BBC News Categorization and Clustering

Introduction

BBC News is an operational business division of the British Broadcasting Corporation (BBC) responsible for the gathering and broadcasting of news and current affairs. The department is the world's largest broadcast news organization and generates about 120 hours of radio and television output each day, as well as online news coverage. The service maintains 50 foreign news with more than 250 correspondents around the world. Moreover, the BBC is required by its charter to be free from both political and commercial influence and answers only to its viewers and listeners.

BBC News online launched in November 1997. It is one of the most popular news websites in the UK, reaching over a quarter of the UK's internet users, and worldwide, with around 14 million global readers every month. The website contains international news coverage as well as finance, entertainment, sport, science, and political news. Many television and radio programs are also available to view on the BBC News online platform which have been available to view 24 hours.

Therefore, Audiences can easily access to read the BBC News on platform online to update the informational diversity all around the globe. The objective in this paper aim to comprehend the key words to classify the dataset. As a result, creating predictive model by using textual analysis as an input and learn the process or key words that essentially classify the types of news.

This paper is organized into 3 parts. The First part illustrates the dataset overview and the data exploration. The second part is about creating a classification model to predict the types of documents. Finally, the clustering algorithms were applied to cluster the News segmentation.

Part 1 Dataset Overview

The dataset was contained Consists of 2,225 documents from the BBC news website corresponding to the stories in five topical areas that consist of Natural Classes: 5 (Business, Entertainment, Politics, Sport, Tech) between 2004 and 2005 from all around the world. The original data came from www.Kaggle.com but was reduced to 900 documents which were randomly chosen 180 documents each subject in order to increase the efficient time of processes. The dataset was downloaded from

https://www.kaggle.com/shivamkushwaha/bbc-full-text-document-classification. The original data contained all information of news in the notepad that was used in this assignment for categorization and clustering. However, this text files were transformed into Excel and use attributes “Information” and ‘Segments’ in order to apply in the data exploration process.

Row No.	No. Documents	Information	Segment
1	001.txt	Ad sales boost Time Warner profit	Business
2	001.txt	Quarterly profits at US media giant TimeWarner jumped 76% to $1.13bn (□□600m) for the three months to December, from $639m year-earlier.	Business
3	001.txt	The firm, which is now one of the biggest investors in Google, benefited from sales of high-speed internet connections and higher advert sales. Ti...	Business
4	001.txt	Time Warner said on Friday that it now owns 8% of search-engine Google. But its own internet business, AOL, had has mixed fortunes. It lost 464,0...	Business
5	001.txt	Time Warner's fourth quarter profits were slightly better than analysts' expectations. But its film division saw profits slump 27% to $284m, helped by...	Business
6	001.txt	TimeWarner is to restate its accounts as part of efforts to resolve an inquiry into AOL by US market regulators. It has already offered to pay $300m to...	Business
7	002.txt	Dollar gains on Greenspan speech	Business
8	002.txt	The dollar has hit its highest level against the euro in almost three months after the Federal Reserve head said the US trade deficit is set to stabilise.	Business
9	002.txt	And Alan Greenspan highlighted the US government's willingness to curb spending and rising household savings as factors which may help to red...	Business
10	002.txt	Worries about the deficit concerns about China do, however, remain. China's currency remains pegged to the dollar and the US currency's sharp fal...	Business
11	003.txt	Yukos unit buyer faces loan claim	Business
12	003.txt	The owners of embattled Russian oil giant Yukos are to ask the buyer of its former production unit to pay back a $900m (□□479m) loan.	Business
13	003.txt	State-owned Rosneft bought the Yugansk unit for $9.3bn in a sale forced by Russia to part settle a $27.5bn tax claim against Yukos. Yukos' owner ...	Business
14	003.txt	Rosneft officials were unavailable for comment. But the company has said it intends to take action against Menatep to recover some of the tax claim...	Business
15	004.txt	High fuel prices hit BA's profits	Business

Figure 1 The original dataset

The dataset was explored using RapidMiner to illustrate the data exploration from Figure 2 to 7. The objective of the initial exploration aims to understand better about basic aspect for this dataset. In figure 2, the bar chart shows an average word count for each segment. Politics and Technology have more words (average above 400 words) in comparison with other categories. On the other hand, Entertainment and Sport are similar and have a smaller number of the words count which are approximately 260 words.

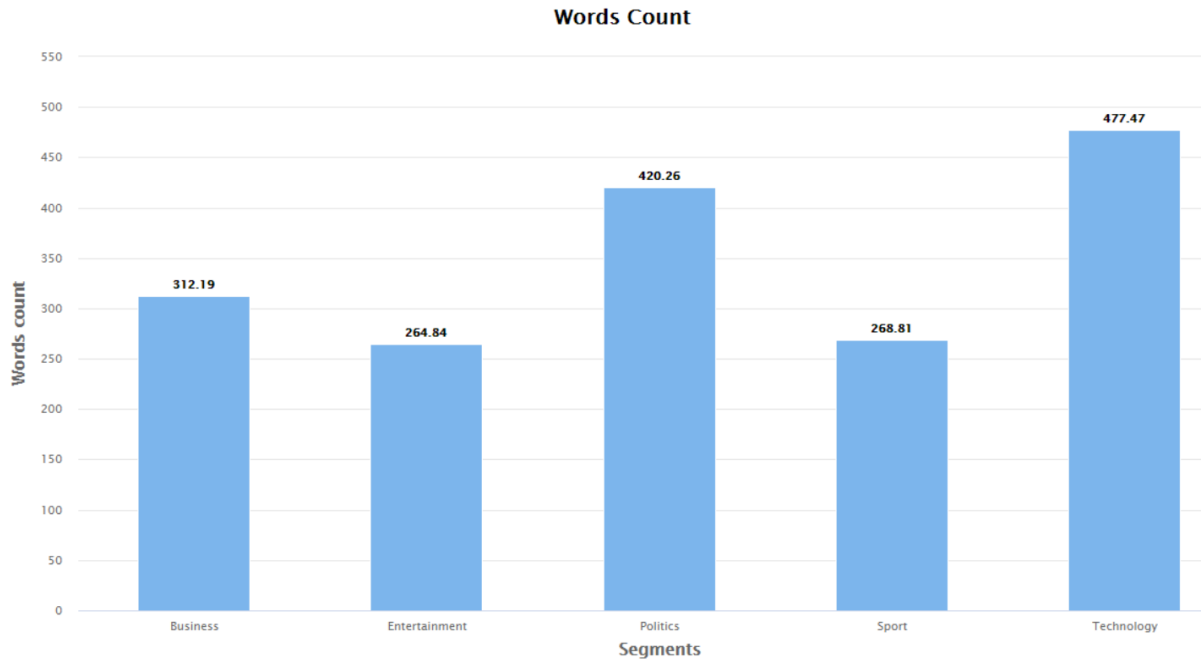


Figure 2 The average words count each segment

In figure 3, the boxplot demonstrates the dispersion of data that shows the interquartile range and Minimum & Maximum data to understand the overall information.

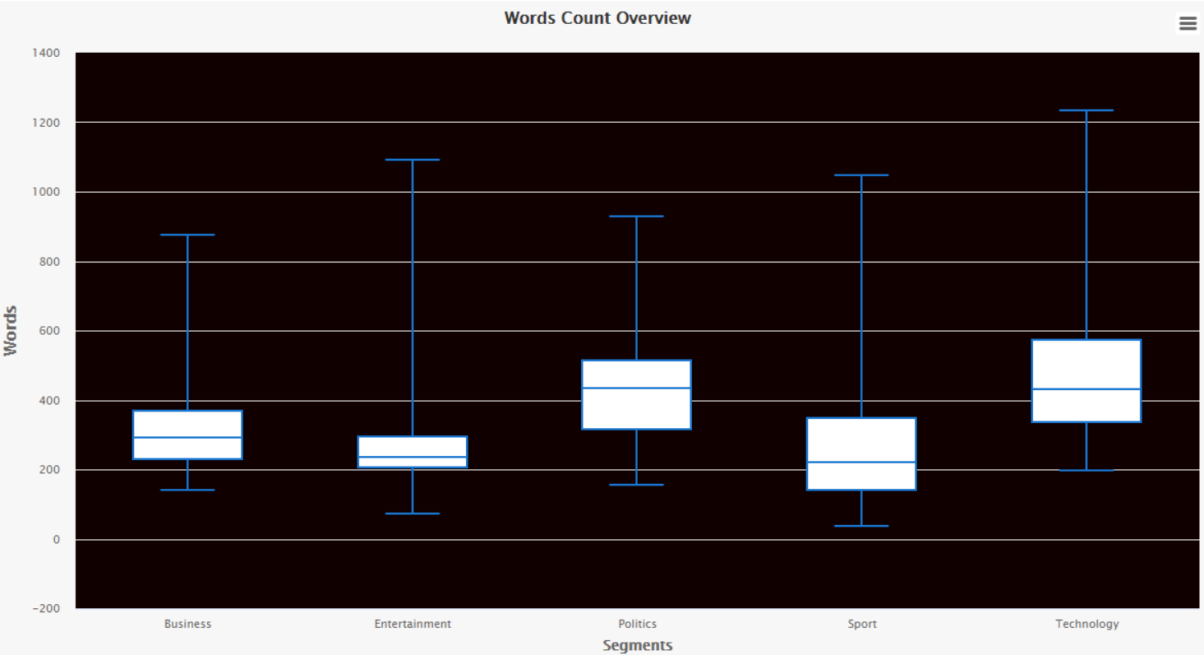


Figure 3 The dispersion of data (words count)

In figure 4, Presenting distribution of words length by histogram is somewhat clarity that the words length of most segment tends to skew to the left.

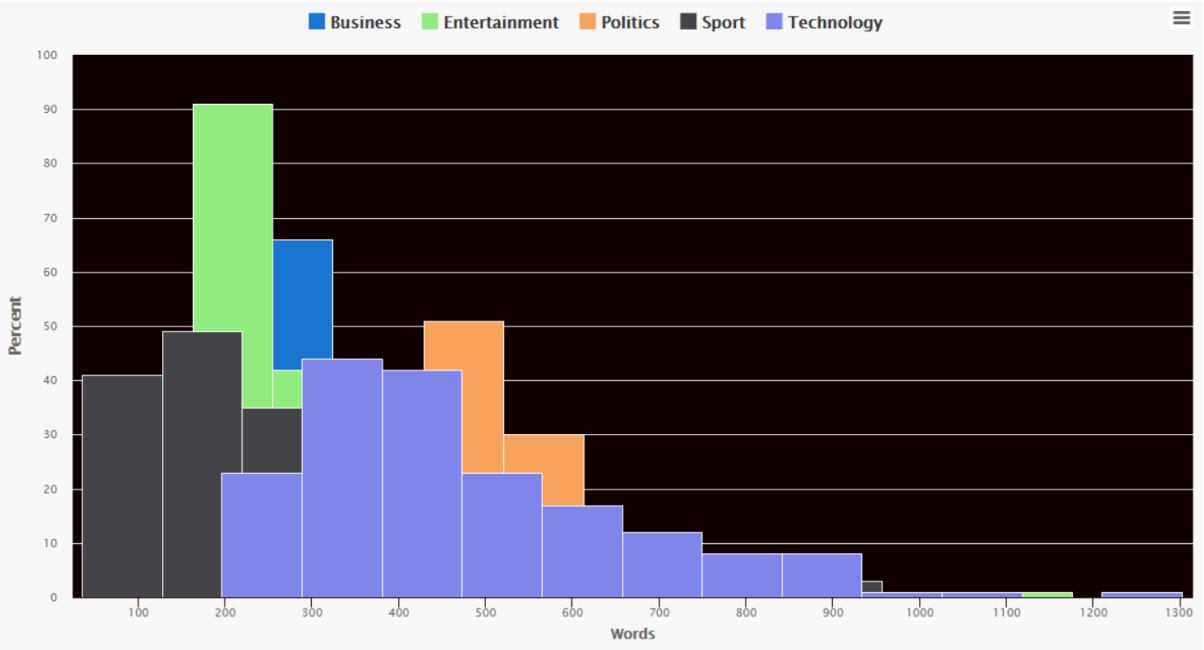


Figure 4 Distribution of words length (words count)

In figure 5, Word Clouds obviously show specific words in every segment, which especially identify the types of news. For example, there are the major proportion of the key words such as "Company", "Product", "Economic", "Profit", "Growth" etc. that indicate the relevant words of the business. However, there are some slightly numbers of Word Cloud that are unable to clearly separate the group. For instance, "Take", "Said", "Time", "World" is appeared more than one category and cannot identify the exact group.

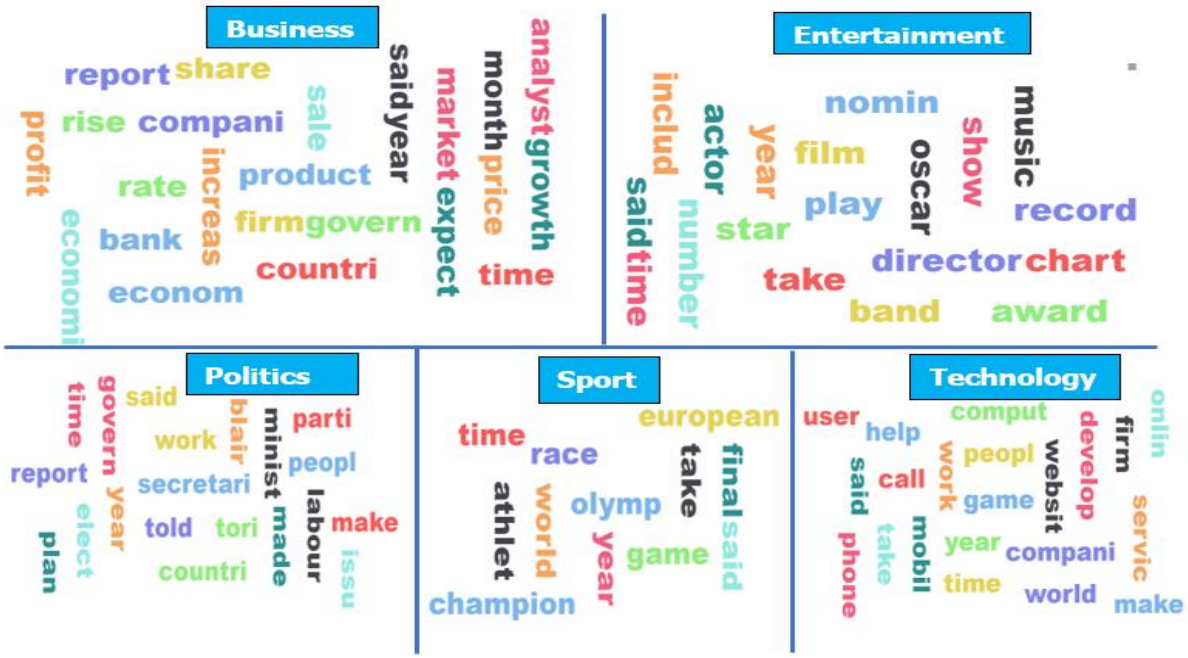


Figure 5 Word Cloud for every category

The maximum TF-IDF of each document was plotted in figure 6. The maximum TF-IDF is 0.911 that is 'Bank'. It is interesting to see that the word 'Bank', 'Load' and 'Game' are able to clearly classify the types of news in figure 7, which have the high TF-IDF point and frequently appeared for specific topic such as the word 'Game' was essentially happened in technology news.

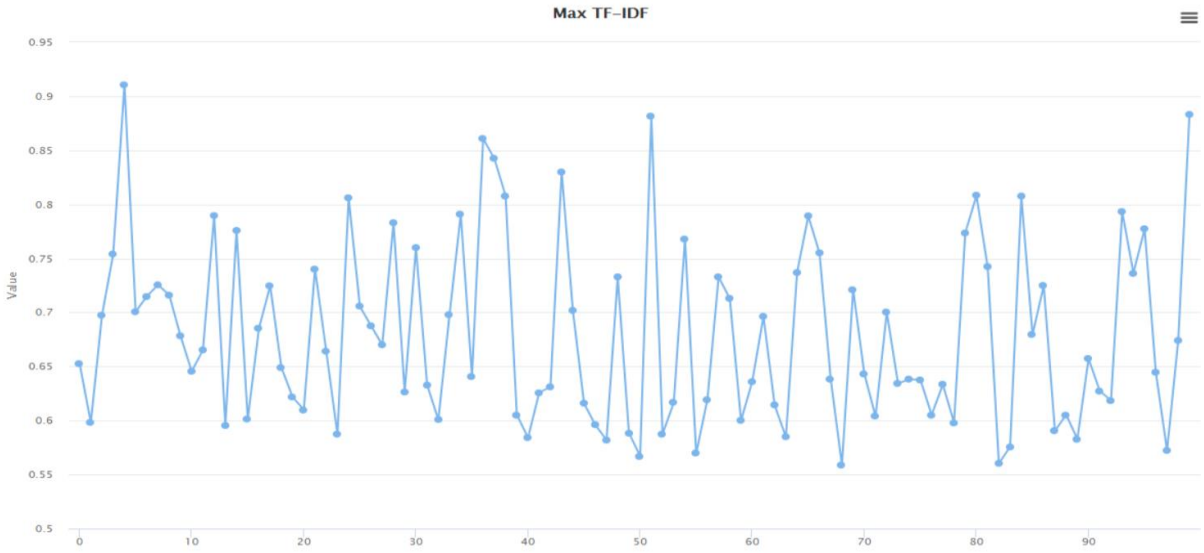


Figure 6 Document Vector and Indexing

Attributes	Max-TF*IDF	Total Occurrences	Document Occurrences	Bus.	Ent.	Pol.	Spo.	Tec.
Bank	**0.911**	198	65	**146**	2	3	2	45
Women	**0.883**	137	56	12	11	**55**	**42**	17
Lord	**0.882**	149	47	3	9	**130**	1	6
Game	**0.842**	598	154	3	14	8	156	**417**
Human	**0.830**	84	45	3	5	**41**	0	**35**

Figure 7 The maximum TF-IDF in top 5 ranking words

Part 2 Classification Processes

2.1 Data preprocessing

The pre-processing method include Transform cases (lowercase), Tokenize, Filter stop words (English), Stem (Porter) and Filter Tokens (by Length) that filter the Minimum and Maximum between 4 and 50 characters.

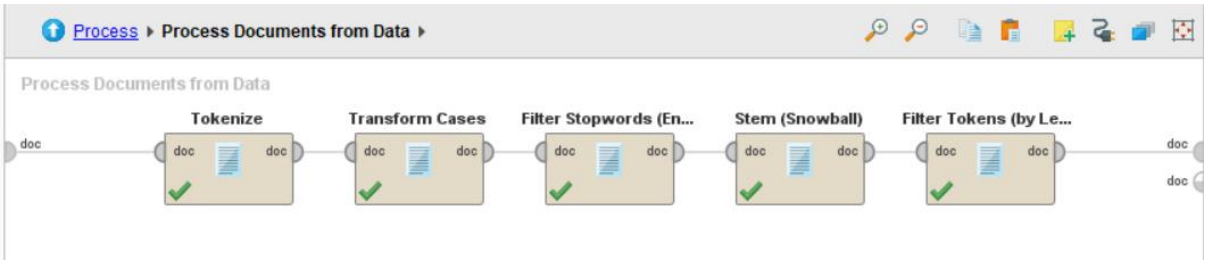


Figure 8 RapidMiner Pre-processing process

2.2 Apply model (Setting Default)

The document vector was created by TF-IDF method with 900 sample documents were used to train and test the model by using Cross Validation operators (number of folds = 10) which contained Naïve Bayes and k-NN inside the operators to compare the accurate predictive models.

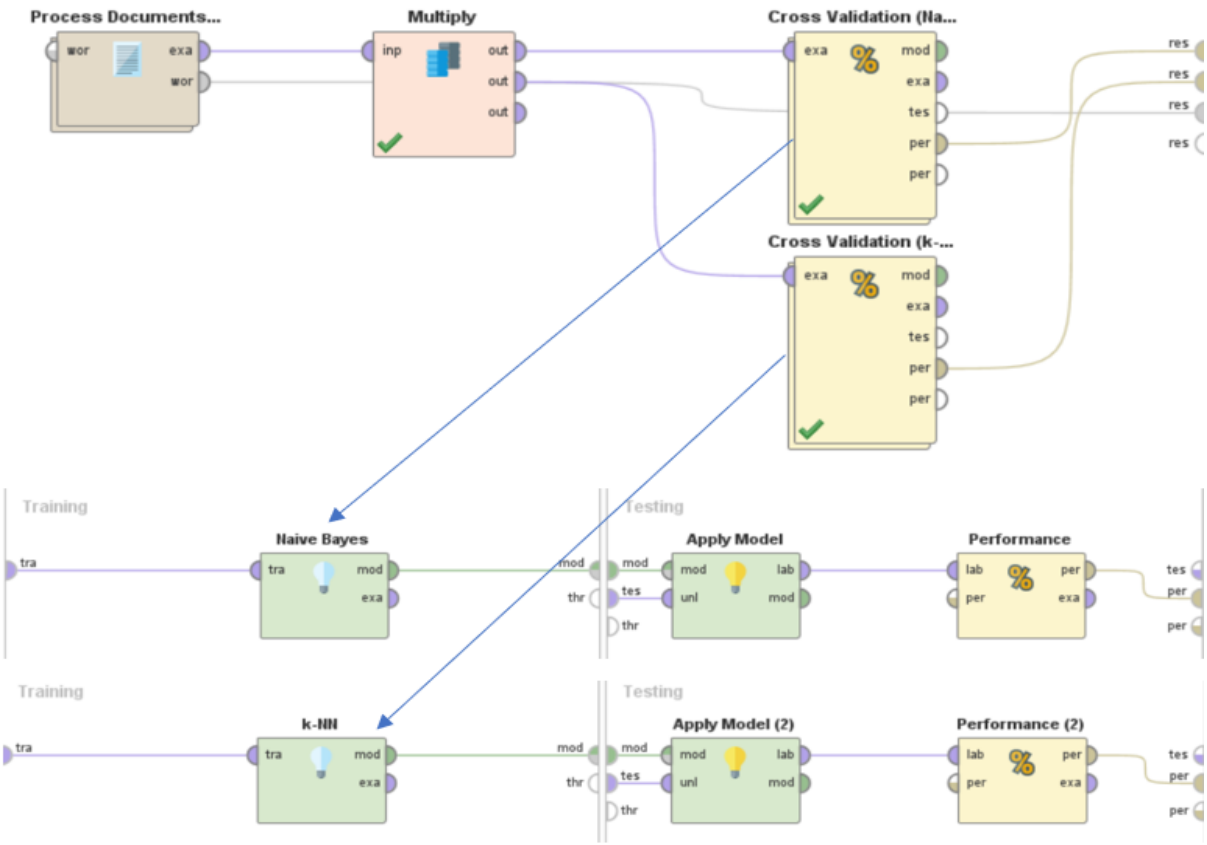


Figure 9 RapidMiner Processes

2.3 Model Evaluation

The result of both models is quite well performance, which Naïve Bayes has accuracy at 87.11%. However, there are some small incorrect predictions that predicted Politics but the truth is Business at 19 documents. Additionally, to predicted Technology but the truth are Business and Entertainment at 20 and 16 documents respectively.

accuracy: 87.11% +/- 3.64% (micro average: 87.11%)

	true Business	true Entertainment	true Politics	true Sport	true Technology	class precision
pred. Business	137	2	7	2	1	91.95%
pred. Entertainment	2	153	7	6	5	88.44%
pred. Politics	19	7	158	2	6	82.29%
pred. Sport	2	2	2	169	1	96.02%
pred. Technology	20	16	6	1	167	79.52%
class recall	76.11%	85.00%	87.78%	93.89%	92.78%	

Figure 10 The result of Naïve bayes model

K-NN has more accuracy than Naïve bayes at 92.23% and there is no significant incorrect prediction was showed in this model.

accuracy: 92.33% +- 3.37% (micro average: 92.33%)

	true Business	true Entertainment	true Politics	true Sport	true Technology	class precision
pred. Business	158	2	3	0	3	95.18%
pred. Entertainment	1	162	5	1	3	94.19%
pred. Politics	11	10	166	2	5	85.57%
pred. Sport	3	1	1	177	1	96.72%
pred. Technology	7	5	5	0	168	90.81%
class recall	87.78%	90.00%	92.22%	98.33%	93.33%	

Figure 11 The result of K-NN model (K = 5)

2.4 Tuning Model

Change vector creation model

In order to improve the efficiency of the model, different options of vector creation were tested. For the previous step, TF-IDF was used to create document vector as a default parameter. For this step, the result from different vector methods such as Binary Occurrence, Term Occurrence and Term Frequency were used to compare the accuracy in each method. Overall, Naïve Bayes has the best accuracy at 92.11% with Binary Occurrence method, while K-NN has the best result at 92.00% With Term frequency Method. Term frequency and TF-IDF have a similarly excellent result, which perform above 90%. However, TF-IDF is more suitable method because it considers the frequency number that was normalized value and the unique words that happen in specific documents. In other words, if the word has high frequent number and occur in less document, it will obtain high TF-IDF score.

Type of Models	Vector Methods		
	Binary Occurrence	Term Occurrence	Term Frequency
Naïve Bayes	92.11%	87.89%	88.89%
K-NN	46.89%	54.11%	92.00%

Figure 12 Comparing accuracy in different vector methods

Change Pruning

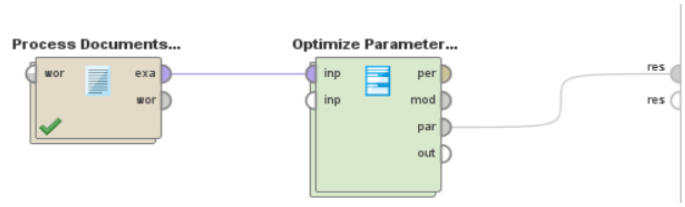
Pruning TF-IDF was performed by setting pruning as a percentage by Creating 3 scenario. As a result of this test, all of three scenarios have a slightly different performance. However, the Second scenario prune by percentual for TF-IDF is lower than 5% and higher than 45%. As one can see, the large number of attributes were dramatically reduced size from 11,670 to 528 attributes but there is able to maintain the accuracy for both models, which increasingly improve an efficient time for running processes.

Type of Models	None	Pruning (3 times)		
		First Time Percentual (3,50)	Second Time Percentual (5,45)	Third Time Percentual (7,40)
Naïve Bayes	87.11%	87.00%	87.78%	87.22%
K-NN	92.33%	90.67%	91.33%	89.33%
Total attributes	11,670	892	528	334

Figure 13 Comparing accuracy in different vector methods

Change Parameters

For K-NN, there is random choice K numbers 10 times to find out the optimum value by using Optimize Parameters (Grid). As a result, the accuracy was slightly adjusted from 91.33% to 93.67% when K numbers change from 5 to 10. It is possible that the data as the same group of the segment are adjacent to one another and it affect the result of accuracy.



Optimize Parameters (Grid) (10 rows, 3 columns)

iteration	k-NN.k	accuracy
1	1	0.898
6	6	0.918
3	3	0.903
2	2	0.899
4	4	0.914
7	7	0.930
8	8	0.930
9	9	0.932
5	5	0.913
10	10	0.937

```
Performance:
PerformanceVector [
-----accuracy: 93.67% +/- 3.23% (micro average: 93.67%)
ConfusionMatrix:
True:   Business        Entertainment   Politics        Sport   Technology
Business:       161     2       3       0       4
Entertainment:  1       163     1       0       3
Politics:       9       9       170     1       2
Sport:  3       0       2       179     1
Technology:     6       6       4       0       170
]
k-NN.k  = 10
```

Figure 14 Adjust parameter in K-NN model

Change model and tuning parameters

Using Decision tree to classify the types of documents because this model easily understands the processing algorithm. For the first time, Decision tree model was set up as default parameters and using almost similar RapidMiner processes as before that show in the figure 15. As a result, Decision tree accuracy is only 45.78% that is obviously ineffective performance.

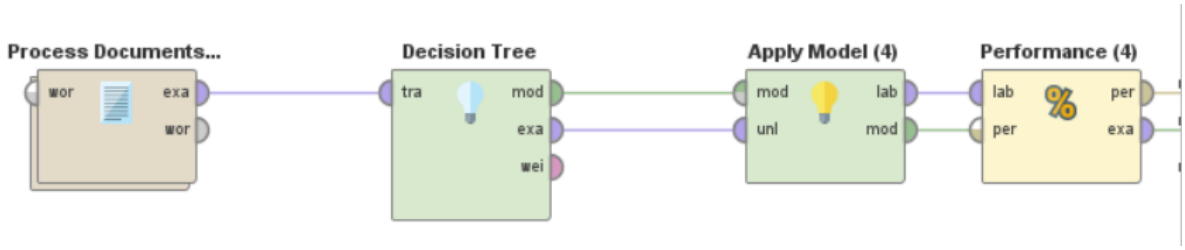
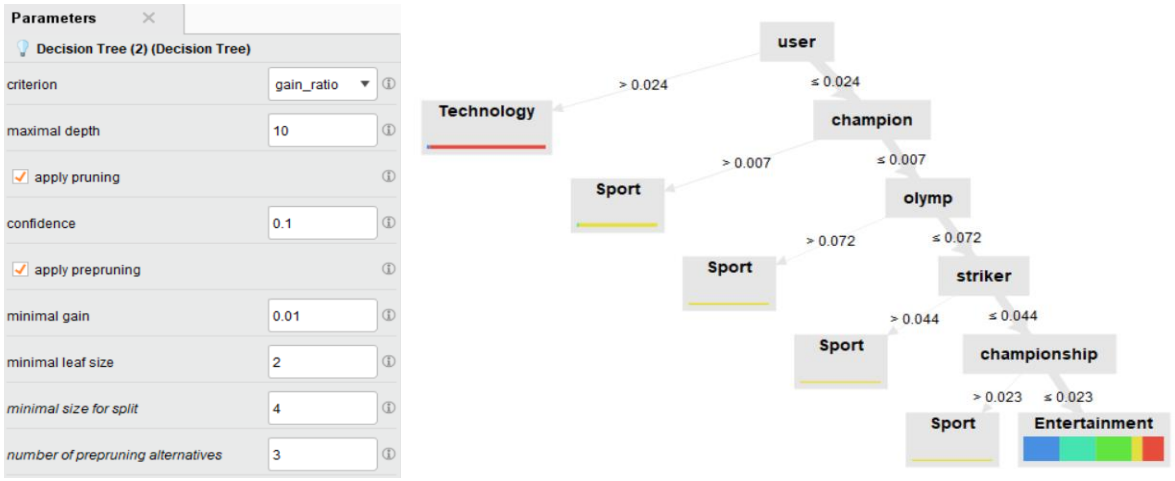


Figure 15 RapidMiner processes for Decision tree model

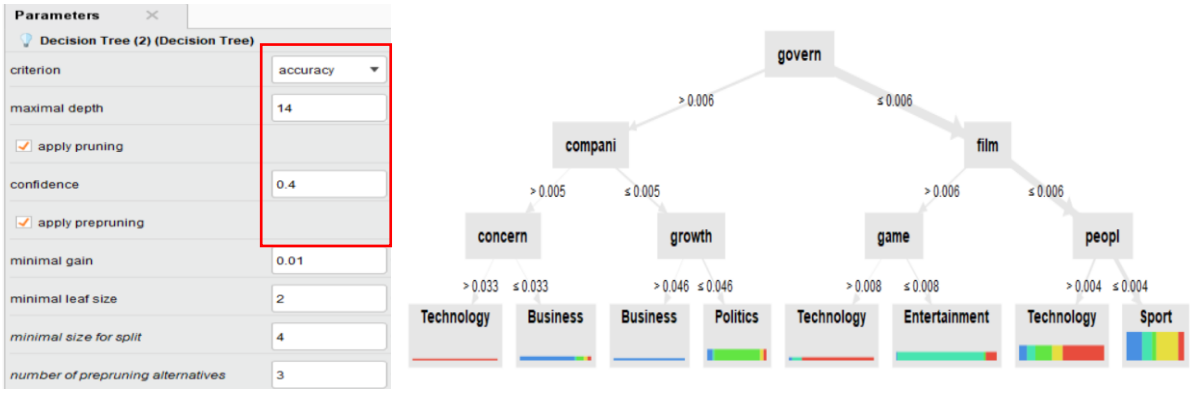


accuracy: 45.78%

	true Business	true Entertainment	true Politics	true Sport	true Technology	class precision
pred. Business	0	0	0	0	0	0.00%
pred. Entertainment	178	179	179	22	105	27.00%
pred. Politics	0	0	0	0	0	0.00%
pred. Sport	0	1	1	158	0	98.75%
pred. Technology	2	0	0	0	75	97.40%
class recall	0.00%	99.44%	0.00%	87.78%	41.67%	

Figure 16 Parameters, Rule Model, Decision tree Accuracy (Default)

Subsequently, Decision tree was adjusted parameters, which use the optimize parameters (Grid) operation to improve the increasing accuracy that change criterion from gain ratio to accuracy, maximal depth and the confidence level. The type of criterion is significant effect to the result because business and politics labeled attributes were ignored to operate in Gain ratio criterion. Whereas, Accuracy criterion consider all of attributes with the most accurate performance. The result after adjustment is high accuracy which substantially increase from 45.78% to 91.67%.



accuracy: 91.67%

	true Business	true Entertainment	true Politics	true Sport	true Technology	class precision
pred. Business	159	2	3	1	2	95.21%
pred. Entertainment	2	157	1	1	6	94.01%
pred. Politics	6	2	168	0	1	94.92%
pred. Sport	4	11	5	173	3	88.27%
pred. Technology	9	8	3	5	168	87.05%
class recall	88.33%	87.22%	93.33%	96.11%	93.33%	

Figure 17 Parameters, Rule Model, Decision tree Accuracy (After tuning parameters)

Part 3 Clustering Processes

3.1 Data preprocessing and Apply model

The pre-processing method use the same operators as Classification Processes which include Transform cases (lowercase), Tokenize, Filter stop words (English), Stem (Porter) and Filter Tokens (by Length) that select the range of characters between 4 and 50 characters. For clustering process, the document vector was created by TF-IDF method and pruning of 5% bottom percentile and 45% top percentile. Moreover, each document vector was applied Single Value Decomposition (SVD) in order for better visualization of the clusters. Creating RapidMiner process by using K-Means with k = 5.

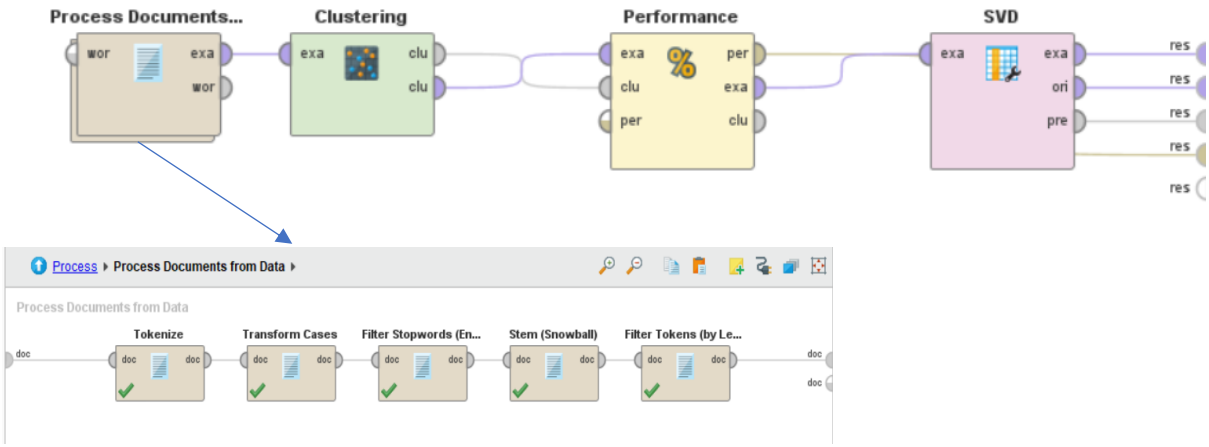


Figure 18 RapidMiner Process

3.2 Model Evaluation

The result was demonstrated in the figure 19 that separate into 5 clusters which clearly segregate with one another except an area of cluster_1 (Dark bule zone) has some green data points that were identified into its.

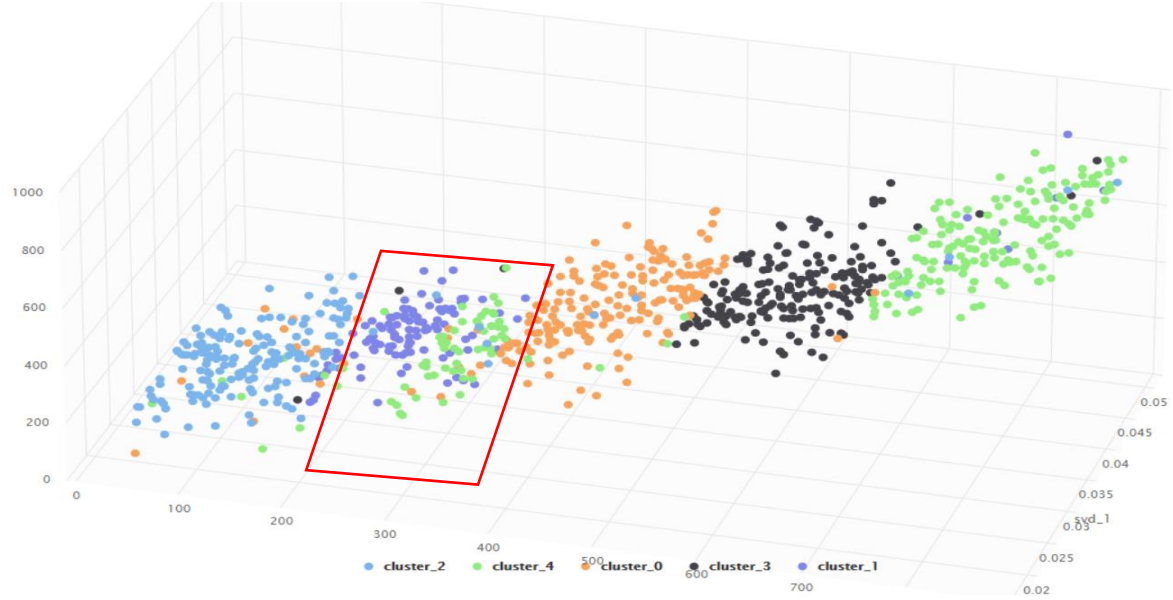


Figure 19 Clustering using K-Means with k = 5

Figure 20 show labeled segments in each cluster which consist of different categories. For instance, the cluster_2 comprise Business at 163 documents as a majority of data and follow by Entertain at 7 documents, Politics at 2 documents and Technology at 4 documents. In other words, in each cluster has significate proportion mix up with the small fraction of the other categories.

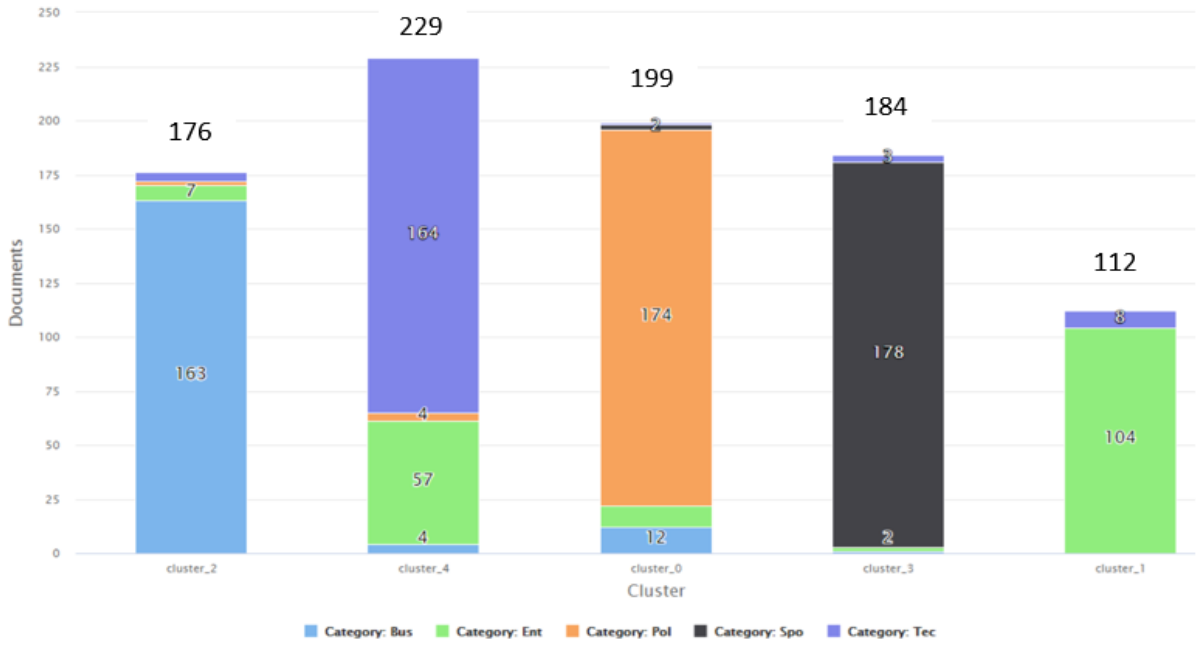


Figure 20 Type of document within each cluster

According to figure 20 that have a major category for each cluster. Therefore, it was assumed that cluster_2 was represented in Business, cluster_4 was represented in Technology, etc. As a result, creating confusion matrix, was showed in table below, has accuracy at 87%. The major incorrect cluster is Cluster_4 that has the number of Technology at 57 documents should be clustered in Cluster_1.

	True					Class Precision
	Business	**Entertainment**	**Politics**	**Sport**	**Technology**	
Pred Business (C_2)	163	7	2	0	4	92.61%
Pred Entertainment (C_1)	0	104	0	0	8	92.86%
Pred Politics (C_0)	12	10	174	2	1	87.44%
Pred Sport (C_3)	1	2	0	178	3	96.74%
Pred Technology (C_4)	4	57	4	0	164	71.62%
Class recall	90.56%	57.78%	96.67%	98.89%	91.11%	
Accuracy: 87.00 %						

3.3 Tuning model

In comparison with the Davies Bouldin's number in this table below, the less Davies Bouldin's number than -0.009 or k = 5 are k = 2 and 4. Therefore, creating K-Means with k = 2 and 4 to see the clustering image in figure 21 and figure 22.

K	Davies Bouldin's index
2	-0.012
3	-0.009
4	-0.010
5	-0.009
10	-0.008

For figure 21, Clustering using K-Means with k = 2

For figure 22, Clustering using K-Means with k = 4

As one can see in figure 21, Business, Politics and Technology were integrated at the first group. On the other hand, Entertainment and Sport are at the second group. It might have probability that key words were used in the business, politics and technology documents are more similar than the key words in the second group. In figure 22, the cluster_3 was contained by Business and Technology documents that have a slightly different amount of both categories and they are a major proportion for this cluster.

Overall, the appropriate number of clusters is k = 5 which obviously see the pattern of data point in the scatter plot. And, in fact the original dataset was come from BBC news online in 5 categories. It might be that each categories had the unique words that can represent a cluster of its own.

Conclusion

In conclusion, BBC news categories were used as a dataset and have been performed with document classification and clustering. the pre-processing method include Transform cases (lowercase), Tokenize, Filter stop words (English), Stem (Porter) and Filter Tokens (by Length) that filter the Minimum and Maximum between 4 and 50 characters.

For classification part, the dataset was transformed by using default and apply to Naïve bayes and K-NN model. The best accurate performance after tuning is K-NN model which prune by percentual for TF-IDF is lower than 5% and higher than 45% and adjust k number from 5 to 10. As a result, the accuracy is 93.67%. Furthermore, the decision tree model was used in this test. After tuning the parameter of the decision tree model, the accuracy was dramatically improved from 45.78% to 91.67%.

The clustering part use the pre-processing method and pruning parameters as same as the classification Process. It was clearly showed the clustering scatter plot, which was segregated into 5 categories that related to the original data. However, the clustering has a good result to cluster categories but it is a bit difficulty to modify model. The result was obviously showed incorrect categories in cluster_4 in figure 19.

As a result, classification model has a greater result at 93.67% and more flexible to adjust the model in comparison with clustering model at 87%.