Very quaffable and great fun: Applying NLP to wine reviews

Iris Hendrickx[1]
[1]Centre for Language Studies
Radboud University, P.O. Box 9103
NL6500 HD Nijmegen, Netherlands
i.hendrickx@let.ru.nl

Els Lefever[2]
[2]Language and Translation Technology Team
Dep. of Translation, Interpreting and
Communication, Ghent University, Belgium
els.lefever@ugent.be

Ilja Croijmans[1], **Asifa Majid**[1,3] **and Antal van den Bosch**[1]
[3]Max Planck Institute for
Psycholinguistics, P.O. Box 310,
6500 AH Nijmegen, The Netherlands
{i.croijmans,asifa.majid,a.vandenbosch}@let.ru.nl

Abstract

We automatically predict properties of wines on the basis of smell and flavor descriptions from experts' wine reviews. We show wine experts are capable of describing their smell and flavor experiences in wine reviews in a sufficiently consistent manner, such that we can use their descriptions to predict properties of a wine based solely on language. The experimental results show promising F-scores when using lexical and semantic information to predict the color, grape variety, country of origin, and price of a wine. This demonstrates, contrary to popular opinion, that wine experts' reviews really are informative.

1 Introduction

Describing smells and flavors is something the average person is not particularly good at. If people are asked to identify familiar smells such as cinnamon and chocolate, they are only able to correctly name the smell around 50% of the time (Cain, 1979; Olofsson and Gottfried, 2015). In comparison to the elaborate vocabulary we have for visual and auditory phenomena, English and other languages spoken in Western societies appear to have few words to describe smells and flavors (Levinson and Majid, 2014; Majid and Burenhult, 2014). Instead, speakers often refer to the source as the name of the smell ('it smells like banana').

Flavor is a complex experience that combines the multisensory sensations of taste, touch and smell. Flavor descriptions contain basic taste descriptors (e.g., sweet, sour, salty, bitter), with metaphorical (e.g., 'elegant') and source-based terminology (e.g., 'it tastes buttery').

The lack of vocabulary for smells and flavors contrasts starkly with the interest people in the West have for flavors and fragrances, and what they are willing to spend on such products. The flavor and fragrance industry is estimated to be worth over $20 billion in 2015[1]. In this context, experts' recommendations are used by the public in order to help them make decisions about purchases. But are the expert recommendations meaningful, given the limitations of language for smells and flavors?

We are interested in the relation between language and sensory information, and how this information is put into words. We focus on descriptions produced by a select group of people who have considerable experience naming smells and flavors, i.e., sommeliers and wine journalists. Through their descriptions, wine experts can influence consumers' purchasing patterns (McCoy, 2006; Horverak, 2009), suggesting their descriptions are written in an informative manner. In this paper we aim to discover whether we can extract the properties of a wine based on the tasting notes written by a wine expert. This should be possible if wine experts are capable of translating their sensory experiences into words in a consistent manner.

Previous experimental studies provide a mixed picture as to whether wine experts' language is consistent. Some studies find similar levels of agreement in smell descriptions generated by wine experts and those generated by novices (Lawless, 1984; Parr et al., 2002), and wine experts use more

[1]http://www.leffingwell.com/top_10.htm

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 306–312,
Berlin, Germany, August 7-12, 2016.

metaphorical descriptions to describe wine (Caballero and Suárez-Toste, 2010; Paradis and Eeg-Olofsson, 2013), which potentially are not as informative about properties of the wine itself. In contrast, others find wine experts use more specific vocabulary (Zucco et al., 2011; Sezille et al., 2014), and find that wine experts are, in fact, more consistent than non-experts, when they describe wines (Croijmans and Majid, 2016).

We examined the following wine properties and aimed to predict these solely on the basis of the review content: color, grape variety, price, and country of origin. The outcomes of this investigation are interesting for two reasons. First, we test the ability of experts to review wines with consistent language using naturalistic materials. Most previous studies about wine experts and their reviewing consistency are performed in experimental settings and cover some dozens of wine reviews (see for example (Gawel and Godden, 2008; Hopfer and Heymann, 2014)). With automatic analysis we are able to scale up to a much larger and more representative set of reviews.

Second, we gather new insights into the specific vocabulary and type of lexical descriptors used to describe smells and flavors, and what words are most distinctive for different wine characteristics. Market analyses (Vigar-Ellis et al., 2015) show that consumers increasingly select wines based on information provided by experts, for example through expert descriptions and recommender systems, and that wine apps become ever more popular. This is a positive development, as research suggests that informed consumers are able to benefit more from the loose relationship between price and quality in wine (Oczkowski and Doucouliagos, 2014).

In the long run, as we are training automatic systems to predict wine properties, we could use such systems for automatic metadata prediction and error correction in wine review databases. These systems are also a first step towards a recommender system for wines based on review content and flavor descriptions. Current recommender systems such as the mobile apps *Vivino*[2] or *Delectable*[3] work with metadata and user-based filtering, i.e. the principle of 'other users also bought . . .'. So there is potential here for content-based recommender systems to be developed.

[2]Vivino: `http://www.vivino.com`

[3]Delectable: `http://www.delectable.com`

2 Related work

The relationship between wines and wine reviews has been studied from many different perspectives, aside from those discussed in the previous section. Economically, the relationship between price, wine quality and wine ratings is interesting as a high rating by a famous wine expert can make a substantial difference to product sales (McCoy, 2006). Goldstein and colleagues (2008) investigated whether a jury of wine experts vs non-experts can taste the difference between expensive and cheap wines, and found while wine experts could distinguish the difference, non-experts could not. Lecocq and Visser (2006) investigated what wine properties determine wine prices. They showed wine experts based their overall wine quality ratings on sensory information, and that expert ratings together with features such as region, vintage and designation explained price differences for a subset of French red wines. The relationship between the chemical substances in a wine and wine quality have also been the focus of research (Chen et al., 2009; Cortez et al., 2009).

Brochet and Dubourdieu (2001) conducted a lexical analysis of four corpora of wine reviews from a cognitive linguistic perspective and concluded wine reviews are not only describing sensory properties of the wine, but also include idealistic and hedonistic information from wine prototypes based on previous experiences. Anthropologists have noted that wine experts form their own discourse community with a particular style and vocabulary (Silverstein, 2006). In this research we aim to discover stylistic and lexical patterns with which we can relate wine reviews to wine properties automatically.

3 Data set

The website `http://www.winemag.com/`, owned by Wine Enthusiast Companies, hosts a substantial catalog of wine descriptions. We downloaded the available reviews[4] and gathered a total of 76,585 wine reviews. The catalog data is structured and contains information about the wine such as the producer, appellation region and country, grape variety, color, alcohol percentage, price, and where to buy it. The expert who writes the wine review also rates the wine by assigning it a score between 80 and 100. The reviews are writ-

[4]Downloading took place in February 2015

ten by 33 different experts, and can be considered concise, with an average length of 39 words.

Wine reviews have a distinct style and vocabulary, which tends to focus on smell and flavor descriptions, as shown in example reviews 1 and 2 from our data. As noted previously, wine experts use creative metaphors to characterize the smell and flavor of a wine, as well as source-based descriptions. The metaphors perhaps add variation to otherwise dull or repetitive descriptions (Paradis and Eeg-Olofsson, 2013; Suárez Toste, 2007).

1. *There is not a great deal of dolcetto grown in the Northwest, but this is the best version I've yet seen. Its vivid, spicy fruit core expresses the soil, the plant and the grape in equal proportion. Sappy flavors of spiced plum and wild berry hold the fort; it's built like a race car, sleek and stylish, with a powerful, tannic frame.*

2. *Here's a fragrant and very aromatic Grillo with cheerful notes of peach, passion fruit and mango. The wine has an easy approach and would pair perfectly with appetizers or finger foods.*

4 Methodology

In our classification experiments, we evaluated the viability of predicting the following four wine properties: color, grape variety, price, and country of origin. Wines can be categorized into three different **colors**: white, red and rosé. The database of winemag.com is not complete in all metadata fields. We excluded reviews with missing metadata from our experiments, and performed this selection separately for each metadata field. For instance, we excluded 5,328 wines without a color label in the color labeling experiment.

For **grape variety** we only considered those wines that were produced from a single grape and for which we had at least 200 reviews in the training set, leading to 33 categories. We disregarded all wines with grape blends, as these can have different ratios of different grape varieties. When different names were used for the same grape, we normalized these to the same category; e.g., *Pinot Gris* (French) and *Pinot Grigio* (Italian) were mapped together manually.

The sample contained wines from 47 different **countries**, ranging from South Korea (3 wines) to USA (31,401 wines). Even though **price** itself is an objective value, a division into cheap and expensive prices is a rather subjective choice. We tested two alternatives: a discretization where cheap wine costs less than $10 and expensive wine at least $100; and a more relaxed version where cheap means less then $15 and expensive at least $50. Wines between these prices were left out in both price experiments.

We pre-processed the data set automatically with the Stanford toolkit (Manning et al., 2014): we tokenized, PoS-tagged and lemmatized the reviews. For the classification experiments, we split the randomized data set into an 80% training and 20% test set. As information sources, we use both lexical and semantic features. A first experimental setup merely uses a bag-of-words (BoW) representation of the wine reviews. To construct these BoW features, we lowercased all lemmas in the review and selected only the content words (PoS-tag *noun*, *verb* or *adjective*) that occurred at least twice in the training set.

As the reviews are short and only contain about 23 content words on average, we decided to also add semantic features to reduce data sparsity. As shown by Kusner and colleagues (2015), semantic representations such as Latent Semantic Indexing and Latent Dirichlet Allocation (LDA) can outperform a BoW representation. For our second experimental setup, we combined our set of BoW features with (1) 100 topics generated with Latent Dirichlet Allocation (Blei, 2012), and (2) 100 clusters based on word embeddings generated with Word2Vec (Mikolov et al., 2013). We ran initial experiments with exemplar-based classification and experimented with different cluster (Word2Vec) and topic (LDA) sizes of 100, 500, 1000, 2000 on the training set. For LDA (McCallum, 2002), we also varied the threshold to assign a topic only to a text when it covered 1%, 2%, or 5% of the text. The best results were obtained with 100 topics and a proportion threshold of 1%. We used these settings throughout our experiments. Two examples of LDA topics are shown here:

LDA42 color rosé strawberry raspberry pink flavor aroma wine light red cherry pale rise dry fresh

LDA49 flavor acidity wine crisp dry clean lime peach citrus lemon fruit pineapple white vanilla

To create the word embeddings we ran Word2Vec on the training corpus, applying the BoW model, a context size of 8, and a word vector dimensionality of 200 features. In a next step, K-means clustering (with $k = 100$) was applied on the resulting word vectors. As an example, we show part of the terms contained by cluster 20, which all have the connotation of "dark/intense":

Property	#test instances	Bag-of-Word features	combined: BoW+LDA+W2V	combined optimised
color	14,213	90.7	94.3	97.6
country	15,317	44.4	58.0	78.2
price big difference	1,135	60.9	61.0	94.6
price small difference	4,922	65.0	80.8	90.6
variety	9,946	30.5	36.6	70.6

Table 1: F-scores per task for Bag-of-Word features, a combination of BoW, LDA and Word2Vec clusters, and combined & optimised LIBSVM parameters.

Word2Vec20 asphalt black-fruit blackness burly dark deep inky masculine muscular purple roasted saturated sun-baked superconcentrated

The Word2Vec clusters were then implemented as binary features, meaning that for each instance containing a word occurring in one of the clusters, the respective cluster is coded by "1" in the feature vector, while the other cluster features are coded as "0".

As a classifier, we used LIBSVM (Chang and Lin, 2011), with the RBF kernel and optimized parameters c and g per prediction task. The parameters for SVM were optimized by means of a Grid search on a randomized subset (5,000 instances) of the training data, resulting in the following parameter settings:

- *color*: $c = 8.0$, $g = 0.0078125$
- *variety*: c =8.0, g =0.0078125
- *country*: c =32.0, g=0.00048828125
- *price big difference*: c =8.0, g=0.03125
- *price small difference*: c =8.0, g=0.0078125

5 Results

Table 1 presents the classification results per wine property for three system flavors: (1) feature vectors including BoW, (2) feature vectors combining BoW features, LDA and Word2Vec clusters, and (3) combined feature vectors trained with an SVM classifier with optimized hyperparameters c and g. The results confirm the initial hypothesis that adding semantic information helps the classifier. In addition, optimizing the c and g parameters for the LIBSVM RBF kernel results in markedly higher classification scores.

To get some insight into what terms are important for these classification results, we computed chi-square feature weights on the training set of examples for the different tasks. The top-10 features with highest chi-square values are shown in Table 2[5].

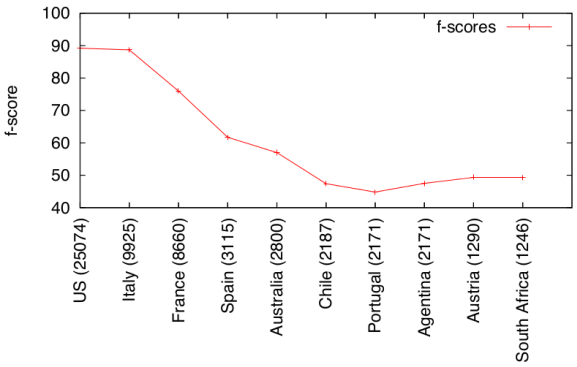

Figure 1: More training material leads to better individual F-scores, as shown for the 10 most frequent country classes.

The classifier for color achieves a rather high F-score, as illustrated in Table 3. The rosé category is the odd one out with a markedly lower F-score. There are two main reasons for this. First, rosé is a low-frequency class compared to the other two classes. Second, rosé wine is made from red grapes, but the grapes are processed in a different way to red wines. Therefore, we expect to find a certain amount of overlap between red and rosé. When we examine the confusion matrix of the classifiers' predictions on the test set, we see that, indeed, most errors are due to misclassifying rosé as red wine.

One could argue color prediction from wine reviews is trivial where the wine color is actually mentioned in the review. Therefore, we also performed an additional experiment with a BoW feature set (with optimized SVM parameters) where the words *red, white,* and *rosé* were removed. This affected the overall F-score by 2.2 points, with the

[5]The variety features are all grape names. For the country features: prokubac is a Serbian grape variety, meoru is a Korean grape that grows at mount Jiri. Yves refers to a French producer. Calatrasus is an erroneous lemma form predicted by the Stanford toolkit for the Italian wine producer Calatrasi.

color	rosé, cherry, tannin, apple, peach, citrus, pear, blush, black, pineapple, chardonnay
variety	aglianico, barbera, prosecco, viognier, moscato, malbec, sirah, carmenère, chenin, zin, franc
country	korea, jiri, rose-like, meoru, morocco, serbian, yves, calatrasus, chocolate-cherry, prokupac
price	year, tannin, age, rich, blackberry, black, vineyard, cellar, currant, vintage, simple

Table 2: Top 10 features based on Chi-Square measures on the training set.

class	#number	prec	recall	f-score
red	9296	97.7	99.0	98.4
white	4582	97.6	97.1	97.4
rose	335	94.5	66.3	77.9

Table 3: Results with optimized and combined SVM for color classification on the test set.

decrease due mostly to the performance drop of the rosé class from an F-score of 76.0 to 31.7.

For the property country we see that more training material has a positive effect on the individual scores, as visualized in Figure 1.

For grape varieties we find individual F-scores varying between 82.4 (Chardonnay grape) and 30.8 (Grenache). The Tempranillo grape, for example, is known for its rather neutral profile, and as a consequence it is often used in blends. The classifier could only distinguish the Tempanillo variety at a moderate rate (F-score 47.5), and the confusion matrix showed it is confused with Cabernet Sauvignon, Malbec, Pinot Noir, and Syrah. Varieties that were relatively easy to predict were Grüner Veltliner (F-score 74.3) and Nebbiolo (F-score 78.6). These grapes are rather strictly bound to geographic areas (Nebbiolo is from the region Piemonte, Italy and Grüner Veltliner is a typical Austrian grape). Cabernet Sauvignon (F-score 68.4) and Syrah (F-score 65.2) are common grapes for which we had many training examples, but they were often wrongly predicted as labels, leading to low precision. We are aware the location of wineries can strongly influence the sensory properties of a wine. The higher scores for grape varieties which are clearly tied to a particular region further confirms this.

With regard to the price, the more relaxed version (*price big difference*) does not seem to benefit from adding semantic features. An analysis of the classification output revealed the trained SVM model nearly always predicts the majority class for both the BoW and combined features, whereas the optimised version predicts both classes with an F-score of 94.6%. In future research, we intend to recast the price classification as a regression task.

6 Conclusions

We have demonstrated that wine experts are capable of describing wines in a sufficiently consistent manner that we can use their descriptions to predict the properties of a wine based solely on its review. Using existing NLP tools and techniques, we were able to produce classifiers that could predict the color, grape variety, price and country of origin of thousands of wines with high F-scores.

This study is a first step in a larger investigation into the relationship between expert language and sensory descriptions. We are particularly interested in lexical descriptors used for smells and flavors, and aim to study the specific terminology at the phrase level. It would also be informative to know to what extent the wines were classified on the basis of smell and flavor descriptions per se, as opposed to other information provided in the reviews, such as vineyard or producer descriptions, for example. The present models cannot address this. In addition, it is interesting to investigate questions of genre and style. For example, we could ask to what extent does the writing style of an author, or the wine ratings, affect these results. Finally, we expect there are differences in the way wines are described in different countries and different languages. Ultimately a multilingual, multinational comparison of wine reviews could uncover further insights into the human linguistic potential for describing complex smells and flavors.

Acknowledgments

Part of this work was funded by The Netherlands Organization for Scientific Research: NWO VICI grant “Human olfaction at the intersection of language, culture and biology”, project number 277-70-011.

References

Davis M. Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Frédéric Brochet and Denis Dubourdieu. 2001. Wine descriptive language supports cognitive specificity of chemical senses. *Brain and Language*, 77:187–196.

Rosario Caballero and Ernesto Suárez-Toste. 2010. A genre approach to imagery in winespeak: Issues and prospects. *Researching and applying metaphor in the real world*, 26:265–288.

William S Cain. 1979. To know with the nose: keys to odor identification. *Science*, 203(4379):467–470.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Deng-feng Chen, Qi-chun Ji, Liang Zhao, and Hongcai Zhang. 2009. The classification of wine based on pca and ann. In Bingyuan Cao, Tai-Fu Li, and Cheng-Yi Zhang, editors, *Fuzzy Information and Engineering Volume 2*, volume 62 of *Advances in Intelligent and Soft Computing*, pages 647–655. Springer Berlin Heidelberg.

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. 2009. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553.

Ilja Croijmans and Asifa Majid. 2016. Not all flavor expertise is equal: The language of wine and coffee experts. *PLoS ONE*.

Richard Gawel and Peter Godden. 2008. Evaluation of the consistency of wine quality assessments from expert wine tasters. *Australian Journal of Grape and Wine Research*, 14(1):1–8.

Robin Goldstein, Johan Almenberg, Anna Dreber, John W Emerson, Alexis Herschkowitsch, and Jacob Katz. 2008. Do more expensive wines taste better? evidence from a large sample of blind tastings. *Journal of Wine Economics*, 3(01):1–9.

Helene Hopfer and Hildegarde Heymann. 2014. Judging wine quality: Do we need experts, consumers or trained panelists? *Food Quality and Preference*, 32:221–233.

Øyvind Horverak. 2009. Wine journalism—marketing or consumers' guide? *Marketing Science*, 28(3):573–579.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Q Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 957–966.

Harry T Lawless. 1984. Flavor description of white wine by "expert" and nonexpert wine consumers. *Journal of Food Science*, 49(1):120–123.

Sébastien Lecocq and Michael Visser. 2006. What determines wine prices: Objective vs. sensory characteristics. *Journal of Wine Economics*, 1(01):42–56.

Stephen C Levinson and Asifa Majid. 2014. Differential ineffability and the senses. *Mind & Language*, 29(4):407–427.

Asifa Majid and Niclas Burenhult. 2014. Odors are expressible in language, as long as you speak the right language. *Cognition*, 130(2):266–270.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Andrew McCallum. 2002. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Elin McCoy. 2006. *The Emperor of Wine: The Rise of Robert M. Parker, Jr., and the Reign of American Taste*. Harper Perennial.

Tomas Mikolov, ILya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Edward Oczkowski and Hristos Doucouliagos. 2014. Wine prices and quality ratings: A meta-regression analysis. *American Journal of Agricultural Economics*, page aau057.

Jonas K Olofsson and Jay A Gottfried. 2015. The muted sense: neurocognitive limitations of olfactory language. *Trends in cognitive sciences*, 19(6):314–321.

Carita Paradis and Mats Eeg-Olofsson. 2013. Describing sensory experience: The genre of wine reviews. *Metaphor and Symbol*, 28(1):22–40.

Wendy V Parr, David Heatherbell, and K Geoffrey White. 2002. Demystifying wine expertise: olfactory threshold, perceptual skill and semantic memory in expert and novice wine judges. *Chemical Senses*, 27(8):747–755.

Caroline Sezille, Arnaud Fournel, Catherine Rouby, Fanny Rinck, and Moustafa Bensafi. 2014. Hedonic appreciation and verbal description of pleasant and unpleasant odors in untrained, trainee cooks, flavorists, and perfumers. *Front. Psychol*, 5:12.

Michael Silverstein. 2006. Old wine, new ethnographic lexicography. *Annual Review of Anthropology*, 35:481–496.

Ernesto Suárez Toste. 2007. Metaphor inside the wine cellar: On the ubiquity of personification schemas in winespeak. *Metaphorik. de*, 12(1):53–64.

Debbie Vigar-Ellis, Leyland Pitt, and Albert Caruana. 2015. Knowledge effects on the exploratory acquisition of wine. *International Journal of Wine Business Research*, 27(2):84–102.

Gesualdo M. Zucco, Aurelio Carassai, Maria Rosa Baroni, and Richard J. Stevenson. 2011. Labeling, identification, and recognition of wine-relevant odorants in expert sommeliers, intermediates, and untrained wine drinkers. *Perception*, 40(5):598–607.