Somm: Into the Model

Shengli Hu
Cornell University
sh2264@cornell.edu

Abstract

To what extent could the sommelier profession, or wine stewardship, be displaced by machine learning algorithms? There are at least three essential skills that make a qualified sommelier: wine theory, blind tasting, and beverage service, as exemplified in the rigorous certification processes of certified sommeliers and above (advanced and master) with the most authoritative body in the industry, the Court of Master Sommelier (hereafter CMS). We propose and train corresponding machine learning models that match these skills, and compare algorithmic results with real data collected from a large group of certified wine professionals. We find that our machine learning models outperform human sommeliers on most tasks — most notably in the section of blind tasting, where both hierarchically supervised Latent Dirichlet Allocation outperforms sommeliers' judgment calls by over 6% in terms of F1-score; in the section of beverage service — wine and food pairing, a modified Siamese neural networks based on BiLSTM achieves better results than sommeliers by 2%. This demonstrates, contrary to popular opinion in the industry, that the sommelier profession is at least to some extent automatable, barring economic (Kleinberg et al., 2017) and psychological (Dietvorst et al., 2015) complications.

1. Introduction and Related Work

Thanks to the Somm documentaries and a general increase in awareness about wine, sommeliers, and the Court of Master Sommeliers, there is now a certain celebrity status, a glamor associated with becoming a sommelier. When encountered with the question — "is the sommelier profession going to be negatively affected by recent advances in machine learning and artificial intelligence?" during informal interviews conducted by authors in the sommelier community, there appears to be a general consensus among professionals that the high standards of hospitality upheld by qualified sommeliers are well beyond the capabilities of machines.

The current study asks the question, to what extent would the sommelier profession be displaced by machine learning algorithms? What aspects of the sommelier profession could be outperformed, and therefore perhaps displaced by what kinds of applications of machine learning?

What makes a qualified sommelier or wine professional? According to the Court of Master Sommelier[1], one of the two organizations held in the highest esteem in the global industry, there are at least three indispensable components as exemplified in the certification exams leading up to the Master Sommelier diploma: theory, blind tasting, and service.

To satisfy the theoretical requirement, sommelier candidates are required to sit on a timed exam of various questions covering a wide range of wine topics including geography, soil, viticulture, laws, history, language, etc. without any officially structured study guides[2]. We argue that this particular task maps to the stream of research concerning *open-domain* Question Answering (hereafter, OQA), where the model is given a question and access to a large corpus (Chen et al., 2017), combining and therefore leveraging both the Information Retrieval (Weinberger et al., 2009) and Machine Comprehension literature (Hermann et al., 2015; Chen et al., 2016). In Section 3, we train an open-domain QA model building upon Chen et al.

[1] "The Court of Master Sommeliers sets the global standard of excellence for beverage service within the hospitality industry with integrity, exemplary knowledge, and humility." — https://www.mastersommeliers.org/

[2] There are indeed a list of recommended references and an unofficial source of study guides popular among candidates: https://www.guildsomm.com, which we use as our training data.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1153–1159
Brussels, Belgium, October 31 - November 4, 2018.

(2017), on a large corpus of wine topics drawn from recommended study resources by CMS. We contrast the machine performance with sommeliers' performance on equivalent test questions.

To satisfy the blind tasting requirement, candidates have to blind taste a flight of wines, precisely describe the wine, and accurately identify the grape varietal, the region (thus the country), the vintage, and the quality level of each. According to wine programs such as CMS or WSET[3], blind tasting consists of two steps — tasting and deduction. Tasting refers to the sensory experience associated with evaluating wines — color, aroma, favor, aftertaste, etc. Proficient candidates are expected be able to detect a wide range of characteristics of the focal wine, and precisely describe the wine with meaningful descriptors accordingly. Deduction is the logical process that leads the candidate to conclude on the identity of the wine given the characteristics he detects in the first step. According to wine educators and master sommeliers such as Geoff Kruth M.S., it is the deduction part of blind tasting that separates great blind tasters from mediocre ones, mostly due to the fact that it requires greater logical thinking and reasoning. We propose that the deduction step maps exactly to the machine learning task of *structured prediction* (Taskar et al., 2005; Belanger and McCallum, 2016; Barutcuoglu et al., 2006; Rousu et al., 2006). In Section 4, we demonstrate that a hierarchical supervised Latent Dirichlet Allocation model (Perotte et al., 2011; Nguyen et al., 2013) trained on a large corpus of textual descriptions of wines of different grape varietals, regions, vintages, and quality levels, outperforms sommeliers in deduction by a large margin.

To satisfy the service requirement, candidates are grilled on questions of wines and spirits, food and wine pairing, salesmanship, and service mechanics in a restaurant setting. In Section 5, we showcase a modified Siamese Neural Network (Yang et al., 2015; Mueller and Thyagarajan, 2016; Neculoiu et al., 2016; Pei et al., 2016; Bertinetto et al., 2016) coupled with Bidirectional Long Short-term Memory Networks (Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997; Zhang et al., 2015) trained on corpora of wine reviews (Hendrickx et al., 2016) and cooking recipes (Tasse and Smith, 2008; Jermsurawong and Habash, 2015) outperforms sommeliers' performance.

[3]Wine & Spirit Education Trust

2. Data Collection and Preprocessing

Our datasets consist of three parts: (1) *Study Resources*: a large corpus consisting of all the recommended resources for sommelier certification by the CMS, for the Question Answering — Theory Component detailed in Section 3; (2) *Wine Reviews*: a massive repository of expert wine reviews with rich meta-data, based on reviews from Decanter, Vinous, Wine Spectator, and Wine Enthusiast, the four widely recognized media outlets in the industry, for the Structured Prediction — Deduction in Blind Tasting Component detailed in Section 4; (3) *Survey Responses* from $1,305$ certified wine professionals, covering topics on theory, deductive tasting, and wine and food pairing, thus providing experts' performance data with which we compare results from our corresponding machine learning models in Section 3, Section 4, and Section 5.

2.1 Preprocessing

The study resource dataset consists of documents of various categories and topics from the Guild-Somm. We treat texts under each sub-category as a document — there are 752 documents in our dataset and the average length of documents is $1,384$ words.

For the wine review dataset, we only consider wines for which we had at least 200 reviews in the training set, leading to $850,119$ reviews combined. When different names were used for the same grape, we normalize these to the same category. For instance, Pinot Bianco (Italy), Pinot Blanc (France), and Weissburgunder (Germany) are mapped together and renamed Pinot Blanc according to the wine grape encyclopedia (Robinson et al., 2013; Robinson and Harding, 2015). We preprocessed all the text data in standard procedures.

2.2 Summary of Datasets

We plot the country and point distributions of our review dataset in Figure 1 , grouped by media outlet. Interestingly, Vinous appears proportionally much more focused on Italian wines and its ratings are more skewed to the right compared to others (a.k.a. greater rating inflation), somehow contrary to the brand image; Wine Spectator is much more focused on Old World whereas Wine Enthusiast is

more evenly distributed across countries.

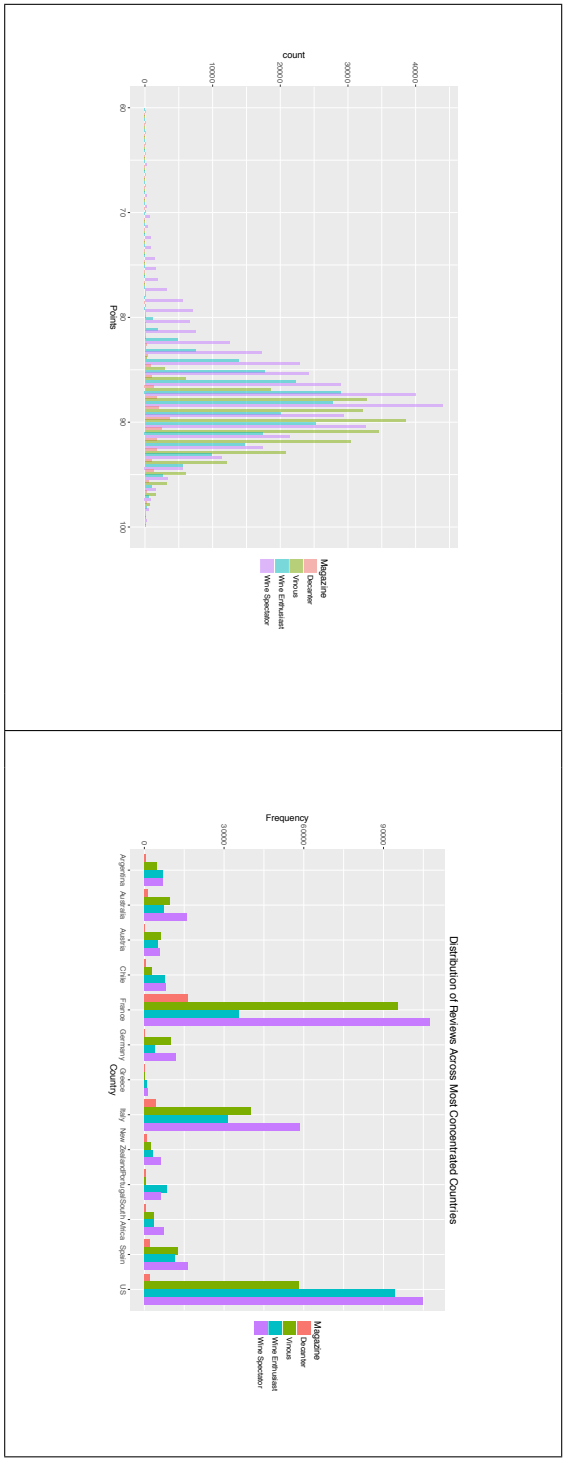

Figure 1: Country and Point Distribution of Wine Reviews

2.3 Survey Details

We administered timed online surveys to wine professionals in several active sommelier communities such as the Guild of Sommeliers, the Society of Wine Educators, etc. Each survey consists of three sections that correspond to the three components respectively — theory, deductive tasting, and pairing. Each section lasts no longer than 15 minutes and consists of 30 questions, randomly drawn from a large pool of practice questions from the Guild of Sommeliers and the Society of Wine Educators.

In the first section, we administered two sets of questions varying difficulty level — one on the level of certified sommelier (CMS level 2), the other on the level of advanced sommelier (CMS level 3). The Question Answering (wine theory) pool consists of $1,400$ questions (700 equivalent to level 2 difficulty, 700 equivalent to level 3 difficulty) from Society of Wine Educators' Certified Specialist of Wine and Certified Specialist of Spirits programs, and $1,480$ questions (740 at level 2 difficulty and 740 at level 3 difficulty) from Guild of Sommelier practice repository. $2,304$ questions were used for training, and the rest for testing. In the second section, we randomly drew 30 textual descriptions from our pool of $850,119$ reviews detailed in Section 2, and asked the subject to deduce the varietal, vintage range, region, and quality level of the wine being described. In the third section, we randomly drew 30 wines (Title and Tasting note) from the repository of wine reviews and 30 recipes from the CMU Recipe Database CURD (Tasse and Smith, 2008), and asked the subject to rate the pairings on a scale from 1 to 5. We circulated our survey to members of the Guild of Sommeliers and Society of Wine Educators communities and received $1,412$ responses. The first section of theory questions serves not only as a dataset compared against QA models, but also as a validation and screening procedure: we removed the responses with fewer than 18[4] correct answers to the 30 questions in section 1, reducing our sample size to $1,305$. Sommelier scores were calculated aggregating all the participants' answers.

3. Wine Theory: open-domain Question Answering

We implemented an open-domain Question Answering system modeled after Chen et al. (2017), consisting of a Document Retriever module and a Document Reader module.

The Document Retriever module finds the three

[4]We choose the cutoff rate of 60% because it is the the pass rate in real sommelier exams both for the certified and advanced.

most relevant documents by comparing documents and questions as TF-IDF weighted bag-of-word vectors that include bigrams. We also adopt the hashing of Weinberger et al. (2009) for mapping bigrams with an unsigned murmur3 hash.

The Document Reader module is essentially a bidirectional Long Short-term Memory Network (BiLSTM) (Hochreiter and Schmidhuber, 1997; Zhang et al., 2015) applied to each paragraph in relevant documents, the predicted answers of which are finally aggregated. For detailed procedures of implementation, we refer readers to Chen et al. (2017). The only differences are, our batch size is 25 and we adopted a dropout rate of 0.1. We document the results in Table 1. Surprisingly, our OQA results converge to the high levels of accuracies achievable by machine comprehension models. We argue that it is because our corpus is relatively small and concentrated on wine-related topics, which results in few complications arising from the integration of large-scale information retrieval and machine comprehension, and therefore more germane to single machine comprehension models. Note that we removed survey results below 60% accuracy, stacking the odds against us because now the sommeliers' performance results are inflated, which could provide partial explanations for OQA being behind. The comparison still looks promising, despite the 4.8% disparity in performance.

Data Generation	Training Set		Test Set	
	Exact Match	F1	Exact Match	F1
Sommeliers	NA	NA	67.1	71.7
OQA System	58.1	67.8	55.7	66.9

Table 1: Evaluation results of OQA in comparison with sommeliers' performance.

Comparing the accuracies across regions, we find sommeliers did much better than DrQA in old world regions while DrQA edged out on most new world regions. It might echo the greater emphasis of sommelier training in real life on the old world, and/or reflect the more complications introduced by French, Italian, German, and Spanish terminologies which we didn't correct for when dealing with the old world wine regions.

4. Blind Tasting: HSLDA

We implemented the Hierarchically Supervised Latent Dirichlet Allocation (HSLDA) model (Perotte et al., 2011) for deduction in blind tasting based on textual descriptions of wines, because of the natural fit in-between — the texts describing wines are hierarchically (from top to bottom: grape varietal, country, region, vintage, quality) and multiply (blends vs. monovarietals)) labeled bag-of-word (simple and performant for reviews) data. For model details, we refer readers to Perotte et al. (2011). We use the subset of wine reviews published in Wine Enthusiast and Decanter for this task. It contains $183,660$ wine descriptions for training and $39,150$ for testing. There are 41.1 terms on average in each document, with a 11.6 standard deviation. There are $12,132$ unique (sub-)categories in the form of "Sangiovese, Italy, Tuscany, 2015, Riserva". We use a Gaussian prior over the regression parameters where a range of values for μ, the mean prior parameter for regression coefficients are evaluated ($\mu \in \{-3, -2.5, -2, \ldots, 1\}$). We set the number of topics to 20 based on small sample testing and CMS tasting grid. Prior distributions of hyperparameters are gamma distributed with a shape parameter of 1 and a scale parameter of $1,000$.

Initial results were less satisfying and most errors occurred because mono-varietals were predicted to be blends and vice versa. Therefore we explored two solutions: (1) we separated our data into mono-varietals (Model 1), and blends (Model 2), and trained HSLDAd separately; (2) we created a smaller yet more balanced training set regarding mono-varietals and blends (Model 3).

In Model 1, we simplified "testable" (i.e., included in Court of Master Sommelier tasting exams) blends in our dataset such as "Southern Rhone red blend", "Marsanne Roussanne blend", "Sangiovese blend", and such were treated the same as the mono-varietals. Model 2 and 3 were trained using the exact blending grape varietals. We believe Model 1 is closest to the decision making processes encountered by sommeliers in CMS certification exams, whereas Model 3 is more likely to resemble sommelier challenges such as Top Somm.

We computed precision and recall of all the categories, yielding a $12,132 \times 12,132$ sparse confusion matrix for Model 3, a $11,672 \times 11,672$ sparse confusion matrix for Model 1, and a 386×386 confusion matrix for Model 2. We then averaged them to get a single real number measurement. Table 2 shows the average F1 scores of different models versus sommeliers' performance.

Likewise, the sommeliers' performance measures represent a conservative(ly higher) estimate since scores lower than 60% in section 1 were removed. We find the HSLDA model, especially of monovarietals, outperforms sommeliers by 6.3%, as measured by F1. In aggregate, sommeliers did significantly better in red and sparkling wines, and in French, German, and Californian wines, whereas Model 1 edged out in white wines, south America wines.

F1 Scores	Training Set	Testing Set
HSLDA1 Monovarietal	71.1	68.4
HSLDA2 Blend	62.5	59.1
HSLDA3 Balanced	59.8	56.4
Sommeliers	NA	62.1

Table 2: Evaluation results of HSLDAs in comparison with sommeliers' performance.

5. Food and Wine Pairing: Siamese Neural Networks with LSTM

For food and wine pairing, we trained an untied and modified version of Manhattan LSTM (Mueller and Thyagarajan, 2016), where we preprocessed the texts differently and applied LSTM-Based Importance Weighting in place of the original simple similarity function coupled with LSTM. We retained from the recipes only ingredients, serving ingredients and essential actions, which were passed to the BiLSTM (Zhang et al., 2015). For a given pair of wine and recipe descriptions, we applied a weight compatibility function $g(h^{(a)}_{T_a}, h^{(b)}_{T_b}) = exp(-||a^T h^{(a)}_{T_a} - b^T h^{(b)}_{T_b}||_1)$, where a^T and b^T are shared network weights applied to BiLSTM representations $h^a_{T_a}$ and $h^b_{T_b}$. For model architecture and other implementation details we refer readers to Mueller and Thyagarajan (2016) and Rücklé and Gurevych (2017).

We obtained our ground-truth labels for wine and recipe pairings on a scale from 1 to 5 using an automated weighting scheme based on wine and food pairing principles (Goldstein and Goldstein, 2006)[5] leveraging the GuildSomm tasting notes of grape varietals and recipe ingredients, under close guidance of a certified sommelier with the CMS. Details of the weighting scheme is included in our online supplementary documents. In the end we simplified our scale to binary — $\{1, 2\}$ converted to 0, $\{3, 4, 5\}$ converted to 1. We document our accuracies in Table 3. Surprisingly, the model edged out by 1.7%.

[5]One of the few recommended resources for certified sommelier candidates on wine and food pairing.

Accuracy	Training Set	Testing Set
Modified MaLSTM	82.3	79.8
Sommeliers	NA	78.1

Table 3: Evaluation results of Modified MaLSTM in comparison to sommeliers' performance.

6. Conclusion and Future Work

We examine how machine learning can be used to understand, assist, and improve human decision-making, echoing recent studies in computational social sciences (Dietvorst et al., 2015; Kleinberg et al., 2017). We dissect sommelier skills into three parts and train ML models for each. We show with our choices of suitable models, collection of valuable datasets and annotations, that ML algorithms outperform sommeliers in essential skills. Future work could improve on:

1. fine-tuning the Open-domain Question Answering for wine knowledge;
2. connecting our HSLDA or hierarchical multi-label classification to robotic sensors to fully mimic the blind tasting task;
3. exploring other simpler and more efficient ML models for pairing tasks;
4. training a joint multi-task model, since it is accepted in the industry that a solid knowledge of wine theory helps immensely in blind tasting and wine service. It would be interesting to quantify the synergy in the learning process;
5. exploring ML applications to other aspects of the service component.

References

Zafer Barutcuoglu, Robert E Schapire, and Olga G Troyanskaya. 2006. Hierarchical multi-label prediction of gene function. *Bioinformatics*, 22(7):830–836.

David Belanger and Andrew McCallum. 2016. Structured prediction energy networks. In *International Conference on Machine Learning*, pages 983–992.

Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. 2016. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer.

Danqi Chen, Jason Bolton, and Christopher D Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. *arXiv preprint arXiv:1606.02858*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114.

Evan Goldstein and Joyce Goldstein. 2006. *Perfect Pairings: A Master Sommeliers Practical Advice for Partnering Wine with Food*. Univ of California Press.

Iris Hendrickx, Els Lefever, Ilja Croijmans, Asifa Majid, and Antal van den Bosch. 2016. Very quaffable and great fun: Applying nlp to wine reviews. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 306–312.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jermsak Jermsurawong and Nizar Habash. 2015. Predicting the structure of cooking recipes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 781–786.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *AAAI*, pages 2786–2792.

Paul Neculoiu, Maarten Versteegh, and Mihai Rotaru. 2016. Learning text similarity with siamese recurrent networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 148–157.

Viet-An Nguyen, Jordan L Boyd-Graber, and Philip Resnik. 2013. Lexical and hierarchical topic regression. In *Advances in neural information processing systems*, pages 1106–1114.

Wenjie Pei, David MJ Tax, and Laurens van der Maaten. 2016. Modeling time series similarity with siamese recurrent networks. *arXiv preprint arXiv:1603.04713*.

Adler J Perotte, Frank Wood, Noemie Elhadad, and Nicholas Bartlett. 2011. Hierarchically supervised latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 2609–2617.

Jancis Robinson and Julia Harding. 2015. *The Oxford companion to wine*. American Chemical Society.

Jancis Robinson, Julia Harding, and José Vouillamoz. 2013. *Wine Grapes: A complete guide to 1,368 vine varieties, including their origins and flavours*. Penguin UK.

Juho Rousu, Craig Saunders, Sandor Szedmak, and John Shawe-Taylor. 2006. Kernel-based learning of hierarchical multilabel classification models. *Journal of Machine Learning Research*, 7(Jul):1601–1626.

Andreas Rücklé and Iryna Gurevych. 2017. Representation learning for answer selection with lstm-based importance weighting. In *IWCS 201712th International Conference on Computational SemanticsShort papers*.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Ben Taskar, Vassil Chatalbashev, Daphne Koller, and Carlos Guestrin. 2005. Learning structured prediction models: A large margin approach. In *Proceedings of the 22nd international conference on Machine learning*, pages 896–903. ACM.

Dan Tasse and Noah A Smith. 2008. Sour cream: Toward semantic processing of recipes. *Carnegie Mellon University, Pittsburgh, Tech. Rep. CMU-LTI-08-005*.

Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. 2009. Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1113–1120. ACM.

Longqi Yang, Yin Cui, Fan Zhang, John P Pollak, Serge Belongie, and Deborah Estrin. 2015. Plateclick: Bootstrapping food preferences through an adaptive visual interface. In *Proceedings of the 24th acm international on conference on information and knowledge management*, pages 183–192. ACM.

Shu Zhang, Dequan Zheng, Xinchen Hu, and Ming Yang. 2015. Bidirectional long short-term memory networks for relation classification. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 73–78.