uOttawa

Faculté de génie
Faculty of Engineering

DTI5125: Data Science Applications

Assignment group 2 (Clustering)

Group 8

Table of Contents

Table of Figures:

Loading books:

Using NTLK, five books were chosen from different categories.

Author Name	Genre
whitman-leaves	Drama
milton-paradise	Poetry
chesterton-brown	Mystery
austen-emma	Romance, Fiction
edgeworth-parents	Children, Fiction

Figure 1 Chosen books

Pre-processing, cleansing, and labelling data:

Two hundred partitions will be chosen from each book. Each partition contains 150 words. The text that will be taken from the book will be cleaned by removing titles, headlines, special characters, digits. Then, stop words will be removed. Finally, the words will be stemmed. Chosen partitions as a string and the list of words in these partitions will be put along with their labels (the book name) and genres in a data frame as shown in the following figure:

	text	list_of_words	book_name	genres
616	yesterday make bed wish might abl buy next win...	[yesterday, make, bed, wish, might, abl, buy, ...	edgeworth-parents	Children, Fiction
5	despair equal allianc address miss smith madam...	[despair, equal, allianc, address, miss, smith...	austen-emma	Romance, Fiction
134	think friend must sorri lose colonel mr campbe...	[think, friend, must, sorri, lose, colonel, mr...	austen-emma	Romance, Fiction
203	corps stretch bed live look upon palpabl live ...	[corps, stretch, bed, live, look, upon, palpab...	whitman-leaves	Drama
864	fenc grey everyth look almost spoke huge figur...	[fenc, grey, everyth, look, almost, spoke, hug...	chesterton-brown	Mystery
857	restrain astonish unlik would shown huge obvio...	[restrain, astonish, unlik, would, shown, huge...	chesterton-brown	Mystery
909	grasp bunch lili valley cri creatur passag tri...	[grasp, bunch, lili, valley, cri, creatur, pas...	chesterton-brown	Mystery
661	ornament bower fate decid excess hot mind enga...	[ornament, bower, fate, decid, excess, hot, mi...	edgeworth-parents	Children, Fiction
863	got properti man use old feudal fabl properli ...	[got, properti, man, use, old, feudal, fabl, p...	chesterton-brown	Mystery
539	good thi great question blasphem without defen...	[good, thi, great, question, blasphem, without...	milton-paradise	Poetry

Figure 2 A sample of 10 partitions from the data

Feature Engineering:

To build a model that can classify text, we need to perform a feature extraction process in which we prepare the input to be in the shape that the model can understand. In this assignment, we used four main techniques of feature extraction: BOW, TF-IDF, LDA, and Word Embedding.

We generalized the code so we could use each method many times with different algorithms without implementing the technique again.

1. BOW:

	0	1	2	3	4	5	6	7	8	9	...	11512	11513	11514	11515	11516	11517	11518	11519	11520	11521
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

Figure 3 A sample from BOW features

As shown in the above figure, the BOW features are 11521 features for each partition.

2. TF-IDF:

	0	1	2	3	4	5	6	7	8	9	...	11512	11513	11514	11515	11516	11517	11518	11519	11520	11521
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	...	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	...	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	...	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	...	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	...	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Figure 4 A sample from TF-IDF features

As shown in the above figure, the TF-IDF features are 11521 features for each partition.

3. LDA:

we have also used LDA as a feature engineering technique to put the data in the shape the algorithm could train on. Still, we had to determine the optimal number of topics so, we used coherence as a measure to determine the optimal value for topic numbers.

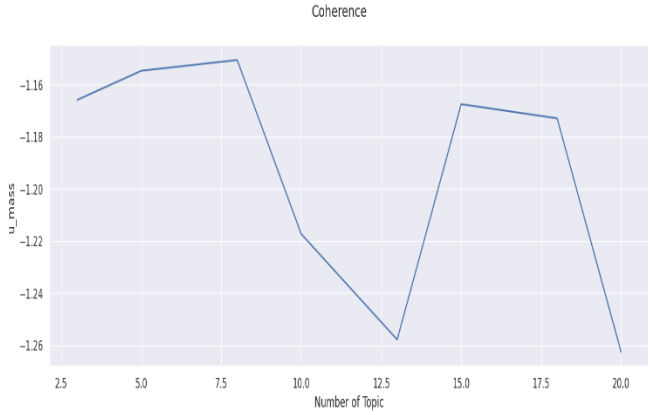


Figure 5 Coherence of LDA features using different number of topics

As the above figure shows the best value is 20 topics.

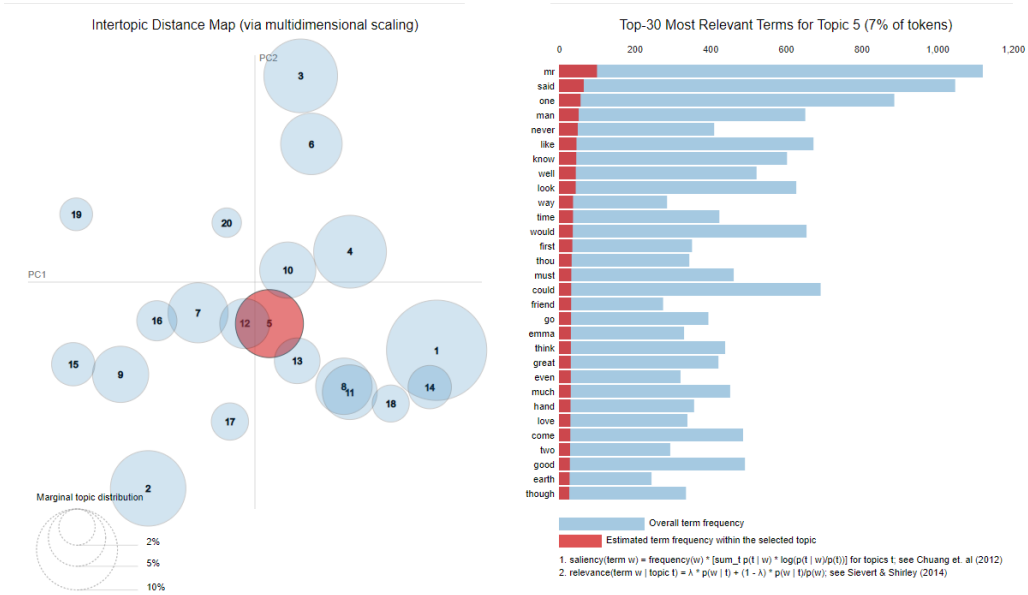


Figure 6 The frequency of top 30 words in topic 5

We visualized the results as shown in the above figure to show the top 30 words in each topic.

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19	T20
0	0.000	0.000	0.000	0.076	0.000	0.000	0.000	0.236	0.000	0.000	0.000	0.000	0.000	0.682	0.000	0.000	0.000	0.000	0.000	0.000
1	0.000	0.000	0.000	0.477	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.515	0.000	0.000	0.000	0.000	0.000	0.000
2	0.393	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.601	0.000	0.000	0.000	0.000	0.000	0.000
3	0.051	0.000	0.000	0.000	0.000	0.000	0.018	0.000	0.000	0.000	0.000	0.000	0.000	0.592	0.000	0.000	0.000	0.334	0.000	0.000
4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.562	0.000	0.000	0.000	0.000	0.000	0.430

Figure 7 A sample from LDA features

4. Word-Embedding:

	0	1	2	3	4	5	6	7	8	9	...	290	291	292	293	294	295	296	297	298	299
0	-0.022	0.067	0.020	-0.038	-0.034	0.049	0.006	-0.025	0.049	0.041	...	-0.034	-0.006	-0.010	-0.002	0.008	0.033	0.030	-0.003	0.008	0.034
1	-0.025	0.076	0.025	-0.046	-0.032	0.047	0.001	-0.026	0.055	0.053	...	-0.040	-0.011	-0.016	0.007	0.009	0.023	0.020	-0.004	-0.002	0.034
2	-0.025	0.075	0.024	-0.047	-0.040	0.044	0.005	-0.026	0.046	0.040	...	-0.041	-0.008	-0.016	0.009	0.010	0.035	0.023	-0.006	0.008	0.036
3	-0.027	0.083	0.019	-0.050	-0.038	0.040	-0.012	-0.033	0.057	0.052	...	-0.047	-0.010	-0.023	0.006	0.020	0.033	0.013	-0.008	-0.006	0.041
4	-0.026	0.069	0.011	-0.036	-0.044	0.045	0.001	-0.027	0.048	0.027	...	-0.041	-0.002	-0.005	0.001	0.007	0.041	0.027	-0.006	0.002	0.040

Figure 8 A sample from word embedding features

We tokenized each partition into 150 words, we used the Gensim library to train our word2vec model on our tokens in all partitions then we convert all our tokens into vectors in each partition, each word is represented with 300 values vectors like the sample in the above figure. Then we used the mean vector Aggregation strategy for all these tokens vectors in each partition, finally, we got 1000 vectors representing our partitions ready to be used in any model in the clustering process.

Models:

In this step, three models (K-means, EM, and Hierarchical) will be used. Each feature engineering method will be used with each model to produce a total of 12 models. The number of clusters in each model is very important hyperparameter to be determined so various metrics will be used to choose it.

1. K-means:

WCSS and Silhouette will be used to determine which k will be better for each feature engineering method.

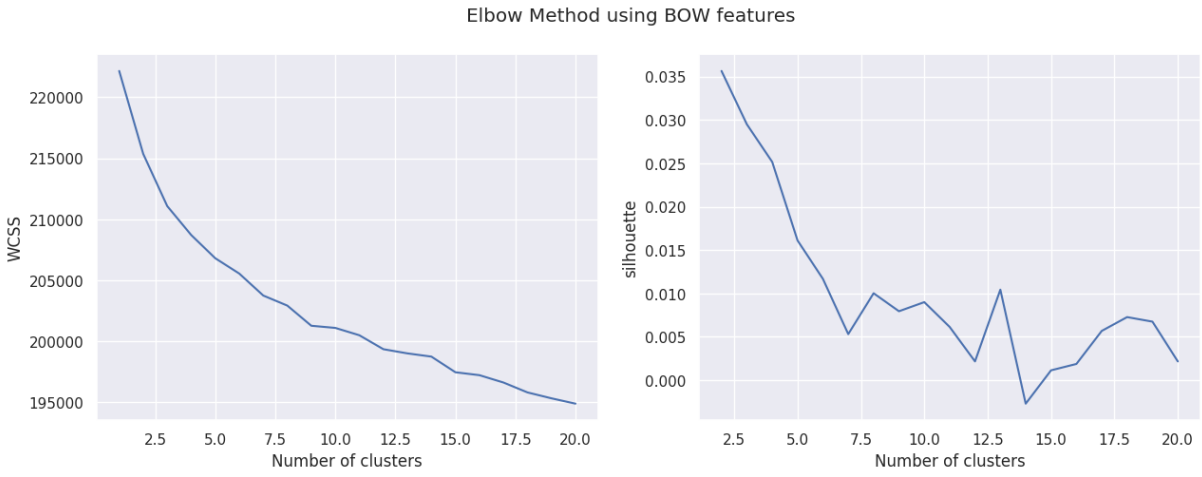


Figure 9 Elbow method using BOW features

From the above figure, the chosen **k will be 13**

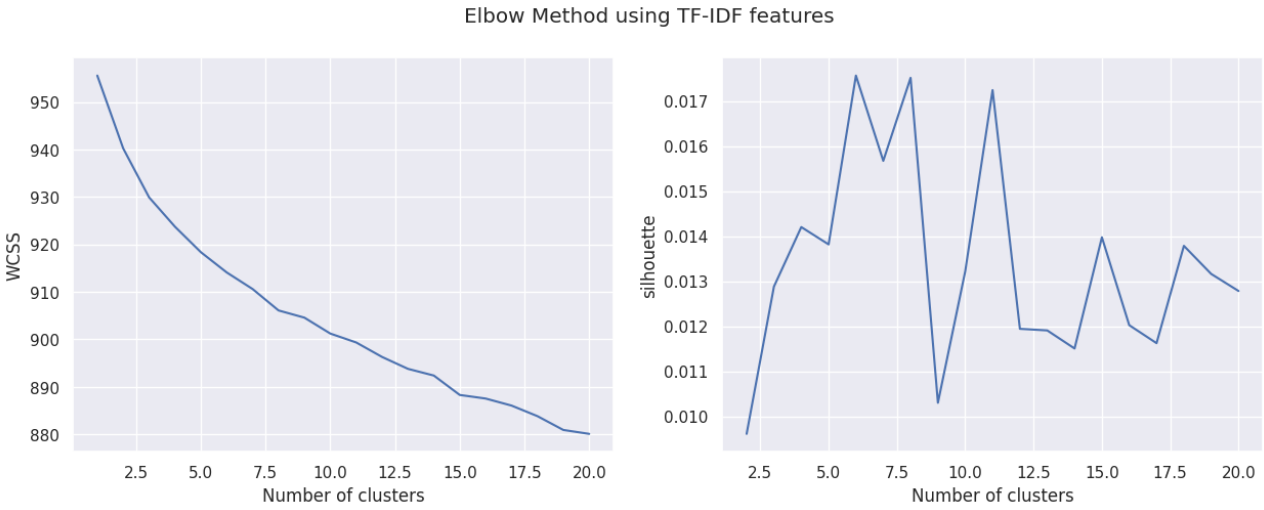


Figure 10 Elbow method using TF-IDF features

From the above figure, the chosen **k will be 6**

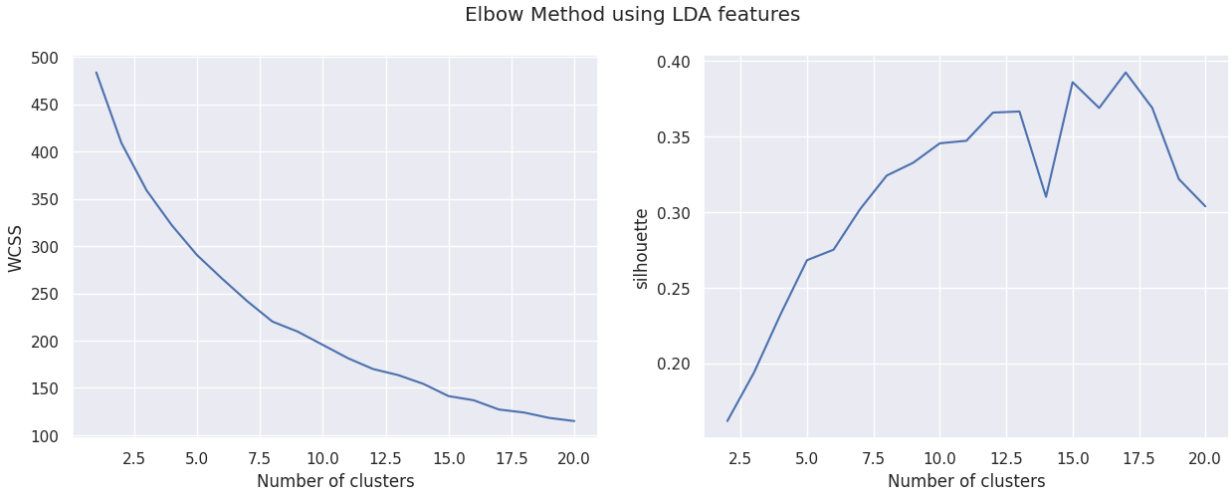


Figure 11 Elbow method using LDA features

From the above figure, the chosen **k will be 15**

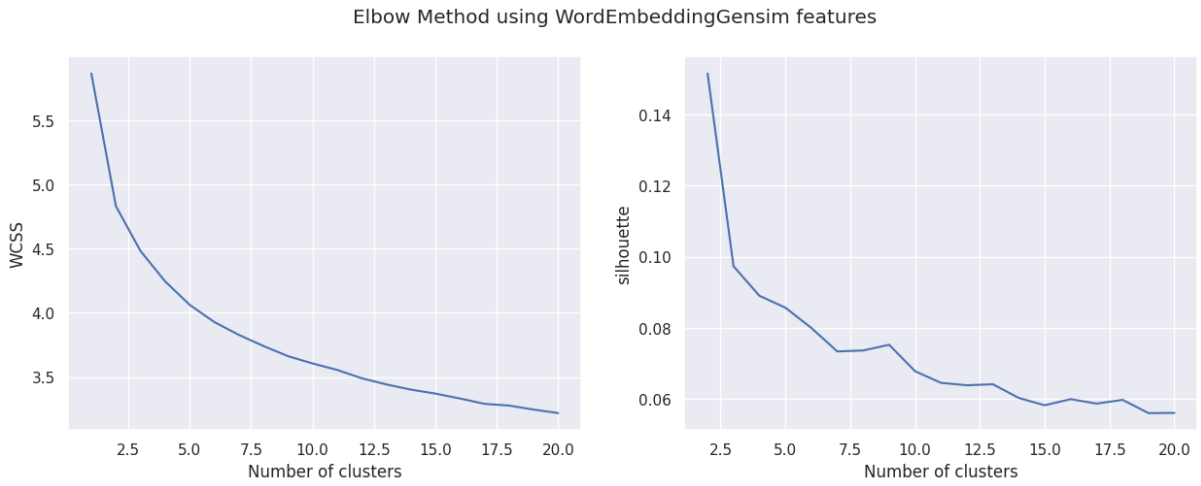


Figure 12 Elbow method using Word Embedding features

From the above figure, the chosen **k will be 5**

2. EM:

AIC, BIC, and Silhouette will be used to determine which k will be better for each feature engineering method.

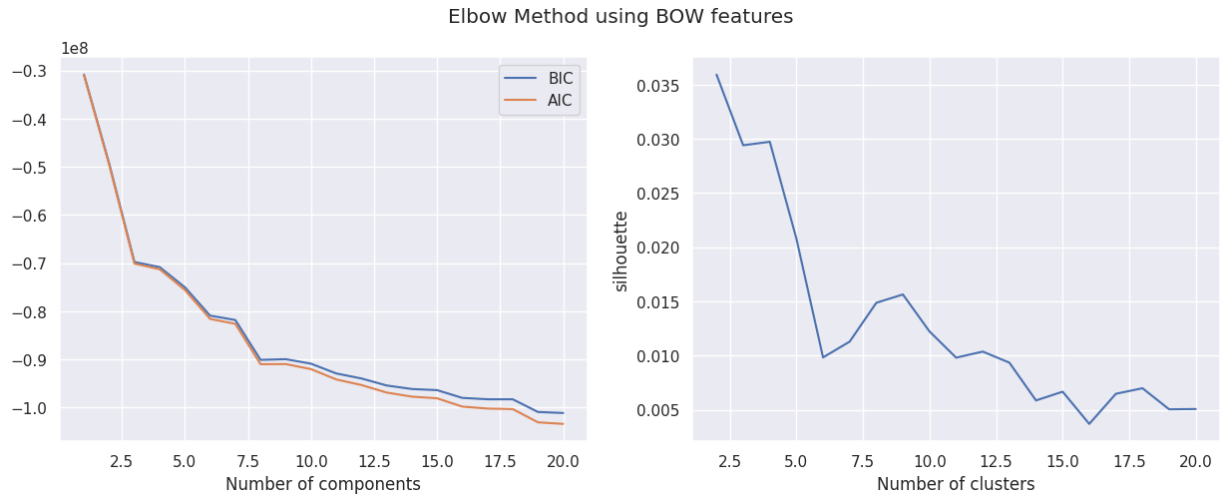


Figure 13 Elbow method using BOW features

From the above figure, the chosen **k will be 8**

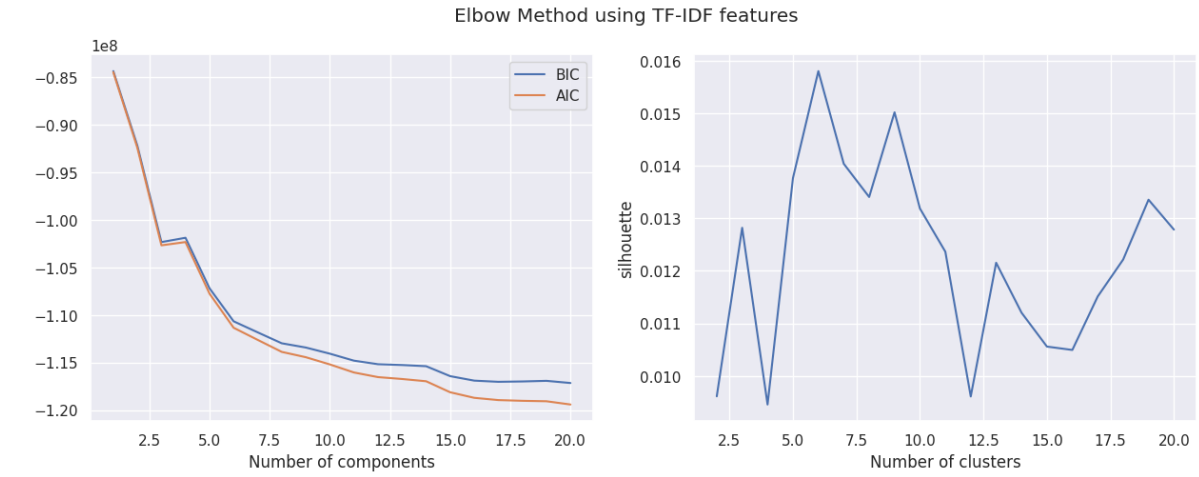


Figure 14 Elbow method using TF-IDF features

From the above figure, the chosen **k will be 6**

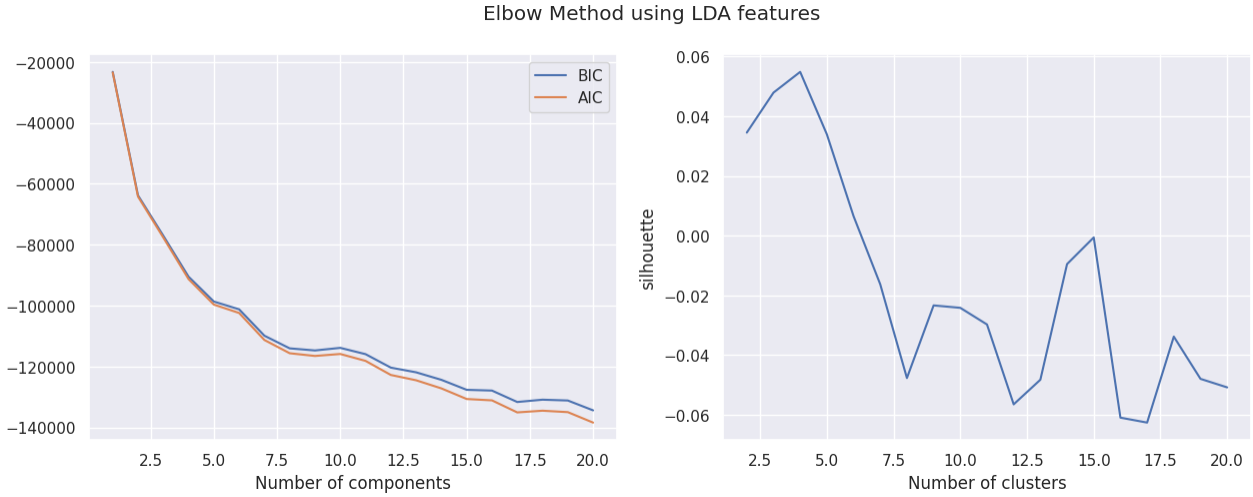


Figure 15 Elbow method using LDA features

From the above figure, the chosen **k will be 5**

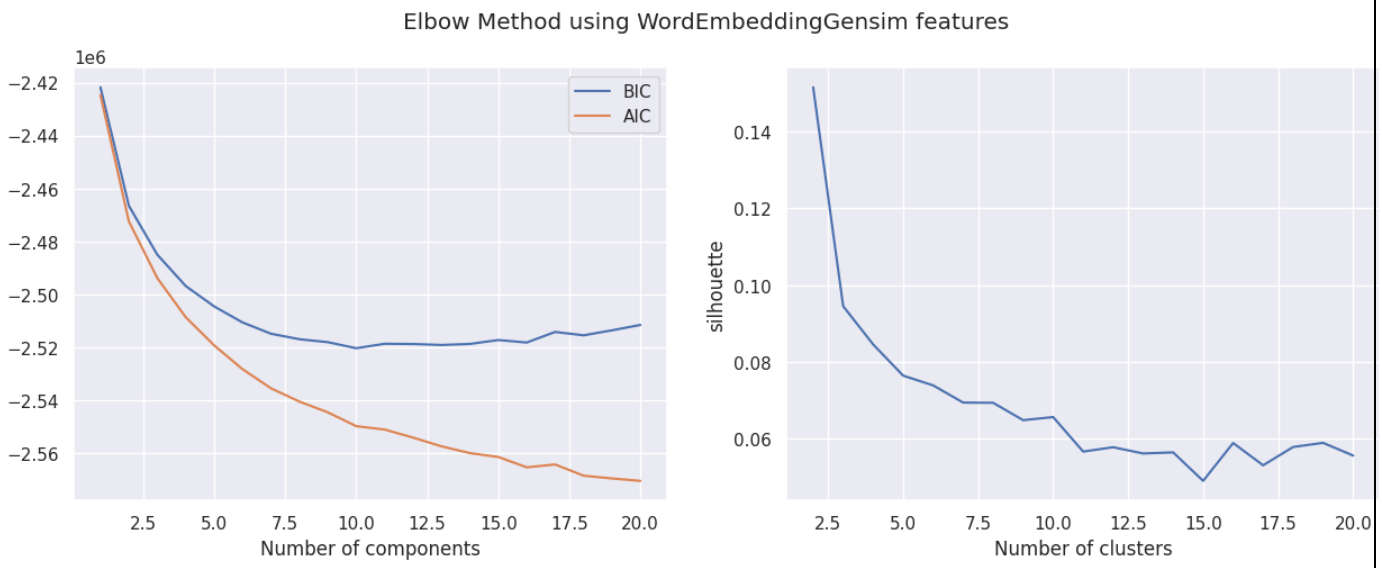
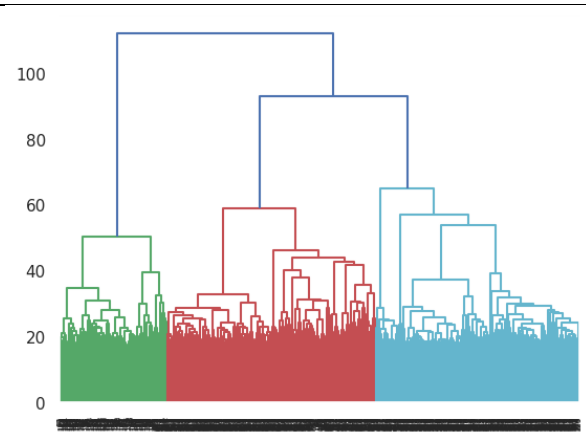
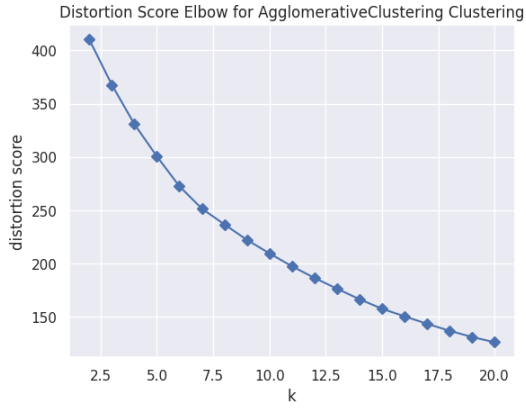
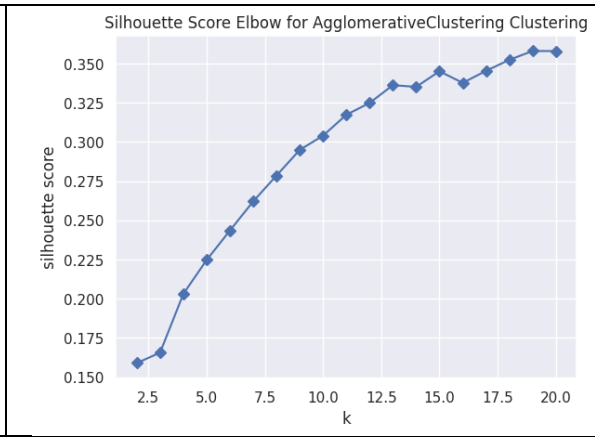


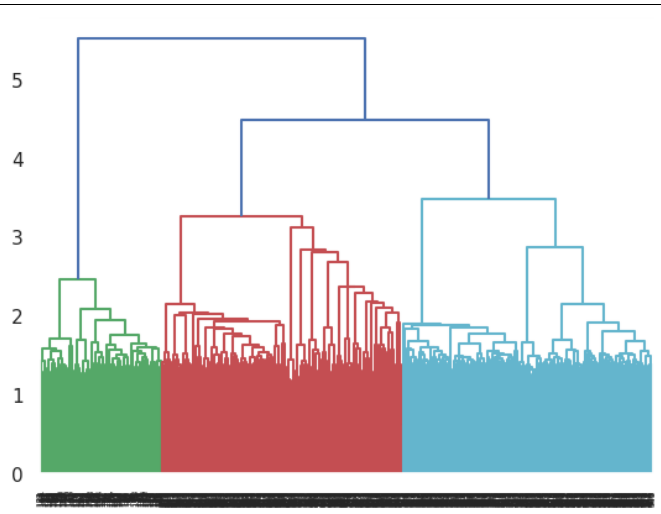
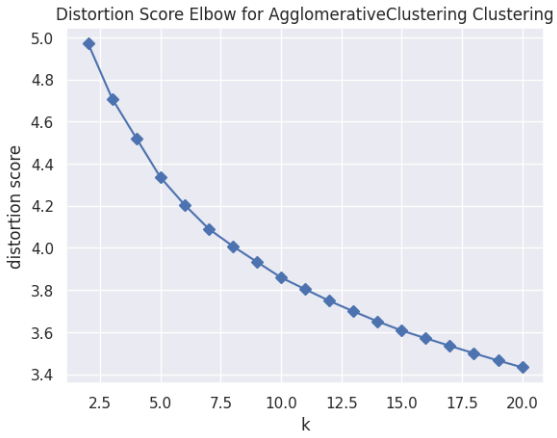
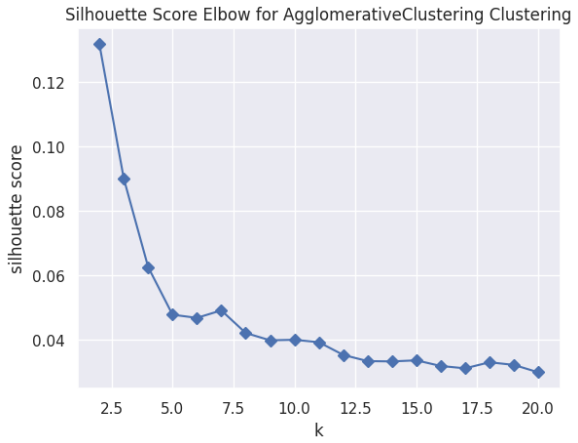
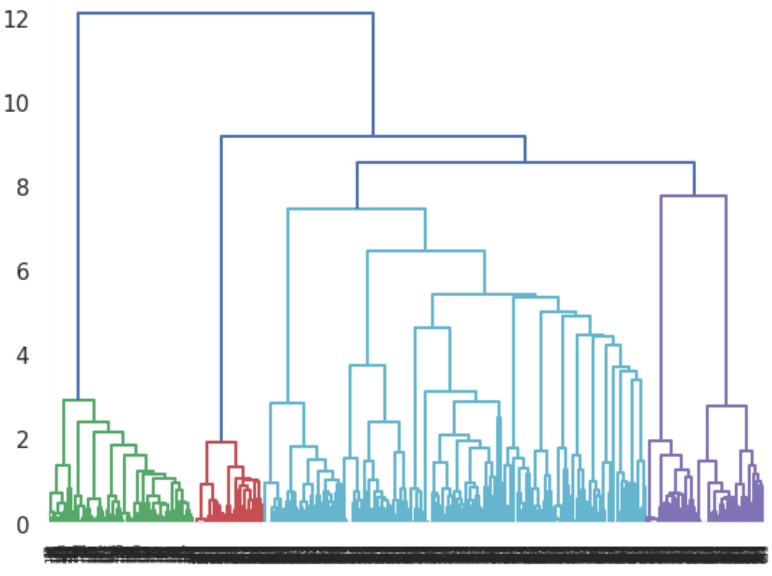
Figure 16 Elbow method using Word Embedding features

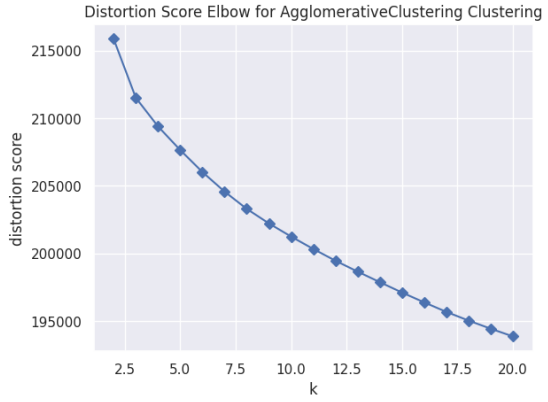
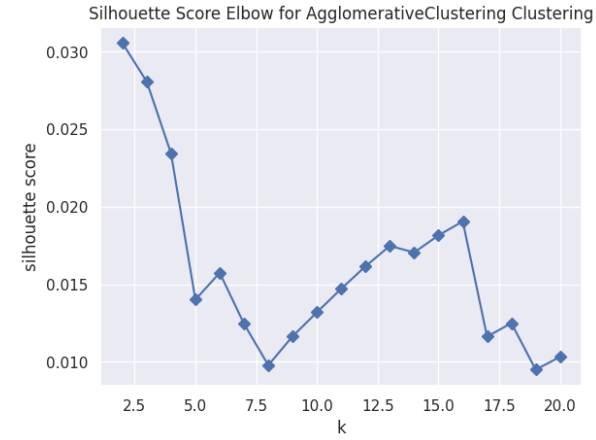
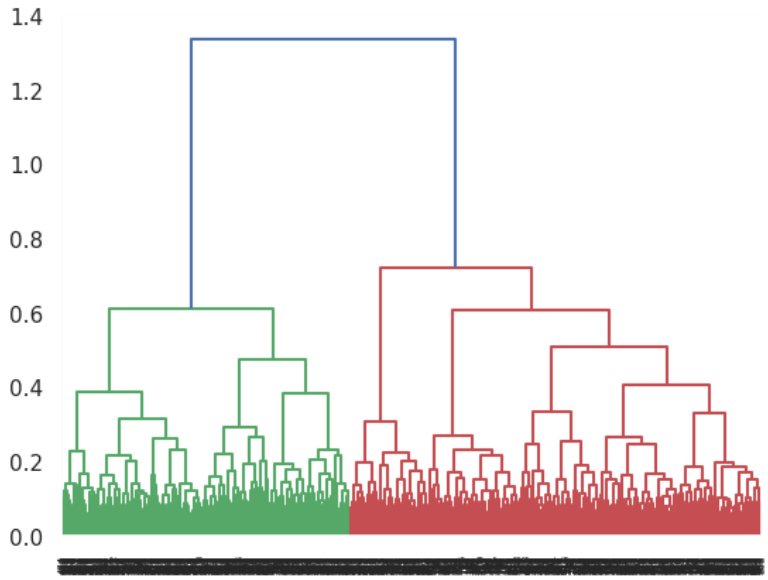
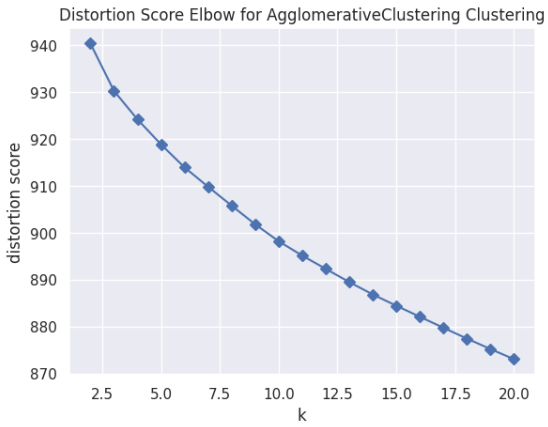
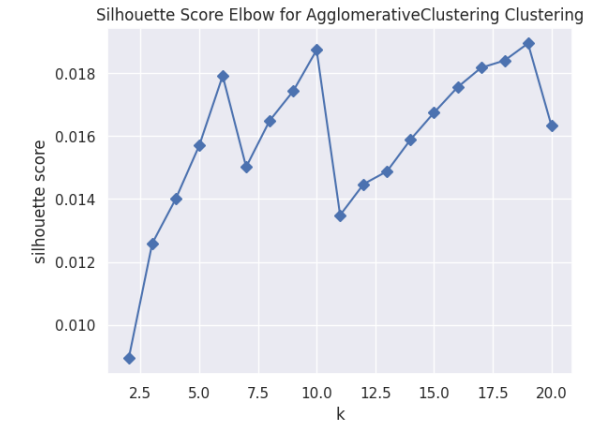
From the above figure, the chosen **k will be 5**

3. Hierarchical model:

Distortion score and Silhouette will be used to determine which k will be better for each feature engineering method. And Dendrogram will be plotted for each feature engineering method.

Dendrogram using BOW features	
Elbow method using BOW features (Chosen **k=15**)	 

Dendrogram using TF-iDF features	
Elbow method using Word Embedding features (Chosen **k=7**)	 
Dendrogram using LDA features	

Elbow method using LDA features (Chosen **k=6**)		
Dendrogram using Word Embedding features		
Elbow method using Word Embedding features (Chosen **k=10**)		

Evaluation and compare and decide which clustering result is the closest to the human labels:

After deciding how many clusters will be good for each model, Silhouette score and V-measure score will be used to compare between those 12 models without knowing any information about the true labels.

To compare between the human labels, first we should map the cluster numbers to labels and this will happen by examining each cluster and assigning a label to this cluster based on the majority of partitions labels in this cluster. Then, Cohen's Kappa will be used to compare between the predicted labels and true labels.

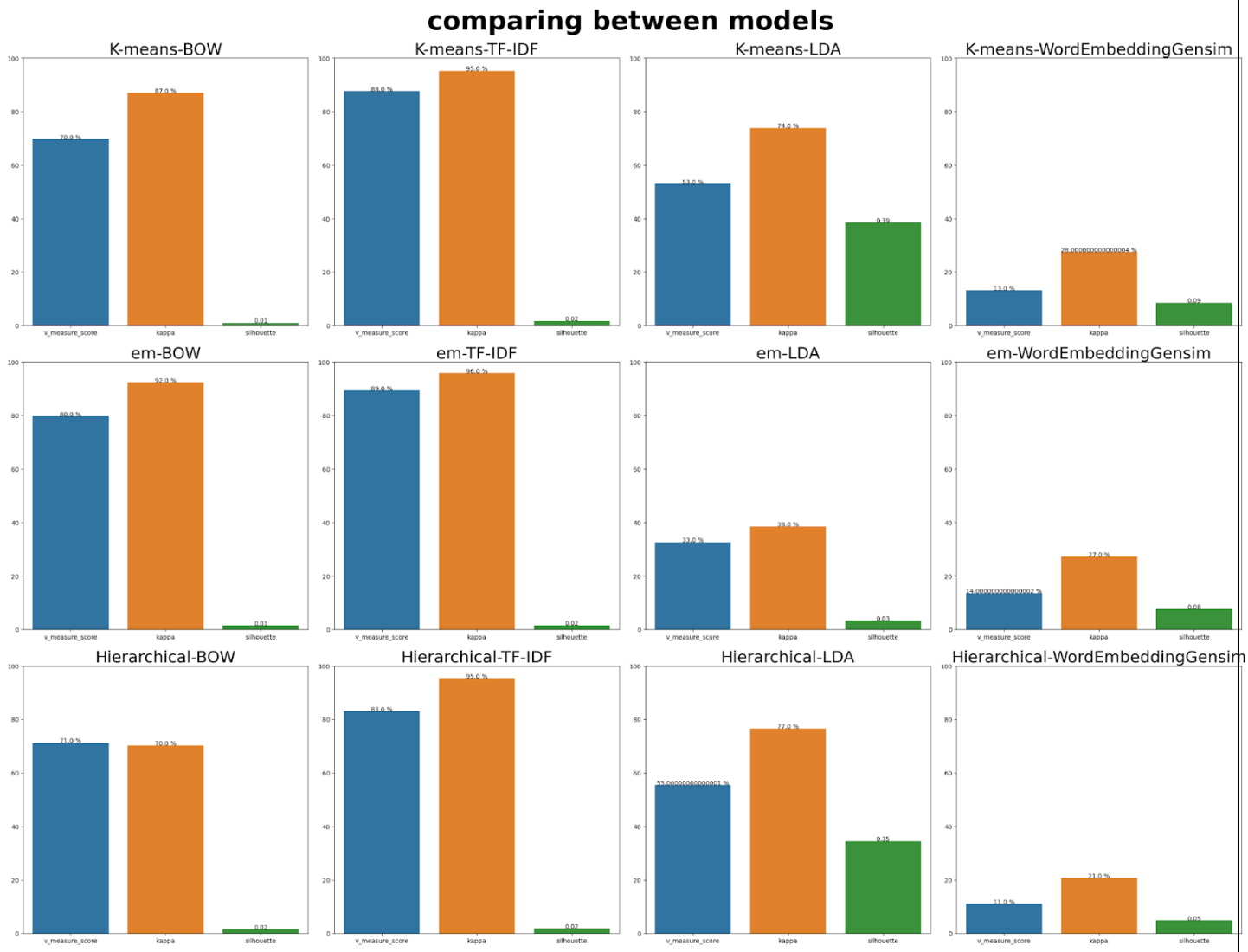


Figure 17 The comparison between models using three different metrics

Choosing the champion model:

From the above figure, **EM with TF-IDF** model gave the closest result to the human labels as the Cohen's Kappa gave **96%** so it will be chosen as our champion model. If Silhouette score is used to decide the best clustering model, then K-means with LDA model will be chosen but it doesn't mean that those clusters are corresponding to the true labels as the Silhouette score measures how well clusters are separated without knowing any information about the true labels. The following figures shows the T-SNE plot of the champion model clusters before and after mapping the clusters to labels.

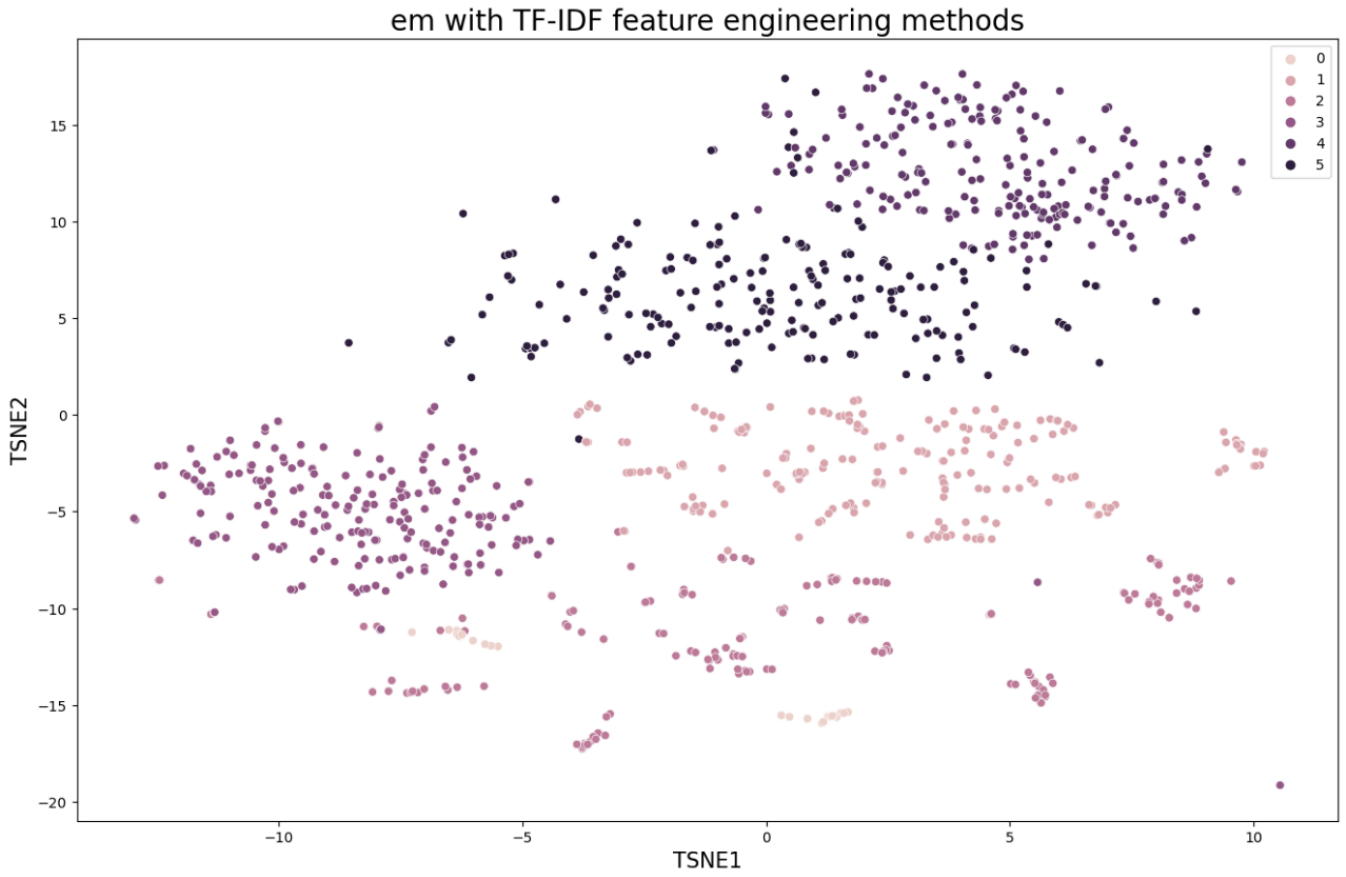


Figure 18 The T-SNE plot of the champion model clusters

Figure 19 The T-SNE plot of the champion model clusters after mapping them to labels

Error Analysis:

Based on the champion model, we compared our predicted target labels with the true target labels by using the confusion matrix, then we found 33 misclassified partitions.

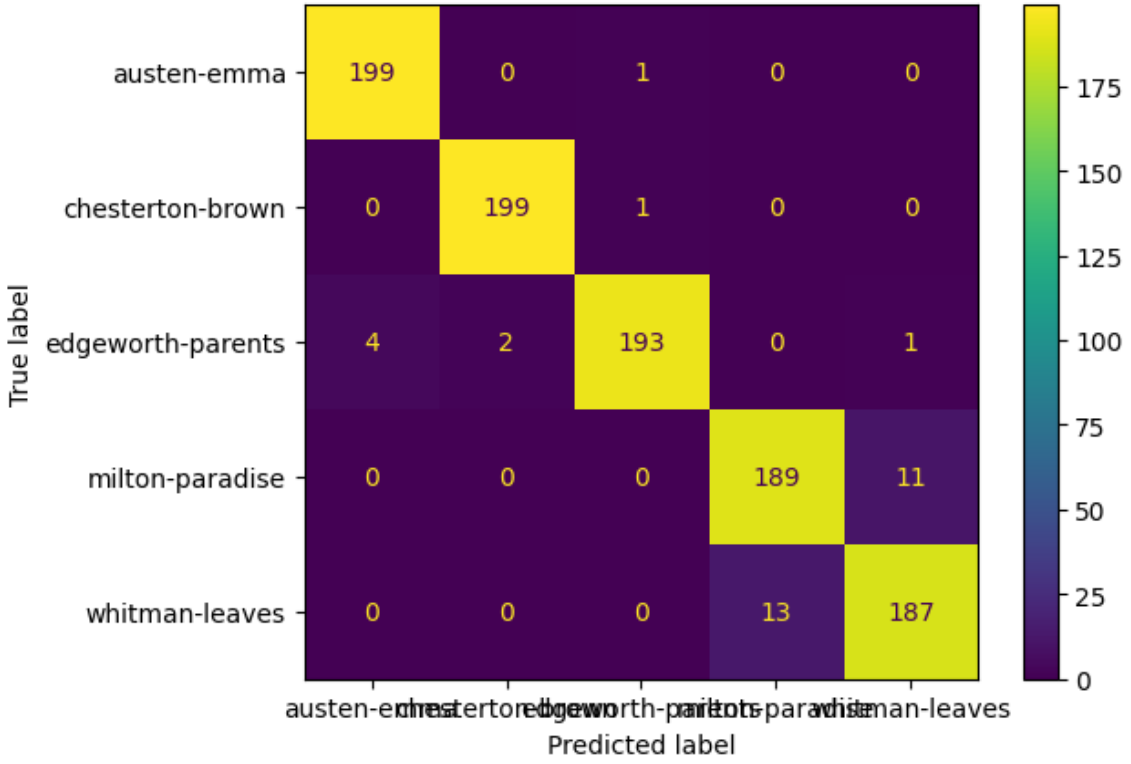


Figure 20 Confusion matrix on the training data

In the following figure, a sample of our miss-classified partitions

	Partitions	edgeworth-parents	milton-paradise	whitman-leaves	chesterton-brown	austen-emma
0	[let, keep, larger, quantiti, usual, year, sen...	--X--	NaN	NaN	NaN	NaN
1	[drawn, thee, fair, strong, good, capabl, hous...	NaN	--X--	NaN	NaN	NaN
2	[stream, tear, sob, tear, throe, choke, wild, ...	NaN	--X--	NaN	NaN	NaN
3	[space, death, like, water, flow, bear, inde, ...	NaN	--X--	NaN	NaN	NaN
4	[shalt, face, thi, fortun, thi, diseas, surmou...	NaN	--X--	NaN	NaN	NaN
5	[mighti, live, present, yet, thou, live, prese...	NaN	--X--	NaN	NaN	NaN

Figure 21 Samples of the misclassified examples

We analysed the misclassified partitions to find the partitions that our model cannot predict their labels, and we said "why?", so we thought that may be because:

1. The wrong predicted book partitions have general words and are not strongly related to that book.
2. The wrong predicted book partitions have similarities in words with the actual book class.

We plot the top 10 frequent words in all our misclassified partitions, and we plot their word cloud.

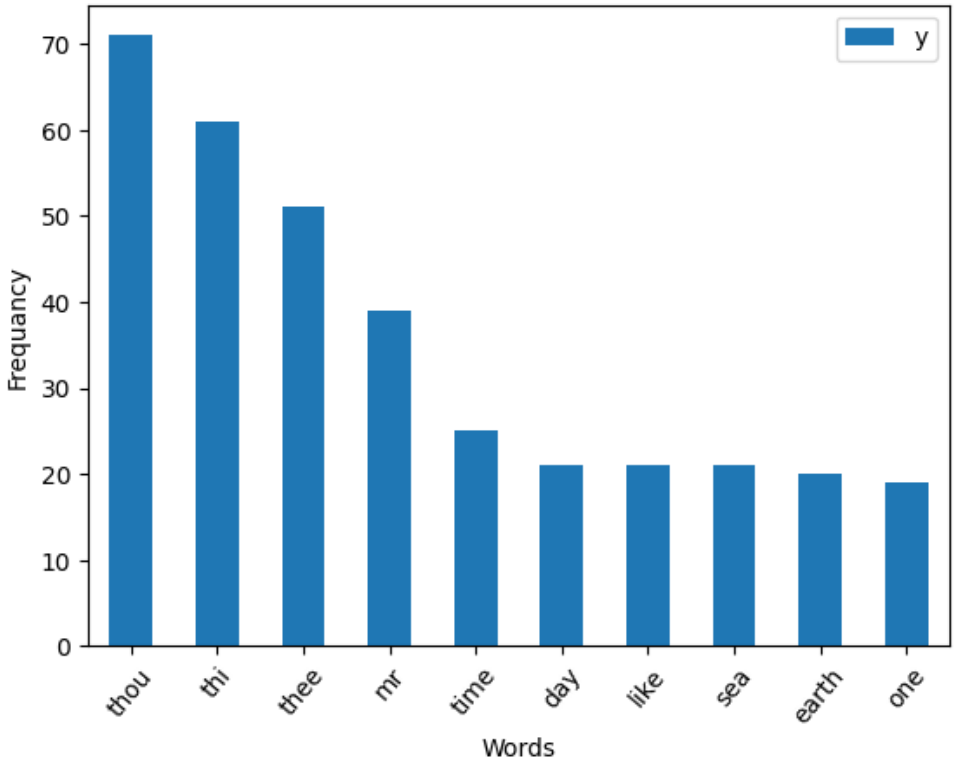


Figure 22 The top 10 words in misclassified partitions

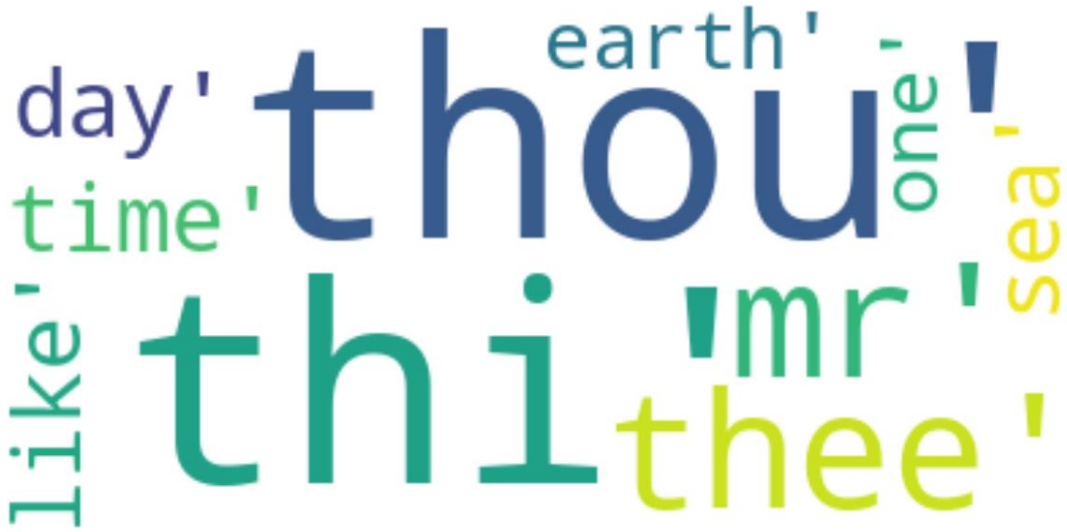


Figure 23 The Word cloud of the top 10 words in misclassified partitions

Regarding the strange words that appear in the previous figures, it's because of the stemming that is used in cleaning process of the data.