Bachelor thesis
Computing Science

Radboud University

Classification and Interchange of Informal and Formal English Text

Author:
Seraph Jin
s1032019

First supervisor/assessor:
prof. dr. ir. A.P. de Vries
A.deVries@cs.ru.nl

Second supervisor:
Agnieszka Szuba
agnieszka@yoast.com

Second assessor:
prof. dr. ir. D. Hiemstra
hiemstra@cs.ru.nl

January 20, 2022

Abstract

The last two decades have seen a growing trend towards machine learning. Document Categorization, the assignment of a document to one or more groups or categories, is one of the data mining problems that cause many discussions. However, as a linguistic problem and sub-category of Document Categorization, formal and informal text classification has received rather little attention despite the fact that a real need exits to resolve the problem. In this paper, we work with machine learning techniques to deliver a model that can distinguish formal and informal texts with JavaScript implement since it needs to run on web browser. The explored classification estimators are decision tree, random forest and logistic regression. The accuracy and proficiency of the estimators are measured to find out which features are crucial in text formality classification. In the end, we discuss how to transform informal text to formal text and vice versa.

Keywords: document classification, Natural Language Processing, JavaScript, Logistic Regression, cross validation, Decision Tree, Random Forest, English Grammar, Formal and Informal text

Contents

Chapter 1

Introduction

The purpose of text categorization [1] is to classify text documents into different predefined classes. There is little published research into the classification of formal and informal (English) text. A systematic study on text formality classification is missing from the literature so far. The most relevant work about formal and informal text classification is done by Fadi Abu Sheikha and Diana Inkpen [2]. They classify formal and informal documents based on the features extracted from the data sets. Some shortcomings of their research are discussed in the following paragraph. Based on analysis and our critical review of this research, we extend the binary classification model from the fore-mentioned research to a multi-class classification model. Instead of predicting a document as either formal or informal, we also work on a model that can predict the formality of a document on a scale from one to three, with degree one being very informal, two being semi-formal/informal, and degree three as very formal. In the end, we discuss which key factors contribute to the formality or informality of a text and provide some ideas or solutions to change the text style. Our objective is to achieve a sustainable and replicable model with a high accuracy with the flexibility to extend it for the future research.

The prior research [2] do not disclose the exact algorithm they used, and they extract the feature using Connexor parser [3] which is not convenient to use locally. The data sets used in the research are not optimal and inconsistent. For example, in the Enron email dataset, even though emails are usually informal in the setting of normal life, many emails in the Enron data sets are quite formal, as expected from a former huge company. Some of the data sets they used include annotations that are difficult to cleanse manually. The rest of the data sets are adequate for the classification, and we also use some of them in our heterogeneous dataset.

This research improves on the prior study and provide an implementation of the text classification models to assist *Yoast* for their future development in that field. Yoast [4] provides a SEO (search engine optimization) plugin for Wordpress and other other similar platforms, which has been used across millions of websites. The function of the SEO

plugin is to help users achieve a high position in search engine results. This includes offering feedback and suggestions for improvements to the content of their websites. In this paper, we focus on the formality side of writing since the styles of writing can influence the readers' attitude towards the content. For example, when reading a fun story, one would expect the creator to make use of abbreviation, slang, colloquial language,etc.. On the other hand, in a newspaper, readers like to see more formal and serious text to build trust in the content of the newspaper. Since text formality and informality contribute greatly to the content of its users, Yoast would like to make the informal and formal classification a tool to assist the process of content creation. This thesis starts with a replication of the text formality classification method from the research [2] done by Fadi Abu Sheikha and Diana Inkpen. Afterwards, we extend the research to multi-class classification on heterogeneous datasets. The concept of the models is well structured to allow for modifications and follow-up research in the future.

Following [2], we extract features from each document and quantify the features as numerals. After the desired features for each documents are extracted, we train and verify the model with stratified k-fold cross validation. Afterwards, we examine the features to see which (subset of) features are essential when distinguishing formal and informal texts with a recursive feature selection method.

In the discussion section we compare several kinds of classifiers (Decision tree classifier, logistic regression classifier and random forest) in the experiment, analyze and discuss the classifiers based on two different datasets, and implement the mock-up classification algorithms. The choices made for the classifiers are limited by the JavaScript implementation. For decision tree model, we only need to distinguish different documents according to the feature splitting suggested by the decision tree model. In other words, the decision tree result could be utilized to create a rule-based algorithm. For logistic regression, we could apply the weight and bias to the features and get an approximate result to determine the type of document. In the end, we also discuss what makes documents formal and informal and how to interchange them.

Chapter 2

Preliminaries

2.1 Binary Classification and multi-classes classification

Both binary classification and multi-class classification are supervised learning tasks. In supervised learning, algorithms take "labelled data" as input and assign the right output tags to the given data variables [5]. Taking our experiment as an example, our algorithm takes a training set of data led with a label "Formal" and "Informal" and predict the input data based on the model learned from the train set. In practice, the labels are usually ordinal i.e. the labels are not coded as "Informal" and "Formal" but as "0" and "1". The difference between binary and multi-class classification is that a binary classification deals with two classes but multi-class classification contains more than two classes. Some machine learning algorithms such as logistic regression do not support multi-class classification. However, there are ways to use a binary classification method on a multi-classes classification problem. This method is explained in section 3.4.2.

2.2 Logistic regression

In logistic regression model, the result of training a model is a sigmoid function that produces a number between 0 to 1. In a binary classifier where the result is either 0 or 1, we classify the result above 0.5 as 1 and otherwise as 0. The function inside the sigmoid function is $w \cdot x + b$ where w is the weight, x is the input feature, b is the bias. The weight w and bias b are trained by each entry from the train set so that our result $sigmoid(w \cdot x + b)$ is close to the label Y.

2.3 Decision Tree

Decision tree [6], as the name suggests, is a tree with many nodes where all nodes together decide the outcome of data sample. We use the mock decision tree in Figure 2.1 to explain this model. The decision tree predicts if a student will be admitted to college. The model predicts the admission by taking the features of the object that we

predict. The first question is if this student graduates from high school. The negative answer immediately results in the student not being admitted, while the positive answer leads to a series of more questions: did they pass English proficiency exam? If they passed, does he have a GPA lower than 3.5? If they did not, are they from England (Passing the English proficiency exam is unnecessary for them)? After going through the leaves and nodes of the decision tree, eventually each object is labelled as either "admitted" or "not admitted". The illustration explains the basic ideas of decision. For

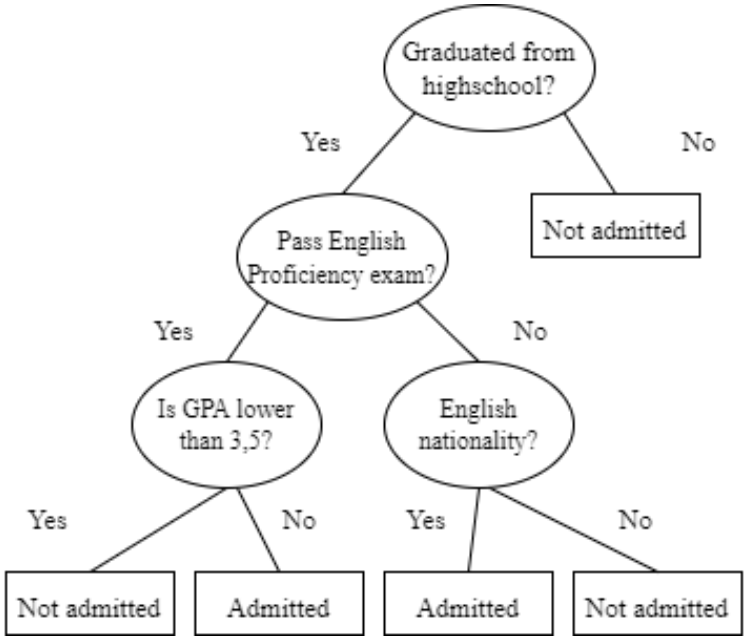


Figure 2.1: A demonstration of a decision tree model.

our model with many more features, the decision tree is larger than the one in Figure 2.1. The algorithm takes the train data set and creates the series of questions similar to the one shown above.

2.4 Random Forest

A random Forest classifier combines a number of decision trees into a single classifier. In this estimator, the dataset is arranged into smaller datasets to fit the decision trees inside of the random forest. During the prediction phase, each decision tree shows an answer, and the model takes the result that is voted by the majority of the decision trees. Random Forests help avoid over-fitting, an error that occurs when the classification shows a high accuracy for specific datasets, and provide better accuracy in general since the model relies on several decision tree models to make the decision instead of one.

2.5 Recursive Feature Elimination Cross Validation

RFECV is a feature selection method in scikit learn Package. Feature selection is important to train the model with the smallest and sufficient number of features since redundant features could mislead and slow down the classification process. The reduction on features is useful when we are not sure about which features are relevant to the classification. The reduction process starts with training the initial feature set and estimating the importance of each feature. The feature that is least relevant to the classification is taken away from the feature set in each reduction process [7]. The process continues until we get the desired number of features, which by default is one. Cross Validation is the estimator that assist the recursive feature elimination. Basically, cross validation can reduce the chance of over-fitting. In our research, we used stratified 10-fold classification, which is explained in section 3.

2.6 Natural Language Processing

In our research, we are dealing with text documents. To extract the features easily from the documents, we have applied Natural language processing techniques to the data set. The Python Standard Library is mostly sufficient for our processing. Since the processing also needs to be done in JavaScript for Yoast users, we choose the functions from python that are similar to or replaceable by the Yoastseo library and JavaScript standard library. We use the standard libraries to split sentences and to cleanse punctuation from the word and such. Besides the standard library, NLTK library and spaCy also contribute to our research. The relevant functions are sentence tokenization and syntax dependency which are described below.

The common way to split sentences in python is by looking for the period, question mark or exclamation mark by the end of the sentence, but there are many exceptions and confusions in English text. For example, the word "U.S." would be separated and recognized as two sentences. Sentence Tokenization helps us avoid these exceptions.

The method to check the voice of a sentence is taken from [8]. The idea is that we check if a sentence has syntax relating to passive voice. The first way of checking is by looking for the 'agent' tag from the sentence. For sentences without an agent, we search for 'nsubjpass'(nominal subject passive) tag in the syntax. The checking is not perfect but is good enough for the passive voice feature we want to extract from the document.

2.7 Evaluation Metrics

Many evaluation metrics can be appropriate for the evaluation of machine learning systems: accuracy, precision, recall, and F1. Our research focuses on accuracy because classifying both informal and formal texts are equally important for us and the size of the data sets of the classes are the same. The accuracy is calculated as:

$$\frac{\textit{The documents that are correctly classified}}{\textit{The total number of the documents}}$$

2.8 Linguistic Terminologies

The following terms are presented for readers without linguistic knowledge.

2.8.1 Agent

Agent is the term that appears in passive voice to describe the subject that carries out the action. In the sentence "The food is eaten by us," "us" is the agent.

2.8.2 Pronoun

A word that serves as a substitute of a noun. For example,"Mr. Smith is here" can be expressed as "he is here, and in that case the pronoun is "he"."

2.8.3 Contraction

Contraction is a word that replaces letters or numbers with apostrophe. For example, the contraction "he's" can be also written out as "he is."

2.8.4 Abbreviation

The purpose of abbreviation is to write words in shorter terms. Abbreviations can be formal or informal. For example, "brb" (be right back) is informal, and "BBC" (The British Broadcasting Corporation) is formal and functions as a name for the organization.

Chapter 3

Research

3.1 Formal and Informal English Texts

Before delving into the definition of formal and informal English, we define what is not in the scope of the current research. In this thesis, we will not focus on the domain of grammar and spelling mistakes. We ignore the fact that grammar and spelling play an important role in formal and informal text classification, because such mistakes are part of another area of study about correct use of English. The writers can effortlessly correct those mistakes by doing a grammar and spelling check. In addition, even some professional academic papers still contain grammar and spelling mistakes, but such unintentional mistakes do not influence the formal nature of the text.

3.1.1 Formal English

Formal English is primarily used in writing styles for formal situations such as newspaper, academic paper or public speech. In the formal writing styles, the norms of grammar are scrupulously adhered to by the authors. The vocabulary in formal English is more complicated than in everyday speech. The formal writing style also tends to employ "fancy" terms such as "scrupulously" that are rarely seen in speech. It also favors multisyllabic words over phrasal verbs i.e. using "establish" instead of "set up". The use of slang and colloquial terminology is also discouraged.[9]

When formal English is employed, sentences are often long and complicated; in Figure 4.1, we observe that for BBC news articles, the average word length in a sentence is between twenty to thirty words. To have a neutral tone, formal English tends to avoid the subjective tones with the informal pronounce such as "me" and "you", and it favors the use of impersonal pronounces such as "we" and "it". In addition, contractions and abbreviations in formal writing are also frequently expanded. Formal English leaves an impression of seriousness, authenticity, profession and distance.

3.1.2 Informal English

Informal English is used for daily life, for example writing an email or letter to someone close to you. Informal English can be perceived as casual and lighthearted, but sometimes it could be vulgar and inappropriate. Though there are no strict grammar rules in informal English, we still find some patterns in informal texts. We can detect many subjective tones from pronounces like "I" and "you". The personal emotion and styles are prominent in informal writing. People also tend to use phrasal verbs to be more understandable. In personal blogs or social media, the exclamation marks and question marks are used more often than in formal texts. [9]

3.2 Data Acquisition

The datasets used to train our model are essential for our research results. In [2], the research presents us with the sources of their datasets; however, after a careful inspection of the datasets, we decide to not use all their proposed datasets for a number of reasons. Firstly, the dataset for informal documents prediction, the Enron Email Database [10], contains emails that should be classified as formal, e.g. promotion emails or company decision letter written in formal styles. Secondly, some datasets are not free of annotations by the publishers and there are some special characters in the texts. To sum up, we take some examples from the datasets they proposed, but additionally, we carefully construct our own dataset.

We start the research with a dataset [11][12] provided by Kaggle, because their dataset is easily accessible. Then, in the second part of the research, we build our own datasets from various sources to add some diversity to the data. These datasets are selected and categorized manually into three classes: formal, informal and semi-formal/informal.

3.2.1 Homogeneous dataset

Our homogeneous formal documents are from the BBC news dataset [11] with five hundred entries, and informal documents are from the blog post dataset [12] also with five hundred entries. Since the news is written by professional writers, they are guaranteed to be formal. The informal blog post dataset consists of blog posts from different domains. A few of the blogs contain non-English words and sentences or some warning message from the website. Since there are very few of them and in real life a blog might also contain several different languages, we decide to not get rid of them. In section 3.5, we do observe that the blog post datasets contain many informal features, like more contractions, more informal pronouns etc.. The one drawback of the homogeneous datasets is that we might risk classifying specific documents instead of classifying two general classes. In other words, our model might only work classifying news articles and blog posts.

3.2.2 Heterogeneous dataset

In the heterogeneous dataset, we select fifty entries for each class (formal, semi-formal/semi-informal, informal) of English texts by hand. We aim to let each class contain an extensive range of documents so that our classification can be more inclusive. For the informal documents, we choose from categories such as diary, movie review, conversations, letters and emails. For formal documents, we obtain articles from different news publishers, academic papers, formal speech, etc. Because the semi-formal documents are also semi-informal documents, we combine them into one category. They contain articles that are not formal enough to be very serious and solemn, but at the same time the articles should be publishable such as a blog that teaches you how to grow plants or entertainment news.

In these datasets, we seek to mimic a real-world situation that not everything is either formal and informal. Then we could take care of some exceptions in the writing. For example, in formal datasets, we have several formal speeches with many informal pronouns while still being formal document. We would like to force our algorithms to take these situations into consideration when building up the model. Admittedly, since the sorting process of the documents is based on subjective opinions, there are possibilities of incorrect classification of the dataset for training and testing. The accuracy is also expected to drop since the features among these classes are not that distinctive anymore.

3.3 Feature Extraction

Based on the descriptions of formal and informal texts in section 3.1, we extract the following features from texts: the average word/letter length per sentence, average word length (of the whole text), average sentence number with passive voice per article, normalized informal and formal pronouns per article, normalized contraction size, and normalized informal and formal word occurrence. The features are chosen based on the result from [2]. To easily extract the features, we process the articles into a list. Then we further divide each article in to a list of sentences. We also create a list of articles in which each article is split into words. Every word and length in the list are free of punctuation except for the punctuation apostrophe since apostrophe is often part of the word to show possessive case and abbreviation (we also need find the occurrence of the abbreviation with the help of apostrophe).

3.3.1 Average Word/Letter Length Per Sentence and Word length

We generate these features by dividing the total word number and the total letter number of the whole article by the sentence number. The word length is the average number of letters for the words in the article, which is done by dividing the total letter number by the total word number. Sentence number and the word number are obtained by measuring the size of the lists we mentioned in the data pre-processing part. The size of a word is obtained by counting letters in each word and then adding them together.

3.3.2 Passive Voice

We obtain the normalized number of passive voice occurrences by checking the fraction of sentences containing passive voice in an article. We use the method proposed in this blog [8] (see the natural language processing section in section 2) to check if a sentence contains passive voice. Since in the normalized result the passive voice feature is negatively correlated with the active voice feature (if $\frac{1}{5}$ are passive then $\frac{4}{5}$ are active), we do not use the active voice number as a feature in the classification process. More correlated features would not improve the model and could instead worsen its performance since more features make the model more complicated. In addition, a lower number of features could speed up the running process of the algorithm.

3.3.3 Informal and Formal Pronouns

We create lists of formal and informal pronouns. For formal pronouns, we choose pronouns associated with the the plural forms of the personal pronouns along with the third personal pronoun "it", and for the informal pronouns we only choose singular first and second personal pronouns.

```
formal_pronoun_list = [ 'we','they','their','theirs','themselves','us','our
    ','ours','ourselves','it','its','itself']

informal_pronoun_list = ['I','you','me','my','mine','your','yours','myself'
    ,'yourself','yourselves']
```

Figure 3.3.3: The formal and informal pronouns lists

First, we split each article into a list of words, and then count the occurrences of formal and informal pronouns from Figure 1 within this list. We normalize this feature by dividing the number of informal or informal pronoun occurrences by the total number of words in an article.

3.3.4 Contraction

Another feature that is prominent in informal writing is the presence of contraction. We use the same approach for informal and formal pronouns to extract the contraction features by using the contraction list (a few examples are listed below) to find all common contraction occurrences and normalizing the feature.

```
aren't
can't
couldn't
didn't
don't
doesn't
```

```
hadn't
hasn't
haven't
```

3.3.5 Informal and Formal Word List

To create this list, we investigate which words are used specifically for formal articles and which words are used frequently for informal texts. The informal list contains phrasal verbs [13] and informal abbreviations [14]. The formal word list contains the word substitute for phrasal verbs and transitional words [15]. Because this category contains many phrasal verbs, we could not use the same method as for the last two features to check how many formal and informal words are present in an article. Our strategy for counting the informal and formal word occurrence is to split an article into sentences, and then look up how many times phrasal words are present in a sentence. We search the word frequency by using the least frequent word(s) in a phrasal verb. For example, for the phrasal verb "make up", if "make" appears three times, and "up" two times in a sentence, then we assume the verb "make up" appears twice in the sentence. Indeed, the algorithm will not be one hundred percent accurate but it does tell us a difference between formal and informal articles as shown in section 3.5.5.

3.4 Research Method

3.4.1 Model Replication

We try to replicate the model from [2], but not all methods are available and feasible for our case. For example, they use the Connexor parser[3] to extract most features, which is not able to run locally. Therefore, we only take inspiration from their reseach but do not follow it completely. All code, datasets and libraries are available in GitHub [16].

3.4.2 Classification Tools

We first explore binary classification of the formal and informal document datasets. We use the extracted features mentioned in the previous section, creating a dataset of X and y with X being the input feature data and y as the output label data. In the binary classification, y can only be "1" or "0" representing whether an article is formal or informal. The dataset contains an equal amount of informal and formal articles (five hundred per each, one thousand in total). To create and train our model, we focus on the following machine learning algorithms: logistic regression, decision tree, and random forest. The choices of algorithms are made because we would eventually like to implement such classification model in JavaScript, with relative straightforward approaches that are easily implemented from scratch. Potentially, we could still do machine learning in JavaScript, but the process is time-consuming and not as efficient as we do it in Python. As we try to avoid directly running the machine learning algorithms in JavaScript, we explain how to achieve it in our chosen classifiers below.

Logistic Regression

In the Logistic Regression model, the algorithm calculates the weight for each parameter (features) and the bias (intercept) of the model. For a JavaScript implementation, we could just use these two values to predict the type of the documents with inputted features as shown in the preliminary section. If the result of an input is above 0.5, then the prediction is that the article is formal otherwise it is informal.

Logistic Regression for Multi-class Classification

The logistic regression classifier is not originally made for multi-class classification, but there are ways to work on multi-class problems with logistic regression. In our experiment, we use One-vs-Rest classification [17]. In the three-class classification, three binary classification models are created:

```
formal text v.s. {semi-formal/informal, informal text}
informal text v.s. {semi-formal/informal, formal}
semi-formal/informal text v.s. {formal, informal}
```

The classifier takes the highest score of the three models as the outcome. This method works for us because the number of classes and the size of the datasets are relatively small.

Decision Tree

The Decision Tree model analyzes the input features and provides information on how one should classify the datasets based on certain rules. This information could be used to create a rule-based algorithm in JavaScript. The idea of random forest is discussed in the section 2; thus, we do not further discuss it here.

Stratified K-fold Cross Validation

We utilize Stratified K-fold cross Validation to test the accuracy of each model. Stratified K-fold Cross Validation is one type of cross validation. To test the accuracy of our model, we need two datasets: a test dataset and a train dataset. The intuition behind Stratified K-fold Cross Validation is that each entry of the dataset is equally represented as training and testing dataset while keeping the classes proportionally distributed. In our case, the datasets are divided into ten parts (one thousand formal and informal documents are divided into ten groups), and these groups are distributed in a way that each group is used in training set and test set. Furthermore, the training and testing datasets have an equal amount of samples from the two classes. The reason we choose stratified K-fold Cross Validation is that each type of documents are equally trained. For example, if we have one hundred fifty entries in the train set, seventy five of them should be from the formal datasets, and the other seventy five are from the informal datasets. If we train the model with one-hundred forty informal documents and ten formal documents, the model is not generalized and unbalanced.

Chapter 4

Visualization

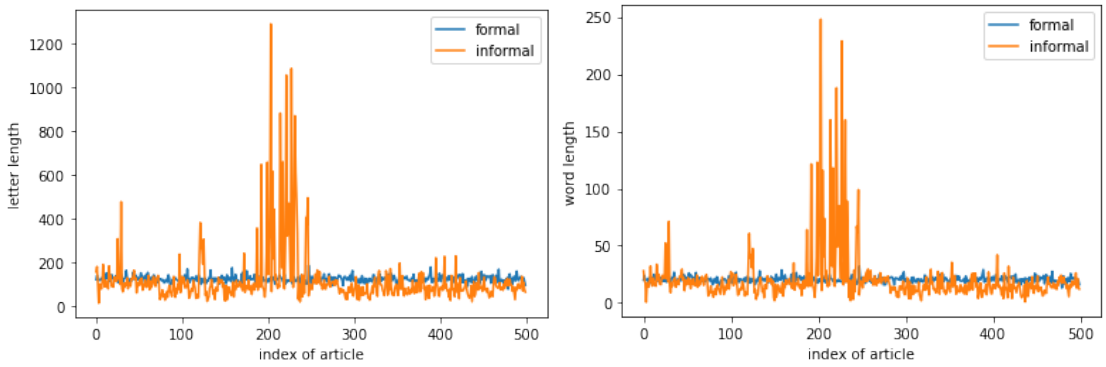
In this section, we demonstrate the distribution of each feature for formal (and semi-formal/informal) and informal documents in the homogeneous dataset and heterogeneous dataset.

4.1 Homogeneous Dataset

4.1.1 Average word per sentence and average letter per sentence

Figure 4.1 compares two graphs of the average word number per sentence and the average letter number per sentence. These two features are correlated and supported by the correlation coefficient function provided in *numpy*. We further discuss whether we should keep both features in the section 5.

```
print(np.corrcoef(avg_word_sentence, avg_letter_sentence))
output:
[[1.         0.95686343]
 [0.95686343 1.        ]]
```



(a) Letter length (b) Word length

Figure 4.1: The comparison of the normalized letter and word lengths on the sentence level between formal and informal English text samples.

We definitely see that formal documents are more consistent with their sentence length; whereas informal documents have no strict limit on length.

4.1.2 Average word length

In Figure 4.2, we observe that formal articles have an average word length between five to eight letter. This feature is consistent for all formal texts. For most informal texts, they have an average word length lower than the one for formal texts.

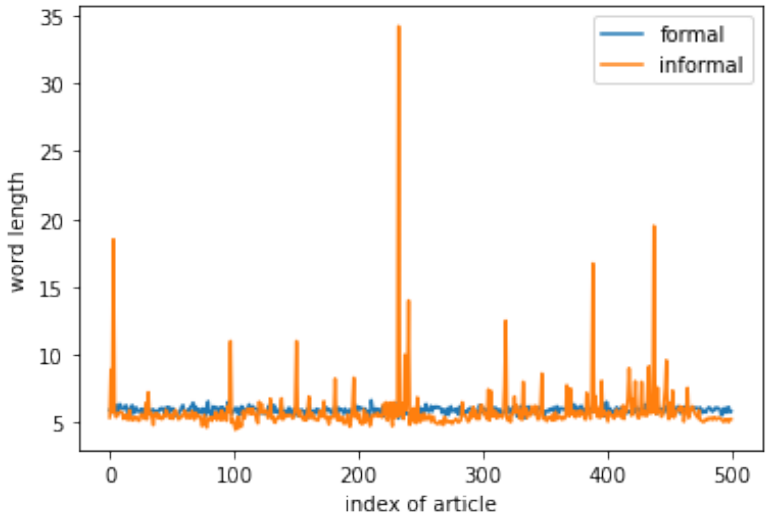


Figure 4.2: The comparison of the normalized word length between formal and informal English text samples.

4.1.3 Passive voice

As mentioned in the sections before, in our sample texts (Figure 4.3), passive voice is used more often in the formal documents than in the informal documents. The majority of the formal documents have more than 20% of their sentences in passive voice, and informal texts do not contain that many passive voice in the sentence.

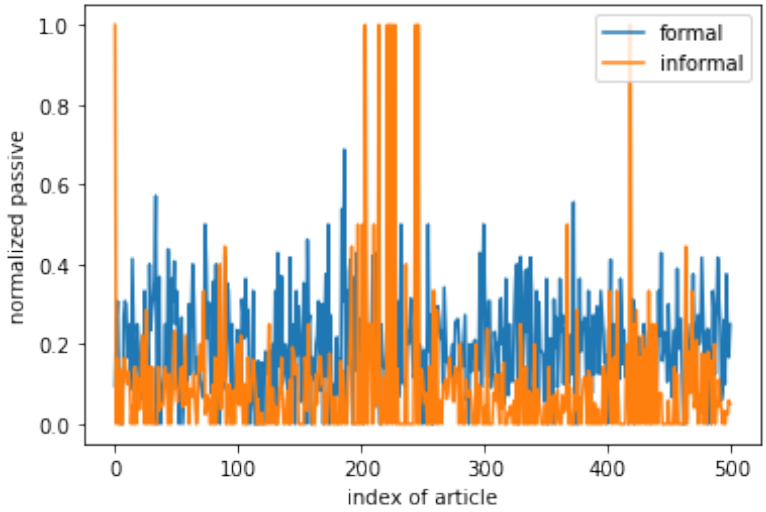
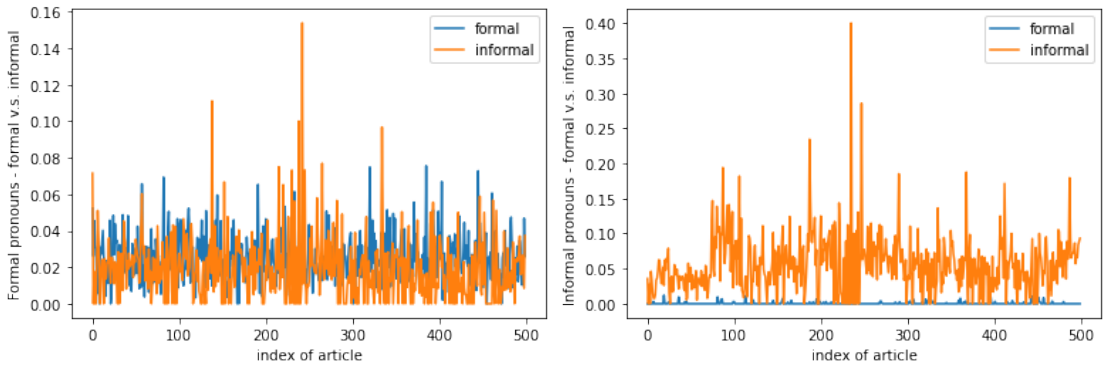


Figure 4.3: The comparison of the normalized passive voice sentence number between formal and informal document

4.1.4 Formal and Informal Pronouns

In Figure 4.4, we observe that it is difficult to tell formal and informal texts apart by simply looking at the formal pronouns. Formal pronouns like “they” and “we” are not exclusive to informal texts. Thus, formal pronouns are not very important during the classification process.

In contrast, the plot for informal pronouns shows a clear distinction between formal and informal documents. Formal documents almost never use informal pronouns “I” and “you”, and this characteristic could effortlessly and perfectly separate the formal and informal English text samples. It is clearly possible to distinguish BBC news articles from blog posts since BBC news almost never uses personal pronouns.



(a) Formal pronouns

(b) Informal pronouns

Figure 4.4: The comparison of the normalized informal and formal pronouns number between formal and informal English text samples.

4.1.5 Contraction

Along with the informal pronouns, contraction is also an indicator of informal documents. Figure 4.5 shows contractions occur very infrequently in formal articles, but are common for informal articles.

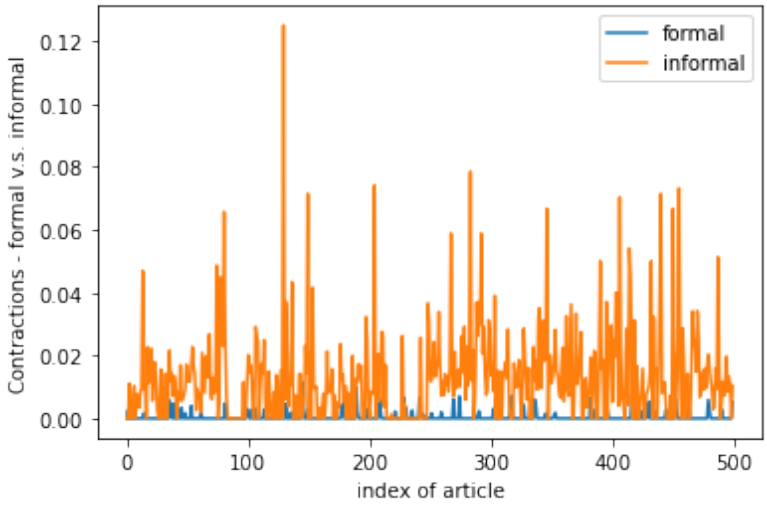
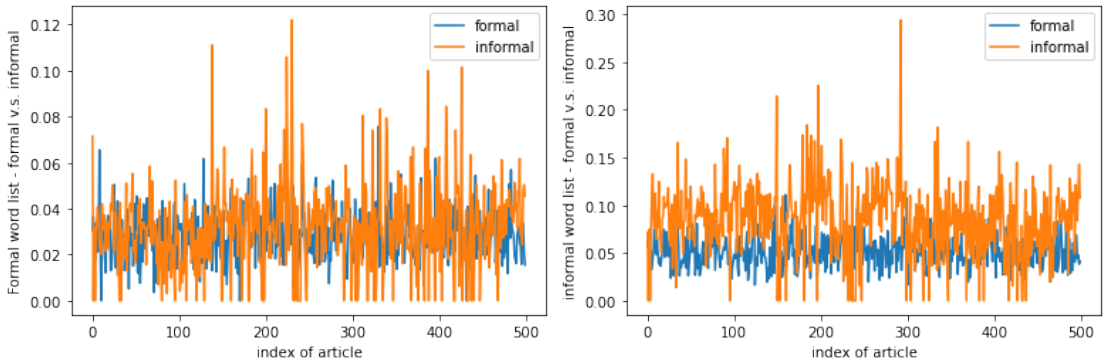


Figure 4.5: The comparison of the normalized contractions between formal and informal English text samples.

4.1.6 Informal and Formal Word List

In Figure 4.6, with formal word list, it is difficult to tell the range of formal and informal datasets apart. The difference is more obvious when using the informal word occurrence

in the articles. In general, informal articles contain more informal words than formal articles.



(a) Formal word list

(b) Informal word list

Figure 4.6: The comparison of the normalized informal and formal pronouns between formal and informal English text samples.

4.2 Heterogeneous Dataset

Despite the small size of the data set and the possibility of very subjective categorization, we would still like to discuss the distribution of the features in our heterogeneous dataset model (Figure 4.7). In Figure 4.7, class one refers to formal dataset, class two refers to semi-formal/informal dataset, and class three refers to informal dataset. In general, formal documents have the same kind of metrics as in the homogeneous dataset. The only differences are that we have some formal documents contain some informal pronouns and words since we include formal letters and the dataset and formal blogs. Informal documents and semi-formal/informal documents from heterogeneous dataset are quite similar to the informal documents from homogeneous dataset. There is no clear distinction between these two classes from the plot.

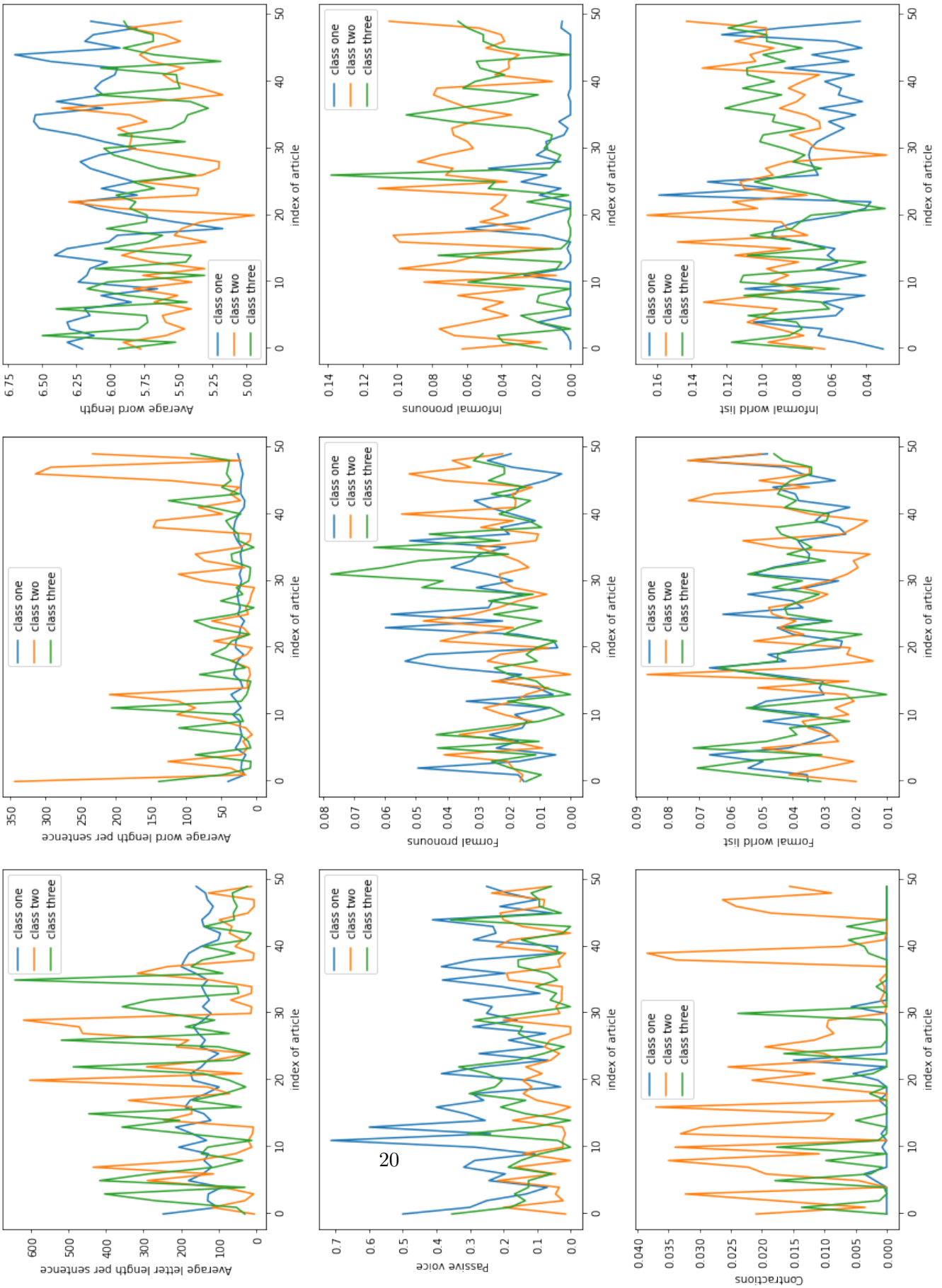


Figure 4.7: The comparison of the English texts from heterogeneous dataset regarding the features

Chapter 5

Results

The overall results are shown in Table 5.1. In the table, we can find how many features the model needs to produce the highest accuracy score, what are the features, the best average accuracy , and the max accuracy scores when we apply all features. In the visualization section, we discussed that the average word per sentence and average letter per sentence are correlated. Usually, we remove correlated features to avoid redundancy, but after removing of the correlated features the accuracy of all models decreased by one percent. Since we do not have a large set of features and both features can make the accuracy score a little higher, we decide to keep both of the features in the classification. In general, the binary classification on homogeneous dataset works better than the multi-class classification on heterogeneous dataset. The details are discussed in the sections below.

	Best number of features	Important features*	best average accuracy	maximum accuracy	minimum accuracy	overall accuracy
decision tree for homogeneous dataset	6	[0,4,5,6,7,8]	98.2%	98%	95%	96.6%
random forest for homogeneous dataset	7	[0,1,3,5,6,7,8]	99%	100%	97%	98.7%
logistic regression for homogeneous dataset	2	[5,8]	91%	93%	82%	85.4%
decision tree for heterogeneous dataset	2	[2,5]	67.33%	93.33%	46.67%	66.67%
random forest for heterogeneous dataset	4	[0,1,2,5]	71.34%	80%	53.33%	70%
logistic regression for heterogeneous dataset	8	[3,5]	52%	60%	40%	50%
*feature index: 0 = the average letter length per sentence, 1 = the average word length per sentence, 2 = the average word length, 3 = the normalized passive voice size, 4 = the average number of formal pronouns, 5 = the average number of informal pronouns, 6 = the average number of contractions, 7 = the average number of the formal words, 8 = the average number of the informal words						

Table 5.1: Heterogeneous and Heterogeneous dataset classification with their results

5.1 Decision Tree

5.1.1 Homogeneous Dataset

In the Decision Tree model (Figure 5.1), we apply Stratified 10-Fold Cross Validation during the process of recursive feature selection to find the relationship between the size of the subset of features and the accuracy. The important features to get the highest average accuracy 98.2% are the average letter length, formal and informal pronouns, contractions, and formal and informal word occurrence. We also find when using only the informal pronouns and informal word occurrence feature, the average accuracy is 96.5%. When using all features as parameters, we get the highest accuracy 100% and the lowest accuracy 95%. On average, the accuracy is 98.1%.

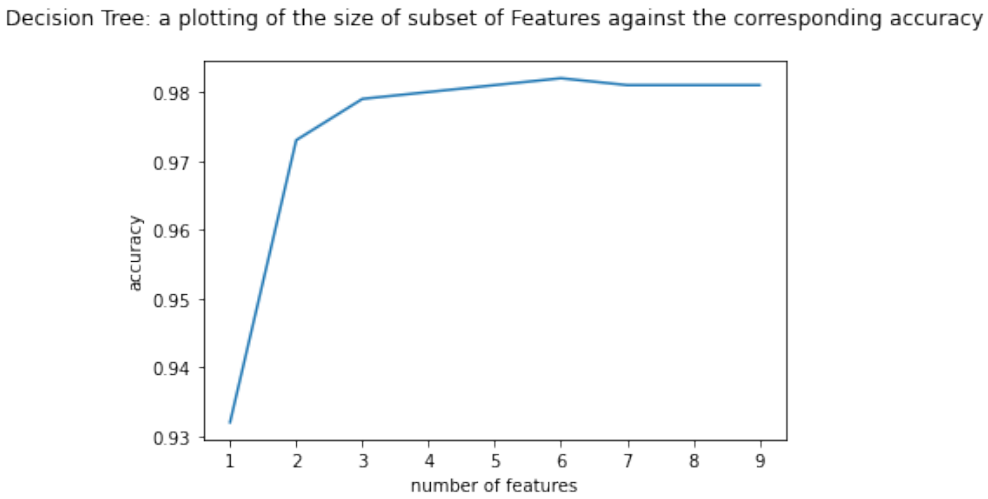


Figure 5.1: Homogeneous dataset: The size of subset of features against the corresponding mean accuracy in Decision Tree Model

5.1.2 Heterogeneous Dataset

In contrast with the homogeneous dataset, decision tree model does not show very promising results for the heterogeneous dataset. The highest average accuracy 67.33% is achieved through two features based on the feature selection: average word length and informal pronouns (see Figure 5.2). When applying all features, the average accuracy drops to 60% with maximum accuracy 73.33% and minimum accuracy 46.67%.

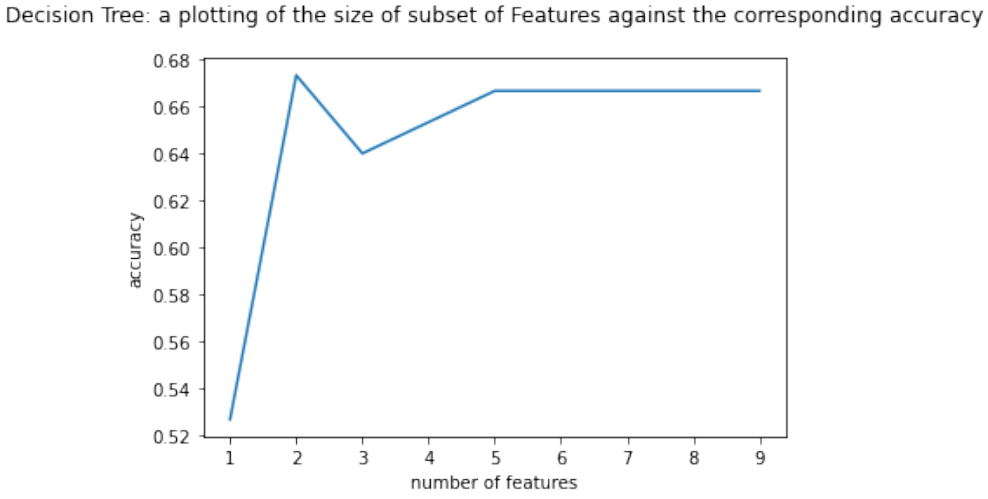


Figure 5.2: Heterogeneous dataset: The size of subset of features against the corresponding mean accuracy in Decision Tree Model

5.2 Random Forest

5.2.1 Homogeneous Dataset

Random Forest achieves the highest average accuracy (99%) using eight features, all features except for average word length. When applying features during the classification, the average accuracy is 98.7%. The maximum accuracy is 100%, and the minimum accuracy is 97%.

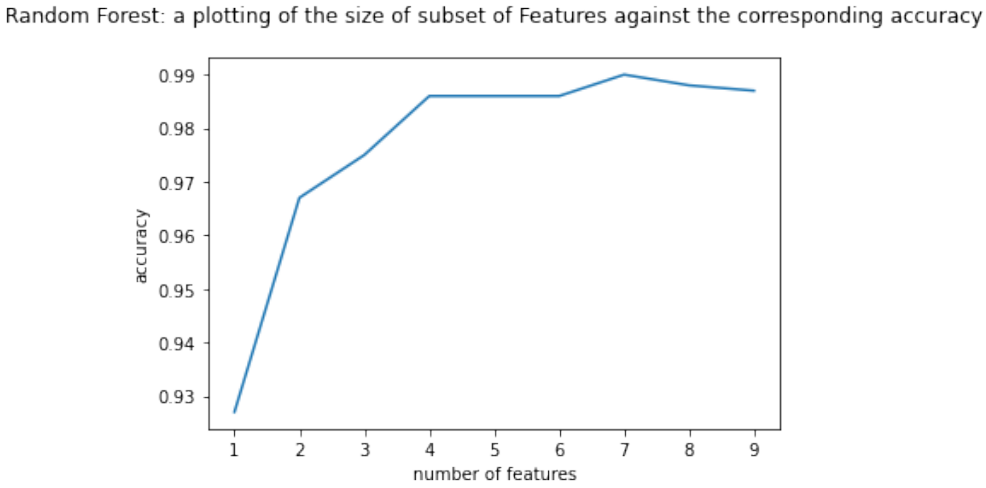


Figure 5.3: Homogeneous dataset: the size of subset of features against the corresponding mean accuracy in Random Forest Model

5.2.2 Heterogeneous Dataset

For the heterogeneous dataset, random forest has achieved a higher average accuracy (71.3%) than the decision tree using four features (Figure 5.4). The important features are average letter per sentence, average word per sentence, average word length and informal pronouns. If we apply all features in the classification, the maximum accuracy is 86.67%, the minimum accuracy is 46.67% and the overall accuracy is 68%.

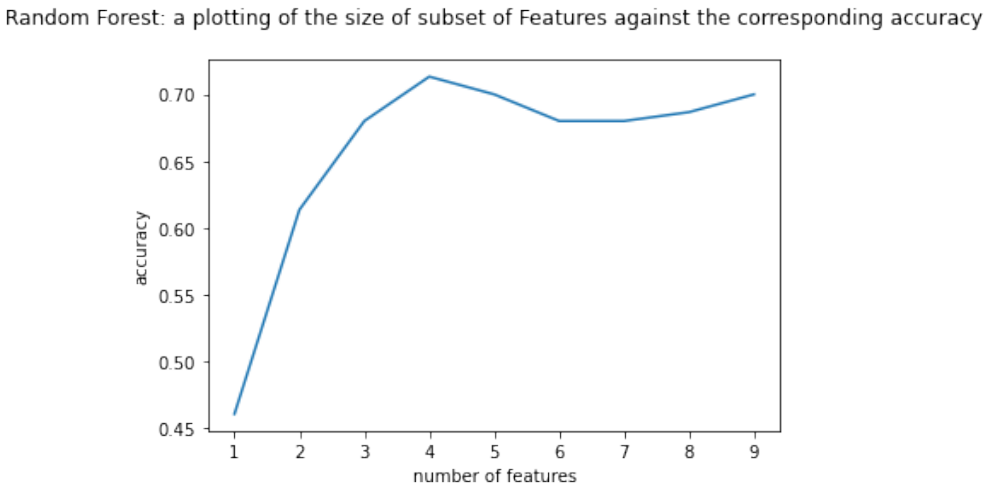


Figure 5.4: Heterogeneous dataset: the size of subset of features against the corresponding mean accuracy in Random Forest Model

5.3 Logistic Regression

5.3.1 Homogeneous Dataset

In Figure 5.5, we could examine the results that we obtain from the recursive feature elimination with stratified 10-fold cross validation. The highest average accuracy 91% is obtained when we use only two features: informal pronouns and informal word occurrence. If we use all nine features in the model, the maximum accuracy in this model is 93%, and the minimum accuracy is 82%. On average, the accuracy is 85.4%.

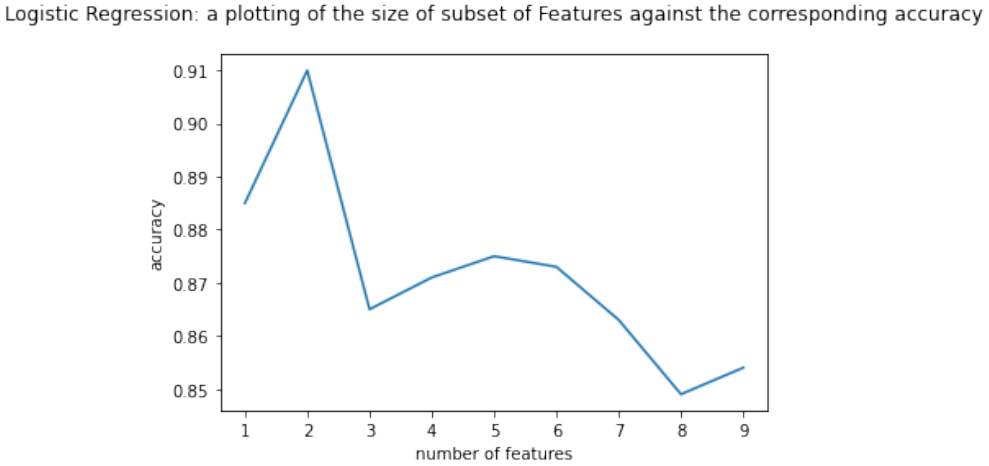


Figure 5.5: Homogeneous dataset: the size of subset of features against the corresponding mean accuracy in Logistic Regression Model

5.3.2 Heterogeneous Dataset

In figure 5.6, we observe again that the accuracy is not as good as the one obtained for the homogeneous dataset. The highest average accuracy achieved by logistic regression is 63.33% with all features except for average letter size per sentence. Taking all nine features into the classification process, the maximum accuracy is 80%, the minimum accuracy is 46.67%, and the average is 62%.

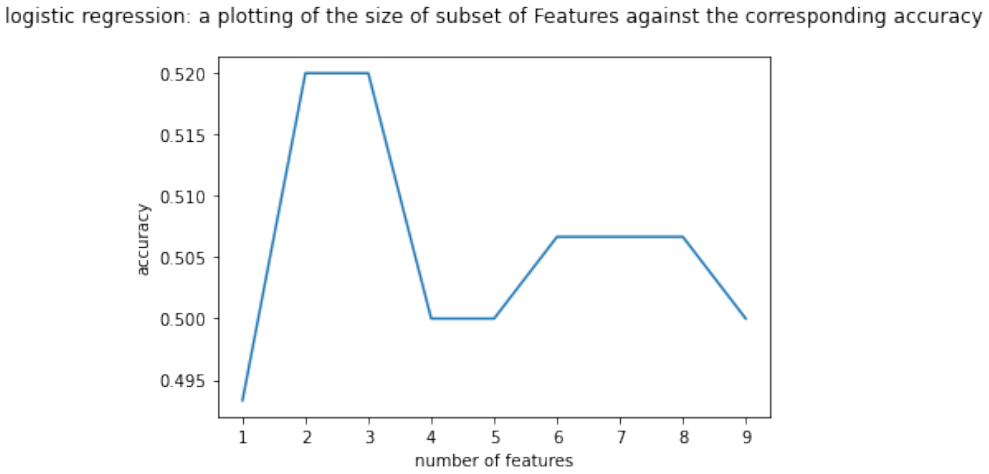


Figure 5.6: Heterogeneous dataset: the size of subset of features against the corresponding mean accuracy in Logistic Regression Model

5.4 Heterogeneous Dataset from Ternary to Binary class

Since we did not get promising results from the heterogeneous dataset with multi-class classification, we combine the semi-formal/informal class and informal class as one class and run another classification between the new dataset and formal dataset. The results are in Table 5.2. We observe a large improvement, and the highest average accuracy of both decision tree and random forest have a higher competitive accuracy.

Heterogeneous dataset	Best number of features	Important features*	best average accuracy	maximum accuracy	minimum accuracy	overall accuracy
decision tree with three classes	2	[2,5]	67.33%	93.33%	46.67%	66.67%
random forest with three classes	4	[0,1,2,5]	71.34%	80%	53.33%	70%
logistic regression with three classes	2	[3,5]	52%	60%	40%	50%
decision tree with two classes	3	[1,2,5]	82%	100%	60%	81.33%
random forest with two classes	5	[0,1,2,5,8]	90%	100%	73.33%	89.33%
logistic regression with two classes	9	[0-8]	75.33	86.67%	66.67%	75.33%
*feature index: 0 = the average letter length per sentence, 1 = the average word length per sentence, 2 = the average word length, 3 = the normalized passive voice size, 4 = the average number of formal pronouns, 5 = the average number of informal pronouns, 6 = the average number of contractions, 7 = the average number of the formal words, 8 = the average number of the informal words						

Table 5.2: Heterogeneous dataset binary and ternary classification with their results

Chapter 6

Implementation and Discussion

6.1 Implementation

Ideally, the implementation of this classification is meant to be done in JavaScript. However, due to the time being and the integration problems with yoastseo library [18], we decide to implement the classification approach in Python. For the purpose of future uses, we demonstrate the classification process in pseudo code below. The idea of this classification is to avoid the use of machine learning on the browser's client side. The implementation is based on the model with five-hundred documents homogeneous data set. The implementation for heterogeneous dataset is similar, but requires further investigation and improvement; thus, we omit the implementation here. We only show the implementation of logistic regression and decision tree since random forest is just about combining multiple decision trees.

6.1.1 Feature extraction

Since the content from the user's side should be seen as private, the feature extraction part also needs to be done on the user's side. Initially, we need to do the same article processing step as we mentioned in the research section: splitting articles into sentences, words, and letters. For the nine features we used in our research, the average letter length per sentence, average word length per sentence and average word length can be easily extracted with the pre-processed data. The Yoastseo library supports passive voice detection which helps us get the normalized passive feature. Obtaining the average number of formal pronouns, informal pronouns, contractions, formal word and informal word is just a matter of going through the articles and looking for certain words. The sorting and counting processes are also implementable in JavaScript.

6.1.2 Logistic Regression

Based on the result from section 5.3.1, we find that best result is achieved by informal word and informal pronoun occurrence. With these two features as input, we get the weight w and intercept b of logistic regression by using the functions provided by [19].

One thing to note is that the model now is trained on the features extracted by Python function. The model will be more accurate if we could use JavaScript to extract the features of train dataset and train our model again in Python to get the results. For the current model, our weight w is a list of two features:

```
[[-9.62491931, -5.79703978]]
```

and the bias is

```
[0.66125403]
```

We use the following function to get the result of the prediction.

$$sigmoid(w \cdot X + b) \text{ where } sigmoid(x) = \frac{1}{1 + e^{-x}}$$

If the result is higher than 0.5 then the document is formal, otherwise the document is informal.

6.1.3 Decision Tree

From the result in section 5.1.1, we know that the decision tree performs the best with the following features: the average letter length, formal and informal pronouns, contractions, and formal and informal word occurrence. After training with these features, we plot the tree for the model. In Figure 6.1, we show the visualization of the decision tree in our model. Based on the leaves and nodes of the decision tree, we can have a rule-based algorithm in JavaScript as:

```
if the average formal pronoun occurrence <= 0.012
    if the average letter length <= 91.458
        if the average letter length <= 179.938
            return formal label
        else
            if the average informal word occurrence <= 0.095
                return formal label
            else
                if normalized contraction number <= 0.013
                    return formal label
                else
                    if the informal word occurrence <= 0.008
                        return formal label
                    else
                        return informal label
    else
        return informal label
else
```

```
if the average formal word occurrence <= 0.059
    return informal label
else
    return formal label
```

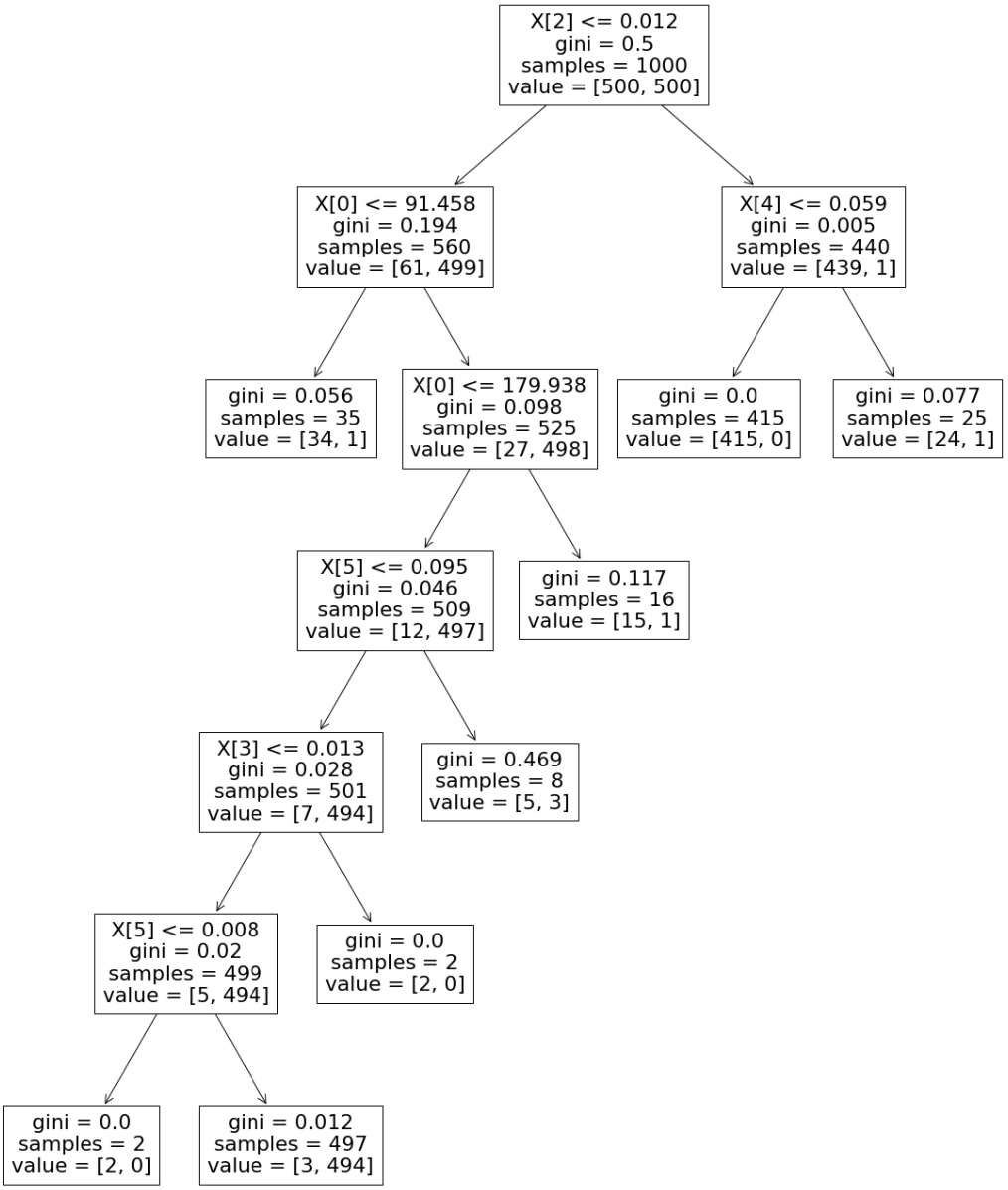


Figure 6.1: Decision Tree Visualization

6.2 Discussion

In this section, we discuss what are the important features when it comes to text classification and provide ideas to interchange formal and informal texts.

Based on the classification process along with the visualization in section 3.5, we discover that it is relatively easy to distinguish formal documents from other types of documents since there are strict rules about formal texts. For informal or semi-formal English texts, the rules are more flexible which leads to more diverse features for the documents. Formal texts have the following characteristics: they have an average of around twenty to thirty words per sentence, they tend to use passive voice, and the informal pronouns, contractions and informal words are also absent for formal writing. In contrast, the features like formal pronouns and formal words do not play a big role during the classification. These two features improve the accuracy of the classifier, but are not the most crucial part when it comes to distinguishing formal and informal texts. The most important features for classification in decision trees and random forest models are informal pronouns, contractions, and informal word occurrences.

For the heterogeneous dataset, we start with a ternary classification, but the categorization is not very optimal. Decision tree and random forest again work better than logistic regression. The necessary features for ternary classification are average word length and informal pronouns. After we combine the non-formal datasets, the classification improves significantly which also confirms that the features for formal documents are more prominent.

When it comes to interchanging informal and formal texts, all features should be taken into account and avoid special cases at the same time. One example of such special case is formal speeches with many informal pronouns "I". To interchange the informal and formal text, we propose the following: We could prompt to the users the average word length per sentence during the process of writing and notify them that a good length for formal writing is around twenty to thirty words. The average sentence is difficult to improve by machines without changing the meaning of the sentences. We should also let the users know that formal documents contain more passive voice in general, but it does not mean they should use passive voice for every sentence since passive voice is not one of the most important features. In fact, excessive use of passive voice is discouraged in the book *On Writing Well* [20] because active voice expresses stronger opinion than passive voice. Excessive use of passive voice also makes the text difficult to read, and writers should take readability into account. For formal writings, we could have an algorithm to suggest to change all informal pronouns "I" and "you" to "we", "one" or "they" for simple modifications. However, for a thorough informal pronoun replacement, we should show several options for the users to replace the pronouns. For example, in the sentence "You are allowed to enter with a student card", the best option is not to change the sentence to "We are allowed to enter with a student card". Possible alternatives could be "Students are allowed to enter with a student card" or "It is allowed

to enter with a student card". Accurate pronoun suggestions or changing active voice to passive voice are something interesting to investigate for the future search; however, it goes without saying that providing a suitable pronoun or changing active to passive are both challenges tasks since it is difficult for the machines to detect the context of the sentence. For informal writing, it is not necessary to do anything with pronouns since formal pronouns are used in all kinds of writings. However, using more personal pronouns does show the characteristics and opinions of the writer. In terms of contractions, we could detect a contraction and write it out for formal writing and the other way around for informal writing. In the experiment, we also made a formal word list and the corresponding informal word list with similar meaning. We could use the word list to suggest word choices to the users according to the type of writing.

Chapter 7

Conclusions and Future Work

In this thesis, we focus on two different classifications on formality of documents with two different data sets. For the homogeneous data set, we adopt binary classification on the BBC news data set and general blog post data set. We achieve very promising results with logistic regression, decision tree and random forest model. In particular, decision tree and random forest have a high average accuracy above 95%. We also discuss the correlations of formality of text with the features we proposed. The most important features in decision tree and random forest models are informal pronouns, contractions and informal word list. However, we believe that all features are important for categorizing a document as formal. Certain features such as informal pronouns might be really good at classification but are not usable for special cases such as formal speeches. Overall, we find the informal features are the key determinant during the classification process. We also give suggestions based on each feature about how to change informal text to formal, and vice versa.

In contrast with the homogeneous dataset, the heterogeneous dataset is more difficult to distinguish when we divide the dataset into three classes. The maximum accuracy score is provided by random forest which is 71.34%. Apart from the subjective data collection and labeling process, this classification is considerably innovative since most other researches do not consider the situation that a document is semi-formal/informal. These types of writings is more common in real life. In this classification, average word length per sentence and informal pronouns become the most significant features. After we change the ternary to binary classification, the performance of our models gets better which again proves formal documents have more rules and distinct features.

7.1 Future Work

One important aspect to successfully classify formal and informal documents is to have well categorized datasets. In the future, we should expand the size of the dataset. The composition of the heterogeneous documents data set is limited by time and monetary constraints. For a more sophisticated research, the classification should be done with

more people (preferably people who have taken English major) for more accurate classification. Then we take the mean or mode rating of each document as its score. We could also categorize the heterogeneous dataset into more classes for example from a range of one to ten instead of three classes.

For the current research, we have a high accuracy on the classification of homogeneous dataset, but the accuracies for heterogeneous datasets are lower for both binary and ternary classification. We should study how to improve the accuracy on the heterogeneous dataset. In further research, we should delve in more real world problems, improve and adjust the algorithm based on users feedback since there are always special situations for different users. During the experimenter, we browse many blogs, news and English texts. All these texts are quite different from one to another. This is thanks to that authors write for a different audience and everyone has a unique style of writing. In different contexts, the definition of "formal" and "informal" also varies. We could do a more in-depth research of formality on different genres such as speech, sport, entertainment news, etc. It is also important that we take other forms of text expression in classification. For examples, the use of emoji and emoticon could be important features of informal blogs. The formal word list and informal word list should also be extended. It will not only help the classification process, but also provide the users with more word choices for their content. We could also check the potential of integrating the readability and formality research on text document. For example, informal document with slang might not be readable for academic writing, and very formal writing with jargon might not be readable for daily conversation.

Bibliography

[1] V. K. Vijayan, K. R. Bindu, and L. Parameswaran, "A comprehensive study of text classification algorithms," in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2017, pp. 1109–1113.

[2] F. Abu Sheikha and D. Inkpen, "Automatic classification of documents by formality," in *Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering(NLPKE-2010)*, 2010, pp. 1–5.

[3] P. B. Fuß, , A. Kurczewska, S. Dyka, R. Mitkov, I. I. Siitonen, , K. K. Ruokolainen, , A. H. Tiedemann, , and J. X. Huang. [Online]. Available: http://www.connexor.com/

[4] T. Yoast, "Yoastseo," Dec 2021. [Online]. Available: https://wordpress.org/plugins/wordpress-seo/

[5] "The 2 types of learning in machine learning: supervised and unsupervised," Jun 2021. [Online]. Available: https://business.blogthinkbig.com/the-2-types-of-learning-in-machine-learning-supervised-and-unsupervised/

[6] P.-N. Tan, A. Karpatne, V. Kumar, and M. Steinbach, *Introduction to data mining.* Pearson, 2020.

[7] "sklearn.feature_selection.rfe." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html

[8] P. byshrinandk, "Exploring python packages – nltk and spacy," Aug 2020. [Online]. Available: https://evrythngunder3d.wordpress.com/2020/07/20/exploring-python-modules-part-2/

[9] D. Park, "Identifying & using formal & informal vocabulary."

[10] "Enron email dataset." [Online]. Available: https://www.cs.cmu.edu/~enron/

[11] H. Gültekin, "Bbc news archive," Jul 2020. [Online]. Available: https://www.kaggle.com/hgultekin/bbcnewsarchive

[12] R. Tatman, "Blog authorship corpus," Aug 2017. [Online]. Available: https://www.kaggle.com/rtatman/blog-authorship-corpus

[13] N. Sultan, "Formal and informal word list," Sep 2021. [Online]. Available: https://grammaran.com/formal-and-informal-word-list/

[14] "Casual abbreviations in english." [Online]. Available: https://poligo.com/en/articles/vocabulary/casual-abbreviations-english

[15] "Academic writing: Linking/transition words." [Online]. Available: https://libguides.staffs.ac.uk/academic_writing/linking

[16] S. Jin, "Seraaphonano/formal_and_informal_english_classification," Jan 2022. [Online]. Available: https://github.com/Seraaphonano/formal_and_informal_english_classification.git

[17] J. Brownlee, "One-vs-rest and one-vs-one for multi-class classification," Apr 2021. [Online]. Available: https://machinelearningmastery.com/one-vs-rest-and-one-vs-one-for-multi-class-classification/

[18] T. Yoast, "Javascript/packages/yoastseo." [Online]. Available: https://github.com/Yoast/javascript/tree/3198b1e848ede555f309a922788f575c5d03f14b/packages/yoastseo

[19] A. Ng, "Logistic regression - neural networks basics." [Online]. Available: https://www.coursera.org/learn/neural-networks-deep-learning/lecture/LoKih/logistic-regression

[20] W. Zinsser, *On writing well*. Harper Paperbacks, 2013.