A Case Study and Qualitative Analysis of Simple Cross-lingual Opinion Mining

Gerhard Hagerer[1][a], Wing Sheung Leung[1], Qiaoxi Liu[1], Hannah Danner[2][b] and Georg Groh[1][c]

[1]*Social Computing Research Group, Department of Informatics, Technical University of Munich, Germany*
[2]*Chair of Marketing and Consumer Research, TUM School of Management, Technical University of Munich, Germany*

Keywords: Opinion Mining, Topic Modeling, Sentiment Analysis, Cross-lingual, Multi-lingual, Market Research.

Abstract: User-generated content from social media is produced in many languages, making it technically challenging to compare the discussed themes from one domain across different cultures and regions. It is relevant for domains in a globalized world, such as market research, where people from two nations and markets might have different requirements for a product. We propose a simple, modern, and effective method for building a single topic model with sentiment analysis capable of covering multiple languages simultaneously, based on a pre-trained state-of-the-art deep neural network for natural language understanding. To demonstrate its feasibility, we apply the model to newspaper articles and user comments of a specific domain, i.e., organic food products and related consumption behavior. The themes match across languages. Additionally, we obtain an high proportion of stable and domain-relevant topics, a meaningful relation between topics and their respective textual contents, and an interpretable representation for social media documents. Marketing can potentially benefit from our method, since it provides an easy-to-use means of addressing specific customer interests from different market regions around the globe. For reproducibility, we provide the code, data, and results of our study[a].

---

[a]https://github.com/apairofbigwings/cross-lingual-opinion-mining

1 INTRODUCTION

Topic modeling on social media texts is difficult, since lack of data as well as spelling and grammatical errors can make the approach unfeasible. Dealing with multiple languages at the same time adds more complexity to the problem which oftentimes makes the approach unusable for domain experts. Thus, we propose a cross-lingually pre-trained deep neural network as a black box with very little textual pre-processing necessary before embedding the texts and forming their clustering and topic distributions.

For our method, we leverage current research regarding multi-lingual topic modeling, see Section 2. We provide an extensive description of a simple method to support domain experts from specific social media domains in its application in Section 3. We qualitatively demonstrate our topic model, its feasibility, and its cross-lingual semantic characteristics

on English and German newspaper and social media texts in Section 4. We aim at inspiring pragmatic ideas to explore the potential for comparative, intercultural market research and agenda setting studies. Unsolved problems and future potential are given in Section 5.

2 RELATED WORK

Topic modeling is meant to learn thematic structure from text corpora. With probabilistic topic modeling methods, such as latent semantic indexing (LSI) (Deerwester et al., 1990) or latent Dirichlet allocation (LDA) (Blei et al., 2003), researchers try to extend the capabilities of topic modeling for application from a single language to multiple languages. Using multilingual dictionaries and translated corpora is an intuitive way to tackle cross-lingual topic modeling problems (Zhang et al., 2010; Vulić et al., 2013). Further examples exist for topic modeling with either dictionaries or translation text collections (Gutiérrez et al., 2016; Boyd-Graber and Blei, 2012; Jagarlamudi and

---

[a] https://orcid.org/0000-0002-2292-0399
[b] https://orcid.org/0000-0001-8387-0818
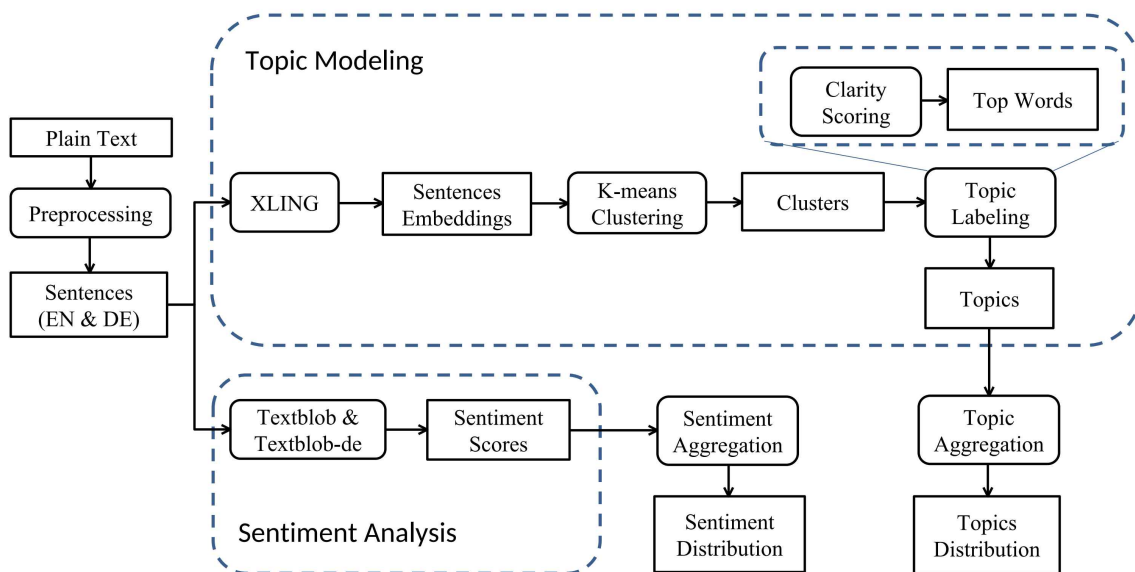[c] https://orcid.org/0000-0002-5942-2297

Figure 1: Plain text is first tokenized into sentences and passed to topic modeling and sentiment analysis. Topic modeling involves (1) converting sentences of both languages into embeddings with XLING, (2) clustering all embeddings with K-means and (3) deriving a topic label of each cluster. Sentiment analysis is performed using Textblob. Topic and sentiment scores are aggregated for the analysis.

Daumé, 2010). However, this puts dependence on the availability of dictionary or good quality of translations. Significant manual labor and verification are required to prevent deteriorating noise.

Recently, methods converting words to vectors according to their semantics are widely adopted (Mikolov et al., 2013). Several studies showed text embeddings improve topic coherence (Bianchi et al., 2020; Srivastava and Sutton, 2017). Regarding multi-linguality, embeddings in word level and sentence level enable text in different languages to be projected to the same vector space (Cer et al., 2018) such that semantically similar texts are clustered together independently of their languages. This favors studies on multi-lingual topic modeling without relying on dictionaries and translation (Xie et al., 2020; Chang et al., 2018). Although providing highly coherent topics, a recreation of word spaces is required when new text corpora are introduced. In our scenario, these limitations are not present.

Regarding the application of topic modeling, various social media corpora are studied by domain experts (Tsur et al., 2015; Ko et al., 2018). They covered on different domains, such as politics, marketing, and public health. Regarding media agenda setting, (Field et al., 2018) studied on how much degree a Russian newspaper related to economic downturn. They also ”introduced embedding-based methods for cross-lingually projecting English frame to Russian” based on Media Frames Corpus. In contrast, we pro-

pose a straightforward topic modeling method without fine-tuning but only clustering necessary on a social media corpus. This enables further investigation on media agenda setting cross-lingually and cross-culturally.

3 TOPIC MODELING METHOD

Figure 1 shows the overall workflow of our topic modeling approach. We aim to conduct simple, cross-lingual topic modeling on user-generated content with no translation, dictionary, and parallel corpus required for aligning the semantic meanings across languages. Our approach solely depends on clustering sentence embeddings for topic modeling. Ready-made sentence representations simplify the approach, since these suppress too frequent, meaningless, and unimportant words automatically without the need to model that part explicitly (Kim et al., 2017).

3.1 Preprocessing

The raw texts of articles and comments are first tokenized into sentences with Natural Language Toolkit (NLTK). Then, URLs, specially for those enclosed with HTML <a> tag, are replaced with string ’url’. After that, sentences with character length smaller than 15 are omitted to minimize noise, since they appear inscrutable and they are only 6.6% out of all sen-
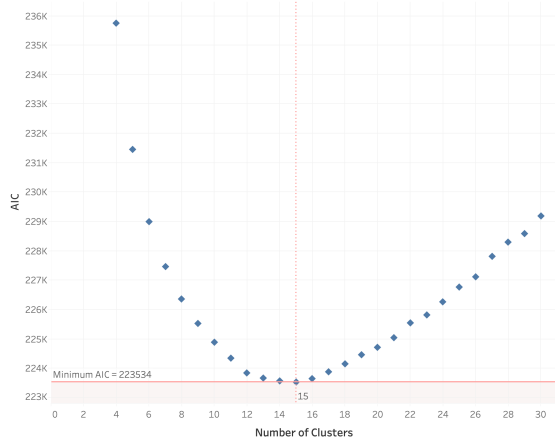
Figure 2: AIC plot indicates $k = 15$ is the global minimum.

tences which is a small portion. After preprocessing, there are 127,464 English sentences and 200,627 German sentences, i.e., total 328,091.

3.2 Cross-lingual Embeddings

In the following paragraph, we provide an explanation of the pre-trained XLING model, which we use for the present work, based on the words of the authors (Chidambaram et al., 2018). XLING calculates *"sentence embeddings that map text written in different languages, but with similar meanings, to nearby embedding space representations"*. Similarity is calculated mathematically as dot product between two sentence embeddings. In order to train the model, the authors *"employ four unique task types for each language pair in order to learn a function g"*, i.e., the eventual sentence-to-vector model. The architecture is based on a Transformer neural network (Vaswani et al., 2017) tailored for modeling multiple languages at once. The tasks on which the model is eventually trained are *"(i) conversational response prediction, (ii) quick thought, (iii) a natural language inference, and (iv) translation ranking as the bridge task"*. The data for training *"is composed of Reddit, Wikipedia, Stanford Natural Language Inference (SNLI), and web mined translation pairs"*.

3.3 Sentence Clustering

K-means clustering algorithm is implemented on both English and German sentence embeddings at the same time. Since XLING provides semantically aligned sentence embeddings of both languages, this joint clustering step helps to establish one topic model for two disjunct datasets irrespective of their language. Clustering is established for a varying number of

clusters k, ranging from 1 to 30. Elbow method is first used for choosing the optimal k but the inertia (sum of squared distances of samples to their closest cluster center) of increasing k decreases rapidly at the beginning and then gently without a significant elbow point. Therefore, Akaike Information Criterion (AIC) is adopted and $k = 15$ is chosen as optimal value as it is the global minimum, see Figure 2. In Section 4.2, further discussion on topic coherence is conducted proving the fact that $k = 15$ resulted semantically coherent topics.

3.4 Topic Labeling

To be able to derive a meaningful topic label for each sentence cluster, the respective top words of each cluster are required. In order to get the top word list, the clarity score is adopted (Cronen-Townsend et al., 2002). According to (Angelidis and Lapata, 2018), it ranks terms with respect to their importance of each cluster c and language l, such that

$$\text{score}_{l,c}(w) = t_{l,c}(w) \log_2 \frac{t_{l,c}(w)}{t_l(w)}, \quad (1)$$

where $t_{l,c}(w)$ and $t_l(w)$ are the l1-normalized tf-idf scores of the word w in the sentences within cluster c and in all sentences, respectively, for a certain language.

Additionally, stopword removal from the top word lists is also a concern when calculating the clarity score. Generally, stopwords are the most frequent words in the documents and sometimes they are too dominant such that they interfere with the result from clarity scoring. Thus, we remove domain-specific high frequency words for each language from corresponding topic top word lists.

Topics are labeled manually based on the English and German top word lists. The results are shown in Table 1 and will be discussed further in Section 4.2 evaluating topic coherence across languages.

3.5 Sentiment Analysis

In addition to topic modeling, we conduct sentiment analysis to investigate the feasibility and meaning of cross-lingual topic-related sentiments in articles and respective comment sections. We make use of Textblob[1] and Textblob-de[2] to assign each of the English and German pre-processed sentences a polarity score. The polarity assignment is first proposed by (Pang and Lee, 2004) and reimplemented by

---

[1] https://github.com/sloria/textblob

[2] https://github.com/markuskiller/textblob-de

Table 1: Top words for all meaningfull topics with $k = 15$ of English and German data.

Topic	English top words	German top words
Environment	pesticide, plant, soil, use, crop, fertilize, pesticide, garden, herbicide, grow	pflanze, pestizid, dunger, boden, gulle, garten, anbau, gemuse, tomate, feld
Retailers	store, whole, shop, groceries, supermarket, local, market, amazon, price, online, discount	aldi, supermarkt, lidl, kauf, laden, lebensmittel, cent, einkauf, wochenmarkt
GMO & organic	gmo, label, gmos, monsanto, product, certificate, usda, genetic, product	produkt, bioprodukt, lebensmittel, gesund, konventionell, biodiesel, herstellung, enthalt, monsanto, pestizid
Food products & taste	taste, milk, sugar, cook, eat, fresh, flavor, fruit, potato, sweet	kase, schmeckt, gurke, essen, analogkase, schmeckt, tomate, milch, geschmack, kochen
Food safety	chemical, cancer, body, acid, effect, cause, toxic, toxin, glyphosat, disease	dioxin, gift, grenzwert, ddt, menge, giftig, toxisch, substanz, chemisch, antibiotika
Research	science, study, scientific, research, gene, scientist, genetic, human, stanford, nature	gentechnik, natur, mensch, wissenschaft, lebenserwartung, genetisch, studie, menschlich, planet
Health & nutrition	eat, diet, healthy, nutritious, health, fat, calory, obesity, junk	lebensmittel, essen, ernahrung, gesund, nahrungsmittel, lebensmittel, nahrung, fett, billig
Politics & policy	govern, public, politic, corporate, regulation, law, obama, vote	politik, skandal, verantwortung, bundestag, schaltet, bestraft, strafe, kontrolle, kriminell
Animals & meat	meat, chicken, anim, cow, beef, egg, fed, raise, pig, grass	tier, fleisch, eier, huhn, schwein, futter, kuh, verunsichert, vergiftet, deutsch
Farming	farm, farmer, agriculture, land, sustain, crop, yield, acre, grow, local	landwirtschaft, landwirt, bau, flache, okologisch, nachhaltig, konventionell, landbau, produktion, ertrag
Prices & profit	price, consume, market, company, profit, product, cost, amazon, money	verbrauch, preis, produkt, billig, qualitat, kunde, kauf, geld, unternehmen, kosten

(De Smedt and Daelemans, 2012). Since the subjectivity assignment is not well-developed in Textblob-de, we filter out sentences with polarity equals to 0 for both English and German sentences in order to derive comparable results.

3.6 Topic and Sentiment Distributions

After assigning a labeled cluster, i.e., a topic, and a sentiment score to each sentence of the corpus, we derive the corresponding distributions.

For topic distributions, all sentences from a document are counted per topic. The distribution is then normalized to be comparable. For sentiment distributions, all sentences from a document are grouped per topic. Topic-wise sentiment distribution is derived based on the sentence-wise polarity scores and the respective median and quartiles. A document in that regard is either an article or all of its comments, i.e., its comment section.

4 TOPIC COHERENCE

In this section, we evaluate the feasibility and semantic coherence of our cross-lingual topic modeling qualitatively. Instead of providing quantitative coherence scores, we aim at a detailed, qualitative analysis of textual examples. We depict representative sentences and words of each topic in subsection 4.2. We investigate to what extent these are semantically coherent, also across languages. We expose the ratio of coherent and incoherent topics and how it develops with increasing number of topics in in subsection 4.3. Eventually, we show the distribution of topics in selected newspaper articles and their respective comment sections to relate the discussed content with our actual topic model on English and German texts.

4.1 Data

The collection of the data used in this study is described in another publication (Danner et al., 2021) as follows. For the analysis we downloaded *"news arti-*
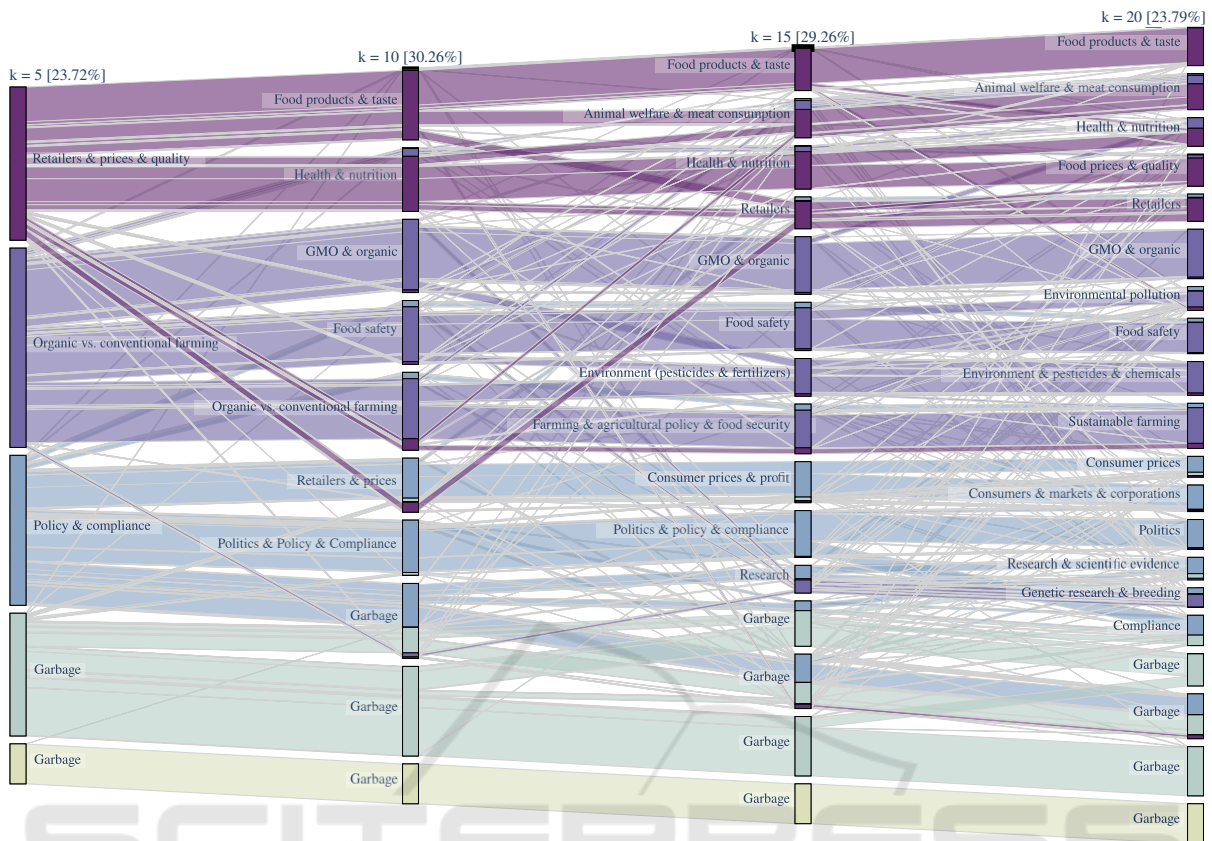
Figure 3: Topic distributions with increasing number of topics *k*. The percentage is the amount of sentences in garbage topics.

cles and reader comments of two major news outlets representative of the German and the United States (US) context", i.e., *spiegel.de* and *nytimes.com*. The creation dates of the texts are *"spanning from January 2007 to February 2020"*. *"Articles and related comments on the issue of organic food were identified using the search terms organic food and organic farming and the German equivalents*. For topic modeling, we utilized *"534 articles and 41,320 comments from the US for the years 2007 to 2020, and 568 articles and 63,379 comments from Germany for the years 2007 to 2017 and the year 2020"*.

4.2 Multi-linguality of Topics

In this section, we evaluate semantic coherence of our cross-lingual topic modeling by depicting the representative sentences and words for each topic and showing the semantic relation. Table 1 shows the first 10 English and German words having the highest clarity scores (see Section 3.4) in each cluster for *k* = 15. Table 2 shows the first 3 English and German sentences whose embeddings have the largest cosine similarity to their corresponding cluster centroids. Both top words and top sentences indicate that the clusters

are grouped reasonably in terms of semantics. For example, this is the case for the topic *Environment (pesticides & fertilizers)* which is indeed related to use of pesticides in planting. Even though this also appears to be the case for the sentences in *GMO & organic* on the first glance, those are actually about organic food and how aspects such as GMO and pesticides relates to the food itself. This and the other representative top words and sentences indicate that clustering on cross-lingual sentence embeddings yield semantically coherent topics.

According to our analysis, top sentences from garbage clusters are always short in length with slightly more than 15 characters. Together with top words (Table 1), these hardly contribute to the organic food domain and corresponding entities. Thus, it is feasible in our case to ignore them.

4.3 Amount of Meaningful Topics

Besides providing coherent cross-lingual topics, our method performs well to distinguish usable from unusable topics, and it provides a constantly high number of relevant topics independently of the number of overall topics. Figure 3 is a Sankey diagram showing

New York Times: *'Major Grocer to Label Foods With Gene-Modified Content'*
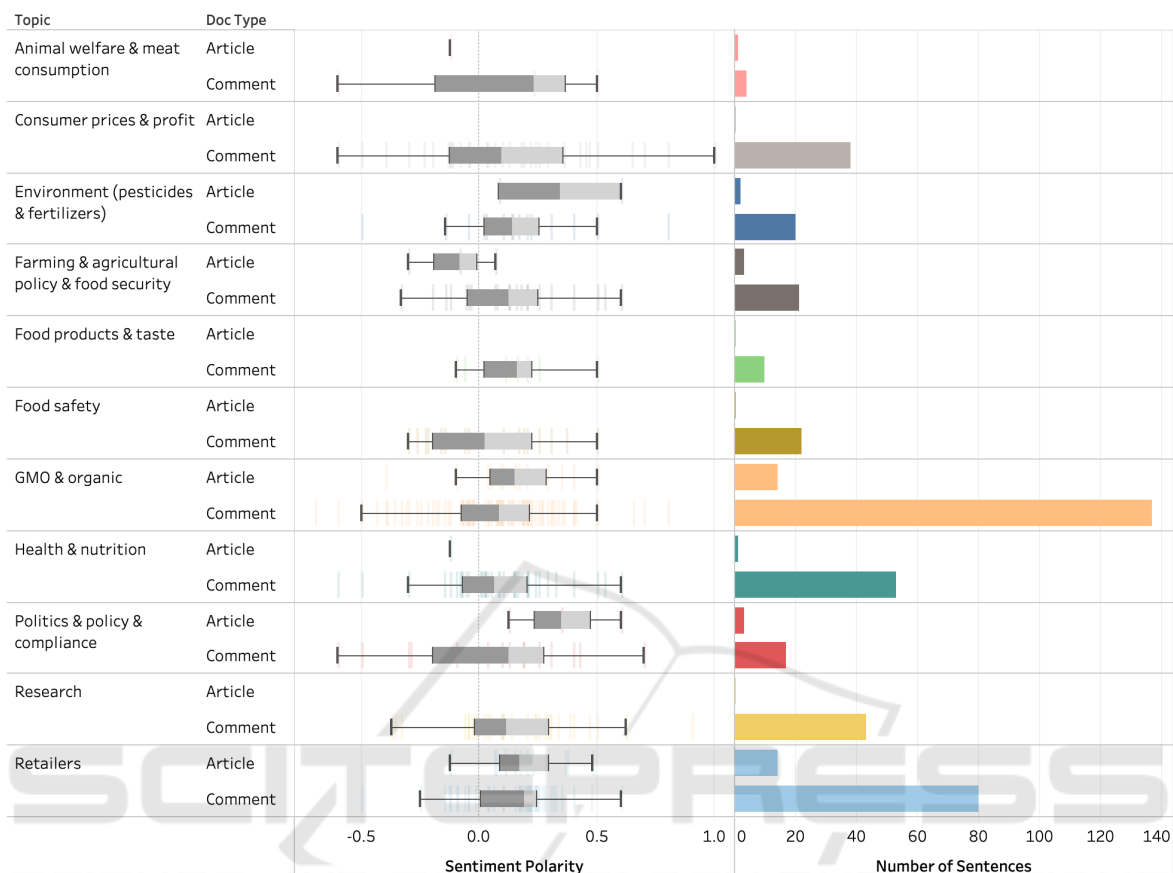

Figure 4: Topic and sentiment distribution for *Grocer*.

the flow of topic assignments for all English and German sentences with increasing number of clusters k. Topic modeling is performed for each k with all preprocessed sentences independently. It can be seen that more specific topics descend from general but related topics as indicated by the colors.

For instance, *GMO & organic*, *Food safety*, *Environment (pesticides & fertilizers)* and *Farming & agricultural policy & food security* for $k = 15$ are derived from *Organic vs. conventional farming* for $k = 5$. *Organic vs. conventional farming* in $k = 5$ generally focuses on advantages brought by organic farming when comparing to conventional farming, such as reducing persistent toxin chemicals from entering to environment, food, and bodies; thus, bio-products are recommended. For $k = 15$, the children topics are more specific. For example, *GMO & organic* shows the aims for having organic food, i.e., avoidance of GMO and poisoning with pesticides. Moreover, *Food safety* in $k = 15$ is further split into *Food safety* and *Environmental pollution*.

To see how the topics relate to their actual sentences, we try to observe the top sentences of each topic, i.e., those sentences of which the embeddings are closest to the centroid. Both English and German sentences are similar and share strong semantic similarity. The *food safety* topic focuses on the toxicity issue of dioxin and other chemicals towards consumers. *Environmental pollution*, which is further splitted from it, for $k = 20$ indeed tells contamination of water resources by chemicals. This shows that fine-grained topics and the way they develop with increasing k have a meaningful relation to ancestor and sister topics.

Sentences without contribution to the organic food domain always remain in garbage clusters in a way that the proportion of usable and unusable clusters does not fluctuate. Thus, the topic model maintains its coherence independent to the number of topics and the despite the fact that k-means is not deterministic in its clustering. This property is helpful, since the number of topics can be chosen as high as necessary to provide a sufficient level of detail for the domain of interest. Moreover, this highlights the meaningfulness
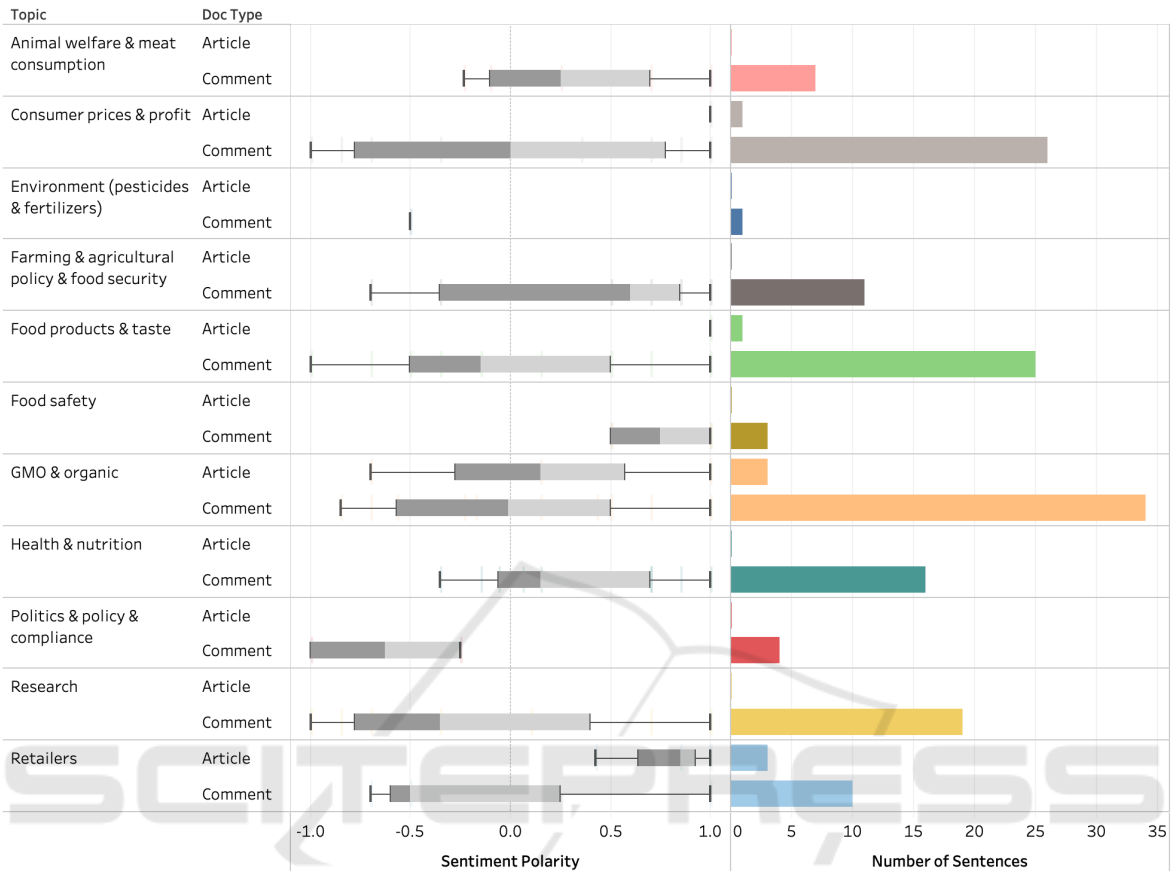
Figure 5: Topic and sentiment distribution for *Öko-Test*.

and robustness of the given sentence representations being able to separate noise from informative data in an unsupervised fashion.

4.4 Validation of Opinion Distributions

In this section, two real text examples are given to evaluate our method qualitatively. The first one is an article from New York Times, titled '*Major Grocer to Label Foods With Gene-Modified Content*'[3], hereafter referred to as *Grocer*. It reported that the first retailer in the United States announced to label all of its genetically modified food sold in its stores. Advocating and opposing stakeholders stated their arguments regarding different aspects. The second example is from Der Spiegel, titled '*"Öko-Test" und Co. – Welche Lebensmittelsiegel wirklich taugen*'[4], below denoted as *Öko-Test*. It reported that number

---

[3]https://www.nytimes.com/2013/03/09/business/grocery-chain-to-require-labels-for-genetically-modified-food.html

[4]https://tinyurl.com/spiegel-lebensmittelsiegel

of food claims, certifications, and seals in Germany were growing as organic labeling was a good promotional strategy indicating high food quality. However, consumers knew little about the details even when the tests for each label were transparent and well-documented. Based on these two summaries, it would be expected that topics related to supermarkets, retailers, and GMO labels are shown to be present in those articles. The *Grocer* article, however, expresses concerns about the consumption of genetically modified food, whereas *Öko-Test* discusses organic food labeling issues from various point of views, among others fair trading and organic fishing.

Topic Distribution. Figures 4 and 5 show the distribution of topics in the overall article sentences. It can be seen that the two topics *Retailers* and *GMO & organic* are mentioned the most in both articles, supporting our hypothesis. The comment section of the *Grocer* article corresponds to the article itself such that most of its sentences also talk about *GMO & organic* and the second most for *Retailers*. How-

ever, the commenters of the *Öko-Test* articles commented more about *GMO & organic* followed with *Consumer prices & profit* and *Food products & taste*. Even though the dominating topics in German differ between article and comments, it can be stated that the topic distribution overall still refers to the actual topics of the given texts and domains. At the same time, differences in the distribution not only between article and comments but also between languages and thus cultures are directly visible, providing a means for clear comparability in several respects.

Sentiment Distribution. Figures 4 and 5 also show the sentiment distribution. Generally, the sentiment of the *Grocer* article spreads out less than that of the *Öko-Test* article. It is observed that, in topic *GMO & organic*, comments score sentiment polarity ranging between 0.50 to −0.70 in *Grocer* and between 1 and −0.85 in *Öko-Test*. This means sentences from *Grocer* show weaker sentiment compared to those from *Öko-Test*. The actual texts indicate that sentiment on our German data indeed has more variance than on English. Thus, the proposed multi-lingual sentiment analysis, Textblob and Textblob-de, appears to represent the data adequately in the given use case. However, it cannot be excluded that the sentiment distribution could be affected given the fact that two different but methodically similar frameworks are used. Different biases and variances could be caused by different models which have differences in the sentiment dictionary size and the subjectivity of human-assigned sentiment scores based on different cultures. Further studies should examine this problem for more robust, domain-independent multi-lingual sentiment prediction.

5 CONCLUSION

This case study shows that our technically simple approach successfully generates an high proportion of relevant and coherent topics for our domain, i.e., organic food products and related consumption behavior based on English and German social media texts. Moreover, the topics display the text contents correctly and support a domain expert in the content analysis of social media texts written in multiple languages.

However, the presented paper did not provide quantitative measurements of topic coherences and comparisons with the state-of-the-art. For mono-language topic modeling, it would be LDA (Blei et al., 2003); for advanced cross-lingual topic modeling, it could be attention-based aspect extraction (He et al.,

2017) utilizing aligned multi-lingual word vectors (Conneau et al., 2017). Several multi-lingual datasets would need to be included for a representative comparison. Since pre-trained models trained on external data are used for the proposed method, it might be relevant for coherence score calculation to include intrinsic coherence scoring methods based on train test splits, such as, UMass coherence score (Mimno et al., 2011), and explore extrinsic methods calculated on external validation corpora, e.g., Wikipedia (Röder et al., 2015).

Regarding multi-lingual sentiment analysis, the difference in the sentiment analysis frameworks for different languages must be considered. For example, since two independent but similar sentiment analysis models are applied for English and German, the sentiment distribution could be affected. Therefore, future studies on developing and evaluating comparable sentiment models should be conducted.

REFERENCES

Angelidis, S. and Lapata, M. (2018). Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. *ArXiv*, abs/1808.08858.

Bianchi, F., Terragni, S., and Hovy, D. (2020). Pre-training is a hot topic: Contextualized document embeddings improve topic coherence.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Boyd-Graber, J. and Blei, D. (2012). Multilingual topic models for unaligned text.

Cer, D., Yang, Y., yi Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder.

Chang, C.-H., Hwang, S.-Y., and Xui, T.-H. (2018). Incorporating word embedding into cross-lingual topic modeling. *2018 IEEE International Congress on Big Data (BigData Congress)*, pages 17–24.

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

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Cronen-Townsend, S., Zhou, Y., and Croft, W. B. (2002). Predicting query performance.

Danner, H., Hagerer, G., Pan, Y., and Groh, G. (2021). The news media and its audience: Agenda-setting on organic food in the united states and germany.

De Smedt, T. and Daelemans, W. (2012). Pattern for

python. *The Journal of Machine Learning Research*, 13(1):2063–2067.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Field, A., Kliger, D., Wintner, S., Pan, J., Jurafsky, D., and Tsvetkov, Y. (2018). Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580, Brussels, Belgium. Association for Computational Linguistics.

Gutiérrez, E., Shutova, E., Lichtenstein, P., de Melo, G., and Gilardi, L. (2016). Detecting cross-cultural differences using a multilingual topic model. *Transactions of the Association for Computational Linguistics*, 4:47–60.

He, R., Lee, W. S., Ng, H. T., and Dahlmeier, D. (2017). An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 388–397, Vancouver, Canada. Association for Computational Linguistics.

Jagarlamudi, J. and Daumé, H. (2010). Extracting multilingual topics from unaligned comparable corpora. In Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S., and van Rijsbergen, K., editors, *Advances in Information Retrieval*, pages 444–456, Berlin, Heidelberg. Springer Berlin Heidelberg.

Kim, H. K., Kim, H., and Cho, S. (2017). Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. *Neurocomputing*, 266:336–352.

Ko, N., Jeong, B., Choi, S., and Yoon, J. (2018). Identifying product opportunities using social media mining: Application of topic modeling and chance discovery theory. *IEEE Access*, 6:1680–1693.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *arXiv preprint cs/0409058*.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.

Srivastava, A. and Sutton, C. (2017). Autoencoding variational inference for topic models.

Tsur, O., Calacci, D., and Lazer, D. (2015). A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Proceedings of*

the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1629–1638, Beijing, China. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Vulić, I., De Smet, W., and Moens, M.-F. (2013). Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora. *Information Retrieval*, 16.

Xie, Q., Zhang, X., Ding, Y., and Song, M. (2020). Monolingual and multilingual topic analysis using lda and bert embeddings. *Journal of Informetrics*, 14(3):101055.

Zhang, D., Mei, Q., and Zhai, C. (2010). Cross-lingual latent topic extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1128–1137, Uppsala, Sweden. Association for Computational Linguistics.

APPENDIX

Table 2: Top sentences of meaningful topics from the whole dataset for $k = 15$ in English and German.

Topic	Top 3 sentences for English and German
Environment (pesticides, herbicides, fertilizers)	Usually, the plant which uses conventional farming will produce the residue of the pesticides. – Some pesticides used in conventional farming, however, may reduce the level of resveratrol in plants. – Also, there is the question of naturally occurring pesticides produced by the plant itself.
	Viele Biopflanzen werden zwar nicht mit Pestiziden behandelt, Ihnen wird jedoch sehr viel mehr Wachstumsfläche zugestanden. – Nicht nur Biobauern benutzen Gülle, und Herbizide und Pestizide werden vor allem in der konventionellen Landwirtschaft eingesetzt. – Zum einen bauen sich die Pestizide und Herbizide relativ schnell ab, nicht zu verwechseln mit Überdüngung durch Gülle oder Belastung mit Schwermetallen.
Retailers	Whole Foods also sells a lot of high quality grocery items that aren't available elsewhere in a lot of places. – Whole Foods executives, however, say their supermarkets can be high quality, organic and natural but also inexpensive. – Larger competitors like Safeway and Kroger have vastly expanded their store-brand offerings of natural and organic products, and they are often cheaper than those at Whole Foods.
	Auch bei ALDI und CO lassen sich hochwertige Lebensmittel erwerben. – Wobei ich feststellen muss, dass andere Supermärkte - zumindest die, die ich frequentiere - auch Wert darauf legen, das gewisse Produkte aus der Region stammen, auch wenn sie konventionell hergestellt wurden. – Der Trend zur Feinkost beschert dem Handel vor allem in den Großstädten steigende Umsätze, wo Bio-Läden hip sind und die kaufkräftigen Kunden beim Einkaufen nicht auf jeden Cent schauen.
GMO & organic	In a nutshell, though, organic means the product meets a number of requirements, such as no GMOs, no non-organic pesticides, etc. – When consumers buy organic, they are guaranteed little more than food that is (in theory at least) produced without synthetic chemicals or G.M.O.'s (genetically modified organisms), and with some attention (again, in theory) to the health of the soil. – Organic food includes products that are grown without the use of synthetic fertilisers, sludge, irradiation, GMOs, or drugs, which already shows how much better it is for health.
	Jeder weiß doch, daß der Vorteil von bio nicht in der erhöhten Aufnahme von Nährstoffen gegenüber konventionellen Produkten liegt, sondern in der Vermeidung, sich mit Pestiziden zu vergiften. – Bio-Lebensmittel genießen einen guten Ruf, weil sie wesentlich weniger Schadstoffe enthalten als konventionell hergestellte Lebensmittel. – Bioprodukte sind kaum gesünder als konventionelle Lebensmittel
Food products & taste	It tastes totally different from your normal vegetables. – It can also be mixed in with the other foods (milk and fruit, oats/rice cooked in milk). – The increased flavor is the result of the food containing more micronutrients.
	Die meisten frischen Zutaten müssen etwas aufbereitet werden, damit sie gegessen werden können. – Es braucht wirklich nur Mehl, Wasser und etwas Salz, keinerlei andere Inhaltsstoffe, Punkt. – Nichts geht über frisch zubereitete Speisen aus gesunden Zutaten.
Food safety	Dioxins are extremely toxic chemicals, and their bioaccumulation in the food chain may potentially lead to dangerous levels of exposure. – Many of the toxins found in non-organic foods are toxins that have a cumulative effect on our bodies. – As proved by various researchers, these chemicals have harmful effects not only on the consumers but also on the environment and farmers.
	Tatsache ist die Dosen um die es bei Nahrungsmittelkontaminationen durch Dioxine geht sind derart gering dass sie auf keinem Wege zu eine signifikanten Gesundheitsgefahr führen. – Dioxine sind UBIQUITÄR und entstehen in nicht unerheblichen Mengen durch natürliche Vorgänge. – Es bedarf erheblicher Dosen um gefährliche Effekte von Dioxin und Lindan nachzuweisen.
Research	Science is not always applied in benign ways, even when we know as much - growth hormones and indiscriminate use of antibiotics in livestock, for example. – The bottom line is that genetically modified organisms have not been proven in any way to be safe, and most of the studies are actually leaning the other direction, which is why many of the world's countries have banned genetically engineered foods. – However, evolution and adaptation, especially to unknown and unnatural substances, takes many generations for humans to achieve.
	Zu wenig verstehen die Wissenschaftler noch von ökologischen und evolutionären Prozessen. – Hinzu kommt dass die Natur ständig neue genetisch veränderte Organismen hervorbringt. – Es geht nicht um die Behauptung von Gentechnik sondern um die Behauptung ihrer Resultate.
Health & nutrition	While the government wants us to eat healthy, it is very true that organic foods are outrageously priced for the small amount of food we recieve. – The health issue with foods lies in our collective wisdom that insists on making foods as cheap as possible. – Of course there are health benefits to eating "organic" food.
	Für die Bevölkerungsschichten die auf günstige Lebensmittel angewiesen sind es komplizierter sich gesund zu ernähren. – Im Vordergrund unserer Lebensmittelwirtschaft steht eben der Profit und nicht die gesunde Ernährung. – Es ist sowieso viel gesünder Lebensmittel zu essen, die einen möglichst geringen Verarbeitungsgrad aufweisen.
Politics & policy & compliance	There are more that "government regulators" involved. – We full well know that the industry in all of its glory takes precedence over the concerns or welfare of the people of this country. – This is all being decided in PRIVATE, There is no involvement by the political or judicial processes that normally make laws in this country.
	Lobbyismus müsste als Straftatbestand angesehen werden und ähnlich schwerwiegend behandelt werden wie Landesverrat. – Das ist das Ergebnis der Lobbyarbeit und unsere Volksvertreter verabschieden solche Strafrahmen nicht versehentlich, sondern ganz bewußt. – Das und ähnliches, ändert nichts an der kriminellen Energie der durch die Politik und Gesetzgebung Vorschub geleistet wird.
Animal welfare & meat consumption	Organically raised animals used for meat must be given organic feed and be free of steroids, growth hormone and antibiotics. – When it comes to meat, again organic is the better option as animals are often treated cruelly and inhumanely to increase production. – Manure produced by organically raised animals wreaks less havoc on the environment, but the meat may still wreak havoc on arteries.
	Diejenigen die noch Fleisch essen nehmen Bio weil diese Tiere etwas weniger gequält werden als wie im konventionellen Bereich. – Im Falle von Fleisch geht das nicht anders, als dass man Tiere quält und dazu mit Dingen füttert, die man kaum noch als Futter bezeichnen kann. – Rinder fressen natürlicherweise kein Zucht-Getreide wegen des hohen Stärke und Fettgehalts.
Farming & agricultural policy & food security	Regardless of capacity the modern crops still need more and more land to feed the more and more people, even if the inefficiencies and failures of corporate agriculture are overcome. – The main problem in organic farming is the availability of adequate organic sources of nutrients (crop residues, composts, manures) to supply crops with all the required nutrients and to maintain soil health. – Without this, the organic farming industry can't sustain economically as most of the Organic food that is produced is bought by the big packaged food companies.
	Daß der Dünger für Bio-Acker nämlich von Nutztieren hergestellt wird und mehr Bedarf daran auch mehr Nutztiere zur Folge hat, gehört nicht zu den Notwendigkeiten, mit denen die Bio-Lobby hausieren geht. – Wegen der Förderung von Biogasanlagen und dem dadurch entstandenen Bedarf etwa an Mais sei Ackerland inzwischen vielerorts so teuer, dass die Bio-Bauern nicht mehr konkurrieren könnten. – Hinzu kommt ein Trend, der immer mehr Landwirte zu Energiewirten werden lässt: Nachwachsende Rohstoffe sind gefragt wie nie.
Consumer prices & profit	Acquisitions such as this takes away consumers' prerogative on where to spend our hard-earned dollars. – The consumer gains nothing from this. – But final cost to consumer is based on supply and demand.
	Beim Verbraucher bleibt so gut wie kein Preisvorteil. – Das man für "bessere" Erzeugnisse mehr zahlen muss, liegt auf der Hand. – Nur müssen diese Billigartikel erst mal produziert werden, bevor der Verbraucher zugreifen kann.