Creating a ‘customer centricity graph’ from unstructured customer feedback

Received (in revised form): 14th December, 2020

Elisabeth Lebmeier
Master’s Student, LMU Munich, Germany

Elisabeth Lebmeier is a master’s student of statistics at LMU Munich, where she is currently working on her thesis about different approaches to aspect-based sentiment analysis. Her research interests include machine learning, deep learning and natural language processing.

Department of Statistics, LMU Munich, Ludwigstr. 33, München 80539, Germany
E-mail: e.lebmeier@gmx.de

Naiwen Hou
Master’s Student, LMU Munich, Germany

Naiwen Hou is a master’s student of statistics at LMU Munich with an economic and social science background.

Department of Statistics, LMU Munich, Ludwigstr. 33, München 80539, Germany
E-mail: hounaiwen@hotmail.com

Korbinian Spann
Managing Director, Insaas, Germany

Korbinian Spann is Managing Director and Head of Data at Insaas, a software company focusing on artificial intelligence and natural language processing. Dr Spann is responsible for the development of dictionaries and the data pipeline.

Insaas GmbH, Floßmannstr. 20, München 81245, Germany
E-mail: korbinian.spann@insaas.ai

Matthias Aßenmacher
PhD Student, LMU Munich, Germany

Matthias Aßenmacher researches natural language processing at LMU Munich, where he is currently studying for a PhD. His main research areas include *inter alia* the comparability as well as possible applications/use cases of large pre-trained language models.

Department of Statistics, LMU Munich, Ludwigstr. 33, München 80539, Germany
E-mail: matthias@stat.uni-muenchen.de

Abstract Certain industries, such as car insurance, do not have many customer touch points and do not offer a great deal of differentiation in the market. Marketers in such industries must therefore analyse vast amounts of customer-generated feedback in order to analyse customer preference in a quantitative manner. At present, this is done via market research or manual work, as an automated tool for summarising unstructured texts is as yet unavailable for certain European languages, including German. This paper discusses how Insaas and LMU Munich have used publicly available feedback on car insurance in Germany to develop a dedicated pipeline for the computation and visualisation of customer opinions. This paper provides an overview of the various steps of the procedure.

KEYWORDS: customer centricity, NLP, AI, dashboard, B2C

INTRODUCTION

Business-to-consumer (B2C) industries, like car insurance companies, have to focus on their customers' needs in order to provide them with the desired product. As touch points between companies and customers are infrequent, companies must get the most they can out of customer feedback. At present, such feedback is mainly found in unstructured texts that are publicly available on the internet, for instance in comparison portals. Star ratings, in particular, are a popular way for consumers to comment on the overall quality of an insurance company. But this is only a very general approach to a complex issue. More differentiating information can be found in the review texts themselves. In order to avoid manual analysis of this vast amount of data, an automated approach for information extraction and visualisation is needed. Working together, Insaas and LMU Munich have developed a multi-step procedure to solve this problem. The solution can detect topics and their polarity and group this information in such a way that customer opinions can be represented in the form of a graph, known as the *customer centricity graph*. This graph helps companies to identify those areas in which areas they perform better than their competitors and where there is room for improvement.

APPROACH

Based on state-of-the-art methods in the field of natural language processing (NLP), Insaas and LMU Munich have developed a pipeline that takes an arbitrary amount of review data as input and compresses that information into a customer centricity graph. This approach is targeted at the car insurance industry and at German review texts. The novelty of this work is the combination of several building blocks to produce a multi-step procedure that can be run automatically.

Its main features are pre-trained German language models from the BERT[1]-family. BERT is a Transformer[2]-based language representation model, ie a model that is trained to represent words in a meaningful way, based on their bi-directional context. As pre-training such models requires massive amounts of computational power (typically Tensor Processing Units) and time, it is common practice to use pre-trained versions of such models. The present research used a variant pre-trained on German texts, so had only to fine-tune the model for the task at hand, ie the classification of reviews with respect to topics or polarities.

In what follows, this paper will provide an overview of the various steps of the procedure (Figure 1). In order to make subsequent explanations easier to understand, two exemplary reviews are added. As a first step, aspects (ie topics) are extracted from the reviews. Aspects highlight different facets

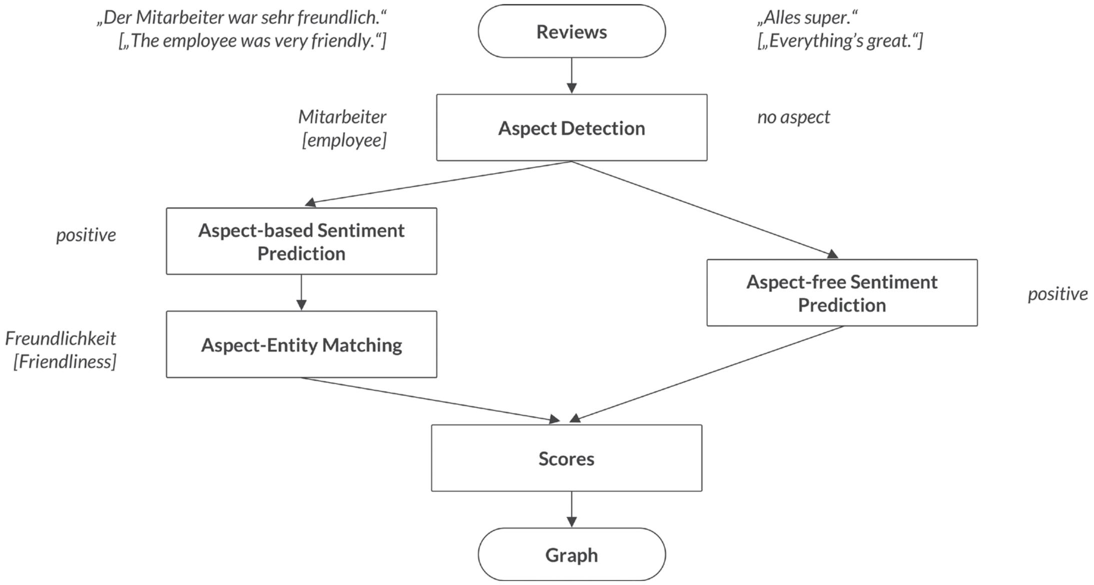

Figure 1: Overview on the steps conducted throughout the pipeline

of a review and may be either explicit or implied; in the latter case, the aspect may be identified from the context. For example, in the case of '*Der Mitarbeiter war sehr freundlich*' ['The employee was very friendly'], '*Mitarbeiter*' ['employee'] is the aspect, while in '*Alles super*' ['Everything's great'], no aspect can be detected. Depending on the results of this first step, the data may be split into two groups: those reviews *with* identified aspects and those *without*.

Aspect-based sentiment analysis is then performed on the data with aspects, meaning that the sentiment, which is basically the emotion or the polarity, is determined separately for each aspect. In the case of '*Mitarbeiter*' ['employee'], the context suggests a positive sentiment for this aspect. Second, the aspects must be matched to their corresponding entities. Entities are categories in which the aspects can be grouped to reduce complexity. They are *a priori* defined to be '*Beratung*', '*Erreichbarkeit*', '*Freundlichkeit*', '*Kompetenz*', '*Qualität*', '*Problemlösung*', '*Preis*' and '*Leistung*', which can be translated as 'guidance', 'availability', 'friendliness', 'expertise', 'quality', 'problem solving', 'price' and 'benefit'. These entities, which can be grouped into either product or service-related, will dominate the definition of the resulting visualisation. In the example, it makes sense to assign '*Mitarbeiter*' ['employee'] to *Freundlichkeit* ['friendliness'], ie entities and sentiments are connected via aspects. For that part of the data without any aspects, a sentiment is predicted for the whole review; this will be called aspect-free sentiment. For the example '*Alles super*' ['*Everything's great*'], this should clearly be positive. All this extracted information is then turned into entity-wise scores by calculating the mean of the sentiments of all aspects belonging to each entity. These are depicted in a radar chart, also known as a spider diagram, with one corner point per entity. For the reviews without any aspects, an aspect-free score is calculated in a similar manner.

THE DATA

The data used for training the different building blocks of the multi-step approach was collected from publicly available web pages such as comparison portals such as Trustpilot (https://de.trustpilot.com/). From the original data set, called the review-wise labelled data, Insaas derived several smaller pieces of data that were targeted for special parts of the pipeline. Short descriptions for these are provided in the following.

Review-wise labelled data

After excluding irrelevant and duplicate reviews, a total of 93,543 samples of data were obtained. Each sample comprised seven variables, namely: feedback, date, source, company, rating, aspect and sentiment. As consumer review text falls under the heading of 'feedback', the present study considers this to be the most important variable. All review texts were written in German and varied in length from just a few words to multiple sentences.

The time stamp in the 'date' variable was initially used solely to identify duplicates. As this paper will discuss, however, including time as an additional dimension in the pipeline can provide an interesting extension to the analysis, hence it was also used to split up reviews by year and thereby create separate graphs.

The variables of 'source' and 'company' are used to indicate the source of the data (ie which comparison portal) and the company being commented upon, respectively. Due to the different sources being used, company names initially differed in spelling and it was necessary to consolidate them in order to group the data by company.

If the source of the review provided star ratings, this information was recorded under the heading of 'rating', on a scale of either 1–5 or 0–10. This information was used by Insaas to correct the predictions of the sentiment model, which predicted the reviews to be 'negative', 'neutral' or 'positive' (stored in the corresponding 'sentiment' column). These sentiments were used as true labels for the aspect-free sentiment prediction.

The 'aspect' variable describes the aspects predicted by the Insaas aspect detection model and serves as the ground truth for aspect prediction.

Aspect-wise labelled data

The second sub data set, which was constructed for the purpose of training one of the building blocks of the pipeline, will be referred to as *aspect-wise labelled data* in order to distinguish it from the *review-wise labelled data*. It was annotated this way because several reviews include more than one aspect. As for aspect-based sentiment classification and aspect-entity matching, the sentiment and the corresponding entity were required for each aspect. This created the need to construct a further data source. Thus, the new labels include aspect-based sentiments and entities for up to three aspects per review. The data comprise a subset of 584 observations of the review-wise labelled data which were manually annotated during the course of the project.

Lemmatisation list

A lemmatisation list was used to efficiently cope with the huge amount of different aspects in the *review-wise labelled data*. Lemmatisation entails grouping inflected words according to their lemma; for example, '*Beiträge*' ['insurance premiums'], '*Beitrages*' and '*Beitrags*' are all assigned to the lemma '*Beitrag*' ['insurance premium'].

Initially, there were over 1,000 different aspects. Not only was this too complex for the model to handle, but there was also the problem of multiple aspects referring to the same underlying construct. To address this, the researchers manually created a list where all aspects were assigned to a so-called *lemmatised aspect*. The approach was extended by grouping words with similar meaning to the lemmatised aspects; for example, aspects like '*Vertragsabschluss*' ['completion of contract'], '*Vertragsformular*' ['contract form'], '*Unterlagen*' ['documents'] and '*Vertragswechsel*' ['change of contract'] were allocated to the lemmatised aspect '*Vertrag*' ['contract']. While this procedure retained the meaning of the aspects, the generalisation made the task less complex. The lemmatisation list held 197 lemmatised aspects, which were utilised during aspect-based sentiment classification and aspect-entity matching.

Entity synonyms

For the task of aspect-entity matching, a list of synonyms for the entities was created. This list was created using ConceptNet,[3] a semantic network that connects potentially related words with one another. For each entity, there were 50–75 synonyms both with respect to meaning, eg '*Hilfsbereitschaft*' ['helpfulness'] for '*Freundlichkeit*' ['friendliness'], as well as spelling, eg '*Qualtiät*' for '*Qualität*'. Also note that some synonyms are not unique for one entity; for example, '*Qualität*' ['quality'] can also be used as a synonym for '*Leistung*' ['benefit'].

PIPELINE BASED ON BERT MODELS

The goal of the project was to create a code pipeline to transform data from one company (serving as input) into a comprehensive visualisation that can be compared with data from other companies. This paper demonstrates the pipeline using data from Allianz and HUK, and will comprehensively discuss the results of each step inside the pipeline.

A total of 10,680 reviews of Allianz and 11,932 reviews of HUK were obtained.

The first step in the pipeline was aspect detection. This entailed training a classifier for so-called multi-label classification, so that reviews may (potentially) be assigned to more than one label (ie aspect). After removing very rare aspects and applying the lemmatisation list, a list of 198 aspects (including a 'no aspect' label) was obtained. For this task, the German DistilBERT[4] model (https://huggingface.co/distilbert-base-german-cased) was employed on a subset of the review-wise labelled data. DistilBERT is a smaller version of BERT that was created to address BERT's memory and computational issues. The key technique for reducing the model size is knowledge distillation, which is discussed in depth elsewhere.[5,6] An in-depth theoretical explanation is beyond the scope of this paper, but the basic idea behind this technique is to train a small(er) *student* model to mimic the predictions of a large(r) *teacher* model.

For the next steps, the data were separated into samples with and without aspects. On the reviews with aspects, aspect-based sentiment classification methods were used to predict one sentiment per aspect. This was necessary as there were reviews like '*Der Mitarbeiter war sehr freundlich, aber der Versicherungsbeitrag zu hoch*' ['*The employee was very friendly, but the insurance premium was too high*', where the sentiments of '*Mitarbeiter*' ['employee'] and '*Versicherungsbeitrag*' ['insurance premium'] contradicted each other. LCF-BERT[7] was selected for this task and trained on the aspect-wise labelled data set, which introduces a local-context-focus (LCF) mechanism. This means that, in order to identify the sentiment of an aspect, an additional focus is set on words that are close to the aspect. The basic BERT model used here was the *bert-base-german-cased* from the huggingface transformers library[8] (https://huggingface.co/bert-base-german-cased). If a review had no predicted sentiment for any aspect, this review was removed from the data set with aspects and added to the one without aspects.

For the task of matching aspects and entities, the list of synonyms for each entity was employed together with FastText[9] embeddings on aspects, entities and entity synonyms. Subsequently, each aspect was paired with all entities as well as entity synonyms. For each pair of embeddings (aspect, entity/entity synonym), the cosine similarity was calculated. In order to obtain an entity for each aspect, the researchers took the ten most similar entities or entity synonyms, looked up the entities of the entity synonyms and chose their mode as the final entity for this aspect. If there was no unique mode, the knowledge from the aspect-entity list that had been extracted from the aspect-wise labelled data set was included. If the entity was still undecided, the entity with the highest similarity was added to the list and another attempt to take the mode was taken. Following this process, there remained two aspects that had no unique entity assigned. For these, the number of entity synonyms was reduced until it was possible to calculate a mode. This list of aspects and entities was used in the pipeline.

If no aspects were found within a review text, it was not possible to employ aspect-based methods for sentiment classification. In such instances, a multi-class classifier was used to predict the review's aspect-free sentiment, thus effectively obtaining the sentiment of the entire review. This was done similarly with aspect detection, but with multi-class instead of multi-label prediction as the target variable contained exactly one label per review. For this task, a German DistilBERT model was fine-tuned on those parts of the review-wise labelled data that were not labelled with any aspects.

To visualise the extracted information, sentiments had to be converted into numbers that could be depicted. The researchers devised multiple scores to deal with special subgroups of reviews and to see which one best showed the results. For reviews without any aspects, aspect-free sentiments were predicted. The absolute values were used to calculate an aspect-free score with the following formula where *review* corresponds to a review without aspects and *N_without_aspects* is the total number of them:

$$aspect\_free\_score = \frac{1}{N\_without\_aspects} \sum_{\substack{review\\without\\aspects}} sentiment(review)$$

Sentiment (review) $\in \{-1; 0; 1\}$ indicates the sentiment of each review, encoded for negative, neutral and positive, respectively. This score is basically the mean of the sentiments with a lower bound of −1 and an upper bound of 1. For Allianz data, the aspect-free scores are 0.5112 and 0.4504 for the years 2016 and 2020, respectively; for HUK data, they are 0.6780 and 0.4268, respectively. This means that the reviews without aspects were more positive in 2016 than in 2020. Comparing both companies, one may observe that HUK obtained a significantly higher value than Allianz in 2016, but that Allianz scored marginally better in 2020.

For the remaining number of reviews with aspects, the researchers calculated a score for each entity. This formula is actually the same as the one for the aspect-free score, but in this case, one takes into account only those sentiments that correspond to the aspects linked to the respective entity. As the connecting point between sentiments and entities is the aspect, the corresponding aspects may be summed as

$$score(entity) = \frac{1}{N\_entity} \sum_{\substack{aspect\\assigned\\to\,entity}} sentiment(aspect)$$

where *N_entity* is the number of aspects assigned to the entity and the *sentiment (aspect)* $\in \{-1; 0; 1\}$ is the sentiment belonging to an aspect of this entity. As these scores do not consider the varying values of *N_entity*, it may also be interesting for future work to include weights to account for this issue in a meaningful way.

THE DASHBOARD

For the final visualisation in the Insaas dashboard, data can be filtered by company

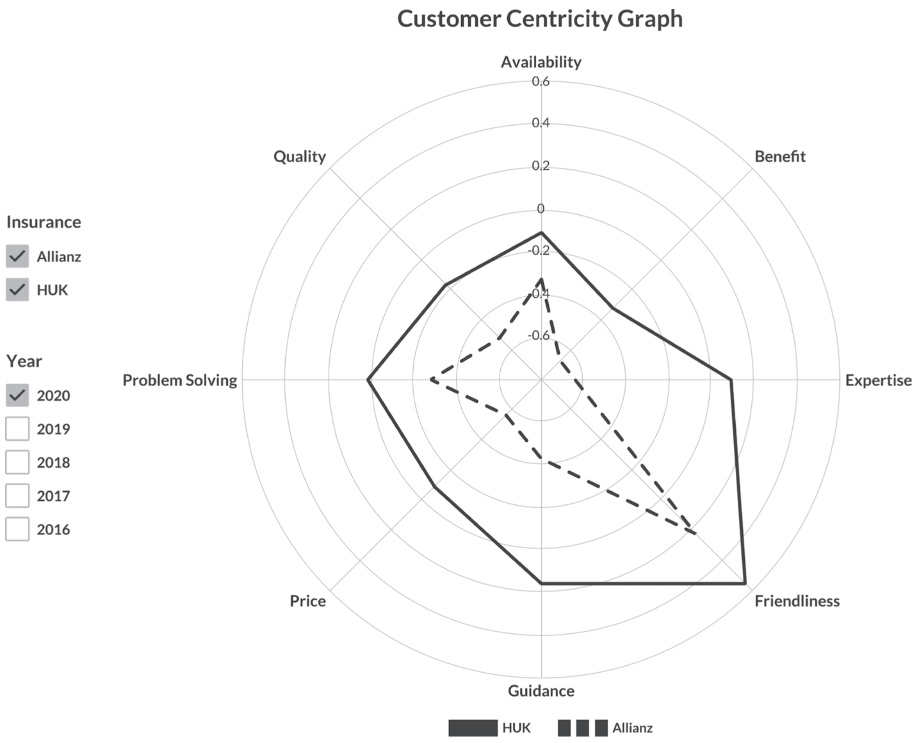

Figure 2: Customer centricity graphs for Allianz and HUK for 2020

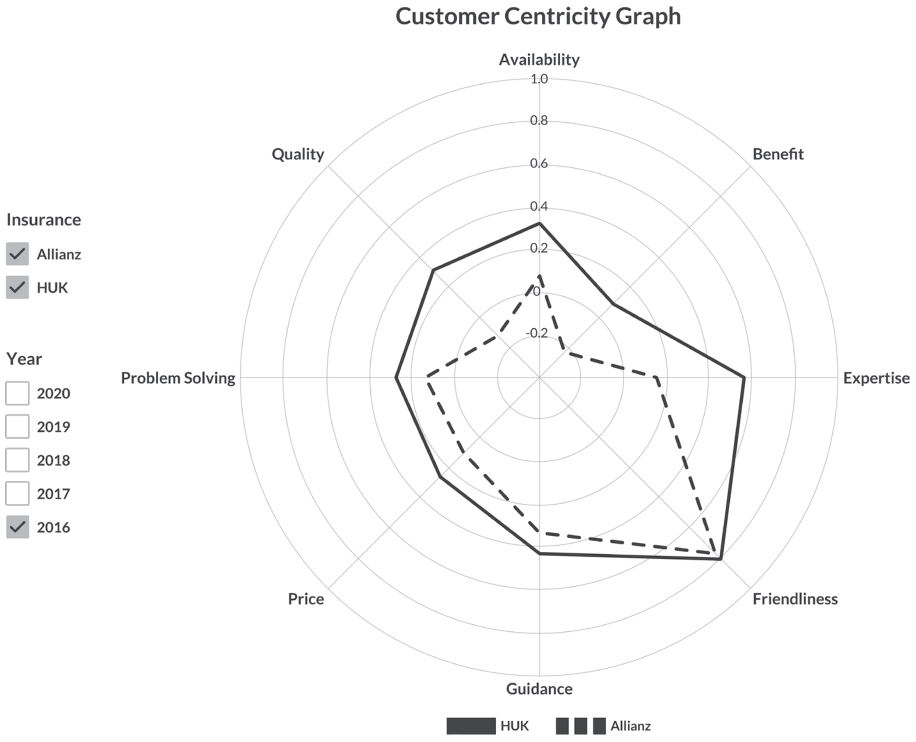

Figure 3: Customer centricity graphs for Allianz and HUK for 2016

and by year. The companies of interest in the present use case were Allianz and HUK. In addition to the entity-wise scores, the dashboard also includes time as a dimension. In Figure 2, which depicts the 2020 customer centricity graphs for Allianz and HUK, one can clearly see that HUK receives higher scores compared with Allianz with respect to all entities. Nevertheless, it is interesting that they both receive the highest values for '*Freundlichkeit*' ['friendliness']. This is also the only entity for which Allianz has reached a positive value, unlike HUK, which obtained a positive value for three entities.

By comparison, Figure 3 shows customer centricity graphs for both companies for the year 2016. Already back then, '*Freundlichkeit*' ['friendliness'] was the highest ranked entity with respect to the sentiment, but besides this, many things appear to be different. Both companies had far better ratings in 2016: while HUK obtained positive sentiment scores for all entities, Allianz did so for all bar two. These year-wise scores can be used to evaluate the impact of certain changes, for example in customer service. Note that the scales differ between 2016 and 2020, as the dashboard automatically adjusts its scaling according to obtained scores.

In Figure 4, absolute frequencies of the sentiments per entity (aggregated over all years) showcase yet another visualisation option of the versatile dashboard. The right side shows values for a company of interest (here, Allianz), while on the left, a so-called 'industry benchmark', consisting here of HUK and Allianz, serves for comparison. Note that the user can configure the composition of the industry benchmark by checking or unchecking the respective boxes. On the x-axis, the total amounts of the predicted sentiments are displayed, scaled to a similar width in order to allow for better visual comparability. Clearly one can see that on both sides '*Kompetenz*' ['expertise'] receives the lowest absolute frequency of sentiments whereas 'Beratung' ['guidance'] is discussed most frequently. These quantities must also be taken into account when interpreting the customer centricity graphs in Figures 2 and 3 as they make entity-wise scores more or less reliable.

CONCLUSION

This study has described the development of an automated approach for the analysis and visualisation of customer opinions from feedback texts that employs state-of-the-art methods from the field of natural language processing. Nevertheless, there are still several issues that could be improved. First of all, each of the steps can potentially perform better. In particular, a context-based approach may be applied to take the aspect-entity matching to the next level. Furthermore, the amount of entities, as shown in Figure 4, could be added to the customer centricity graph, for example by adjusting the angles of the entities according to their proportion of all entities. Another way of visualisation could be to take the height as a new dimension of the radar chart. The higher a score is placed in this dimension, the more entities it is based on.

Despite these issues, the researchers have established a working infrastructure for extracting valuable and differentiating information from review data. As the pipeline is built in a modular fashion, its building blocks can be easily modified or improved without the need to change everything else.

With respect to developing this research, future studies could integrate the quality of each review into the scores. Following the hypothesis that reviews with good grammar and spelling show a more fine-grained and reliable opinion, this might obtain interesting new results. This information could be added in the form of weights. Emojis and emoticons are also related to the style of writing. These can be used to further improve the sentiment predictions.

Another idea would be to extend the set of entities, as the given set of eight

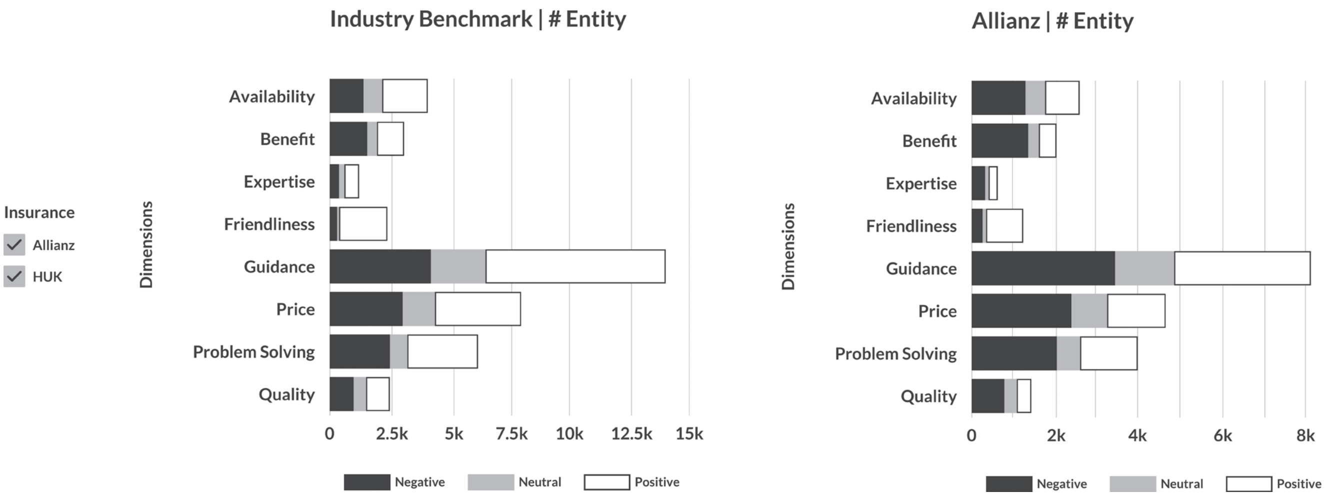

Figure 4: Sentiment frequencies per entity for Allianz data versus an industry benchmark built from HUK and Allianz data; numbers are summed up over all years

entities is not always sufficient to categorise all the various categories that people discuss. As such, it might be of benefit to add new entities, for example from the field of marketing and sales. Generalising these entities may also make the approach applicable to other business sectors.

References

1. Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019) ‘BERT: Pre-training of deep bidirectional transformers for language understanding’, in ‘Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, 2nd–7th June’, Vol. 1, pp. 4171–4186.
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. (2017) ‘Attention is all you need’, *Advances in Neural Information Processing Systems*, Vol. 30, pp. 5998–6008.
3. Chen, M., Tian, Y., Yang, M. and Zaniolo, C. (2016) ‘Multilingual knowledge graph embeddings for cross-lingual knowledge alignment’, arXiv preprint arXiv:1611.03954.
4. Sanh, V., Debut, L., Chaumond, J. and Wolf, T. (2019) ‘DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter’, arXiv preprint arXiv:1910.01108.
5. Buciluǎ, C., Caruana, R. and Niculescu-Mizil, A. (2006) ‘Model compression’, in ‘Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, 20th–23rd August’, pp. 535–541.
6. Hinton, G., Vinyals, O. and Dean, J. (2015) ‘Distilling the knowledge in a neural network’, arXiv preprint arXiv:1503.02531.
7. Zeng, B., Yang, H., Xu, R., Zhou, W. and Han, X. (2019) ‘LCF: A local context focus mechanism for aspect-based sentiment classification’, *Applied Sciences*, Vol. 9, No. 16, p. 3389.
8. Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q. and Rush, A. M. (2020) ‘Transformers: State-of-the-art natural language processing’, in ‘Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 16th–20th November’, pp. 38–45.
9. Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. (2017) ‘Enriching word vectors with subword information’, *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 135–146.