Specialty Coffee Origin Classification

Alan Cacique Tamariz
Linköping University / Linköping, Sweden
alaca734@student.liu.se

Abstract

Coffee reviews, specifically "Blind Assessments," are a type of review where the tester, a Q-Grader, does not know anything about the coffee being graded, including his origin. They have been essential in describing specialty coffees to consumers, from coffee 'aficionados' to green coffee bean buyers. In this project, the main idea was to prove whether the Text Classification Models can predict coffee origin based on the Blind Assessment or not. The data was gathered through web scrapping, and 1954 reviews from 31 countries were obtained. Dummy Classifier (random), Multinomial Naive Bayes (MNB), and NuSVC versus Distil-Bert were applied and compared using a balanced and imbalanced dataset. In both contexts, NuSVC performed better than the rest, particularly regarding Distil-Bert, which is considered the state of the art. It was found that the Blind Assessment has some bias derived from the taste of the coffee professional taster, contributing to the overall complexity of the task and consequently affecting the scores of the methods.

1 Introduction

Approximately during the last 23 years, specialty coffee has grown at a fast pace. The Specialty Coffee Association (SCA) organized the first Barista competition in 2000 (KUDAK, 2019), alongside a few coffee discipline additions such as the Brew and Latte art competition.

The rise of specialty coffee has brought innovation in different areas, not only in methods but also in describing the coffee flavor. This gave origin to Q grading and Q-graders, who are responsible for grading the coffee and providing a professional description that includes a sensory analysis.

These professional descriptions sometimes come as "Blind Assessments" because the current grader knows nothing about the coffee that is about to be tested. But how is this related to Text Mining and its applications?

Text classification is part of the conceptual framework of Text Mining, which falls under the categorization part (Zhai and Massung, 2016). Qian Li et al. (Li et al., 2022) define text classification as the procedure of designating pre-defined labels for text, which is an essential and significant task in many Natural Language Processing applications like topic labeling. From 1960 to 2010, Naive Bayes, KNN, and SVM models (traditional) were the most used for classification; from 2010 until today, methods based on deep learning are used more (Li et al., 2022).

Origin or geographical growing location, among volatile and non-volatile compounds (chemical composition), processing, roasting, and cup preparation contribute to the flavor of the coffee, which is considered the final output (Sunarharum et al., 2014). Taking this into account, the relationship between Coffee and Text Mining, particularly Text Classification, relies on another question: Can the origin of a Coffee be predicted just based on its "Blind Assignment"?

This project focuses on applying different traditional and contemporary Text Classification methods. It compares them and analyzes whether a straightforward description is enough to differentiate the origin of distinct single-origin coffees or if other factors must be considered.

2 Theory

2.1 Specialty Coffee

Specialty coffee refers to coffees that scored 80 points or more on a scale from 0 to 100, graded according to the Specialty Coffee Association (SCA). It has been performed by a certified Q-Grader(Quality Grader). The bottom limit of 80 points means it is a very good coffee, and the upper limit of 100 is outstanding.

The evaluation involves testing a sample of 350 grams of coffee in two phases. In the green phase, the beans are analyzed before roasting by a visual inspection whose principal objective is to find the number of Primary and Secondary defects, among which are black beans and broken beans, respectively. Then, the coffee is roasted according to the SCA standards, and the cupping phase starts with brewing it by pouring 150 ml of water per 8.25 grams of coffee at 200° C. The sensory analysis of different properties such as fragrance/aroma, flavor, aftertaste, acidity, body, balance, sweetness, clean cup, uniformity, overall aspect, and defects (SCA).

The Q-Graders are the only professionals authorized and certified by the Coffee Quality Institute to perform the evaluation. To become one, several courses about taste, flavors, aroma, and green coffee must be taken and passed. All the procedure from the tester to the grader is standardized internationally, which helps to give repeatability and achieve high standards (CI-).

2.2 Multinomial Naive Bayes

Multinomial Naive Bayes is a supervised probabilistic learning method described by the equation in Figure 1. The main objective is to compute the probability of document d being in class c; in other words, to obtain the most likely best class or maximum a posteriori (Manning et al., 2008).

$$P(c|d) \propto P(c) \prod_{1\leq k\leq n_d} P(t_k|c)$$

Figure 1: Multinomial NB

Logarithms are introduced to each term to compute the maximum a posteriori class, and an addition can be made. This is to avoid floating point underflow due to the multiplication of probabilities. Then, the following formula is applied (Manning et al., 2008).

$$c_{map} = \underset{c\in\mathbb{C}}{\arg\max}\,[\log \hat{P}(c) + \sum_{1\leq k\leq n_d} \log \hat{P}(t_k|c)].$$

Figure 2: Maximum a posteriori class

$P(t_k|c)$ refers to the conditional probability of the term t_k occurring in a document of class c. $P(c)$ is the probability of a document occurring in class c. t_k is a token, which is part of the tokens in the document d, that are considered part of the vocabulary used for the classification $< t_1, ..., t_{n_d} >$. The following formulas are applied to compute the terms within.

$$\hat{P}(c) = \frac{N_c}{N},$$

Figure 3: Prior probability

$$\hat{P}(t|c) = \frac{T_{ct}+1}{\sum_{t'\in V}(T_{ct'}+1)} = \frac{T_{ct}+1}{(\sum_{t'\in V} T_{ct'})+B'},$$

Figure 4: Conditional Probability

To compute the prior probability, we divide the number of documents in class c N_c by the sum of all documents from all classes N. This is the relative frequency. For the conditional probability, the relative frequency of the term t in documents from class c. Where T_{ct} is the number of times the term t is in the documents from the training that are from class c. In addition, a "+1" is added or a Laplace smoothing ($B = |V|$ =total number of terms in the vocabulary) to eliminate zeros. This is because the MLE is zero for some combination of term-class not represented in the training data set.

2.3 Support Vector Machine NuSVC

The Support Vector Machine is a supervised learning binary classifier whose primary objective is to find a linear boundary that helps separate data points from different classes in a certain space.

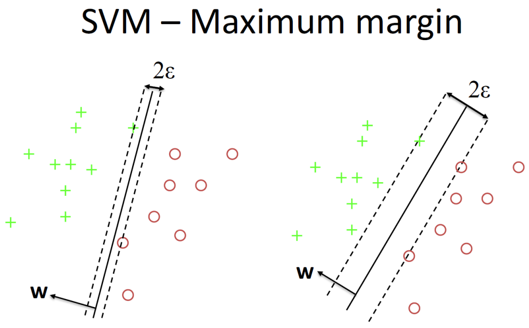


Figure 5: SVM-Margin (Borga, 2023)

As Figure 5 shows, a middle boundary has a margin (ε) to each side. These margins can be supported in one or more data points, these are the so-called support vectors.

$$\min_{\mathbf{w},w_0,\xi_i}\left\{||\mathbf{w}||^2 + \overset{\text{user-defined trade-off parameter}}{C}\sum_{i=1}^{K}\xi_i\right\}$$
$$\text{subject to } y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 - \overset{\text{slack variable}}{\xi_i}$$

Figure 6: SVM Loss function (Borga, 2023)

To obtain the greatest number of correct samples predicted, the model maximizes the margin by minimizing the parameter (W) that is the weight vector. The penalty term C defines how much penalty a misclassification receives and, consequently, how narrow the margin is and the lower number of support vectors. The term ζ is the slack variable that indicates the quantity of data points that can be on the incorrect side of the margin. These terms give flexibility to the model to allow a better fit to the data. If a data sample x_i multiply by W^T and added by W_0 is bigger then 1-ζ, it means that it is outside the margins (Borga, 2023).

In this project, SVM is called SVC, Support Vector Classifier. The difference between SVC vs. Nu-SVC is the change of the penalty term C for ν. This term controls the number of support vectors and margin errors; it is a lower bound of the fraction of support vectors and an upper bound on the fraction of margin errors (SchÃ¶lkopf et al., 2000).

2.3.1 One vs One

The SVC (and Nu-SVC) is a binary classifier. The One vs One method is applied to the scikit learn implementation of Nu-SVC to adapt this model to a multi-class classifier. What it does is that several one-classifiers per pair of classes are built; the number of classifiers is defined by the formula in Figure 7 (Skl).

$$\frac{nclasses*(nclasses-1)}{2}$$

Figure 7: Number of classifiers (Skl)

To make a prediction given a data sample x_i, the sample is the input of each binary classifier, and each one makes a prediction. In the end, the final output is the class with the most votes or the most predicted by the classifiers (Skl).

2.4 Distil-Bert

To better understand Distil-Bert, "the teacher" BERT should be introduced first. BERT is the acronym for Bidirectional Encoder Representations from Transformers, a Natural Language Processing model, which is trained for tasks such as sentiment analysis, text classification, named entity recognition, etc. (Muller, 2022)

The core of BERT and another LLM is the Transformer. The base form of a Transformer is the Encoder and the Decoder, which can be seen in Figures 8 and 9. The innovation behind this is using self-attention" to compute representations of its input and output without using a sequence-aligned RNNs or convolution" (Vaswani et al., 2017). Self-attention is a mechanism that allows one to relate different positions of a single sequence, a sentence in this case, to compute the representation of the sequence. BERT only uses the encoder block.

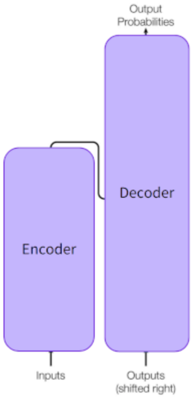


Figure 8: Transformer Encoder/Decoder (Muller, 2022)

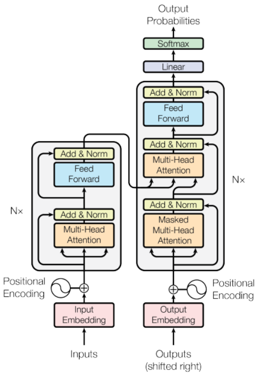


Figure 9: Transformer Layers (Vaswani et al., 2017)

The first key factor in BERT's operation is using a masked language model pre-training objective. What it does is that it takes a sentence in its tokens format, and it randomly masks or hides some of the tokens, with the "objective to predict the original vocabulary id of the word based only on its context" (Devlin et al., 2019). The second key factor is that BERT learns about relationships between two sentences using the next sentence prediction task, which pre-trains text-pair representations. In other words, BERT learns to predict if one sentence follows the previous sentences (Devlin et al., 2019).

2.4.1 Knowledge distillation

Distil-Bert is a smaller version of BERT (Figure 10) trained with a compression technique called Knowledge distillation. The main idea of this technique is that a smaller model called the student (Distil-Bert) is trained to reproduce the behavior of the bigger model called the teacher (BERT) or an ensemble of multiple models (Sanh et al., 2020). The student has the same architecture as BERT, but the token-type embeddings and the pooler are discarded and have fewer layers by a factor of 2. Another significant differentiation is that Distil-BERT uses dynamic masking but does not have a next-sentence prediction objective. All these changes combined resulted in a version of BERT that, according to its creators, is 40% smaller, 60% faster, and has 97% of BERT language understanding capabilities(Sanh et al., 2020).

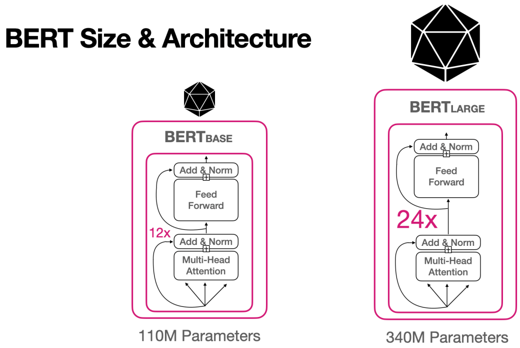
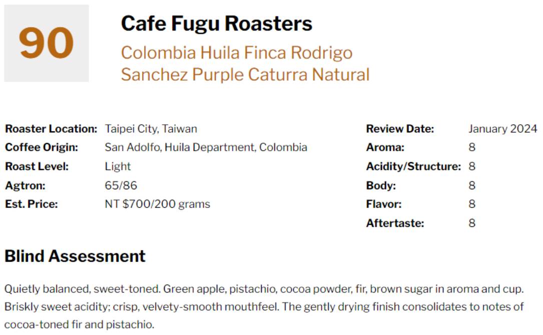


Figure 10: BERT vs Distil-BERT size comparison (Muller, 2022)

3 Data

The data used for this project was extracted from the "Coffee Review" website [1]. This website contains over 3000 specialty coffee reviews, with a grade of 90 points or higher. Each review is structured as shown in Figure 11; the parts extracted were the Blind Assessment and the Coffee Origin, specifically the country.



90

Cafe Fugu Roasters

Colombia Huila Finca Rodrigo
Sanchez Purple Caturra Natural

Roaster Location:	Taipei City, Taiwan	**Review Date:**	January 2024
Coffee Origin:	San Adolfo, Huila Department, Colombia	**Aroma:**	8
Roast Level:	Light	**Acidity/Structure:**	8
Agtron:	65/86	**Body:**	8
Est. Price:	NT $700/200 grams	**Flavor:**	8
		Aftertaste:	8

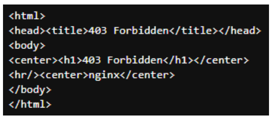
Blind Assessment

Quietly balanced, sweet-toned. Green apple, pistachio, cocoa powder, fir, brown sugar in aroma and cup. Briskly sweet acidity; crisp, velvety-smooth mouthfeel. The gently drying finish consolidates to notes of cocoa-toned fir and pistachio.

Figure 11: Coffee review (Vaswani et al., 2017)

Before performing the web scrapping, the /robots.txt[2] website was reviewed to ensure that I could extract information about the data as a student. However, it is known that even if it is not prohibited, some sites have implemented security against web scrapping.

BeautifulSoup [3] and Trafilatura [4] were used to do web scrapping. The reason why both libraries were applied is that when trying to do the extraction with Beautifulsoup, the task was being blocked (Figure 12); by test and error, it was found that if the Trafilatura library was used just for the extraction, the task was not blocked. In addition, a VPN to change the IP address was added after noticing some extra CAPTCHAs in Google search when entering a new search. Because of these situations, to not be banned entirely from the website and to not overload the site, the decision was only to extract 2000 reviews.



```
<html>
<head><title>403 Forbidden</title></head>
<body>
<center><h1>403 Forbidden</h1></center>
<hr/><center>nginx</center>
</body>
</html>
```

Figure 12: Web Scrapping (Vaswani et al., 2017)

Phases of extraction:

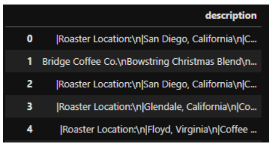
1. The hyperlinks of the first 100 pages of reviews were created by emulating the structure of the first hyperlink 'https://www.coffeereview.com/review/page/%s/'.

2. All the links that contain reviews were extracted by accessing all the hyperlinks from each review page hyperlink created. File saved in "links_reviews.txt".

3. The extraction of HTML raw text from each review was performed by accessing each hyperlink, appending it to a pandas Data Frame, and save to the file 'raw_reviews.csv'.

3.1 Preprocessing

Figure 13 shows the text structure obtained in the pandas Data Frame; this was the starting point of the preprocessing.



	description
0	|Roaster Location:\n|San Diego, California\n|C...
1	Bridge Coffee Co.\nBowstring Christmas Blend\n...
2	|Roaster Location:\n|San Diego, California\n|C...
3	|Roaster Location:\n|Glendale, California\n|Co...
4	|Roaster Location:\n|Floyd, Virginia\n|Coffee ...

Figure 13: Raw reviews

To obtain the desired data frame structure of having one column with the Blind Assessment and one from the origin, the following steps were followed:

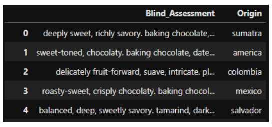
1. Create a new pandas Data frame with two columns: "Blind_Assessment" and "Origin"

[1] https://www.coffeereview.com/review/
[2] https://www.coffeereview.com/robots.txt
[3] https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[4] https://trafilatura.readthedocs.io/en/latest/

2. For each raw review, find the Origin and Blind Assessment of the coffee by finding the word before and after. For origin, the words were 'Origin:' and '|Roast'; for Blind Assessment, they were 'Assessment' and 'Notes.'

3. Then convert to lowercase and remove numbers or special characters.

4. Append respectively to the columns on the reviews Data Frame created on point 1. (Figure 14)



	Blind_Assessment	Origin
0	deeply sweet, richly savory. baking chocolate,...	sumatra
1	sweet-toned, chocolaty. baking chocolate, date...	america
2	delicately fruit-forward, suave, intricate. pl...	colombia
3	roasty-sweet, crisply chocolaty. baking chocol...	mexico
4	balanced, deep, sweetly savory. tamarind, dark...	salvador

Figure 14: Reviews pandas Data Frame

After the steps followed above and reviewing in detail the data obtained, the modifications made were the following to clean the data set:

1. Remove coffees with blends(multiple origins) or undisclosed origin.

2. Remove coffees with only one review.

3. change the origin labels 'rica' to 'costarica', 'rico' to 'puertorico' and 'republic' to 'dominicanrepublic'.

The result data frame is in the file "single_origin_coffees.csv". All the files mentioned in this report are available on the GitLab repository [5]. This data frame contains 1954 reviews from 31 different countries.

Figure 15 shows that more than 25% of the reviews come from Ethiopia, followed by Colombia with 15%, Guatemala with 8.9%, Kenya with 7.5%, and Costa Rica with 5.9%. By taking these countries as the Top 5 Countries Origin with the most reviews, from Figure 16, it can be concluded that these five countries represent 57% of the reviews. Furthermore, considering the countries from the Top 8, these represent 75% of reviews, and taking the top 16 countries reflects 90% of the reviews. Considering that the data set has blind assessments from 31 countries, it can be concluded that it is imbalanced.

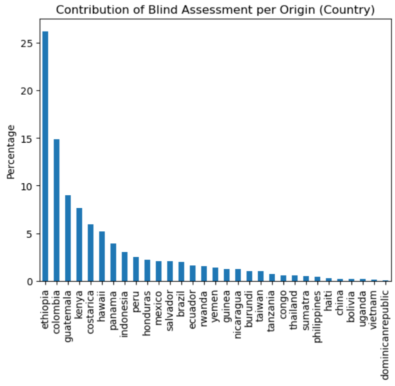
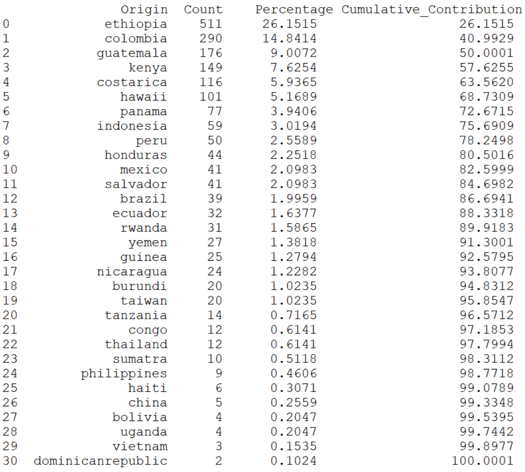


Figure 15: Contribution of reviews per Country



```
               Origin  Count   Percentage Cumulative_Contribution
0            ethiopia    511    26.1515                  26.1515
1            colombia    290    14.8414                  40.9929
2           guatemala    176     9.0072                  50.0001
3               kenya    149     7.6254                  57.6255
4           costarica    116     5.9365                  63.5620
5              hawaii    101     5.1689                  68.7309
6              panama     77     3.9406                  72.6715
7           indonesia     59     3.0194                  75.6909
8                peru     50     2.5589                  78.2498
9            honduras     44     2.2518                  80.5016
10             mexico     41     2.0983                  82.5999
11           salvador     41     2.0983                  84.6982
12             brazil     39     1.9959                  86.6941
13            ecuador     32     1.6377                  88.3318
14             rwanda     31     1.5865                  89.9183
15              yemen     27     1.3818                  91.3001
16             guinea     25     1.2794                  92.5795
17          nicaragua     24     1.2282                  93.8077
18            burundi     20     1.0235                  94.8312
19             taiwan     20     1.0235                  95.8547
20           tanzania     14     0.7165                  96.5712
21              congo     12     0.6141                  97.1853
22           thailand     12     0.6141                  97.7994
23            sumatra     10     0.5118                  98.3112
24        philippines      9     0.4606                  98.7718
25              haiti      6     0.3071                  99.0789
26              china      5     0.2559                  99.3348
27            bolivia      4     0.2047                  99.5395
28             uganda      4     0.2047                  99.7442
29            vietnam      3     0.1535                  99.8977
30  dominicanrepublic      2     0.1024                 100.0001
```

Figure 16: Single Origin Data Frame statistics

For the models, two datasets were applied for training and testing and are mentioned by the following names in the code df1 and df1_2. df1 contains the same data as mentioned in this section, and df1_2 has only the reviews from the top 8 countries with the most reviews, which are the single-origin coffees with more than 50 reviews.

4 Method

The general workflow of the Text Classification project is presented in Figure 17. The traditional text classification methods applied were a Dummy Classifier (baseline), Naive Bayes (sklearn-multinomialNB()), SVM (sklearn-Nu_SVC), and the contemporary deep learning method was DistilBert.

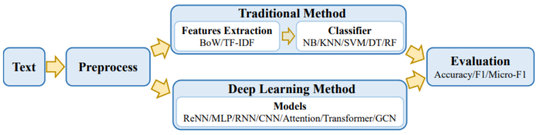


Figure 17: Flowchart of Text Classification(Li et al., 2022)

[5] https://gitlab.liu.se/alaca734/text-mining.git

The methods were trained with the two datasets df1 and df1_2. df1 is the most imbalanced because it considers 31 countries and some of them only have two reviews. In addition, the methods trained with this data set have a standard split of train and test data sets. In comparison, methods trained with df1_2 have a stratified split, and this dataset contains just the top 8 Single Origin Coffees with the most reviews. These changed the structures of the test and training sets.

Therefore, a direct comparison of score performances between methods trained with df1 and df1_2 cannot be done. In other words, we can only say one method is better or worse related to one another within the same dataset these were trained. However, it can be compared if the best classifiers for each dataset are the same.

The project plan changed as the methods were applied; the first idea was to compare the methods with the original/imbalanced dataset. However, when the performance metrics showed poor performance, the decision was to stop and try with a smaller dataset. The main reason behind having two datasets is to know if the imbalance is affecting the classifier's performance and by how much or if there are other factors.

Consequently, Distil-Bert was not trained with the imbalance dataset because it was likely to perform poorly.

4.1 Text

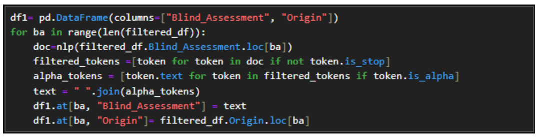
This refers to the data set obtained from web scraping, which is explained in point 3.

4.2 Preprocess

The preprocessing can be considered a two-step stage. The first stage was to clean the raw HTML text obtained from the Coffee Reviews website, and the outcome was the text.

4.2.1 Traditional Methods

In the second stage, the stop words and the non-alphanumeric characters were removed from each Blind Assessment with the help of the Spacy[6] library (Figure 18). Then the TfidfVectorizer() function was applied in the Pipeline(see Figure 19)

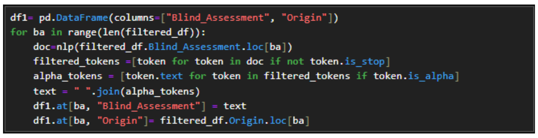


```
df1= pd.DataFrame(columns=["Blind_Assessment", "Origin"])
for ba in range(len(filtered_df)):
    doc=nlp(filtered_df.Blind_Assessment.loc[ba])
    filtered_tokens =[token for token in doc if not token.is_stop]
    alpha_tokens = [token.text for token in filtered_tokens if token.is_alpha]
    text = " ".join(alpha_tokens)
    df1.at[ba, "Blind_Assessment"] = text
    df1.at[ba, "Origin"]= filtered_df.Origin.loc[ba]
```

Figure 18: Preprocess

4.2.2 Contemporary Deep Learning Methods

The code written for training Distil-Bert is an adaptation of the example available in Huggin Face[7]. First, two dictionaries have to be created. The first one contains as keys the unique labels(Origins) and as items an assigned number, and the second is the inverse version of the first one. These are needed for the model in the AutoModelForSequenceClassification() function. Then, the dataset has to be transformed from a pandas DataFrame to a 'Dataset' object type; the 'Dataset' is divided into two dictionaries, training and validation (see Figure 19).



```
df1= pd.DataFrame(columns=["Blind_Assessment", "Origin"])
for ba in range(len(filtered_df)):
    doc=nlp(filtered_df.Blind_Assessment.loc[ba])
    filtered_tokens =[token for token in doc if not token.is_stop]
    alpha_tokens = [token.text for token in filtered_tokens if token.is_alpha]
    text = " ".join(alpha_tokens)
    df1.at[ba, "Blind_Assessment"] = text
    df1.at[ba, "Origin"]= filtered_df.Origin.loc[ba]
```

Figure 19: Preprocess

4.3 Training

4.3.1 Traditional Methods

The dataset was first split, training 65%, and testing 35%. Then, within a Pipeline, the vectorizer and the model, DummyClassifier(), MultinomialNB(), and NuSVC(), were defined. The parameters for each model were determined, and after this, the pipeline and parameters were passed to a GridSearchCV to perform cross-validation. The CV number for this was 2 for df1.

For the same models with the data set df1_2, the training and test split was done using the function StratifiedShuffleSplit(). The Pipeline() and GridSearch() were applied as mentioned above, but the CV number for MultinomialNB() was 16 and for NuSVC() 30. It cannot be bigger than two before because some Origins on df1 have only two reviews.

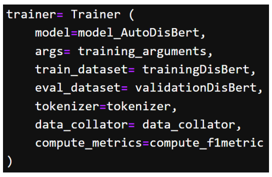
4.3.2 Contemporary Deep Learning Methods

The data collator function DataCollatorWithPadding() was used to prepare the batches of data for the training. It added padding to all the data to have the same length, and the tokenizer from Distil-Bert is passed through this function. Then, the F1 score metric was loaded from the evaluate package, and the F1 macro specified. The AutoModelForSequenceClassification defined the model, and the number of labels and the dictionaries of labels and

[6] https://spacy.io/universe/project/spacy-cleaner

[7] https://huggingface.co/docs/transformers/tasks/sequence_classification

numbers from point 4.2.2 were introduced. Afterwards, the training arguments were defined, such as learning rate, batch size, number of training epochs, and weight decay. Finally, with the help of the Trainer function, all the functions declared within this point were united(see Figure 20).



```
trainer= Trainer (
    model=model_AutoDisBert,
    args= training_arguments,
    train_dataset= trainingDisBert,
    eval_dataset= validationDisBert,
    tokenizer=tokenizer,
    data_collator= data_collator,
    compute_metrics=compute_f1metric
)
```

Figure 20: Trainer

4.4 Confusion Matrix and TFIDF

After getting low numbers for the F1 Macro score, a Confusion Matrix was plotted to determine what Origins were the most wrongly predicted. A visual method was applied to the top 10 tokens or terms with the highest TFIDF using a Word Cloud plot with n-grams of 1. These measurements are to understand the complexity underlying the task and to check if there is a reason why it is too difficult for the models to predict the Origins correctly.

The plot "Common words with highest TFIDF values between Origins" (Figure 30) is to have a better understanding of how similar the words are in the word cloud plots. It shows what percentage of the top 10 words with the highest TFIDF value for each country, are common to the Origins: Ethiopia, Kenya, and Panama.

5 Results

The following chart presents the F1 Macro Average Scores for each method trained with the two datasets. NuSVC obtained the highest score with dataset d1_2, and Multinomial NB trained with df1 obtained the lowest score.

F1 Macro Average Score

Data set	**Dummy**	**MN NB**	**Nu-SVC**	**Distil-Bert**
df1	0.03	0.07	0.08	N/A
df1_2	0.14	0.23	0.31	0.30

The confusion matrix from Figure 21 contained the True labels vs the Predicted labels from the NuSVC models trained with the df1_2 dataset, which is the combination with a higher F1 score.

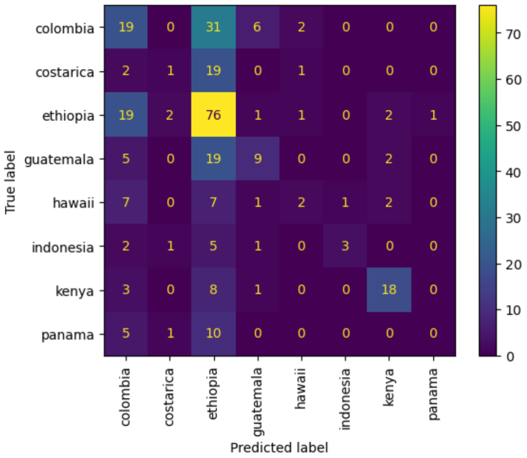


Figure 21: Confusion Matrix

In Figures 22, to 29, the words with the highest TFIDF value from each origin are visualized with a word cloud plot.

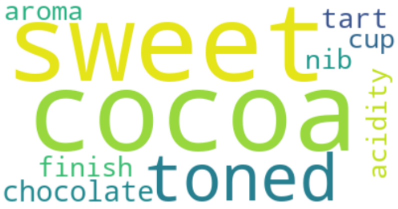


Figure 22: Ethiopia Word Cloud Higher TFIDF

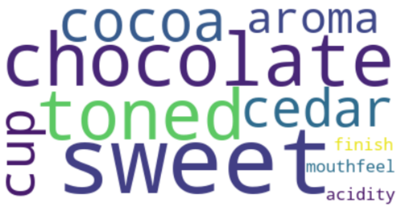


Figure 23: Guatemala Word Cloud Higher TFIDF

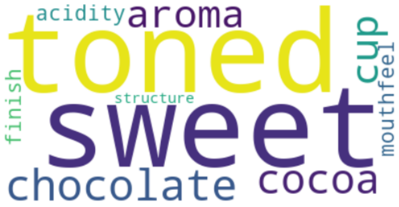


Figure 24: Colombia Word Cloud Higher TFIDF

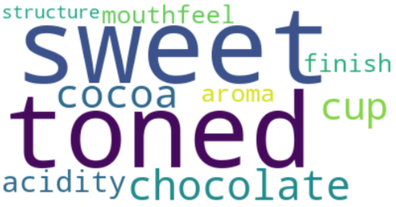


Figure 25: Hawaii Word Cloud Higher TFIDF

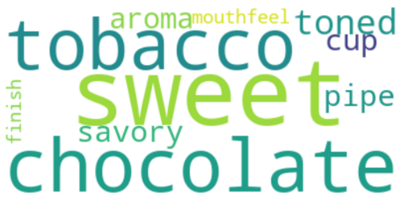


Figure 26: Indonesia Word Cloud Higher TFIDF

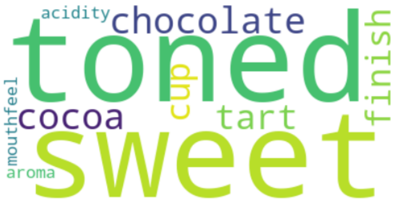


Figure 27: Costa Rica Word Cloud Higher TFIDF

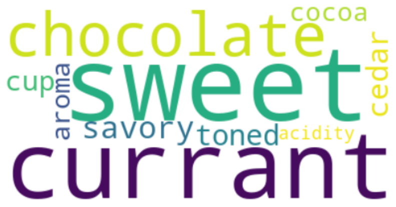


Figure 28: Kenya Word Cloud Higher TFIDF

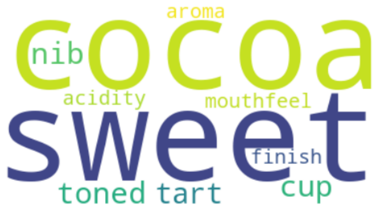


Figure 29: Panama Word Cloud Higher TFIDF

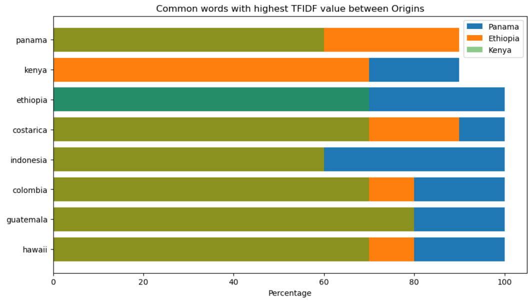


Figure 30: Common words between Origins

6 Discussion

The F1 Macro Average Chart shows that all models within each dataset (imbalanced and balanced) performed better than the baseline (DummyClassifier()), which is based on random guesses. For the imbalance dataset (df1), the Nu-SVC classifier was better than the Multinomial NB. For the balanced dataset (df1_2), the Nu-SVC was also better than the Multinomial NB and also better than the state-of-the-art model Distil-Bert just by 0.01. In both contexts, the highest scores were from the NU-SVC model.

The confusion matrix from the best model NuSVC was plotted among the models compared in the project to understand better why the models performed badly. It can be seen that Ethiopia, which is the origin with the most reviews, is also the one with the most correctly predicted Origins based on the Blind Assessments and also the origin that is most frequently wrongly predicted when another Origin is the true label. Also, origins such as Colombia, Guatemala, and Kenya, which are the top 2, 3, and 4 with the most reviews, have more correctly predicted labels, which enforced the hypothesis that the imbalanced dataset is one of the causes producing the bad performance.

By examining which origins had the most significant quantity of wrongly predicted Ethiopia. Colombia and Guatemala were found. This could be obvious since these also have the most reviews to be predicted, but there could also be something else. The top 10 words with the highest TFIDF were plotted to take a closer look. Common words between the origins are sweet, cocoa, chocolate, and acidity, among others. To know how much similarity there is between these words, the plot in Figure 30 was drawn. As mentioned in point 4.4, each origin is compared with Ethiopia, Kenya, and Panama. Ethiopia was selected because it is the origin that is the most incorrectly predicted. Kenya because it has a high number of correctly predicted origin labels, and Panama to understand why it does not have any labels correctly predicted. Figure 30 reflects that Panama has 100% similarity or the same top 10 words with the highest TFIDF as 6 of 7 Origins, and with the Origin remaining Kenya, which has 90% similarity. Making it difficult for the models to differentiate. In contrast, Kenya, which has a higher number of correctly origin labels predicted, has less similarity across all origins, whereas Guatemala has the highest similarity with 80%. Ethiopia's similarity is between these two origins; its highest similarity is with Panama at 90%, and the lowest is Indonesia at 60%. This is another insight, that the models cannot differentiate correctly, besides the imbalance, because the blind assessments are too similar.

This leads to the following finding: could it be that even with standardized procedures for evaluating coffee, which includes tests of every part of the coffee production process and certified Q-Graders, the Blind Assessments (or reviews) have a particular bias? An article from Mateus Manfrin (Manfrin Artêncio et al., 2023), a researcher from the University of Sau Paulo, found that the sensory perception of professional coffee tasters is affected by their knowledge about the geographical indication (origin) of the coffees they were testing. Furthermore, it found that when the professionals knew the story about the origin of the coffee, they judged the acidity and overall flavor as more intense—in contrast, nutty, cacao, and caramel notes had higher scores when they did not have any information about the origin of the coffee. These are some words with the highest TFIDF plotted on the Word Clouds.

7 Conclusion

The present project has applied, tested, and compared different Text Mining methods to know if predicting the Origin of a Coffee is possible based on his "Blind Assessment." A partial conclusion can be drawn from the data and methods used. Partial because the model's performance can be improved with better data, in other words, more 'Blind Assessments' and reduce the imbalance of it; some further suggestions could be data augmentation or web scraping of all the reviews from the Coffee Reviews website. Within the limitations of the data, it was found that a traditional method, NuSVC, outperformed Distil-Bert, with respective scores of 0.31 versus 0.30 in F1 macro, both trained with the balanced dataset. From the models trained with an imbalanced dataset, Nu-SVC also performed better, with a score of 0.08. NU-SVC imbalance performed 2.6 times better than its baseline (Dummy), and Nu-SVC balance was 2.1 better than its baseline (Dummy).
Another meaningful insight is that it is probable that the "Blind Assessments" have certain biases that are provoking the models not to learn any significant features that help to differentiate between origins. It was found there is a grade of relation between the common flavors that professional tasters described in the coffee when they do not know about the flavor, and these match with the most common highest TFIDF words from origins. Which origin predicted was mostly mistaken by Ethiopia.

References

1.12.1.3. onevsoneclassifier.

Protocols Best Practices — Specialty Coffee Association.

Quality evaluation.

Magnus Borga. 2023. Lecture 2 notes in supervised learning -linear classifiers.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

KELSEY KUDAK. 2019. Calling the shots: 20 years of the world barista championship - 25 magazine, issue 9.

Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S Yu, and Lifang He. 2022. A survey on text classification: From traditional to deep learning. *ACM Trans. Intell. Syst. Technol.*, 13(2):1–41.

Mateus Manfrin Artêncio, Alvaro Luis Lamas Cassago, Renata Kelly da Silva, Fabiana M. Carvalho, Fernando Batista Da Costa, Marina Toledo Lourenção Rocha, and Janaina de Moura Engracia Giraldi. 2023. The impact of coffee origin information on sensory and hedonic judgment of fine amazonian robusta coffee. *Journal of Sensory Studies*, 38(3):e12827.

C.D. Manning, P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. An Introduction to Information Retrieval. Cambridge University Press.

Britney Muller. 2022. Bert 101 state of the art nlp model explained.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Bernhard SchÃ¶lkopf, Alex J. Smola, Robert C. Williamson, and Peter L. Bartlett. 2000. New Support Vector Algorithms. *Neural Computation*, 12(5):1207–1245.

Wenny B Sunarharum, David J Williams, and Heather E Smyth. 2014. Complexity of coffee flavor: A compositional and sensory perspective. *Food Res. Int.*, 62:315–325.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

ChengXiang Zhai and Sean Massung. 2016. *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*, volume 12. Association for Computing Machinery and Morgan & Claypool.