Aug 10th, 12:00 AM

Convolutional Neural Networks for Survey Response Classification

Nikolai Stein
Julius-Maximilians-University, nikolai.stein@uni-wuerzburg.de

Felix Oberdorf
Julius-Maximilians-University, Felix.Oberdorf@uni-wuerzburg.de

Jonas Pirner
Julius-Maximilians-University, jonas.pirner@web.de

Convolutional Neural Networks for Survey Response Classification

Completed Research

Nikolai Stein
University of Wuerzburg, Germany
nikolai.stein@uni-wuerzburg.de

Felix Oberdorf
University of Wuerzburg, Germany
felix.oberdorf@uni-wuerzburg.de

Jonas Pirner
University of Wuerzburg, Germany
jonas.pirner@stud-mail.uni-wuerzburg.de

Abstract

Artificial Intelligence reveals great potential for enterprises e.g., intelligent services. However, small and medium enterprises struggle with Artificial Intelligence due to limited resources. Especially tasks such as survey response classification are yet not investigated. We address this research gap by means of a data science study. In particular, we analyze several baseline classification pipelines leveraging logistic regression, random forests, and linear support vector machines against wide headed CNN architectures with one-hot encoding or character embedding inputs. We find that the SVM model outperforms all other evaluated models in the setting at hand. In addition, we analyze the different predictions of the models and show typical prediction errors by means of a chord diagram of commonly misclassified brands.

Keywords

survey classification, open answer, CNN

Introduction

The advent of Artificial Intelligence has led to a variety of intelligent services, reinvented business models, and re-engineered business processes across all major industries. Yet, especially small and medium-sized enterprises (SMEs) often struggle to utilize AI due to a lack of monetary resources and knowledge.

Today, most AI applications rely on the increasing availability of large image and sensor data-sets from manufacturing processes (Stein et al. 2018), social graphs (Zhang et al. 2018), ERP and CRM systems (Najafabadi et al. 2015), and text data (Zhang et al. 2015). Traditionally, large amounts of text data are collected by conducting surveys. In the past, the resulting data helped institutions, such as governments, to investigate ongoing social problems in the population (Groves et al. 2004). Nowadays, survey data is used more broadly to formulate market strategies, to evaluate the performance of new and existing products (Rea and Parker 2012), and for scientific research. The proliferation of crowdsourcing and survey platforms, such as SurveyMonkey and Amazon MTurk, significantly lowers the cost of conducting surveys and leads to a rapid increase of available survey data (Bentley et al. 2017; Mullinix et al. 2015). However, most surveys are performed by highly specialized companies that lack the manpower to process and analyze the data manually[1]. Therefore, researchers, as well as practitioners, have to find efficient algorithms to leverage the increased availability of survey data. Algorithms can be used to partially or fully automate individual stages of the analytics process while or after the survey has been conducted.

In this study, we focus on the question of customer brand awareness, a topic of high relevance for SMEs. Agostini et al. (2015) show by quantitative analysis that high brand awareness (e.g. corporate trademarks)

[1]Many survey solution companies are SMEs (www.quirks.com/articles/25-top-survey-solutions-companies)

has the potential to influence SMEs' sales positively. To assess their brand awareness, companies typically carry out open-ended surveys (Alimen and Cerit 2010; Marquardt et al. 2011) to capture all possible customer responses. However, open answers are not unambiguously as people will have typos and use different abbreviations for the same brands. A common way of dealing with this problem is to use human coders. After the survey has been conducted and the results have been saved, coders will try to assign the answers to predefined classes, a process referred to as survey-coding. This process induces high costs as brand awareness studies are typically conducted multiple times (e.g., before and after a marketing campaign).

The goal of this study is to develop and evaluate an information system that assists human coders (as visualized in Figure 1) and significantly reduces the manual effort required to perform brand awareness studies. Therefore, we cooperate with a medium-sized (~150 employees) Swedish market research company. This company provides us with historic open answer survey data with predefined human coded classes. We use these classes as labels and the open answer survey data as input for comparing multiple classification algorithms. To this end, we compare state of the art machine learning algorithms[2] with convolutional neural network (CNN) models. Within the presentation of the regarding automated framework, we especially highlight prediction certainty results as well as commonly misclassified brands by a chord diagram.

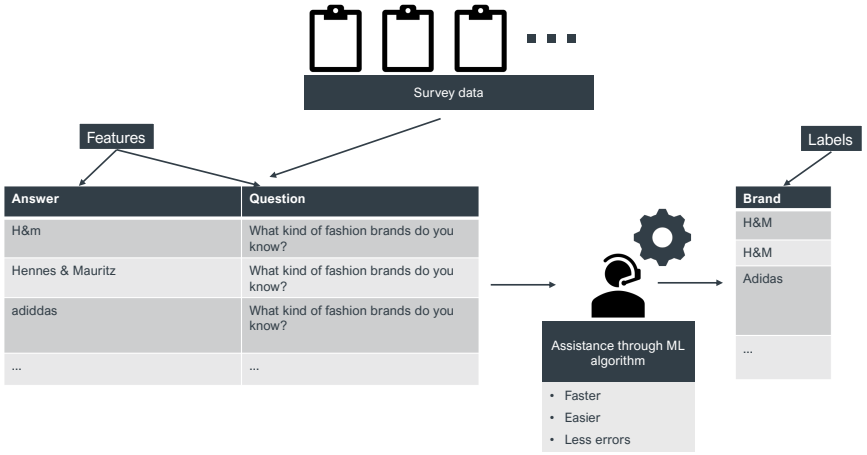


Figure 1. Survey response classification

Related Work and Research Methodology

Due to its high relevance, the problem of automating the survey coding process has been the subject of research for over two decades. Thus, different methods have been established, which can mainly be divided into rule-based systems and machine learning based approaches.

Rule-based systems: Initially, authors relied on rule-based systems to automate the coding process. Viechnicki (1998) use a pattern-matching strategy as well as a probabilistic method to tackle the problem at hand. Macchia and Murgia (2002) suggest to manually define a dictionary to check if answers contain words that associate them with the code. The drawbacks of such rule-based methods are that they do not scale well and require manually updated rules as soon as there are changes in the coding scheme (e.g., new code categories, new questions).

Machine learning: In recent years, the advent of machine learning in general and deep learning, in particular, has enabled ample new possibilities to automate the survey coding process and overcome the drawbacks of rule-based systems. Giorgetti and Sebastiani (2003) test a Naïve Bayes and a SVM classifier on a corpus of social surveys carried out by the US National Opinion Research Center (NORC). They used the same data-sets as Viechnicki (1998), and their experimental results showed significant improvements (+18% for Naïve Bayes and +26% for SVM on average per data-set). Similar approaches have been developed and

[2]Logistic regression, Random Forest, or Support Vector Machine

applied by various researchers (Mantecon et al. 2018; Schonlau and Couper 2016; Spasic et al. 2018). While deep neural networks have outperformed other machine learning approaches in almost all fields of natural language processing (Conneau et al. 2016; Devlin et al. 2018; Lai et al. 2015; Wang et al. 2017), they have–to the best of our knowledge–not yet been used in the field of survey response classification.

We want to address this research gap utilizing a data science study following the guidelines for applying big data analytics put forward by Müller et al. (2016). Correspondingly, we structure the following analysis along with the three phases:

Data collection: We use a large data-set provided by a Swedish market research company. The data-set contains over 150,000 question-answer pairs collected across multiple brand awareness surveys. Besides, the data-set provides information on the respondents (e.g., age, gender).

Data analysis: We develop distinct feature engineering pipelines to preprocess the questions as well as the answers. Subsequently, we train and evaluate multiple machine learning models to classify survey responses. Leveraging the features, we train logistic regression, random forest, and support vector machine models as baseline models. Additionally, we train two different two-headed (question-head and answer-head) convolutional neural networks on the raw data.

Result interpretation: We evaluate and compare the performance of the different models. To this end, we find mixed results. On the one hand, the two-headed CNNs are not able to outperform the best baseline models in terms of accuracy, precision, and recall. However, the CNNs provide more confident classifications in terms of predicted probabilities. While the baseline models often predict classes with low confidence, the CNNs usually provide high certainty levels.

Data Collection and Exploratory Data Analysis

The data-set is obtained by multiple surveys performed by the cooperating Swedish market research company. The objective of the surveys was to measure brand awareness among the respondents. By conducting the survey, the interviewee can collect a wide variety of data that will later be used for analysis, especially when open-ended questions are included. This data may be directly related to the participants (e.g. age of the participant) or relate to the behavior of the persons during the survey (e.g. average response time).

We apply an exploratory data analysis (Tukey 1977) and especially focus on textual data, with the proposed survey questions and corresponding answers as features as well as the–correct–brand as the label. An example of the data-set, which consists of about 150k observations, is presented in Table 1.

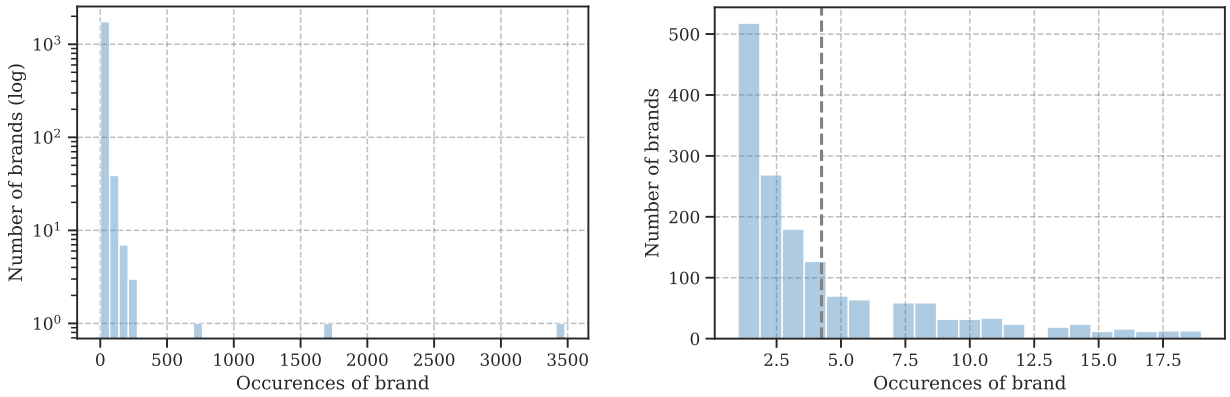
Question	Answer	Brand
Do you know where on the Internet you book trips with MTR Express?	Mtrexpress.se	MTR Express
Which banks do you know?	Länsförsäkringar	Länsförsäkringar
What broadband providers have you noticed during the last 30 days?	Never bought	No // general code
Which companies that issue consumer loans do you know of or have heard of?	icabanken	ICA Banken

Table 1. Data-set sample with question and answer (input features) as well as brand columns (labels).

The first column consists of 423 unique questions, with the focus of exploring customer's brand association with a given subject. The particular subject ranges from fashion brands to airlines. In the answer column, the respondent's input is stored, with 9318 unique entities in contrast to 1806 unique brand labels. The significantly higher number of unique values in the answer column shows that customer answers are not unambiguous and emphasizes the key challenge for classification.

A statistical analysis of the data-set reveals that 84% of the answers are duplicates. This is attributed to the question-answer pairs that contain common and right spelled answers. Probably, the high amount of duplicates affects the classification task by limiting the neural network's generalization. In this study, our focus is a general approach for survey response classification; hence we drop duplicate values. As our objective is the correct brand classification, thereby arises a multiclass (1806 brand labels) classification problem, which implies frequently recurring classes.

For the model training, we remove duplicates to avoid problems associated with high-class imbalances. However, this does not guarantee that each answer for a brand only occurs once (Figure 2). Cases, where the same answer is given under different questions, are not affected by duplicate removal. A brand could thus end up in both train and test set and lead to overfitting. Hence we introduce the out-of-vocabulary (OOV) set. The OOV set only contains the observations whose answers occur exclusively in the test data-set. For these 1600 occurring observations, the model must find general properties of the class to perform a correct classification.



(a) with outliers **(b) without outliers**

Figure 2. Occurrences for classes with outliers (a), which correspond to classes that are used to summarize undesired answers (such as answers of type "I don't know") and a cleaned version without outliers (b).

Models for Brand Classification

Baseline Model

We construct the baseline model within a sklearn pipeline (Figure 3). Therefore, the pipeline sequentially applies a list of transformations as preprocessing, feature transformation[3], modeling and regularization. The last element of the pipeline is a final estimator. To this end, we use logistic regression (LR), random forest (RF), and a linear support vector machine (SVM). This approach results in a modular machine learning pipeline that is comprehensible and transparent.

We leverage the pipeline to execute an individual hyperparameter optimization for each estimator. As visualized in Figure 3, we find that in the answer pipeline tokenization of characters performs better than tokenization of words, for all evaluated estimators. We assume that this is because answers often contain typos. In contrast, the questions are spelled correctly, and we can use word tokenization in the question pipeline. The relatively small initial vocabulary sizes allow us to consider several n-gram variations without causing significant memory problems. We find that using all n-grams in the range of 1-4 yields the best results for all models. Additionally, we find that different parameters have to be chosen depending on the estimator for the remaining transformations (count method and transformation method).

[3]We integrate a bag-of-words approach, one-hot encoding, term frequency-inverse document frequency as well as word embeddings as feature transformations.

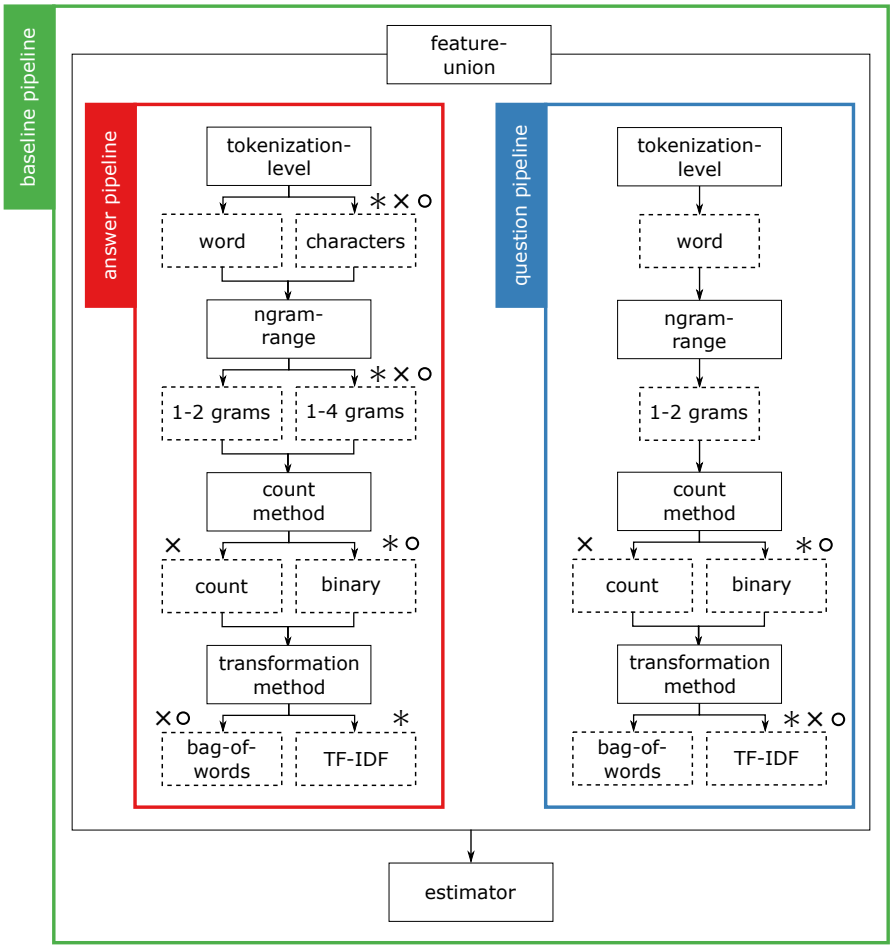


Figure 3. Hyperparameter tuning of baseline pipelines with highlighted best results for logistic regression (∗), random forest (×) and SVM (○).

CNN Model

We developed our CNN architecture within an explorative neural network modeling process and find that a **wide** CNN-structure (Wd-CNN) outperforms **hierarchically** and **dilated** network structures. Again, we apply hyperparameter tuning to find well suited model parameters (solid boxes)–the number of filters and the kernel size (Zhang and Wallace 2015))–to the Wd-CNN with the respective hyperparameter options (dashed boxes). To account for additional non-linearity, we follow Rao and McMahan (2019) and included kernels with size 1 as network-in-network connections. After the Wd-CNN and concatenation layers, the fully connected (200 dense units) and output (369 dense units) layers, are frozen and not changed during the training. For the training, we apply a learning rate of 0.0001 and an ADAM optimizer for a maximum of 2000 epochs. In addition to the hyperparameter optimization, we compare one-hot encoding (OHE) and pre-trained character embedding (CE) as input types and highlight each model's best hyperparameter combination (Figure 4).

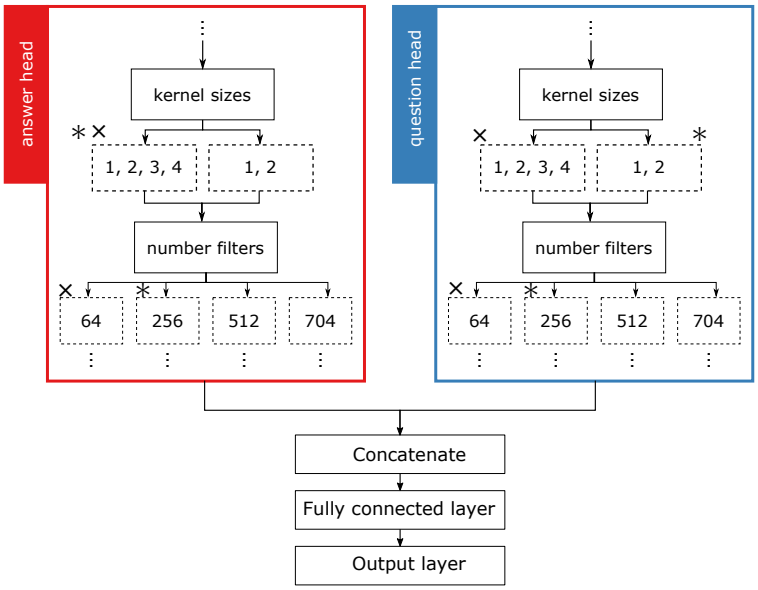


Figure 4. Hyperparameter tuning for the wide CNN architectures with highlighted best results for answer head character embeddings (∗) and one-hot encoding (×).

Evaluation

For the evaluation, we separate the data-set in train- and test-data (training 77%, test 33%) and perform a 3-fold cross-validation. Based on this data-sets, we train the models and compare the accuracy, precision, and recall metrics. Subsequently, we evaluate the certainty of the predictions.

Model performance

The results for the baseline pipeline (Table 2) reveal that the chosen baseline models are by far more powerful than techniques, only predicting the most common class. The **Base SVM** pipeline outperforms the other ones for all observed metrics; hence it serves as a natural candidate to benchmark our CNN approaches.

	Accuracy		Precision		Recall	
	test (%)	OOV (%)	test (%)	OOV (%)	test (%)	OOV (%)
Base LR	93.1	87.5	94.6	82.6	87.9	77.9
Base RF	95.3	89.8	96.7	86.3	93.6	83.0
Base SVM	**96.2**	**91.7**	**97.0**	**89.4**	**95.9**	**87.2**

Table 2. Baseline results for Logistic regression (LR), Random forrest (RF), Support vector machine (SVM) and test data scenarios.

Besides the baseline pipeline, we evaluate the Wd-CNN for both one-hot encoding (OHE) and character embedding (CE) input types. The results are shown in Table 3, where we depict that the two CNN architectures are not able to outperform the Base SVM. The only metric that could be improved by the Wd-CNN CE is the precision. Especially on the OOV set, the Base SVM provides much better results, indicating that it generalizes better than any of the CNN architectures.

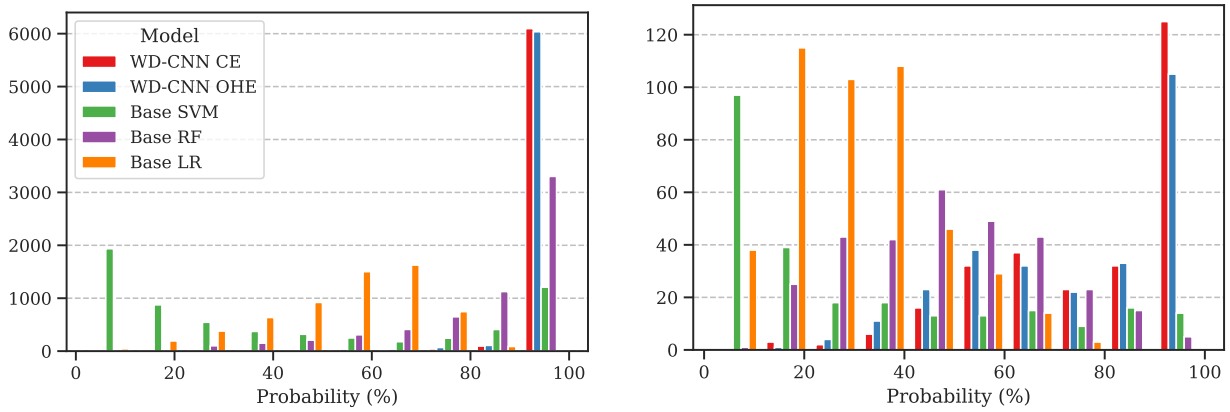
	Accuracy		*Precision*		*Recall*	
	test (%)	OOV (%)	test (%)	OOV (%)	test (%)	OOV (%)
Base SVM	**96.2**	**91.7**	97.0	**89.4**	**95.9**	**87.2**
Wd-CNN OHE	95.9	89.2	97.0	85.8	95.1	84.8
Wd-CNN CE	95.8	88.7	**97.2**	85.5	94.5	82.9

Table 3. Final test results for wide CNN for one-hot encoding (OHE), character embedding (CE) and test data scenarios compared to the Base SVM.

Model certainty

Next, we discuss prediction certainty and the context of commonly confused brands for both baseline models and CNN architectures based on the respective predictions. We determine each prediction by the highest class probability, which is seen as a measure of prediction confidence.

We analyze the respective frequencies for right and wrong predictions (Figure 5) with the result of the CNN architectures adequately expressing high certainties. Whereas the baseline models have correct right predictions, the probability is usually under-confident.



(a) right predictions **(b) wrong predictions**

Figure 5. Prediction certainty of baseline models and CNN architectures for right (a) and wrong (b) predictions on test set.

By additionally analyzing the wrong prediction, we depict that CNN's confidence comes at a price. Even for wrong predictions, the CNNs are confident about the decisions. Actually, the number of those observations is rather low, and we investigate this by analyzing commonly confused brands in a chord diagram (Figure 6).

In the chord diagram, we present the predictions of WD-CNN OHE with the finding that for many brands, the class *No // general code* ("nej // generell kod") is mispredicted. This is because the class is a summary of denial answers. Besides, we see some brands which usually can only be distinguished by context as *Norwegian* vs. *Norwegian bank* or *ICA* vs. *ICA banken*, where the brand is just specified more precisely. Another form of confusion happens if the brands are conceptual similar as *TV10* vs. *Kanal 11* or *British Airways* vs. *American Airlines*. A few occasions include brands with two different labels for the same brand, which indicates a fault of data management.

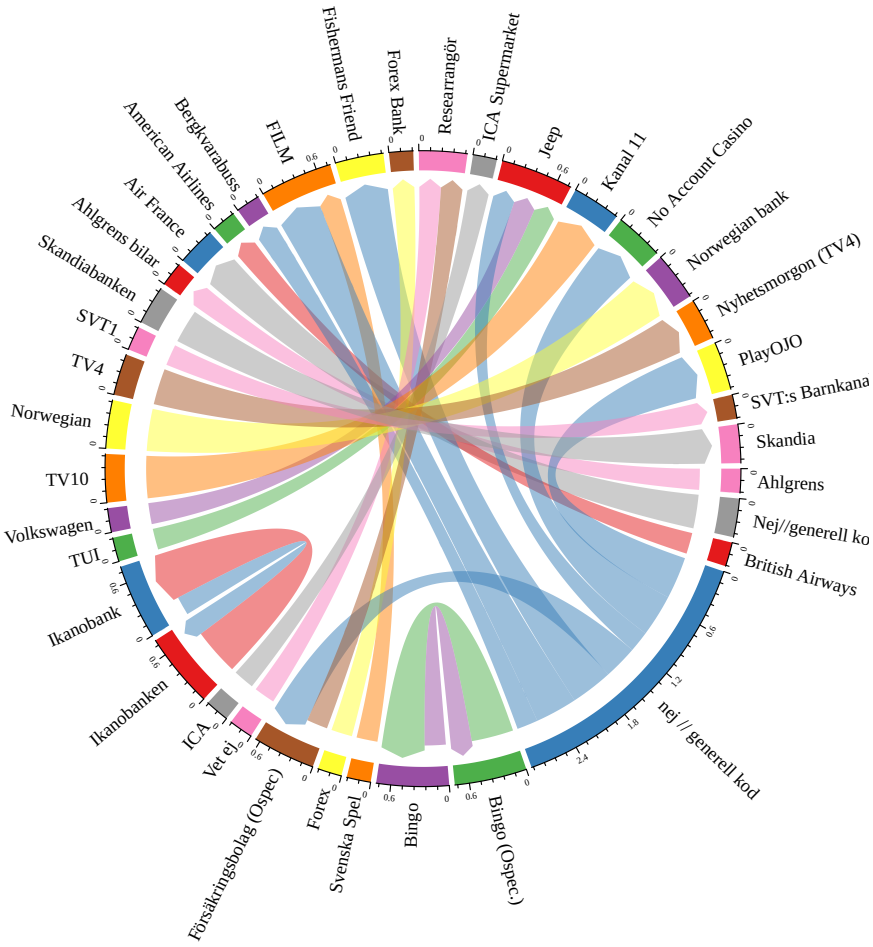


Figure 6. Chord diagram of commonly misclassified brands

Finally, we want to mention inherent ambiguity as a source of mispredictions. The reason therefore is in the nature of open-ended survey questions as participants could answer duplicated letters as airlinesss. Obviously, this highlights that survey inconsistencies lead to inconsistent results, and hence clear guidelines are necessary.

Conclusion and Outlook

SMEs' limited resources are one reason for their struggle with the utilization of AI. Especially in market research and open-ended surveys, AI reveals massive potential in automated survey-coding.

Our study's goal is the development and evaluation of an information system that assists human coders by reducing the manual effort required in brand awareness studies. Therefore we apply deep neural networks in the field of survey response classification. For addressing this research gap, we structure our research by

the guidelines for applying big data analytics put forward by Müller et al. (2016).

Based on the exploratory data analysis, we compare several baseline pipelines leveraging logistic regression, random forest, and linear support vector machine classifiers against wide headed CNN architectures with one-hot encoding or character embedding inputs. We tune a subset of hyperparameters of each model and evaluate them by comparing accuracy, precision, and recall on train-test-splitted data-sets as well as the introduced OOV-data-set. Finally, we depict that the Base SVM model outperforms all models–including CNN architectures–for most of the metrics. We further analyze this by comparing the prediction certainties for right and wrong predictions, where we depict that CNN architectures have a high certainty, for right as well as for wrong predictions. In addition, we present and analyze a chord diagram of commonly misclassified brands. Thereby limitations as data inconsistencies are discussed.

We showcase the development and the application of a pipeline-based approach for reducing manual effort for open answer survey classification by means of a single case study. Future work should evaluate the different estimators implemented in our pipeline on new data-sets to assess the robustness of our findings. However, we are confident that the proposed pipeline can be leveraged in new settings due to its dynamic approach. Additionally, future work should analyze how the human coders leverage an information system supporting their work.

Our research contributes to methodology by introducing a comprehensible and transparent approach for survey response classification. Thereby, we particularly highlight the outperforming results of the Base SVM vs. CNN-architectures, and present approaches for detailed prediction certainty and misclassification analysis. To this end, our research allows the utilization of AI in research and enterprises–especially SMEs.

REFERENCES

Agostini, L., Filippini, R., and Nosella, A. (2015). "Brand-Building Efforts and Their Association with SME Sales Performance," *Journal of Small Business Management* (53:S1), pp. 161–173. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jsbm.12185.

Alimen, N. and Cerit, A. G. (2010). "Dimensions of brand knowledge," *Journal of Enterprise Information Management* ().

Bentley, F. R., Daskalova, N., and White, B. (2017). "Comparing the reliability of Amazon Mechanical Turk and Survey Monkey to traditional market research surveys," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems,* pp. 1092–1099.

Conneau, A., Schwenk, H., Barrault, L., and Lecun, Y. (2016). "Very deep convolutional networks for natural language processing," *arXiv preprint arXiv:1606.01781* (2).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805* ().

Giorgetti, D. and Sebastiani, F. (2003). "Automating survey coding by multiclass text categorization techniques," *Journal of the American Society for Information Science and Technology* (54:14), pp. 1269–1277.

Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E., and Tourangeau, R. (2004). *Survey Methodology,* Wiley Series in Survey Methodology. Wiley.

Lai, S., Xu, L., Liu, K., and Zhao, J. (2015). "Recurrent convolutional neural networks for text classification," in *Twenty-ninth AAAI conference on artificial intelligence,*

Macchia, S. and Murgia, M. (2002). "Coding of textual responses: Various issues on automated coding and computer assisted coding," in *Proceedings of JADT-02, 6th international conference on the statistical analysis of textual data, St-Malo, FR,* pp. 471–482.

Mantecon, J. G. A., Ghavidel, H. A., Zouaq, A., Jovanovic, J., and McDonald, J. (2018). "A Comparison of Features for the Automatic Labeling of Student Answers to Open-Ended Questions." *International Educational Data Mining Society* ().

Marquardt, A. J., Golicic, S. L., and Davis, D. F. (2011). "B2B services branding in the logistics services industry," *Journal of Services Marketing* ().

Müller, O., Junglas, I., Brocke, J. v., and Debortoli, S. (2016). "Utilizing big data analytics for information systems research: challenges, promises and guidelines," *European Journal of Information Systems* (25:4), pp. 289–302.

Mullinix, K. J., Leeper, T. J., Druckman, J. N., and Freese, J. (2015). "The generalizability of survey experiments," *Journal of Experimental Political Science* (2:2), pp. 109–138.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., and Muharemagic, E. (2015). "Deep learning applications and challenges in big data analytics," *Journal of Big Data* (2:1), p. 1.

Rao, D. and McMahan, B. (2019). *Natural Language Processing with PyTorch: Build Intelligent Language Applications Using Deep Learning,* O'Reilly Media.

Rea, L. and Parker, R. (2012). *Designing and Conducting Survey Research: A Comprehensive Guide,* Wiley.

Schonlau, M. and Couper, M. (2016). *Semi-automated categorization of open-ended questions. Doi. Org*.

Spasic, I., Owen, D., Smith, A., and Button, K. (2018). *Closing in on open-ended patient questionnaires with text mining*.

Stein, N., Meller, J., and Flath, C. M. (2018). "Big data on the shop-floor: sensor-based decision-support for manual processes," *Journal of Business Economics* (88:5), pp. 593–616.

Tukey, J. W. (1977). *Exploratory Data Analysis,* Addison-Wesley.

Viechnicki, P. (1998). "A performance evaluation of automatic survey classifiers," in *Grammatical Inference, 4th International Colloquium, ICGI-98, Ames, Iowa, USA, July 1998, Proceedings,* vol. 1433. Lecture Notes in Artificial Intelligence. Springer, pp. 244–256.

Wang, J., Wang, Z., Zhang, D., and Yan, J. (2017). "Combining Knowledge with Deep Convolutional Neural Networks for Short Text Classification." in *IJCAI,* pp. 2915–2921.

Zhang, X., Zhao, J., and LeCun, Y. (2015). "Character-level convolutional networks for text classification," in *Advances in neural information processing systems,* pp. 649–657.

Zhang, Y. and Wallace, B. C. (2015). "A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification," *CoRR* (abs/1510.03820). arXiv: 1510.03820.

Zhang, Z., Cui, P., and Zhu, W. (2018). "Deep learning on graphs: A survey," *arXiv preprint arXiv:1812.04202* ().