Deep Learning Wine Grape Variety Recognition Proposal

APS360 Applied Fundamentals of Machine Learning
Instructor: Yani Ioannou

Prepared by:

Group 18

Chen, Ryan	Deng, Sara	Tian, Yunying (Kyra)	Wang, Yining
1003912992	1003218109	1005975768	1005728134

June 2, 2021

Word Count:1356/1400

Contents

1 Introduction

Canadians have a liking for wine in general compared to other alcohols including beer and cider, with the retail sales of wine forecasted to exceed 13.2 billion U.S. dollars in 2022 [1]. There is an expected increase in wine demand from consumers and a bounce for overall sales post-COVID as hospitality, cruise lines, sports and concerts reopen, and deferred events such as weddings take place [2], and consequently, so does the demand for technologies that support customers' user experience. As channel shifting became the byword for 2020 in the wine industry, customers started to adapt and go online for purchases, so there is a remarkable direct-to-customer channel movement [3]. Opportunities for marginal growth will be found by investing in e-commerce and digital strategies that help people, especially customers, to extract basic key information from perplexing online wine reviews [3]. The goal of the project is to envision grape variety from the wine taste descriptions. Considering wine is produced dominantly by fermenting grapes and sometimes other fruits or plants [1], the scope of this project is narrowed to focus on wine fermented with grapes only to ensure a list of standard and predictable output categories.

Machine learning, especially supervised deep learning for classification, is considered a reasonable approach for this grape variety recognition task in the following aspects. The problem involves complex and large datasets, while deep learning outperforms other techniques in general if the data size is relatively large [4]. Machine learning grants direct learning from a rich experience, i.e. thousands of wine records with associated taste descriptions (as input) and grape types (as output) are collectable from the Wine Enthusiast website [4], [5]. Besides, no clear predetermined solution to the problem is available, while machine learning is capable of executing without relying on pre-established equations [6]. Moreover, uncertainty exists in terms of how the taste reviews are phrased and organized. Thus, it is suitable to apply supervised learning that makes predictions based on evidence in the presence of uncertainty [6].

2 Background & Related Work

A discovered prior work is a paper named *Convolutional Neural Networks for Sentence Classification* [7]. It investigated how the Convolutional Neural Network (CNN) can be applied to a series of experiments to train on top of pre-trained word vectors for sentence-level classification tasks [7], [8].

A simple CNN is trained with one layer of convolution on top of word vectors acquired from an unsupervised neural language model [7]. 100 billion words of Google News were taken to train the vectors [8]. Despite little tuning of hyperparameters [9], the simple CNN with one layer of convolution performed remarkably well on multiple benchmarks [7]. These include detecting positive/negative reviews from movie comments with one sentence per comment, which is in nature close to detecting grape variety from wine descriptions. The results suggest that the pre-trained vectors are good, universal feature extractors and can be utilized across datasets for various classification tasks [7]. These results proved that pre-trained word embedding models can be utilized with CNNs to classify sentences into keywords as output which serves as a feasible approach for the project.

3 Data Processing

The raw data will be collected from Wine Enthusiast which includes nearly 310,000 bottles of wine with information on vintage, variety, country, description, etc. [10]. A Python website scraping script would be used to download and store the raw data in a csv file with only two fields: variety and description, based on the design of the system [11]. The description field would be further transformed from sentences to lists of keywords that are comprehensible for the Natural Language Processing model [12]. First, the description (an entire string) would be formatted into lowercases and then use nltk.tokenize to split sentences into individual words (a list of strings). Then, numerical values and stopwords from nltk.corpus would be removed from strings (list of NLTK stopwords [13]). As a result, an updated dataset would be stored into a csv file with variety in string and description in list. However, based on the diagram of grape variety from an example dataset on Kaggle, the distribution of variety is unbalanced: 10% and 9 % of the wine are made from

Chardonnay and Pinot Noir with a total of 632 varieties [14]. Thus, when selecting input data, this would be adjusted by randomly deselecting some data from Chardonnay and Pinot Noir.

4 Architecture

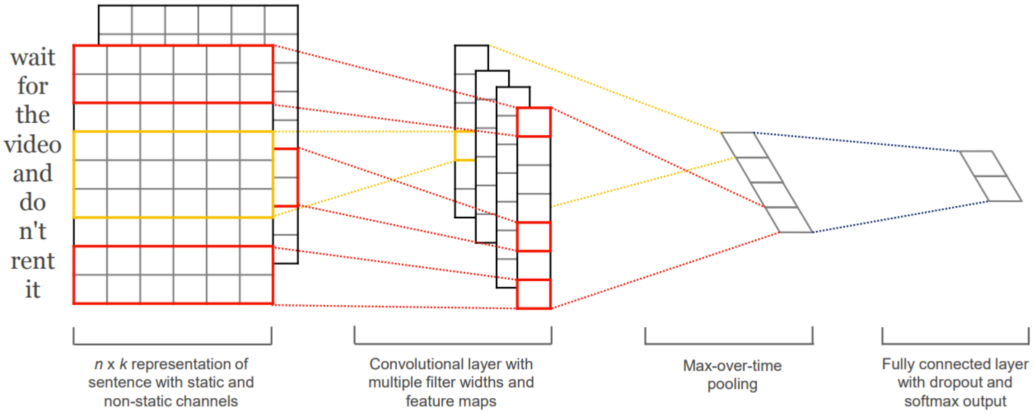
wait
for
the
video
and
do
n't
rent
it
n x k representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

Figure 1: Rough model architecture from previous work by Kim [7].

The architecture chosen to accomplish this supervised learning task is a CNN. The CNN will include a word embedding layer to transform the input sentences to tensors, convolutional and pooling layers to learn specific features, as well as fully connected layers for classification. An illustration of the model architecture is shown in Figure 1.

4.1 Input

As previously described in the Data Processing section, the input to the deep learning model will be processed wine reviews.

4.2 Word Embedding

The input texts will be transformed into tensors through the word embedding layer utilizing embedding techniques. The most popular type of approach for this transformation is also neural network based. While it is an option to train our own word embedding model for this task, pre-trained models such as word2vec [8], GloVe [15] or fastText [16] could be implemented as well. For simplicity and reliability, the GloVe method is chosen to perform text to tensor transformations.

4.3 Output

The output of the deep learning model will be one-hot encoded labels of grape varieties. The class with the highest probability will be chosen as the grape variety recognized for the given wine description input.

4.4 Backup Option

We are aware that a Recurrent Neural Network (RNN) could also be utilized to tackle this problem, but due to our limited knowledge of this particular architecture at the current stage, we will propose a CNN approach at this time. If the option to utilize a RNN is deemed to be more promising after gaining more insight into the architecture, it is a possibility to explore that path instead of the CNN method.

5 Baseline Model

The baseline model chosen to compare the neural network against is a support-vector machine (SVM). SVMs are supervised learning models based on the Structural Risk Minimization principle. They are also proven to be well-suited for text classification problems prior to the prevalent usage of deep learning for the task [17]. Thus, a SVM can be utilized as a reasonable baseline model to gauge the performance of the neural network.

6 Ethical Considerations

Based on the source website, this system is only valid for English since the description of wines are in English and the Natural Language Processing model that was used only works for English according to the stopwords and the logic behind tokenization texts. In addition, the system is limited to the grape varieties that exist on source website. If a wine with a new variety appeared, the output of the system would not be accurate.

7 Project Plan

Table 1: Project Task Schedule

Task	Member(s) Responsible	Deadline
Data Collection	Kyra	Jun 16
Data Cleaning	Kyra	Jun 18
Data Splitting (into training, validation, and testing)	Yining	Jun 20
Model Construction and Training	Yining	Jun 22
Meeting 1 (target at analyzing and evaluating the training outcome)	Together	Jun 23
Model Validation and Hyperparameter Tuning	Ryan	Jun 25
Model Testing	Sara	Jun 27
Meeting 2 (reflect on whole process and begin writing report)	Together	Jun 28
Project Progress Report Writing	Together	Jul 2
Video Presentation Preparation and Final Report Writing	Together	Jul 31

8 Risk Register

In case of any emergencies the team may face during the process, several possible risks are listed below.

8.1 Absence of Team Member

If one of the team members has an emergency situation and could not complete his/her task, his/her work would be splitted equally to three parts and distributed to the other members. Also, all of the members have guaranteed that if he/she has a private issue, he/she would convey his/her thoughts to others as soon as possible.

8.2 Loss of Data and Failure of Combining Separate Works

The team considered the failures when managing data and combining work together. The loss of data can be avoided by the team members regularly collecting and downloading the data on their own computers. For the issue of code collaboration, the team has decided to use GitHub to collaborate, which would allow the members to compile different versions of the software and commit their work to the remote team repository. Git would prevent potential conflicts when merging from different versions(branches).

8.3 Biases of Dataset

Since the dataset is large and complex, it is very easy to have biases when collecting the data. For instance, the different amount of grape types can generate biases to the project (the most frequent types of grape are Chardonnay and Pinot Noir [14], while the least frequent type is Portuguese Red). In order to avoid such biases, the team decided to randomly deselect data from the most prevalent grape types for the training dataset. When facing other biases, taking the average can also be a potential solution.

9 GitHub

GitHub will be utilized for version control and project collaboration purposes. The link to the project repository is: `https://github.com/saradeng/APS360.git`

References

[1] J. Conway, "Wine market in canada: Statistics and facts," Dec 2020. [Online]. Available: https://www.statista.com/topics/2996/wine-market-in-canada/#:~:text=Retail%20sales%20of%20wine%20in,in%20Nunavut%20purchase%20the%20least

[2] "Wine industry trends and report 2021," May 2021. [Online]. Available: https://www.svb.com/trends-insights/reports/wine-report

[3] "Wine industry expectations for 2021 to 2022," Apr 2021. [Online]. Available: https://estatevineyard.com/tag/wine-industry/

[4] S. Mahapatra, "Why deep learning over traditional machine learning?" Mar 2018. [Online]. Available: https://towardsdatascience.com/why-deep-learning-is-needed-over-traditional-machine-learning-1b6a99177063

[5] "Wine enthusiast: 318, 187 search results for wine reviews," Jun 2021. [Online]. Available: https://www.winemag.com/?s=&amp;drink_type=wine.

[6] "Machine learning: How it works, techniques applications." Jun 2021. [Online]. Available: https://www.mathworks.com/discovery/machine-learning.html.

[7] Y. Kim, "Convolutional neural networks for sentence classification." Jun 2021. [Online]. Available: https://arxiv.org/pdf/1408.5882.pdf.

[8] M. Etal, "Google code archive for pre-trained vectors." Jun 2021. [Online]. Available: https://code.google.com/archive/p/word2vec/.

[9] M. D. Zeiler, "Adadelta: An adaptive learning rate method," Dec 2012. [Online]. Available: https://code.google.com/archive/p/word2vec/.

[10] "Best wine ratings amp; reviews online: Wine enthusiast magazine."

[11] Z. Thoutt and V. Mukhtarulin, "Zack thoutt - wine deep learning," July 2020. [Online]. Available: https://github.com/zackthoutt/wine-deep-learning

[12] G. Bedi, "Simple guide to text classification(nlp) using svm and naive bayes with python," July 2020. [Online]. Available: https://medium.com/@bedigunjit/simple-guide-to-text-classification-nlp-using-svm-and-naive-bayes-with-python-421db3a72d34

[13] S. Bleier, "Nltk's list of english stopwords," 2010. [Online]. Available: https://gist.github.com/sebleier/554280

[14] Z. Thoutt, "Wine reviews," Nov 2017. [Online]. Available: https://www.kaggle.com/zynicide/wine-reviews?select=winemag-data_first150k.csv.

[15] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” Aug 2014. [Online]. Available: https://nlp.stanford.edu/projects/glove/

[16] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “fasttext: Library for efficient text classification and representation learning,” Aug 2016. [Online]. Available: https://fasttext.cc/

[17] T. Joachims, “Text categorization with support vector machines: Learning with many relevant features,” Jun 2005. [Online]. Available: https://link.springer.com/chapter/10.1007/BFb0026683