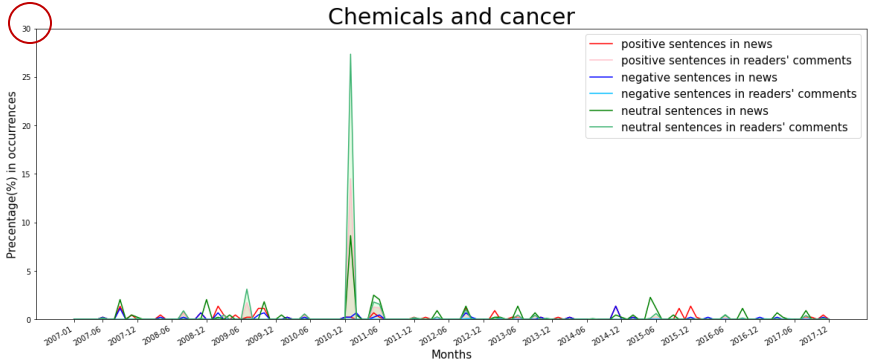


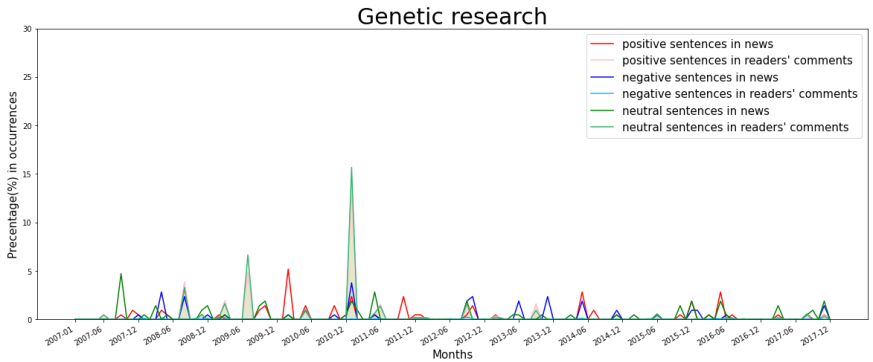


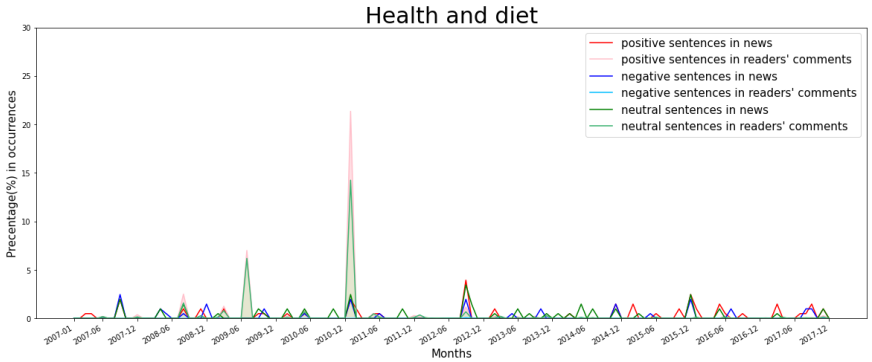


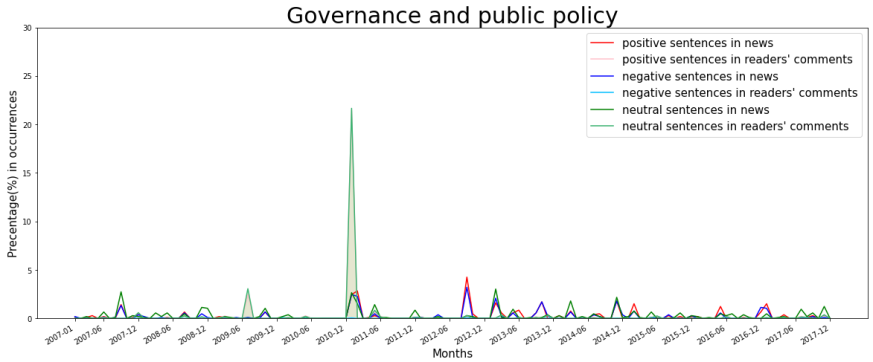
Appendix: Change of *sentiment* occurence rate

Der Spiegel



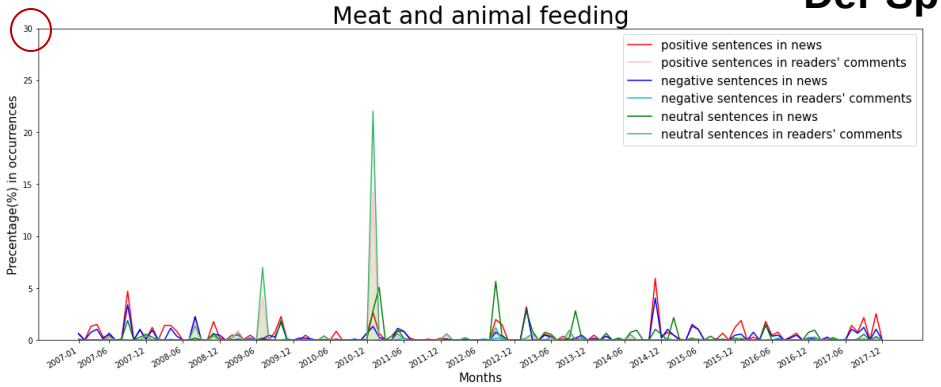


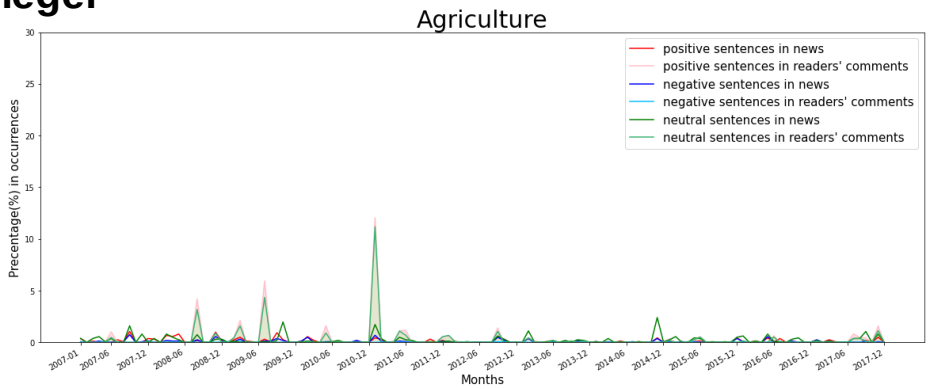


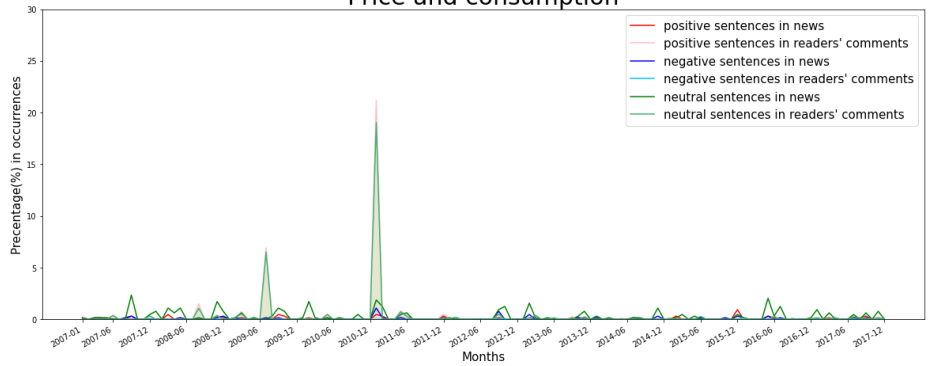


Appendix: Change of *sentiment* occurence rate

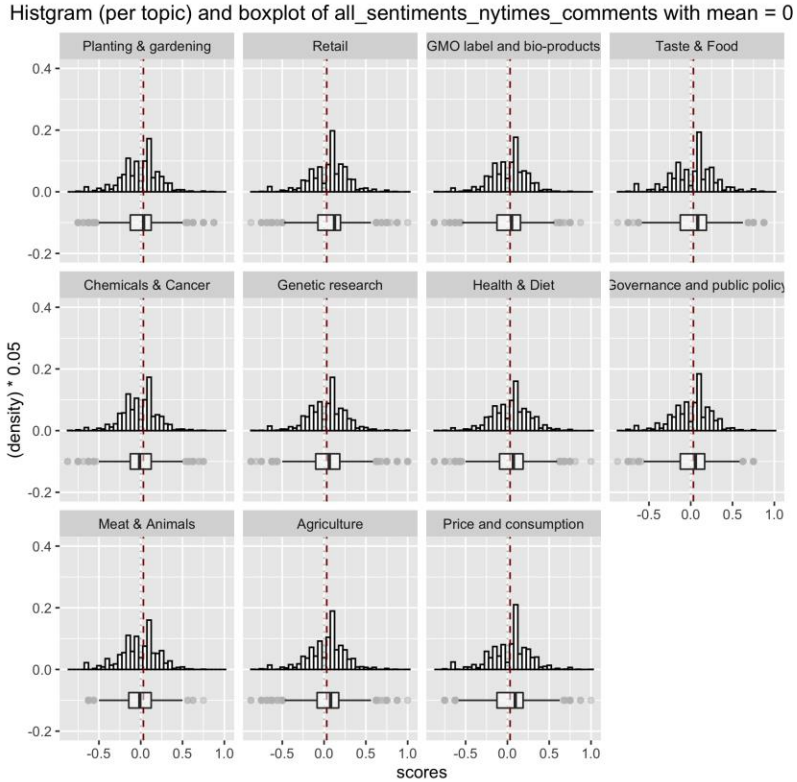
Der Spiegel

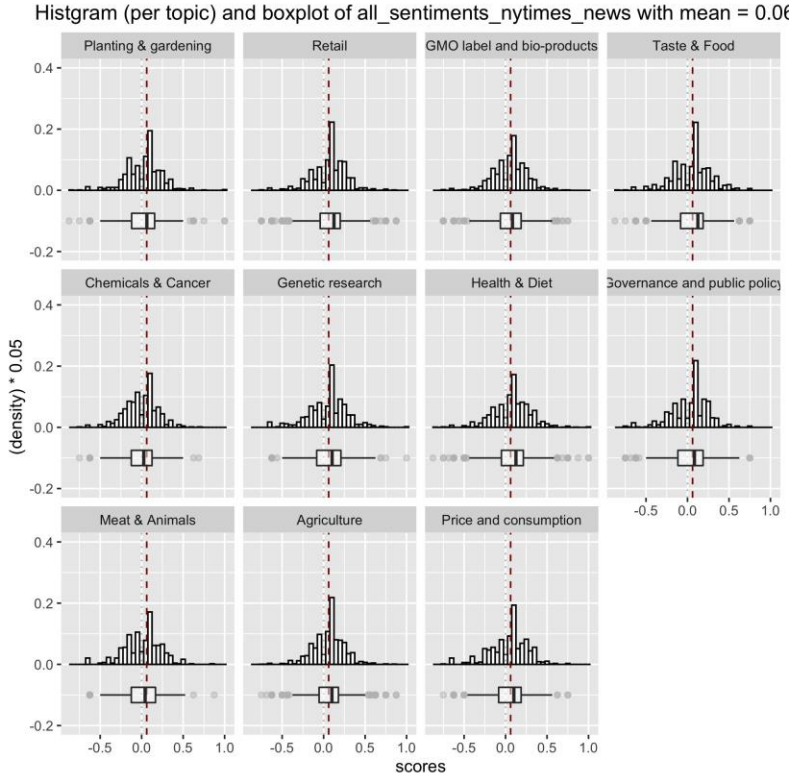




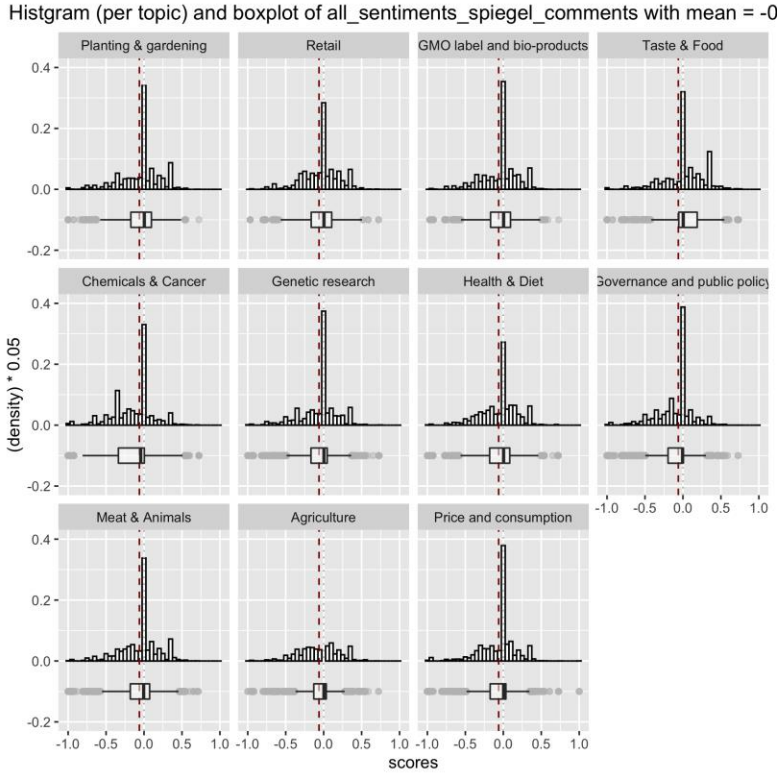


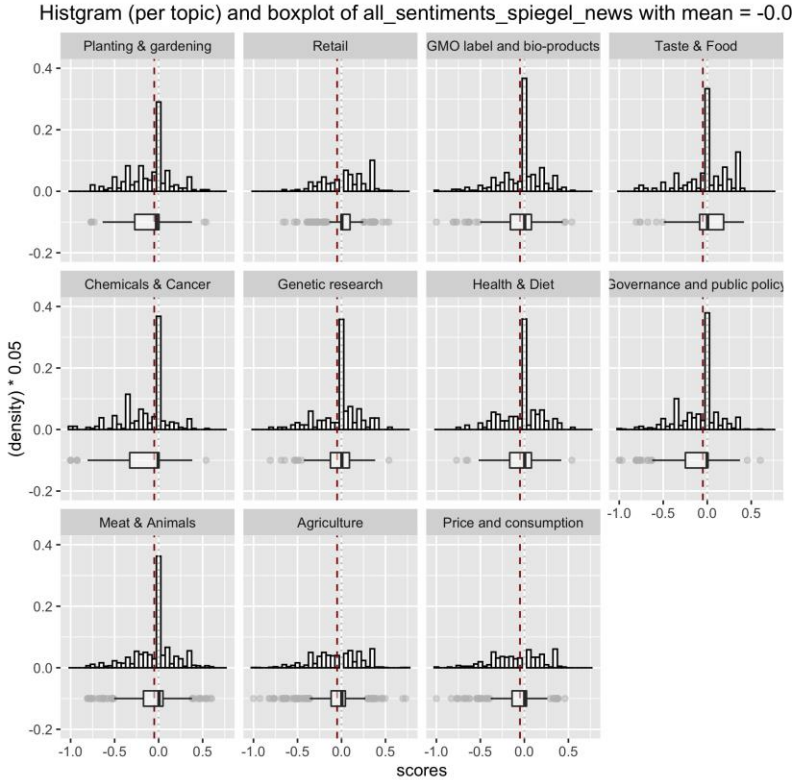
15 sentiment score per topic distribution (NYTimes)





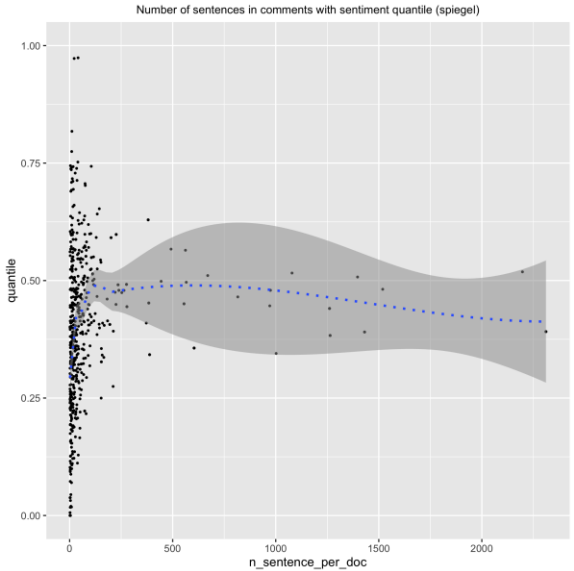
15 sentiment score per topic distribution (Spiegel)

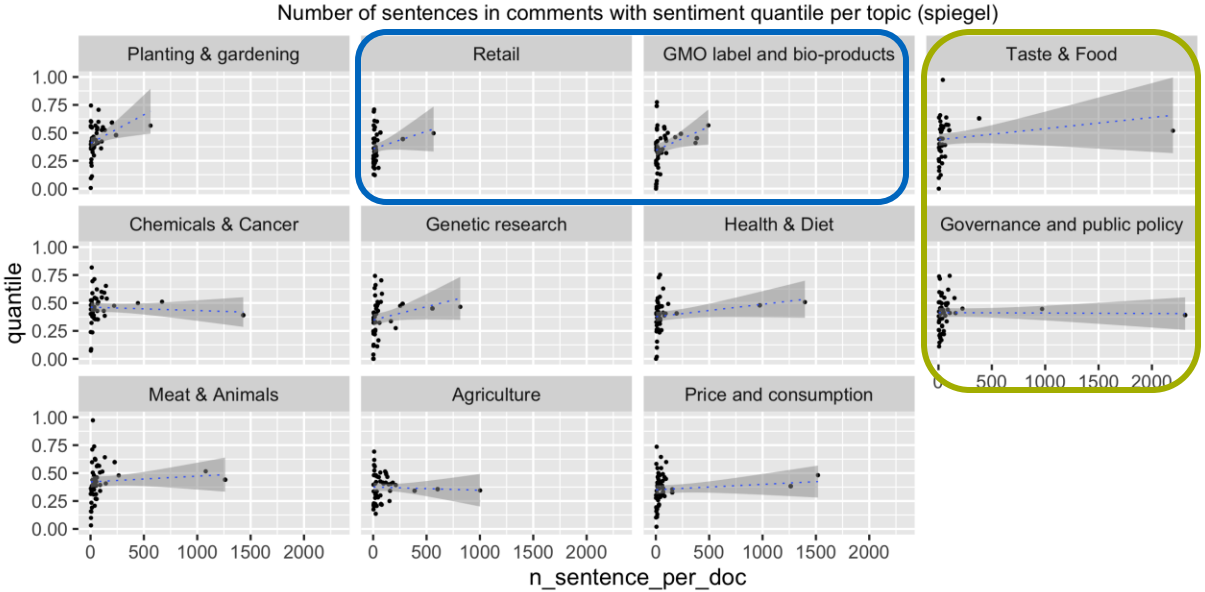




Global View: Culture affects the length or language ?

Spiegel comments (609 points, 61 news, also avg. 10 topics are mentioned in comments of one news)





Global View: The hottest topic? How related? (Spiegel)

- "color" --> polarity of sentiments (per topic) --> mean of sentiment scores --> interval (-1,1)
- "width" --> dominanceof a topic --> weight (num_s in one topic / total_num_s) in an article --> interval (0,1)
- "height" --> fluctuation of sentiments (per topic) --> quantile of sentiment scores --> interval (0,2)
- "area of the box" --> hotness of a topic --> dominance * fluctuation