Which Wine is that? A NLP Approach to Wine Variety Detection

Nirogi Surya Priyanka
Geethanjali College of Engineering and Technology, Hyderabad, India

Saumya Borwankar
Insitute of Technology, Nirma University, Gujrat, India

Jaynil Shah
Sardar Vallabhbhai Patel National Institute of Technology, Surat, India

Sai Rohith
Manipal institute of technology, Manipal, India

Deepankur Kansal
Indian Institute of Technology, Kanpur, Meerut, India

N Ganesh Rohit
Mahatma Gandhi Institute of Technology, Telangana, India

Jai Prakash Verma
Insitute of Technology, Nirma University, Gujrat, India
Corresponding author: Saumya Borwankar Email: 17bec095@nirmauni.ac.in

Detecting wine variety based on the country, year of origin and the review alone is a very difficult task. This problem is converted into a classification problem using the combination of Natural Language Processing (NLP) and machine learning.

2021. In Rahul Srivastava & Aditya Kumar Singh Pundir (eds.), *New Frontiers in Communication and Intelligent Systems*, 137–149. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-81-95502-00-4-17

A public dataset that consisted of wine variety corresponding to its country, review, designation, province, winery and year was used for analysis of proposed approach. Natural Language Processing is used as a preprocessing step in our approach along with neural network to build a classifier. For more robust training k-fold cross validation technique was also looked and implemented for the same. The evaluation metric chosen to analyse the performance was accuracy. 98% accuracy was attained on the evaluation set with our model. To our best knowledge, this is the first attempt to exploit NLP and neural networks to deal with prediction of wine variety.

Keywords: Wine prediction, Neural Networks, Deep Learning, Machine Learning, Text Analytics.

1 Introduction

A wine steward or sommelier is a trained and learned wine expert who specializes in all aspects related to wine, for example wine and food pairing or wine detection. Becoming a sommelier takes many years and is a very difficult course at the same time. Sommeliers play an important role in getting a hold of the best goods available to fulfil restaurant customers' expectations as wine becomes more popular and consumers become more informed [1]. Sommeliers have a significant impact on the sale of wine in restaurants, particularly in smaller establishments and fine dining establishments. When picking wines and suggesting wines to consumers, respondents focused on value for money, vineyard reputation, kind of variety, and tracking customer preference, among other things [1].

The use of wine has grown in recent years because it has a good relationship with health, particularly in terms of controlling heart rate variability [2]. Because of the growing consumption of wine, the wine industry was obliged to undergo certification and certain quality assurance testing. Along with this, they are conscious of the cost, which is an essential consideration in preserving the wine's quality. Diverse varieties of wine serve variable functions, and the chemical content of each variety of wine is likewise different.

It is essential to comprehend the contribution of different chemical characteristics utilized in different types of wine in order to retain quality at a lower cost [3]. Due to a lack of technological resources in the past, it was impossible for most companies to categorize wines based on reviews since it costed too much and took too long. With the use of machine learning, it is now feasible to categorize wines.

One essential part of human behaviour is the study of language and the ability to talk, write, and communicate. Natural language processing began after

the notion of conversing with non-human technologies was studied as the study of human languages progressed. Natural language processing is the process of developing and scheming computer systems that can analyse, comprehend, and synthesise natural human languages. Natural language is a branch of artificial intelligence which aims to comprehend and generate meaningful human-language utterances. Speech recognition, language translation, information retrieval, text summarization, and other applications of natural language processing have been developed over time.

Before we get into the intricacies, let's take a look at the stages of NLP, or the steps that a phrase goes through before a parse tree is created. NLP is divided into numerous steps depending on the application, but we will focus on three of them here: language modelling, parts-of-speech tagging, and parsing.

Any NLP application's first objective is to create a parse tree for a sentence from that language's set of sentences. However, in order to create a parse tree, one must first determine the class where all of the words belong, i.e., whether a word is an adjective, a verb, or something else. The language model is used to accurately determine the class to which a given word belongs.

Henceforth, the above expressed activities are backward order and rely upon one another as delineated in the chart. Note that the chart is explicit to the methodology determined in this review that is measurement language demonstrating, POS labeling and parsing. Certain methodologies like neural organizations, may not affirm to this ordered grouping.

The attentiveness has been grown in the wine industry in recent years which demands extension in this industry. Therefore, companies are investing in new technologies to improve wine prediction and selling. In this direction, wine prediction plays a role. This paper explores machine learning techniques to correctly predict wine quality from various reviews and descriptions. The experiments show that the prediction was more accurate if only impact-full features are used in prediction instead of all the features.

The major contributions from this work involve introduction of natural language processing in wine prediction along with careful data selection for the task of classification of wine variety.

2 Paper Organization

The paper is partitioned into 5 parts, first being the introduction, secondly the dataset, third being proposed research work, fourth being the results and lastly followed by conclusion and future work.

3 Relevant Work

NLP developments have moved forward on functions such as information retrieval, machine translation, question answering, text summarization, topic modelling, information extraction, and more recently, opinion mining since its start in the 1950s. The majority of early NLP research concentrated on syntax, partly because syntactic processing was clearly required, and partly because the notion of syntax-driven processing was implicitly or explicitly endorsed. One of the inborn issues that raises its head a few times in NLP is the issue of uncertainty. Scientists need to manage uncertainty in pretty much every period of processing.

There are comparable issues in different stages. Truly, the language processing applications worked by making a standard based programming that inspected the construction of sentence to check whether it fits the design determined. Rule based methodologies before long become unmanageable for enormous principles. With a little more than 100 guidelines, the collaboration between these principles turns out to be excessively intricate.

The sheer amount of information and important decisions rendered these approaches obsolete before long. Late methods, on the other hand, employ tactics that take advantage of the deluge of data available to construct language models. At the end of the day, continuing approaches to language processing rely on information-driven approaches to achieve language acquisition goals.

Mittal et al. [4] presented an extractive based Text Summarization approach based on Genetic Algorithms. They expressed the single document as a Directed-Acyclic-Graph in this article. Each DAG edge is given a weight based on a schema described in the article. They utilise an Objective function to define the summary's quality in terms like readability (readability factor), cohesiveness (cohesion factor), and subject relevance (topic relation factor).

In [5], Tandel et al. presented a multidocument summarising approach that allows customers to condense important material from many texts supplied as a single input. This approach has the capability to save a large amount of time while also increasing efficiency. They were motivated by existing techniques such as Cluster-based, Topic-based, and Lexical Chain-based at the time. LexRank prohibits the maximising of score for sentences that are unrelated to the document's main subject.

In [6], Modi and Oza go through three single document approaches and two multi-document techniques in depth. In [7], Jain et al. developed a methodology for extractive text summarization that employed Word Vector Embedding. Ac-

cording to their article, there are four major issues to consider while retrieving data. They are identifying the most important sentences from the text, deleting superfluous material that is not essential to the paper's topic, minimising the details, and compiling the first extracted useful information into a condensed and structured report. They suggested a Word Vector Embedding technique to extract the prominent, then employed a Neural Network for Extractive Summarization utilising Supervised Learning to solve the aforementioned problems.

In [8], Nithin et al. presented an overview of recent studies on abstractive text summarization. Summarization methods are of two major types: abstractive and extractive. According to Jain et al [7], the Extractive technique will choose the document's most important sentences and create a summary from them while keeping sentence coherence and sticking to the document's subject.

The relevance of summarising and categorising product reviews was highlighted by Pawar et al in [9]. SVM and Nave Bayes were utilised as hybrid classifiers. They also came to the conclusion that as the number of classifiers grows, so does the accuracy. [10] proposes "Query-based Summarization utilising subject background knowledge" (2017).Essentially, a query-oriented method means that the summary is created depending on the query provided as an input. The query-based paradigm is ineffective since most inquiries do not contain semantic details or information. Summarization, according to Ziyan in [11] cannot produce reliable findings when a term has many meanings. As a result, specific domain expertise of the document's core subject is also required. This highlights the importance of text summarization. However, the issue comes when the referencing is done incorrectly. As a result, this study presents a coreference resolution technique for resolving this issue and delivering correct results. In a similar vein, Paul Gigioli, Nikhita Sagar, Anand Rao, and Joseph Voyles [12] added a deep reinforced abstractive summarization method capable of going through biomedical paragraphs and summarising them into a short summary.

There are wide range of methods for calculating the alikness of two documents, including TFIDF (Term frequency-inverse document frequency), LSA (Latent Semantic Analysis) and Cosine Similarity. Manhattan distance and Euclidean distance. Every technique uses a different process to determine similarity [13], [14].

The underlying meaning or notion of the document file is discovered using LSA approaches. When each word refers to a single format and each context is described using only one word, the strategy is useful. And LSA is merely mapping the document's words to concepts [15]. A two-dimensional approach in nature since it calculates the distance between two dimensions (horizontal and vertical) [16]. Jai et al. [17] in their paper discusses about computer-assisted evaluation

techniques are sometimes referred to as objective questionnaires. Text analytics considers evaluation based on subjective answers to be a problem, because the text answer will be compared to the accessible correct text answer. This paper emphasises the problems with computer-assisted automated evaluation and proposes a paradigm for dealing with them.

To detect sentiments in text, there are two primary approaches. Symbolic approaches and Machine Learning techniques [18] are the two types. The usage of available lexical resources is used in a lot of unsupervised sentiment categorization studies employing symbolic techniques. For sentiment analysis, Turney [19] employed a bag-of-words technique. Relationships between individual words are ignored in this technique, and a document is a bunch of words. The total sentiment is calculated by determining the sentiments of each word and combining those values using some aggregation methods.

The lexical database WordNet [20] was utilised by Kamps et al. [21] to determine the emotional content of a word along several dimensions. They used WordNet to create a distance metric and assess the semantic orientation of adjectives. The WordNet database is made up of words that are linked through synonym relationships.

Baroni et al. [22] developed a method that uses word space model formalism to solve the lexical substitution problem. It depicts a word's local context as well as its overall dissemination. EmotiNet, a conceptual representation of text that stores the structure and semantics of real events for a certain domain, was introduced by Balahur et al. [23]. Emotinet identified emotional responses produced by events using the notion of Finite State Automata.

4 Dataset

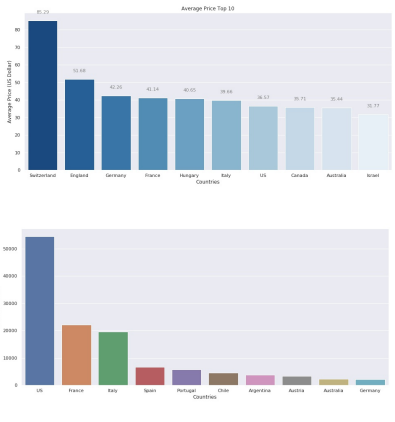
The dataset is publically available on Kaggle [24]. The dataset has around 130,000 reviews and 10 columns containing country, description, designation, points, price, province, region1, region2, taster-name, taster-twitter-handle, title, variety, winery.

1. Country: Represents the country of wine.
2. Description: Tells us about the wine's smell, taste, look and feel.
3. Designation: Shows us where the wine's grape were grown.
4. Points: Tells us about the total number of points The wine was evaluated

on a scale of 1-100 by a Sommelier (albeit they claim to only post ratings for wines with a score of >=80).

5. Price: Tells us what the bottle of wine costs.
6. Province: The state from where the wine is from.
7. Region1: The area in that province.
8. Region2: More accurate region withing the area.
9. Taster-name: The person's name who reviewed and tasted the wine.
10. Taster-twitter-handle: The twitter handle of the taster.
11. Title: The title of the review and usually contains the vintage.
12. Variety: The grapes that were used to make the vine
13. Winery: The winery that made the wine.

Figure 1(a) tells us about the average price of the top 10 countries and Figure 1(b) tells us about the number of wines tasted according to countries. According to figure 1(a) wine from Switzerland is the costliest, and according to figure 1(b) US has the most wine tasted.



[a]

[b]

Figure 1: (a) Average Price of Top 10 wines and where they come from.
(b) Number of wine tasted according to countries

5 Proposed Research Work

The implementation was carried out on a machine having 8GB DDR4 RAM, a 4Gb Nvidia GTX1050 Graphics Processing Unit (GPU), and an Intel Core i7-7700HQ Central Processing Unit (CPU) 2.80GHz which is a 64-bit processor. An flow diagram of our approach is given in figure 2. For our proposed work we have described each component in the following part:

1. Data Cleaning: We extract useful information such as years from review-title column and make a new column named 'data-vector'. Along with the year we concatenate rest of the columns except winery column. So this makes our final data column which has all the information regarding the wine variety.

2. Data pre-processing: Removing stop words from the 'data-vector' column. After which Term frequency inverse document frequency (TF-IDF) takes place for the list of word obtained. Resulting in a final vector. Along with this the wine variety column is label encoded to correctly fit into the dimensions of our neural network.

3. Dividing data: The data is then split into 5 folds to perform 5-fold cross-validation.

4. Training: For training the final vector is passed through our proposed neural network described in section 5.1.

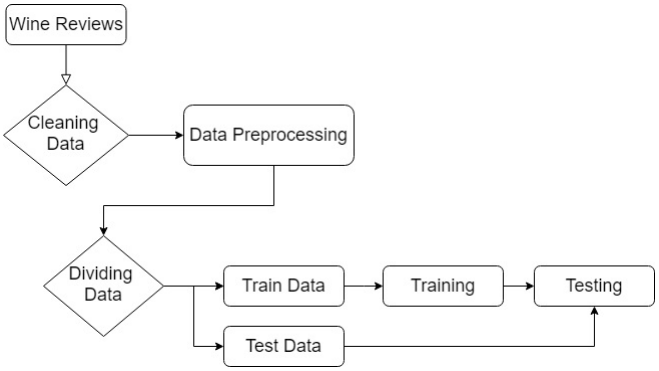


Figure 2: Flow diagram for our approach.

Data cleaning was done which involved removing redundant information and importance of related features was studied along with the preprocessing of the

data which involved dividing the raw data into various categories, to train the neural network. NLP techniques were used to preprocess the data which involved removing stop words and using term frequency–inverse document frequency for every feature before feeding it to the neural network. The neural network was then trained, which had features like country, review etc. and the target feature as wine variety. The model was able to perform efficiently.

5.1 Neural Network Architecture

For our neural network architecture we have a sequential model with just 2 dense layers. There was no need to have deep layers as the shallow network was able to perform satisfactorily. The first dense layer has 100 neurons and output layer has the total number of unique wine variety. The loss function used is sparse categorical crossentropy with 'adam' optimizer and the accuracy metric was used to evaluate the model. The approach was also validated with the help 5 fold cross validation.

5.2 Stop words and Term frequency inverse document frequency

Stop words usually are words that are often employed in a language. Stop words in English include "a," "the," "is," "are," and others. Stop words are frequently used in NLP to remove terms that are so widely used that they contain little relevant information.

Term Frequency Inverse Document Frequency is abbreviated as TF-IDF. This is a popular technique for converting text into a meaningful numerical representation, which is then used to fit a machine learning algorithm for prediction. The TF-IDF is a subtask of information retrieval and extraction that tries to represent the significance of a word to a document that is part of a corpus (a collection of documents).

6 Results

To evaluate our approach, accuracy as a metric was chosen. Accuracy in this case will make us aware how well the model is performing in predicting wine variety based on its review. The results are mentioned in table 1. We could not compare it to other methods as this is the initial approach to such problem to our knowledge. With this approach normal people can easily find the variety of the wine with just review. This can be incredibly useful for people wanting to get started with being a sommelier.

Table 1: Experiment results

Approach	Accuracy	Precision	Recall	F1 score
Our method	96.92%	97.3%	96.4%	97.78%

7 Conclusion

Being able to detect wine from a review can be a hard task and so can be the detection of wine variety from blind tasting of the wine. This paper includes the power of machine learning and NLP to overcome this problem and in turn help people leverage this method to correctly identify wine variety based on the review of the wine. In future, other datasets may be looked at for training while improving the complexity of the architecture of the model for more accurate prediction.

References

[1] B. B. Dewald, "The role of the sommeliers and their influence on US restaurant wine sales", *Int. J. Wine Business Res.*, 2008.

[2] I. Janszky et al., "Wine drinking is associated with increased heart rate variability in women with coronary heart disease", *Heart*, vol. 91, no. 3, pp. 314–318, 2005.

[3] V. R. Preedy, "Electronic noses and tongues in food science", Academic Press, 2016.

[4] N. Chatterjee, A. Mittal, and S. Goyal, "Single document extractive text summarization using genetic algorithms," in 2012 *Third Interna- tional Conference on Emerging Applications of Information Technology*, IEEE, 2012, pp. 19–23.

[5] A. Tandel et al., "Multi- document text summarization-a survey," in *International Conference on Data Mining and Advanced Computing*, IEEE, 2016, pp. 331–334.

[6] S. Modi and R. Oza, "Review on abstractive text summarization techniques (atst) for single and multi documents," in *International Conference on Computing, Power and Communication Technologies*, IEEE, 2018, pp. 1173–1176.

[7] A. Jain, D. Bhatia, and M. K. Thakur, "Extractive text summarization using word vector embedding," in *International Conference on machine learning and data science* IEEE, 2017, pp. 51–55.

[8] N. Raphal, H. Duwarah, and P. Daniel, "Survey on abstractive text summarization," in *International Conference on Communication and Signal Processing*, IEEE, 2018, pp. 0513–0517.

[9] R. Boorugu and G. Ramesh, "A survey on NLP based text summariza- tion for summarizing product reviews," in *Second International Conference on Inventive Research in Computing Applications*, IEEE, 2020, pp. 352–356.

2021. In Rahul Srivastava & Aditya Kumar Singh Pundir (eds.), *New Frontiers in Communication and Intelligent Systems*, 137–149. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-81-95502-00-4-17

[10] Y. Wei and Y. Zhizhuo, "Query based summarization using topic background knowledge," in *13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, IEEE, 2017, pp. 2569–2572.

[11] Z. Shi, "The design and implementation of domain-specific text sum- marization system based on co-reference resolution algorithm," in *Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, vol. 5. IEEE, 2010, pp. 2390–2394.

[12] P. Gigioli et al., "Domain-aware abstractive text summarization for medical documents," in *IEEE International Conference on Bioinformatics and Biomedicine*, 2018, pp. 2338–2343.

[13] "Web content available on dated: 29-08-2016 on the link: http://www.irfacility.org/scoring-and-ranking-techniques-tf-idf-term-weighting-andcosine-similarity," 2016.

[14] P. Verma, B. Patel, and A. Patel, "Big data analysis: recommendation system with hadoop framework," in *IEEE International Conference on Computational Intelligence Communication Technology*, 2015, pp. 92–97.

[15] "Web content available on dated: 29-08-2016 on the link:https://technowiki.wordpress.com/2011/08/27/latent-semantic-analysislsa-tutorial/," 2016.

[16] "Web content available on dated: 29-08-2016 on the link: http://www.cuttheknot.org/pythagoras/distanceformula.shtml," 2016.

[17] N. Dave, H. Mistry, and J. P. Verma, "Text data analysis: Computer aided automated assessment system," in *3rd International Conference on Computational Intelligence Communication Technology* IEEE, 2017, pp. 1–4.

[18] E. Boiy et al., "Automatic sentiment analysis in on-line text." in *ELPUB*, pp. 349–360, 2007.

[19] P. D. Turney, "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews," arXiv preprint cs/0212032, 2002.

[20] C. Fellbaum, "Wordnet: An electronic lexical database (language, speech, and communication), illustrated edition edn," 1998.

[21] J. Kamps et al., "Using wordnet to measure semantic orientations of adjectives." in *LREC*, vol. 4. Citeseer, pp. 1115–1118, 2004.

[22] D. Pucci et al., "Unsupervised lexical substitution with a word space model," in *Proceedings of EVALITA work- shop, 11th Congress of Italian Association for Artificial Intelligence* Citeseer, 2009.

[23] A. Balahur, J. M. Hermida, and A. Montoyo, "Building and exploiting emotinet, a knowledge base for emotion detection based on the appraisal theory model," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 88–101, 2011.

[24] "Kaggle dataset. Available at https://www.kaggle.com/zynicide/wine-reviews", 2017.