Textual Descriptions Used for Classification of Oaked vs Unoaked Wines

Ronald Böck[1], Siddarth Venkateswaran[1], Thi Nguyen[2] and Dominik Durner[2]

Abstract: Winemaking and grapegrowing are sciences with a long tradition dealing with one of the most complex beverages in the world. This complexity stems from the winemaking process itself as well as the characteristics of the final product. Wine's aroma is often described through scalar assessments, though here we are focusing on textual descriptions, transferring methods from the natural language processing (NLP) community to the wine domain, in particular to analyse the statements of human panellists. Textual descriptions were used for the classification of oaked versus unoaked wines as an initial demonstration of NLP in the wine domain. We achieved significant discrimination results of 0.79 F1-score comparing BERT and Naïve Bayes classifiers. This shows that more natural textual (and potentially spoken) descriptions of wine, being later combined with classical scalar assessments, can provide more flexibility to human panellists.

Keywords: Wine Descriptors, Classification, BERT, NLP

1 Introduction

Viticulture and oenology, the sciences of grapegrowing and winemaking, have a long tradition dealing with one of the most chemically and sensorially complex beverages in the world. The many volatile compounds contributing to wine aroma can be measured by instrumental means, such as gas chromatography, though an accurate prediction of how wine will actually smell based on its chemical composition is not yet possible [FdS22]. Therefore, the responsibility of describing wine aroma currently lies with human panellists taking part in formal sensory evaluations [BGT21].

For the PINOT[3] project, we are developing a multimodal, artificial intelligence (AI)-based approach to make assessments of wine aroma more objective. As can be seen in Figure 1, there are many points along the wine production and supply chain where AI can potentially assist (also in terms of sustainability), ranging from the prediction of wine quality based on the starting grape material, to matching wines to consumers' tastes and expectations. Here, we are focussing on the application of AI toward describing the finished product, assisting in the sensory evaluation of wine aroma. As a first step, we sought to determine whether AI

[1] Genie Enterprise, Research Division, Donnersbergweg 1, 67059 Ludwigshafen, Germany {rboeck,venkat}@genie-enterprise.com
[2] Weincampus Neustadt, Breitenweg 71, 67345 Neustadt an der Weinstraße, Germany {thi.nguyen,dominik.durner}@dlr.rlp.de
[3] https://pinot-ai.com

can be used to extract a common "understanding" of wine's properties based on textual descriptions generated by panellists during sensory evaluation.

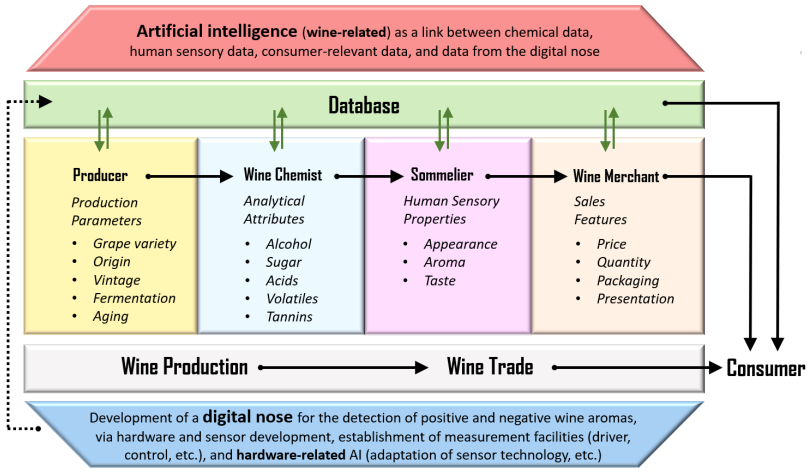

Fig. 1: Scope graphic highlighting the production chain in wine making. Additionally indicating the particular process' stakeholders. This current core is being framed by the options of AI support on an either hardware- or wine-related perspective.

1.1 Motivation and Research Aspect

The purpose of this investigation was to transfer Natural language processing (NLP) techniques to the winemaking domain and to evaluate the performance of two approaches to text-based wine assessment, relying on large neural networks (Bidirectional Encoder Representation from Transformers (BERT)) and on classical machine learning (Naïve Bayes). Their ability to distinguish between oaked and unoaked wines based on textual descriptions collected during sensory evaluation by wine professionals was compared, aiming for a possible combination with classical scalar assessments, thereby providing more flexibility to the human panellists.

1.2 Related Work

Until recently, sensory evaluation has conventionally relied on quantitative methods resulting in numerical descriptions of flavour; considered by many as "gold standard" in sensory evaluation, descriptive analysis yields intensity values for various aroma, taste, and mouthfeel properties, as rated by trained panellists on scales. Alternatively, check-all-that-apply methodology generates the counts/frequencies at which certain descriptors are chosen by panellists to describe the product [BGT21]. While such data is relatively simple to

process and analyse, panel training (to reach consensual understanding of descriptors) is often time-consuming and costly. There is also the risk of forgetting to include important attributes, and the risk of panellists feeling imposed/limited by the list of attributes [Vi23]. However, this landscape is changing, with more rapid, cost-effective, and consumer-oriented methods being developed, involving open questions and allowing free responses [Pi15]. Such methodologies have only recently been implemented in the field of oenology for wine analysis [La13; Ma20; Vi23]. Panellists are asked to describe wines in their own words, and while the text generated is much more natural and richer in information, the manner in which it must be pre-processed for analysis is far from straightforward [Ma21]. Advancements in technology, particularly in machine learning, have facilitated the automation of a once tedious process [Vi20].

Classical text classification techniques involve multiple pre-processing steps, creating an embedding on these cleaned texts using approaches like Term Frequency Inverse Document Frequency (TF-IDF), and feeding them to a classification algorithm (e.g. [IKT05]). One drawback of this method involves the curse of dimensionality, which arises due to an increase in the amount of data and features over a period of time. To handle this issue, feature reduction techniques like Global Vectors for Word Representation were implemented (e.g. [Si22]). Further issues with classical Machine Learning (ML) approaches involve the lack of consideration of context in which a phrase is used (e.g. [SMH22]). Recently, the ability to train large neural networks led to the development of models like BERT [De18] which takes into account the semantic nature of texts, thereby attaining better classification performances as compared to traditional ML techniques [GG20]. In this paper, we compare performances of large neural networks [De18] and classical ML techniques [Le07], distinguishing oaked and unoaked wines, based on textual descriptors, collected from sensory evaluation by trained wine professionals.

2 Data Collection and Preparation

Classification is based on more than 1,200 textual descriptions generated during a wine tasting conducted in January 2023 at Weincampus Neustadt. Sixteen wines (four white, four red, with and without oak) were served in both black and clear glasses to a panel of 16 wine experts, who were asked to provide textual descriptions of each wine's smell and taste.[4]

For experiments, we used a stratified sample approach, considering multiple data splits naturally given by tasting conditions. These will be referred to as scenarios for the classification experiments conducted in this paper, described in Table 1. Furthermore, we compared classification based on text only to classification based on a combination of text and meta-information giving additional details about the wine, namely its vintage, grape variety, country of origin, and vineyard. The baseline in general, but also specifically for the scenarios 2 to 7, is scenario 1 where all texts were taken into consideration.

[4] We are only able to share the data on request.

Scenarios	Data Selection	Train/Test Breakup per Class
1	Taking All Texts Into Consideration	600 - 40
2 & 3	Considering texts specific to black or clear glass sessions respectively	300 - 20
4 - 7	Considering texts fine-grained to either smell or taste sessions, based on black or clear glasses	150 - 10

Tab. 1: Data breakup across scenarios

3 Methods and Experimental Setup

3.1 Classification Algorithms and Data Embedding Techniques

Since the manuscript's aim is to demonstrate capabilities of NLP-techniques in the wine domain, we selected two approaches for the classification experiments: a BERT-based Sentence Classifier [De18] and the Naïve Bayes Classifier [Le07]. BERT (cf. [De18]) is a neural network applied commonly in the field of NLP. It is trained in an unsupervised manner to predict texts surrounding a specific word, making them bi-directional in nature. Considering the limitations of preceding NLP-based, recurrent-based neural networks which could process textual representations in a fixed sequence (either left to right, or vice-versa), the bi-directional nature of BERT made it a ground-breaking implementation for all downstream NLP tasks (e.g. sentence classification). The Naïve Bayes Classifier is based on the Bayes' Theorem that assumes the effect that particularly the presence of a word in a text is independent of the value of other words predicting a class (cf. [Le07]). Given the fact that the Naïve Bayes Classifier is considered as a baseline classification algorithm, and given the ground-breaking abilities of the BERT model for tasks related to NLP, these methods were chosen for all experiments in this research.

Furthermore, two different textual embedding techniques were employed, namely German-based BERT embeddings as well as TF-IDF embeddings. The German BERT models were used in a pre-trained version of [De20], being adapted to the current needs in the wine domain. Considering their bi-directional nature, the BERT embeddings are contextual by definition. In contrast, TF-IDF embeddings measures the frequency of occurence of a word in a given set of sentences, neither considering the position of a word in a sentence, nor its surrounding words (cf. [SMH22]), this being non-contextual.

In our experiments, the BERT classifier was fine-tuned on BERT embedded texts, while the Naïve Bayes Classifier was trained from scratch on the TF-IDF embedded texts. These techniques were chosen to compare the model performances of large neural networks, which have attained state-of-the-art performances in the field of NLP, against classical ML techniques (cf. e.g. [GG20; SMH22]), in the wine domain.

3.2 Experimental Setup

The BERT Sentence Classifier experiments were based, in particular, on the Flair framework [Ak19]. For matter of reproducibility, the network and training parameters were set as follows, given the predefined structure of BERT according to [Ak19; De18]: a learning rate of $5e^{-6}$, a mini batch size of 4, each for a total of 10 epochs.
Experiments using the Naïve Bayes approach were implemented using the default setting in the Scikit-Learn package [Pe11].
In addition, we conducted separate experiments per classifier type and scenario, considering either purely textual content or textual content merged with wine-metadata. To account for generalisation, at least 5-folds for each condition were trained. For presentation of results, we used the F1-Scores per scenario, being averaged across respective folds.

Additionally, for a better rating of the achievements, non-parametric statistical evaluations were performed to check if the differences between the textual descriptors across different scenarios, mentioned in Section 2, were significant. Further, intra-scenario significance values were computed to check if adding meta-data to the textual descriptors yielded any differences in the model performances. For both investigations, we applied the non-parametric Kruskal-Wallis-Test (cf. [KW52]), using two significance levels of $p < 0.05$ (significant) and $p < 0.01$ (highly significant), usually compared to the baseline.

4 Results

Table 2 compares the model performances across the scenarios. Considering the baseline (scenario 1), an F1-score ≥ 0.73 was achieved using the BERT-classifier with and without the addition of wine metadata. It can also be seen across all scenarios that BERT-classifiers outperformed the Naïve Bayes counterparts.
Regarding the other scenarios, the best performance was achieved on the black glass setting (highlighted in Table 2). A first interpretation might be that the human participants were rather focused to the wine's characteristics. However, this is a matter of further investigations.
Using wine-metadata improved the classification performance just slightly (cf. Table 2).
With respect to the motivation in Section 1.1, we can state that a textual description-based classification of wine characteristics, especially considering oaked vs unoaked wines, is possible. Furthermore, we saw that there is some discriminative power in the textual descriptions provided by the human panel to estimate wine characteristics.

On delving into the non-parametric statistical evaluations at an inter-scenario level, BERT-classifiers observed smaller p-values, although no statistical significance was reached (mean[5] p-value of 0.15) as compared to Naïve Bayes classifiers (mean p-value of 0.46).
Further, at an intra-scenario level we found the following: Considering the BERT results, we can state that significant difference are given between baseline (scenario 1) and multiple

[5] Mean value was calculated as average across all folds and scenarios.

Scenario	Only Textual Descriptions		Textual Descriptions + Wine Metadata	
	BERT	Naïve Bayes	BERT	Naïve Bayes
1	0.73	0.59	0.78	0.58
2	**0.79***	0.60	0.78	0.59
3	0.77	0.58	0.77*	0.60
4	0.68	0.61	0.74	0.61
5	0.65**	0.50	0.67**	0.48
6	0.74*	0.64	0.75**	0.64
7	0.63**	0.53	0.64**	0.54

Tab. 2: Comparison of BERT Sentence Classifier fine-tuned on German BERT embeddings, and Naïve Bayes Classifier trained on TF-IDF embeddings across different scenarios in which the wine descriptions were collected. Significance is indicated by * ($p < 0.05$) and ** ($p < 0.01$).

other scenarios, especially scenario 2 ($p = 0.04$), 5 ($p = 0.00$), 6 ($p = 0.01$), and 7 ($p = 0.00$). Same observations are holding true in the setting of metadata, where scenario 1 has significant difference to scenario 3 ($p = 0.02$), 5 ($p = 0.00$), 6 ($p = 0.00$), and 7 ($p = 0.00$). However, there was no significant difference observed across scenarios utilising the Naïve Bayes classifier (with an mean p-value of 0.92).

In a broader sense, the usage of large neural networks outperforms classical ML-techniques, irrespective of the size of the corpus. They have the ability to learn the semantic and the syntactic nature of the texts, due to which they tend to find differences across classes. This benefit can be used also in the wine domain to estimate particular characteristics more objectively, aiming for a support in the entire production process as mentioned in Section 1.

5 Discussion and Conclusion

This paper transferred NLP-techniques to the wine domain and evaluated the performance of BERT against Naïve Bayes models for the task of classifying wines (oaked/unoaked), based on the textual descriptions provided by trained human panellists. While BERT-classifiers outperformed Naïve Bayes classifiers across all scenarios (cf. Table 2), they tended to perform best when trained with texts related to the glass' colour (i.e. black/white, sessions 2 and 3). Also regarding the limited availability of data, and a 50% chance level of a 2-class task, F1-scores greater than 0.63 (worst performance) were already achieved across all scenarios, with and without the addition of metadata. This is showing the usability of NLP methods also in the wine domain.

As mentioned in Section 1, our work is also related to the support of wine makers and stakeholders in oenology. Using a combination of sensor-based measures of wines and trained NLP models on panel-based assessments, we are aiming in future work for an AI-based assistance in quality assessment of wines.

Acknowledgement

We acknowledge support by the PINOT project funded by the German Federal Ministry of Food and Agriculture (BMEL) under grant number 28DK107A20 and 28DK107C20.

Bibliography

[Ak19] Akbik, A.; Bergmann, T.; Blythe, D.; Rasul, K.; Schweter, S.; Vollgraf, R.: FLAIR: An easy-to-use framework for state-of-the-art NLP. In: NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations). Pp. 54–59, 2019.

[BGT21] Barbe, J.-C.; Garbay, J.; Tempére, S.: The Sensory Space of Wines: From Concept to Evaluation and Description. A Review. Foods 10/6, 2021, ISSN: 2304-8158.

[De18] Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805/, 2018.

[De20] Deepset: Open Sourcing German BERT Model, Accessed on May 4, 2023, 2020, URL: https://www.deepset.ai/german-bert.

[FdS22] Ferreira, V.; de la Fuente, A.; Sáenz-Navajas, M. P.: 1 - Wine aroma vectors and sensory attributes. In (Reynolds, A. G., ed.): Managing Wine Quality (Second Edition). Second Edition, Woodhead Publishing Series in Food Science, Technology and Nutrition, Woodhead Publishing, pp. 3–39, 2022, ISBN: 978-0-08-102067-8.

[GG20] González-Carvajal, S.; Garrido-Merchán, E. C.: Comparing BERT against traditional machine learning text classification. arXiv preprint arXiv:2005.13012/, 2020.

[IKT05] Ikonomakis, M.; Kotsiantis, S.; Tampakas, V.: Text classification using machine learning techniques. WSEAS transactions on computers 4/8, pp. 966–974, 2005.

[KW52] Kruskal, W. H.; Wallis, W. A.: Use of Ranks in One-Criterion Variance Analysis. Journal of the American Statistical Association 47/260, pp. 583–621, 1952, visited on: 01/12/2023.

[La13] Lawrence, G.; Symoneaux, R.; Maitre, I.; Brossaud, F.; Maestrojuan, M.; Mehinagic, E.: Using the free comments method for sensory characterisation of Cabernet Franc wines: Comparison with classical profiling in a professional context. English, Food Quality and Preference 30/2, pp. 145–155, 2013.

[Le07] Leung, K. M. et al.: Naive bayesian classifier. Polytechnic University Department of Computer Science/Finance and Risk Engineering 2007/, pp. 123–156, 2007.

[Ma20] Mahieu, B.; Visalli, M.; Thomas, A.; Schlich, P.: Free-comment outperformed check-all-that-apply in the sensory characterisation of wines with consumers at home. Food Quality and Preference 84/, ed. by Ltd., E. S., p. 103937, Sept. 2020.

[Ma21] Mahieu, B.; Schlich, P.; Visalli, M.; Cardot, H.: A multiple-response chi-square framework for the analysis of Free-Comment and Check-All-That-Apply data. Food Quality and Preference 93/, p. 104256, Oct. 2021.

[Pe11] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12/, pp. 2825–2830, 2011.

[Pi15] Piqueras-Fiszman, B.: 12 - Open-ended questions in sensory testing practice. In (Delarue, J.; Lawlor, J. B.; Rogeaux, M., eds.): Rapid Sensory Profiling Techniques. Woodhead Publishing Series in Food Science, Technology and Nutrition, Woodhead Publishing, pp. 247–267, 2015, ISBN: 978-1-78242-248-8.

[Si22] Singh, K. N.; Devi, S. D.; Devi, H. M.; Mahanta, A. K.: A novel approach for dimension reduction using word embedding: An enhanced text classification approach. International Journal of Information Management Data Insights 2/1, p. 100061, 2022.

[SMH22] Subakti, A.; Murfi, H.; Hariadi, N.: The performance of BERT as data representation of text clustering. Journal of big Data 9/1, pp. 1–21, 2022.

[Vi20] Visalli, M.; Mahieu, B.; Thomas, A.; Schlich, P.: Automated sentiment analysis of Free-Comment: An indirect liking measurement? Food Quality and Preference 82/, p. 103888, 2020, ISSN: 0950-3293.

[Vi23] Visalli, M.; Dubois, M.; Schlich, P.; Ric, F.; Cardebat, J.-M.; Georgantzis, N.: Relevance of free-comment to describe wine temporal sensory perception: An application with panels varying in culture and expertise. Food Quality and Preference 105/, p. 104785, 2023, ISSN: 0950-3293.