3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11632

Combining content analysis and neural networks to analyze discussion topics in online comments about organic food

Hannah Danner[1], Gerhard Hagerer[2], Florian Kasischke[2], Georg Groh[2]
[1]TUM School of Management, Technical University of Munich, Germany, [2]TUM Department of Informatics, Technical University of Munich, Germany.

Abstract

Consumers increasingly share their opinions about products in social media. However, the analysis of this user-generated content is limited either to small, in-depth qualitative analyses or to larger but often more superficial analyses based on word frequencies. Using the example of online comments about organic food, we investigate the relationship between qualitative analyses and latest deep neural networks in three steps. First, a qualitative content analysis defines a class system of opinions. Second, a pre-trained neural network, the Universal Sentence Encoder, analyzes semantic features for each class. Third, we show by manual inspection and descriptive statistics that these features match with the given class structure from our qualitative study. We conclude that semantic features from deep pre-trained neural networks have the potential to serve for the analysis of larger data sets, in our case on organic food. We exemplify a way to scale up sample size while maintaining the detail of class systems provided by qualitative content analyses. As the USE is pre-trained on many domains, it can be applied to different domains than organic food and support consumer and public opinion researchers as well as marketing practitioners in further uncovering the potential of insights from user-generated content.

Keywords: *deep neural networks; natural language processing; consumer research; content analysis; social media; organic food.*

1. Introduction

Novel communication technologies sparked the desire of users to publicly share opinions on online platforms (Ziegele et al., 2014). These developments provide an increasing amount of user-generated content, such as online user comments, which can be exploited by marketing and consumer research to gain insights into consumer thinking (Balducci & Marinova, 2018). Beginning with Kozinets' (2002) netnography of online communities, social scientists have increasingly analyzed textual user-generated content with established methods such as content analysis (Krippendorff, 2019). However, due to time and human resources required, such qualitative analyses are limited to small data samples. More recently, advances in automated text analysis and data collection enable consumer researchers to efficiently analyze larger datasets in a short amount of time and facilitate the detection of patterns, and compare measurements over time or between datasets. For an overview of methods see Berger et al., 2020). Frequently employed methods are dictionary-based approaches (e.g., LIWC, Tausczik & Pennebaker, 2010) relying on word frequencies. Researchers using automated text analysis have started to incorporate methods from the field of natural language processing (NLP, such as of data-mining, data-preprocessing, simple classifiers, and topic models (Latent Dirichlet Allocation, Blei, 2012) (for an overview see Vidal et al., 2018). However, to the best our knowledge, there has been little research on how qualitative and NLP methods can be combined fruitfully. Latest advances in NLP are neural networks that account for the semantic context of words, i.e., word embeddings (Mikolov et al., 2013), or sentences, i.e., sentence embeddings (Cer et al., 2018). In this paper, we explore how such embeddings particularly lend themselves to be combined with qualitative text analysis by matching the analysis-depth of the latter with the scope of pre-trained sentence embeddings. In three steps, we present a novel approach for how a qualitative content analysis can be combined and enhanced with deep neural networks for semantic similarity.

We apply the approach to the case of organic food. Not only is a growing share of consumers aware of and buys organic food (Hemmerling et al., 2015)—making it an increasingly important consumer research topic—, consumers also voice their opinions about organic food online (Danner & Menapace, 2020; Meza & Park, 2016; Olson, 2017). The analysis of online user-generated content can thus deliver valuable insights into which product attributes and related topics matter to consumers and what could be potential purchase drivers and barriers.

2. Methodology

In step 1 of our approach, a qualitative text analysis is conducted to develop a class system and manually classify a dataset of interest. In Step 2, we use semantic features from pre-trained neural networks to investigate the semantic characteristics and the respective frequencies for each class. Step 3 presents criteria to combine results of both methods.

2.1. Step 1 – Qualitative Analysis

To exemplify the approach, for step 1, we draw on a recent qualitative content analysis by Danner and Menapace (2020) of online comments about organic food. They manually extracted and classified consumer opinions (referred to as beliefs) about organic food to understand consumers' perception of organic. The authors collected 1069 online comments about organic food from high-coverage US news websites (e.g., nytimes.com, washingtonpost.com) and forums (e.g., reddit.com, quora.com). The 1069 comments consisted of 5510 sentences. Among these 5510 sentences, the two coders identified 1065 containing belief statements about organic food and subsequently classified those belief statements into 64 belief classes and 21 superordinate topics. For example, the sentence stated by a commenter *organic farming is better for nature* was attributed to the belief class *organic farming protects the environment*, which in turn was attributed to the topic class *environment*. By counting the frequencies of belief statements per category, the authors presented a detailed picture of topics salient to the online commenters in the data.

2.2. Step 2 – Universal Sentence Encoder

Using the same data and class system as in step 1, we find similar sentences for each class using the Universal Sentence Encoder (USE). USE is a recent advance in NLP and deep learning (Cer et al., 2018). Its architecture is based on the widely adopted Transformer architecture (Vaswani et al., 2017). USE is a deep neural network model pre-trained on large scale text corpora from many domains. From there, the statistical knowledge in terms of generalizable, intermediate, semantic vector representations, which are also referred to as features or embeddings, can be used to quantify the semantics of specific domains, here organic food. USE works on sentence level providing sentence embeddings. The semantics of a given sentence are expressed by its vector representation. When compared to other sentences, the cosine similarity ranges between 1 (similar) to -1 (dissimilar).

We applied USE to automatically find semantically similar sentences for each of the 64 beliefs identified by Danner and Menapace (2020) (e.g., *organic farming protects the environment*) (Table 1). First, USE transformed each of the 64 beliefs and the 5510 sentences into an embedding. Second, USE measured the cosine similarity, i.e. the angular distance, between the embedding of each of the 64 beliefs (also referred to as seed sentences) and each of the 5510 sentences. When choosing a low threshold level for cosine similarity (i.e., the closer to -1), many sentences are considered as similar, whereas at high levels fewer sentences are considered as similar.

2.3. Step 3 - Evaluation

Eventually, we determine the appropriate level of semantic similarity, i.e., the respective cosine similarity threshold level which yields similar frequencies compared to the qualitative

content analysis as reference. To this end, we inspect the thresholding results for cosine similarity levels from 0.7 to 0.84 based on the following criteria. (1) In the content analysis, 1065 sentences were relevant as in containing beliefs about organic food. A meaningful sentence filtering should yield a similar amount of relevant sentences. (2) The number of sentences assigned into the different classes should be similar for both methods. Therefore, we inspected the relative class frequencies and also calculated the Pearson correlation between the class frequencies for different cosine similarity levels. Figure 1 displays a trade-off between semantic similarity and class frequencies: the lower the cosine similarity (i.e., the less similar the sentences), the higher the correlation between the two methods. (3) Manual inspection should confirm the semantic cohesion between the manually and the automatically assigned sentences. Note that we performed the evaluation at topic level (21 topic classes) as the 64 belief classes are very detailed and in part semantically too similar (e.g., *organic farming is better for the environment* and *conventional farming harms the environment*).

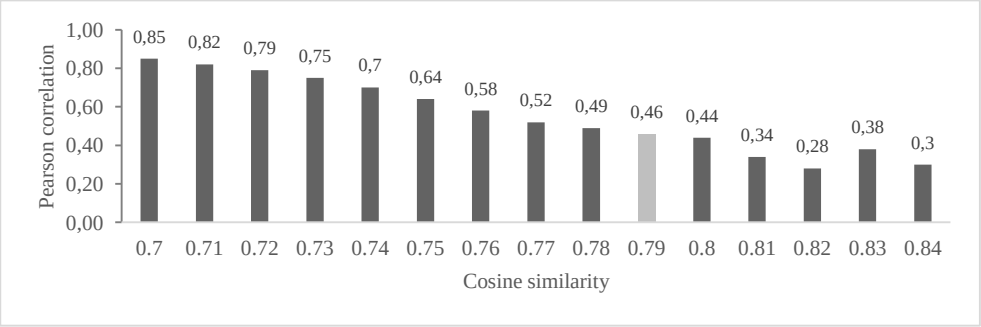

Figure 1. Pearson correlation of class frequencies (21 topic classes) between content analysis and USE. Source: own illustration.

3. Results

Applying the aforementioned evaluation criteria, the thresholding performed best at a cosine similarity of 0.79. (1) At this level of similarity, USE found 1376 relevant sentences, which roughly corresponds to the 1065 relevant sentences identified in the manual analysis. (2) As highlighted in Figure 1, for cosine similarity of 0.79, both methods yielded similar class frequencies, indicated by a correlation of r = 0.46. However, class frequencies do not match perfectly. Looking at the relative class frequencies for each of the 21 topic classes in Figure 2, we find that the class frequencies for both methods are more similar for some topics than for others. For example, the topic *environment* accounts for 11% of sentences in the content analysis and 18% in the similarity thresholding. The most frequent topics in the content analysis were *system integrity*, *food safety*, *environment*; the most frequent topics in USE were *environment*, *system integrity*, *farmer welfare*.

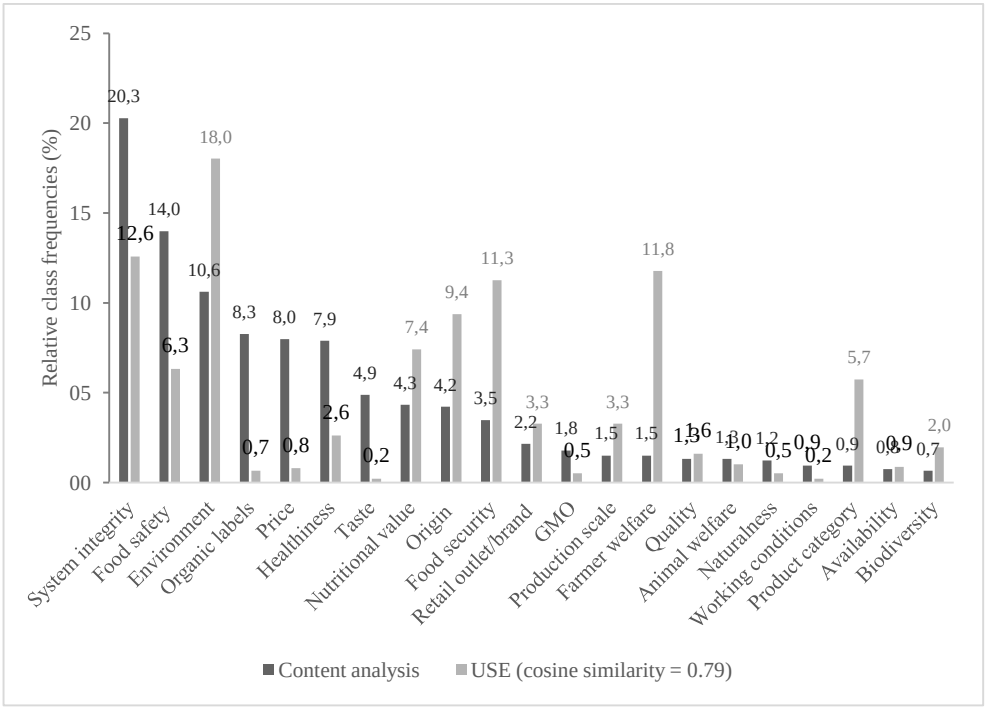

Figure 2. Relative class frequencies (21 topic classes) in content analysis and USE.
Topics are ordered in descending frequency according to the content analysis. Source: own illustration.

Table 1. A seed sentence from content analysis and the 11 sentences identified as similar by USE (cosine similarity = 0.79).

seed	Organic farming protects the environment.
1	Organic farming can help to preserve our environment for future generations.
2	The depletion of the soil and monoculture is what causes factory farming produce to be less nutritious than organic.
3	Mythbusting 101: Organic Farming > Conventional Agriculture
4	A lot of what I've read has said that organic farming is not better for the environment.
5	Organic is for the environment.
6	And from this we hear that organic farming is "devastating" to the environment.
7	Organic farming is much closer to the way Mother Nature farms.
8	GMOs can be super beneficial - to the consumer, the farmer, the environment.
9	Organic farming is greener
10	Besides delivering health benefits, organic farming is better for the environment.
11	Organic is for the environment.

Source: own illustration.

(3) For cosine similarity of 0.79, manual inspection showed very high semantic cohesion between the seed sentences per topic and the sentences identified as similar by USE. Table 1 displays the 11 sentences that USE found to be similar to the belief *organic farming protects the environment* at a cosine similarity of 0.79. All 11 are concerned with the effect of organic farming on the environment. However, sentences 3, 4, and 6 carry negative and thus the sentiment opposite to the seed sentence. Thus, while USE correctly identifies the topic, the sentiment is not always correctly classified, which is one reason why comparisons at topic level were chosen for this study. In addition, the manual inspection of the sentences classified by both methods proved that both methods classified largely the same sentences in the respective classes.

4. Discussion

USE appears to be an effective and easy to use method to analyze large text corpora by searching for sentences that are semantically similar to seed sentences of interest. Seed sentences can originate, for instance, from a small-scale qualitative study—here the belief classes identified by Danner & Menapace (2020). Provided a manually developed class system, it can analyze any unseen dataset, —here 5510 sentences on organic food—,

according to semantic similarity. In the present example, a human researcher selected the required level of similarity by evaluating the features generated by USE based of descriptive statistics and manual inspection. We suggested several criteria to select the appropriate similarity level as an alternative to training a classifier. Training a reliable classifier to classify fine-grained classes as complex as 64 different organic food beliefs requires large amounts of labeled data, which often exceed the resources of common research projects in the field of consumer and opinion research, and as it also applied to the presented example.

The selected similarity threshold was valid as the filtered sentences were widely coherent with the qualitative content analysis. In a subsequent step, USE could be applied to filter a larger unseen data set on organic food. Thus, the potential of the suggested approach lies in its scalability. We can extrapolate the detail of insight characteristic of qualitative research to analyze class frequencies in a larger data set of user-generated content.

Being still in an early phase, our approach bears potential for further refinements. We used a very large class system with 64 belief classes grouped into 21 topics, which also contained classes semantically very similar to each other. Using fewer and more distinct classes could thus improve the coherence between a manual classification and automatic classification based on USE. Furthermore, USE reliably finds the sentences containing similar topics, but does not always correctly distinguish positive and negative sentiment regarding the topic. Therefore, while suitable for topic classification, its use for sentiment analysis is bound to the manual control of a human researcher and domain expert. The imperfect match between manual classification and automatic filtering may also originate from the selection of the unit of analysis, a well-discussed issue in qualitative research (Campbell et al., 2013). The unit of analysis in USE are sentences, whereas in the content analysis, the unit of analysis could also stretch beyond a single sentence, and qualitative researchers can use domain knowledge for understanding and classifying text.

5. Conclusion

In a three-step approach, we suggested how a topic classification of a qualitative content analysis—here of online comments about organic food—can be combined with neural networks like USE to find similar sentences. We proved that embedding techniques largely fit the results of qualitative analysis and point out their methodological potential. USE considers the semantic coherence between words and sentences and delivers in-depth insights by providing the original consumer phrasings (see Table 1) instead of abstract word lists and word frequencies as in more simple approaches of automated text analysis, such as dictionary-based approaches or LDA topic modeling.

Additional potential lies in cross-lingual applications using multilingual USE: Researchers can use the same seed sentences in one language and analyze data sets in different languages

to make cross-country comparisons. Analyzing user-generated content, consumer researchers can learn about which product attributes and topics salient to consumers and potentially serve as purchase drivers or barriers. Based on this, consumer typologies and clusters can be derived. An improved understanding of consumers' opinions can support the design of organic products as well as labeling policies. Another application of USE lies in using items of established scales from survey research as seed sentences and analyze their similarity and prevalence in social media data. In addition, the suggested approach could be promising for market monitoring based on the targeted detection of social media content. For example, social media managers can observe the prevalence and development of certain opinions over time.

References

Balducci, B., & Marinova, D. (2018). Unstructured data in marketing. *Journal of the Academy of Marketing Science*, *46*(4), 557–590. https://doi.org/10.1007/s11747-018-0581-x

Berger, J., Humphreys, A., Ludwig, S., Moe, W. W., Netzer, O., & Schweidel, D. A. (2020). Uniting the Tribes: Using Text for Marketing Insight. *Journal of Marketing*, *84*(1), 1–25. https://doi.org/10.1177/0022242919873106

Campbell, J. L., Quincy, C., Osserman, J., & Pedersen, O. K. (2013). Coding In-depth Semistructured Interviews. *Sociological Methods & Research*, *42*(3), 294–320. https://doi.org/10.1177/0049124113500475

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St. John, R., . . . Kurzweil, R. (2018). Universal Sentence Encoder. *ArXiv*. Retrieved from http://arxiv.org/pdf/1803.11175v2

Danner, H., & Menapace, L. (2020). Using Online Comments to Explore Consumer Beliefs Regarding Organic Food in German-Speaking Countries and the United States. *Food Quality and Preference*, *83*(103912). https://doi.org/10.1016/j.foodqual.2020.103912

Hemmerling, S., Hamm, U., & Spiller, A. (2015). Consumption behaviour regarding organic food from a marketing perspective—a literature review. *Organic Agriculture*, *5*(4), 277–313. https://doi.org/10.1007/s13165-015-0109-3

Kozinets, R. V. (2002). The Field Behind the Screen: Using Netnography for Marketing Research in Online Communities. *Journal of Marketing Research*, *39*(1), 61–72. https://doi.org/10.1509/jmkr.39.1.61.18935

Krippendorff, K. (2019). *Content analysis: An introduction to its methodology* (Fourth edition). Los Angeles, London, New Delhi, Singapore: SAGE.

Meza, X. V., & Park, H. W. (2016). Organic products in Mexico and South Korea on Twitter. *Journal of Business Ethics*, *135*(3), 587–603. https://doi.org/10.1007/s10551-014-2345-y

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. Retrieved from http://arxiv.org/pdf/1301.3781v3

Olson, E. L. (2017). The rationalization and persistence of organic food beliefs in the face of contrary evidence. *Journal of Cleaner Production*, *140*, 1007–1013. https://doi.org/10.1016/j.jclepro.2016.06.005

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention Is All You Need. *ArXiv*. Retrieved from http://arxiv.org/pdf/1706.03762v5

Vidal, L., Ares, G., & Jaeger, S. R. (2018). Chapter 6 - Application of Social Media for Consumer Research. In G. Ares & P. Varela (Eds.), *Woodhead Publishing Series in Food Science, Technology and Nutrition. Methods in consumer research* (pp. 125–155). Duxford, United Kingdom: Woodhead Publishing.

Ziegele, M., Breiner, T., & Quiring, O. (2014). What Creates Interactivity in Online News Discussions? An Exploratory Analysis of Discussion Factors in User Comments on News Items. *Journal of Communication*, *64*(6), 1111–1138. https://doi.org/10.1111/jcom.12123