Note: The comments in the R scripts submitted explain why each step of analysis was taken.

TASK 1

1.1 Introduction

It is no secret that natural disasters are a recurring concern for societies and governments worldwide due to the devastating aftermath and consequences that they bring. According to the Centre for Research on the Epidemiology of Disasters (2023), natural hazards and disasters have caused 30,704 deaths and affected 185 million people worldwide in 2022 alone. Not only do these natural disasters ruin lives, but they also disrupt the economies of countries, accumulating to 223.8 billion US dollars of economic losses. We can see how conducting studies on natural disasters can be a main concern for many. In this modern technological era with the abundance of information overload, online news articles are a great source to gain information regarding natural disasters. Extracting information from big amounts of unstructured text data regarding natural disasters however, can be a daunting task. Therefore, in this short analysis, Latent Dirichlet Allocation (LDA) topic modelling is applied on 35 news articles regarding natural disasters to get a rough idea of the categories of text data available. This will then provide a more comprehensive understanding of the prevailing themes surrounding natural disasters that are portrayed by the media. LDA can also help in organising the data, where the assignment of topics to news articles is good for creating structured disaster databases in which information retrieval can be made easy.

As such, 35 news articles on natural disasters have been extracted from three news sources, namely New Straits Times, Reuters and The Star, each with 12, 11 and 12 news articles respectively. All the news articles are quite recent, that is from the year 2021 onwards and thus emphasises on disasters that have taken place not too long ago. News articles on many kinds of disasters are extracted with these disasters occurring in many different parts of world such as Malaysia, China, Australia, the United States of America, Mexico, Japan and other countries too. All the text files are named based on the type of natural disaster and area where it took place. The types of disasters analysed are explained in Table 1 below. Different aspects of topic modelling and text analysis are explained in the following subsections.

Table 1 Types of natural disasters, their source of hazard and description

Source of Hazard	Natural Disaster	Description
Geological hazard	Avalanche	Torrential descent of snow down a hill or mountain.
	Landslide	A number of mass wasting processes that could involve a variety of ground motions, including rockfalls, mudflows, debris flows and shallow or deep-seated slope failures.
	Earthquake	Seismic waves that are the result of a rapid release of energy in the Earth's crust. At the surface of the Earth, earthquakes show themselves by making the ground shake, vibrate and sometimes move.
	Volcanic eruption	The force from a volcanic eruption and falling rocks can cause danger. In general, after cooling, volcanic ash may form a cloud and settle densely nearby, which is harmful. Volcanic activity can also cause tsunamis.
Water hazard	Floods	Land is submerged by an excess of water.
	Tsunami	A sequence of waves in a body of water brought on by the movement of a significant amount of water, as in an ocean or a sizable lake.
Extreme weather hazard (hot and dry conditions)	Drought	Unusual soil dryness brought on by continuous periods of precipitation that are much below average.
	Heatwave	An extended period of abnormally hot weather.
	Wildfire	Massive flames that frequently begin in wilderness areas that can be caused by lightning or drought.
Extreme weather hazard (cold weather events)	Blizzard	Harsh winter storms with high winds and lots of snow.
Extreme weather hazard (strong winds)	Storm	A severe atmospheric disturbance that typically includes heavy winds, rain, thunder and lightning.
	Thunderstorm	A storm that produces thunder, lightning and typically heavy rain.

	Cyclone/Typhoon	Wind blowing circularly in the direction of a region with low air pressure.

1.2 Per-Topic-Per-Word Probabilities

It is important to understand that every topic is a mixture of words, but in topic modelling, hard clustering is not done. Words could overlap in a few different topics, being shared and not solely belonging to a single topic. In this section, a four-topic LDA model is created followed by the extraction of per-topic-per-word probabilities, also known as beta. The most common terms within each of the four topics are shown in Figure 1 below.

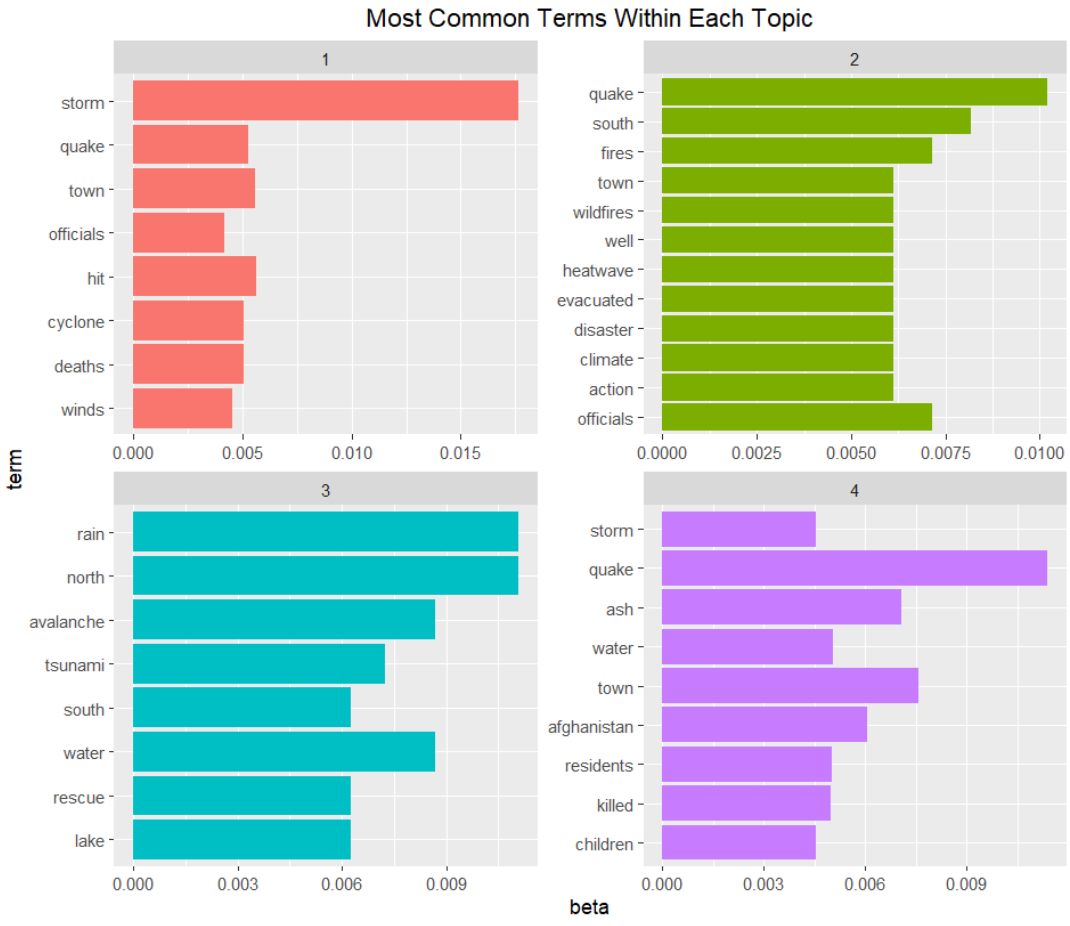

Figure 0 Most common terms within Topic 1, Topic 2, Topic 3 and Topic 4

The x-axis represents the beta, that is the probability for the particular words to be generated from their respective topics, indicating their relative importance in the topic. Notably,

certain words such as “storm”, “quake, “town”, “officials”, “south”, and “water” are common in more than one topic. However, it is still possible for us to observe which topic they are most likely to be generated from. The words connected to each topic can be used to determine its meaning. The most common words in Topic 1 include “storm”, “cyclone” and “winds”, indicating natural disasters caused by extreme weather hazards, namely strong winds. The word “storm” especially, has a very high probability of 0.018 as compared to the other words, portraying its significance in Topic 1. For Topic 2, the words “fires”, “wildfires” and “heatwave” signify a strong association to natural disasters caused by hot and dry conditions. Words such as “action” and “officials” indicate that there is an effort by authorities to deal with the situations in this topic. Next, we look at Topic 3 suggesting natural disasters related to water hazards from the words “rain”, “tsunami”, “water” and “lake”. Essentially, avalanches (“avalanche”) can also be caused by water hazards in that the bonding at layer boundaries can be greatly weakened with the presence of liquid water in the snowpack (WSL Institute for Snow and Avalanche Research SLF n.d.). Lastly, the final topic represents words such as “quake” and “ash” that convey a specific type of natural disaster caused by geological hazards, that is an earthquake. The word “killed” is also commonly associated with earthquakes since many people end up dead because of this natural disaster. Earthquakes usually involve many people since it covers wide areas, as is seen through the words “town”, “residents” and “children”. All in all, Table 2 below shows human-added topic labels by self-judgement for each of the four topics.

Table 2 Four topics determined by LDA model and their human-added labels

Topic	Human-added Topic Label
1	Natural disasters due to strong winds
2	Natural disasters due to hot and dry conditions
3	Natural disasters due to water hazards
4	Earthquakes

1.3 Beta Spread

Beta spread is the dispersion of beta values across a pair of topics, which reveal how differently words are distributed throughout the topics. They represent how much similarity or distinction there is between the words connected to each topic. Since beta spreads are graphed between two

topics, the beta spread between Topic 2 and Topic 4 is chosen for this subsection out of all six pairs of beta spreads. These two topics are chosen because of the interesting and clearly distinctive words that appear for each topic, as can be seen in Figure 2 below. The log ratio values represent how much the beta for Topic 4 prevails over the beta for topic 2. A higher log ratio magnitude would then denote a stronger correlation between the term and the corresponding topic.

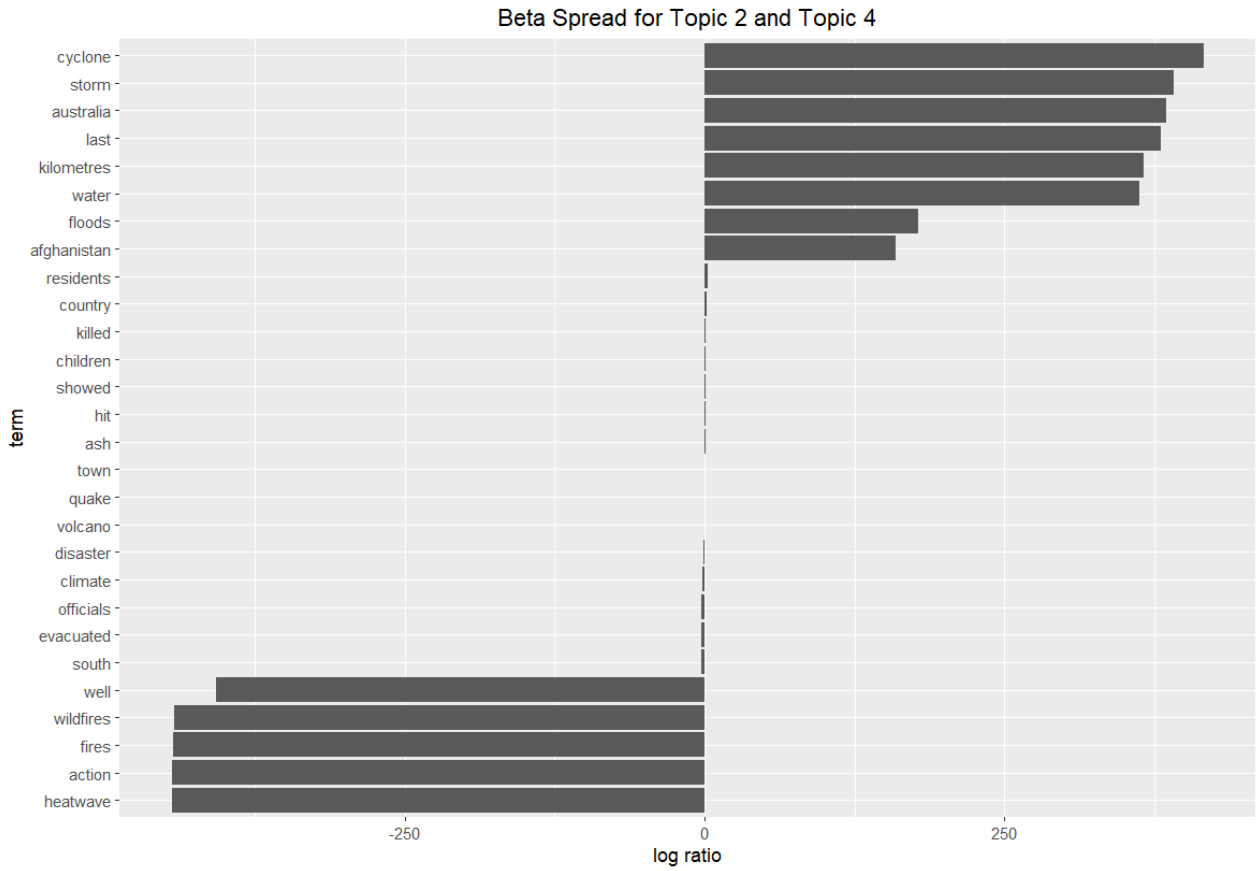

Figure 2 The beta spread for Topic 2 and Topic 4

From Figure 2, we see that even the most relevant words (beta greater than 0.006 for Topic 2 and beta greater than 0.004 for Topic 4) comprise of only a few terms have a really great difference in beta between Topic 2 and Topic 4. The most significant difference can be seen between the terms “cyclone” and “heatwave”, with log ratio values of 417 and -445 respectively. This shows that the two words are clearly separated for Topic 4 and Topic 2, meaning that “cyclone” is the most unlikely to appear in news articles in Topic 2 whereas “heatwave” is the most unlikely to appear in Topic 4. Words like “quake” and “volcano” on the other hand are not

clearly separated. Even though they have slightly higher probabilities in Topic 4 and Topic 2 respectively, they could easily appear in news articles from the other topic too, as is shown in Figure 1 where “quake” has high beta values in both topics.

The words more common in Topic 2 include natural disasters caused by hot and dry conditions such as “heatwave”, “fires” and “wildfires”. This confirms the human-added topic label from the earlier subsection for Topic 2. However, topic 4 is more characterised by words such as “cyclone”, “storm”, “water” and “floods”, meaning that these natural disasters caused by strong winds and water hazards also seem to appear in Topic 4. Nevertheless, words like “residents”, “killed”, “children”, “ash”, “town” and “quake” do still make an appearance in topic 4. Worth noting, the word “afghanistan” also has a high log ratio, since there were earthquakes and floods taking place over there. This confirms that our human-added topic labels from before are on the right track, even if there is some overlapping. Only a few words are clearly associated to the particular topics.

1.4 Per-Document-Per-Topic Probabilities

Besides beta probabilities, there are also gamma probabilities that are the per-document-per-topic probabilities. It is important to understand that every news article or document is a mixture of four topics in our case. Each news article may have specific proportions of words from the four topics. In Figure 3 below, each gamma value shows the estimated proportion of words from the particular news article that is generated from that topic.

	document	topic	gamma
	<chr>	<int>	<dbl>
1	AvalancheCopenhagen.txt	1	0.000210
2	AvalancheCopenhagen.txt	2	0.000210
3	AvalancheCopenhagen.txt	3	0.999
4	AvalancheCopenhagen.txt	4	0.000210
5	AvalancheGuwahati.txt	1	0.0000656
6	AvalancheGuwahati.txt	2	0.0000656
7	AvalancheGuwahati.txt	3	1.00
8	AvalancheGuwahati.txt	4	0.0000656
9	AvalancheMuzaffarabad.txt	1	0.000146
10	AvalancheMuzaffarabad.txt	2	0.000146
11	AvalancheMuzaffarabad.txt	3	1.00
12	AvalancheMuzaffarabad.txt	4	0.000146
13	AvalancheParis.txt	1	0.000263
14	AvalancheParis.txt	2	0.000263
15	AvalancheParis.txt	3	0.999
16	AvalancheParis.txt	4	0.000263
17	BlizzardNewYork.txt	1	1.00
18	BlizzardNewYork.txt	2	0.000131
19	BlizzardNewYork.txt	3	0.000131
20	BlizzardNewYork.txt	4	0.000131
21	CycloneBlantyreMaputo.txt	1	1.00
22	CycloneBlantyreMaputo.txt	2	0.0000397
23	CycloneBlantyreMaputo.txt	3	0.0000397
24	CycloneBlantyreMaputo.txt	4	0.0000397
25	CycloneDhaka.txt	1	1.00
26	CycloneDhaka.txt	2	0.0000930
27	CycloneDhaka.txt	3	0.0000930
28	CycloneDhaka.txt	4	0.0000930
29	CycloneSydney.txt	1	0.0000767
30	CycloneSydney.txt	2	0.0000767
31	CycloneSydney.txt	3	0.0000767
32	CycloneSydney.txt	4	1.00
33	DroughtAddisAbaba.txt	1	0.000155
34	DroughtAddisAbaba.txt	2	0.000155
35	DroughtAddisAbaba.txt	3	1.00
36	DroughtAddisAbaba.txt	4	0.000155
37	DroughtBangkok.txt	1	0.000131
38	DroughtBangkok.txt	2	0.000131
39	DroughtBangkok.txt	3	1.00
40	DroughtBangkok.txt	4	0.000131
41	EarthquakeGardez.txt	1	0.0810
42	EarthquakeGardez.txt	2	0.0000948
43	EarthquakeGardez.txt	3	0.0000948
44	EarthquakeGardez.txt	4	0.919
45	EarthquakeMexico.txt	1	1.00
46	EarthquakeMexico.txt	2	0.0000962
47	EarthquakeMexico.txt	3	0.0000962
48	EarthquakeMexico.txt	4	0.0000962
49	EarthquakeTaipei.txt	1	0.0000578
50	EarthquakeTaipei.txt	2	0.0000578
51	EarthquakeTaipei.txt	3	0.0000578
52	EarthquakeTaipei.txt	4	1.00
53	EarthquakeTurkiyeSyria.txt	1	0.000112
54	EarthquakeTurkiyeSyria.txt	2	1.00
55	EarthquakeTurkiyeSyria.txt	3	0.000112
56	EarthquakeTurkiyeSyria.txt	4	0.000112
57	FloodsAfghanistan.txt	1	0.0000981
58	FloodsAfghanistan.txt	2	0.0000981
59	FloodsAfghanistan.txt	3	0.0000981
60	FloodsAfghanistan.txt	4	1.00
61	FloodsHaiti.txt	1	0.000187
62	FloodsHaiti.txt	2	0.000187
63	FloodsHaiti.txt	3	0.999
64	FloodsHaiti.txt	4	0.000187
65	FloodsNZ.txt	1	0.000125
66	FloodsNZ.txt	2	0.000125
67	FloodsNZ.txt	3	1.00
68	FloodsNZ.txt	4	0.000125
69	FloodsSeoul.txt	1	0.000129
70	FloodsSeoul.txt	2	0.000129
71	FloodsSeoul.txt	3	1.00
72	FloodsSeoul.txt	4	0.000129
73	FloodsSydney.txt	1	0.0000830
74	FloodsSydney.txt	2	0.0000830
75	FloodsSydney.txt	3	0.0000830
76	FloodsSydney.txt	4	1.00
77	HeatwaveKualaLumpur.txt	1	0.0000966
78	HeatwaveKualaLumpur.txt	2	1.00
79	HeatwaveKualaLumpur.txt	3	0.0000966
80	HeatwaveKualaLumpur.txt	4	0.0000966
81	LandslideBeijing.txt	1	0.000307
82	LandslideBeijing.txt	2	0.000307
83	LandslideBeijing.txt	3	0.999
84	LandslideBeijing.txt	4	0.000307
85	LandslideKarbala.txt	1	0.000156
86	LandslideKarbala.txt	2	0.000156
87	LandslideKarbala.txt	3	1.00
88	LandslideKarbala.txt	4	0.000156
89	LandslideMaracay.txt	1	0.000118
90	LandslideMaracay.txt	2	0.000118
91	LandslideMaracay.txt	3	0.000118
92	LandslideMaracay.txt	4	1.00
93	LandslideRome.txt	1	0.999
94	LandslideRome.txt	2	0.000215
95	LandslideRome.txt	3	0.000215
96	LandslideRome.txt	4	0.000215
97	LandslideTokyo.txt	1	0.000105
98	LandslideTokyo.txt	2	0.000105
99	LandslideTokyo.txt	3	1.00
100	LandslideTokyo.txt	4	0.000105
101	StormPortAuxBasques.txt	1	1.00
102	StormPortAuxBasques.txt	2	0.0000525
103	StormPortAuxBasques.txt	3	0.0000525
104	StormPortAuxBasques.txt	4	0.0000525
105	StormUS.txt	1	1.00
106	StormUS.txt	2	0.000101
107	StormUS.txt	3	0.000101
108	StormUS.txt	4	0.000101
109	ThunderstormSibu.txt	1	0.000106
110	ThunderstormSibu.txt	2	0.000106
111	ThunderstormSibu.txt	3	0.000106
112	ThunderstormSibu.txt	4	1.00
113	TyphoonPhilippines.txt	1	1.00
114	TyphoonPhilippines.txt	2	0.0000957
115	TyphoonPhilippines.txt	3	0.0000957
116	TyphoonPhilippines.txt	4	0.0000957
117	VolcanicTsunamiNukuAlofa.txt	1	0.0000535
118	VolcanicTsunamiNukuAlofa.txt	2	0.0000535
119	VolcanicTsunamiNukuAlofa.txt	3	1.00
120	VolcanicTsunamiNukuAlofa.txt	4	0.0000535
121	VolcanoJakarta.txt	1	0.000143
122	VolcanoJakarta.txt	2	0.000143
123	VolcanoJakarta.txt	3	0.000143
124	VolcanoJakarta.txt	4	1.00
125	VolcanoMexicoCity.txt	1	0.000114
126	VolcanoMexicoCity.txt	2	1.00
127	VolcanoMexicoCity.txt	3	0.000114
128	VolcanoMexicoCity.txt	4	0.000114
129	VolcanoPetropavlovskKamchatsky.txt	1	0.0000834
130	VolcanoPetropavlovskKamchatsky.txt	2	0.0000834
131	VolcanoPetropavlovskKamchatsky.txt	3	0.0000834
132	VolcanoPetropavlovskKamchatsky.txt	4	1.00
133	WildfireRome.txt	1	0.000157
134	WildfireRome.txt	2	1.00
135	WildfireRome.txt	3	0.000157
136	WildfireRome.txt	4	0.000157
137	WildfireSantiago.txt	1	0.0000690
138	WildfireSantiago.txt	2	1.00
139	WildfireSantiago.txt	3	0.0000690
140	WildfireSantiago.txt	4	0.0000690

Figure 3 Per-document-per-topic-probabilities

For example, the LDA model estimates that 99.9% of the words in *AvalancheCopenhagen.txt* were generated from Topic 3 whereas only 0.021% of the words were generated from each of the other three topics respectively. This practically means that *AvalancheCopenhagen.txt* was drawn almost entirely from Topic 3 since its gamma is almost zero for all the other topics. For easier observance and interpretation, a stacked bar chart is plotted in Figure 4 to show the composition of gamma probabilities in each topic for each of the 35 news articles.

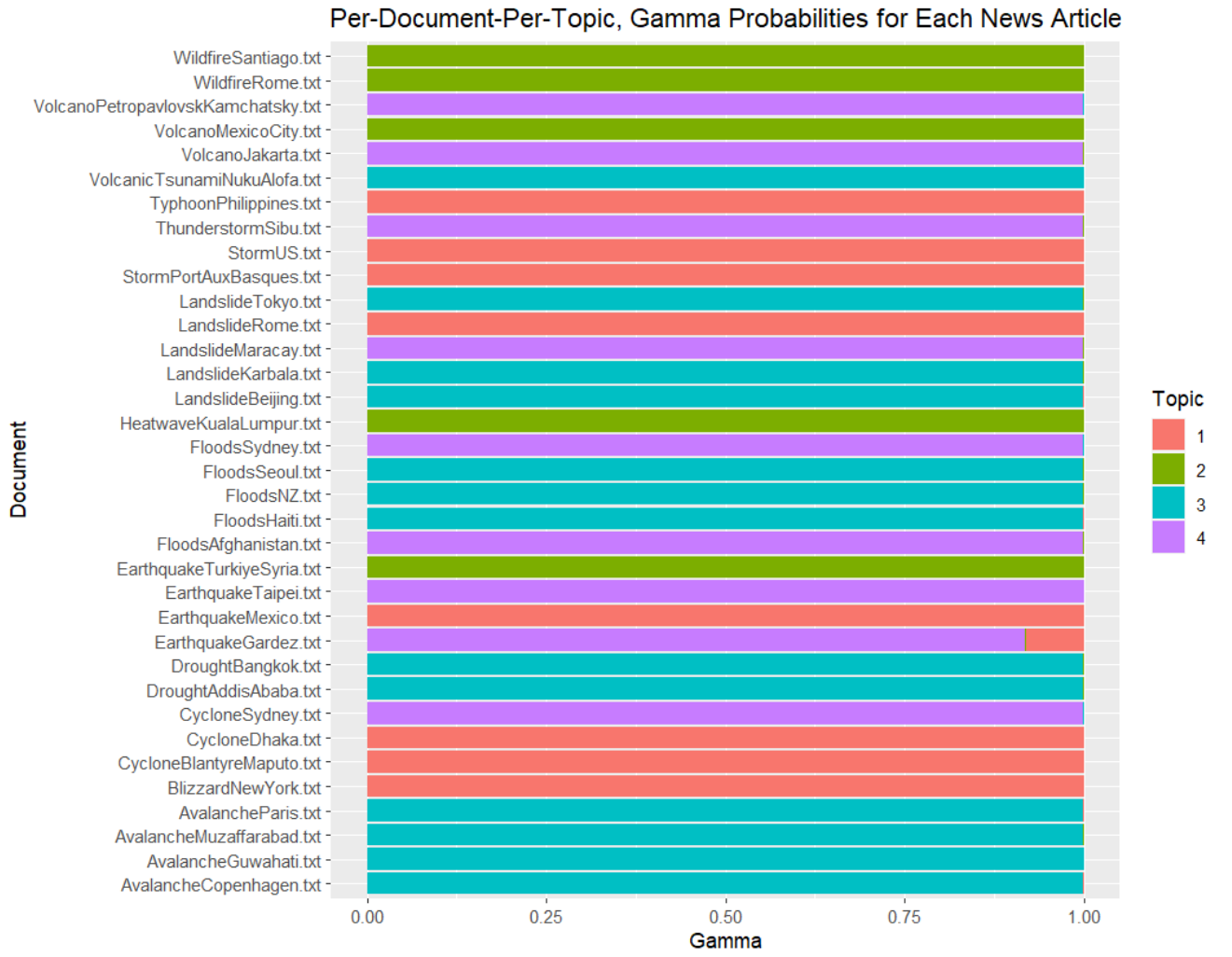

Figure 4 Per-document-per-topic, gamma probabilities for each of the 35 news articles

As can be seen from Figure 4, nearly all the news articles were drawn almost entirely from one particular topic with no mixtures. Only a very small proportion goes to other topics (cannot be

seen clearly in the stacked bar chart due to really small value). However, there is one news article, namely *EarthquakeGardez.txt* that was drawn from a mixture of all four topics, particularly 91.9% in Topic 4 and 8.1% in Topic 1. It still shows a high leaning towards Topic 4. In order to verify this grouping, the most common words in *EarthquakeGardez.txt* is shown in Figure 5.

	document	term	count
	<chr>	<chr>	<dbl>
1	EarthquakeGardez.txt	quake	9
2	EarthquakeGardez.txt	afghanistan	8
3	EarthquakeGardez.txt	struck	4
4	EarthquakeGardez.txt	country	3
5	EarthquakeGardez.txt	disaster	3
6	EarthquakeGardez.txt	killed	3
7	EarthquakeGardez.txt	takeover	3
8	EarthquakeGardez.txt	taliban	3
9	EarthquakeGardez.txt	affected	2
10	EarthquakeGardez.txt	afp	2

… with 152 more rows

Figure 5 Most frequent terms appearing in *EarthquakeGardez.txt*

According to the most frequent terms appearing in the news article, it seems to be about an earthquake that has taken place in Afghanistan, to the point that people were killed. This is similar to what we have labelled Topic 4 from before, indicating that the LDA model did a good job in placing this news article in the right topic as an earthquake related news article.

1.5 Word Cloud and Word Association

Other than looking at the news articles one by one to get the topics and probabilities, we can also observe all 35 news articles as a whole by generating a word cloud. This will help us to grasp the main themes, topics and issues discussed by the media regarding natural disasters in just a glance. The most prevalent natural disasters portrayed by the media can also be determined through the identification of words that appear most frequently throughout the whole 35 news articles. Moreover, the most pressing issues can thus be addressed by the authorities for further analysis and improvements. The word cloud in Figure 6 below involves words that were stemmed and then cleaned further to only represent fully spelled out words that are relevant and bring meaning. Words with similar frequencies are plotted with the same colour.

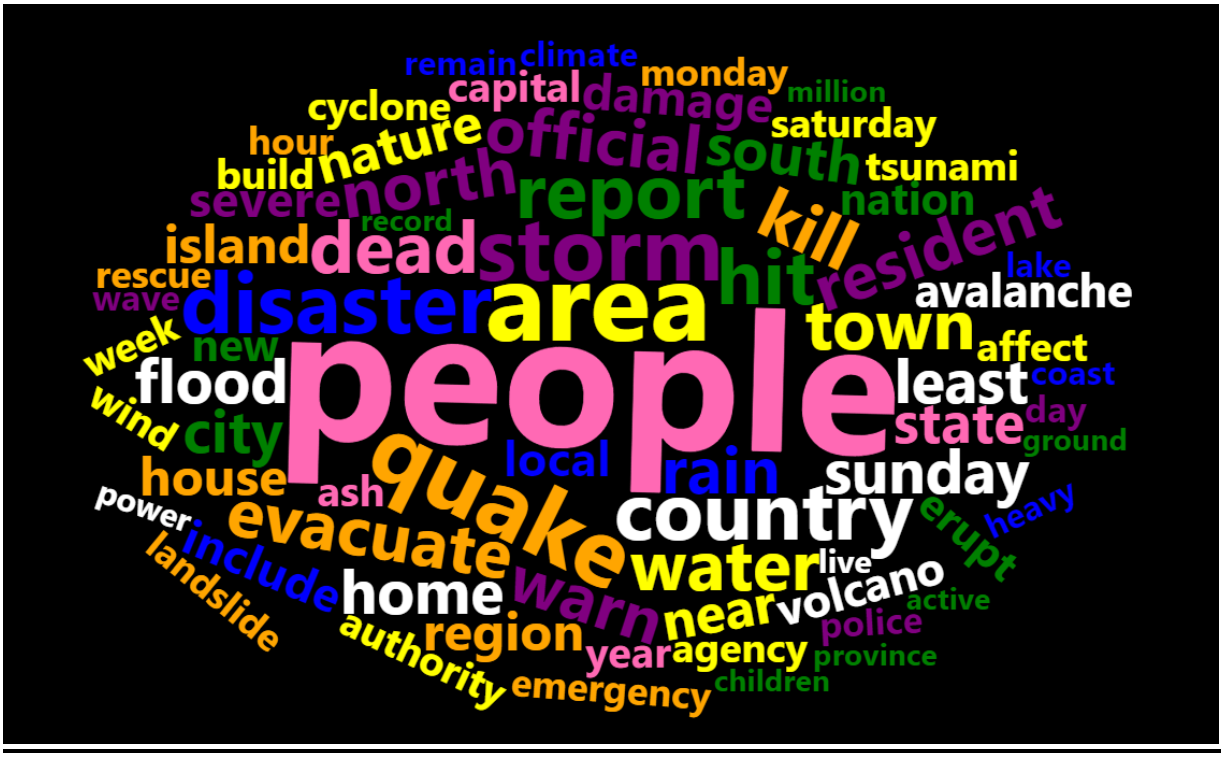

Figure 6 Word cloud of 35 news articles regarding natural disasters

From the word cloud above, we observe that the word "people" occurs most frequently, making 95 appearances. This is a logical occurrence seeing as natural disasters involve many parties and tend to impact the lives of many. The usage of the word "people" could mean that the media discusses the impact of these disasters on human lives, their well-being and any actions taken to help them. The term "area" could refer to the localities affected, portraying how the media talks about the impact on particular areas too. To no one's surprise, the words "quake" (48), "storm" (44) and such, appear in the text as it is mainly about natural disasters. Of course, these words will be mentioned in the text. Other natural disasters such as "avalanche" (23), "tsunami" (20) and "cyclone" (20) do not appear as frequently, indicating how earthquakes and storms are more prevalent in the news articles gathered. Words like "region" (26), "city" (29), "coast" (17), "north" (33), "south" (29) and so on are geographic indicators, where the spatial extent of the disasters is discussed.

Importantly, terms like "emergency" (19), "kill" (34), "damage" (25), "rescue" (19) and "evacuate" (34), all appear to convey the severity of the aftermath of the natural disasters. These words should be analysed further so that the relevant parties can work towards enhancing the safety

of many and reducing bad impacts. Without doubt, the authorities and relevant parties are doing their jobs to help with the effects of the natural disasters as words like “authority” (20), “official” (33) and “agency” (20) appear in the text. The context of time is also important in natural disaster news as the words “saturday” (30), “sunday” (20), “monday” (19), “week” (20) and “hour” (19) appear in the text. The major terms and their importance to natural disasters are identified through this word cloud. Apparently, the media is very comprehensive in their coverage of natural disasters as they discuss areas, time and the impacts on people. Some disasters are more prevalent than others, which could be due to more news articles on earthquakes and storms being gathered or maybe these disasters were very extreme in the impacts they caused.

Following this, a simple word association analysis is carried out to view how the top three most frequent words, “people”, “area” and “quake” co-occur with certain words. As such, a high correlation limit of 0.55 is set to only filter for word associations that have a very strong correlation to each other.

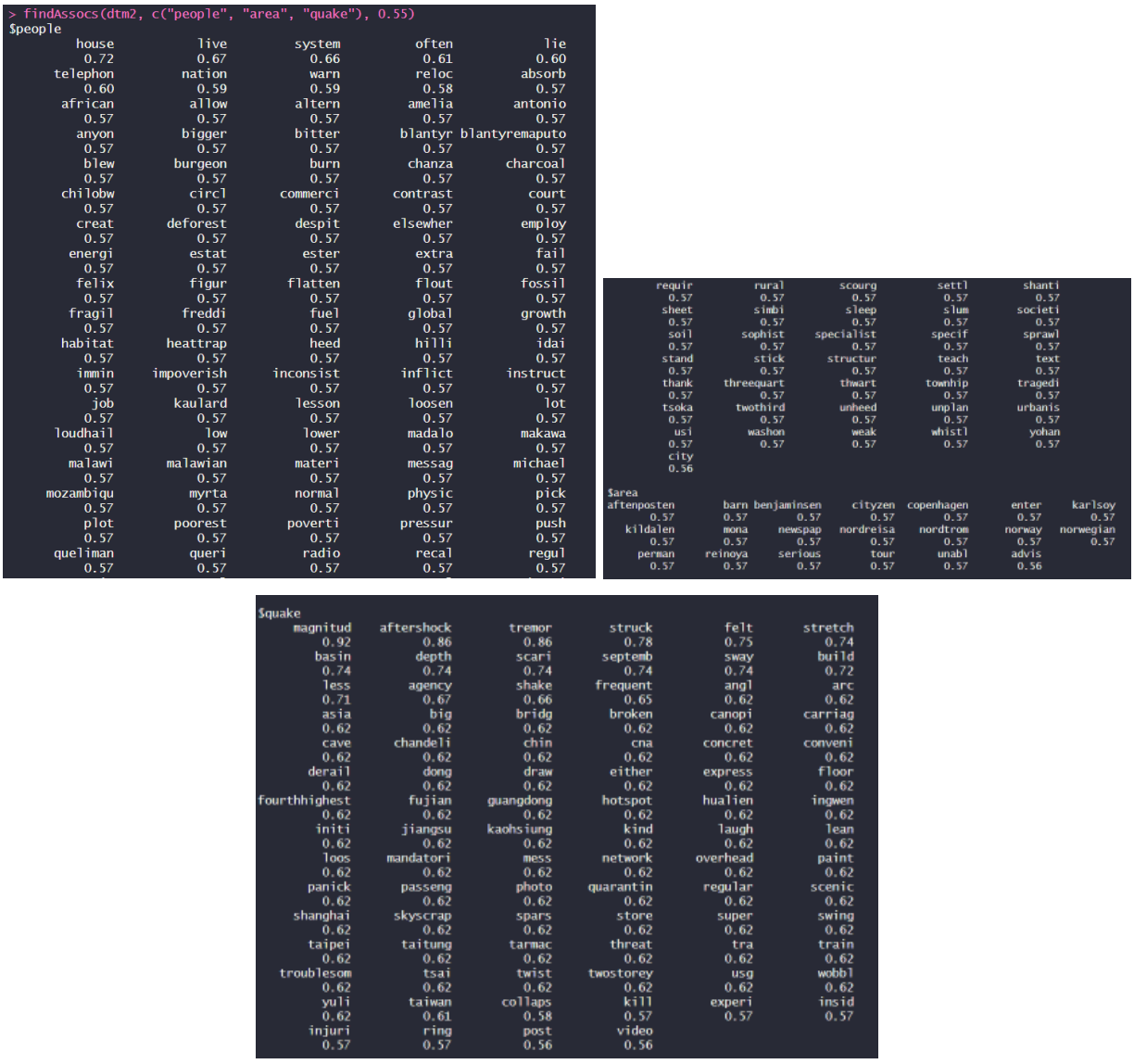

Figure 7 Word Associations for the top three most frequent terms in the news articles

It is observed that the terms “people” and “quake” have strong associations with so many other terms. Noteworthily, the term “quake” is very strongly correlated to the terms “magnitud” (magnitude), “aftershock” and “tremor”, with the highest correlation values of 0.92, 0.86 and 0.86 respectively. This makes sense as those are terms that often occur when talking about earthquakes. We also see it being associated with “panick” (panic), “kill and “injuri” (injury), depicting the impact of the earthquakes that take place, to the point it ruins lives. “People” is most associated with the term “house”, which might indicate how the media discusses the impact of natural disasters on their homes.

Interestingly, the term “area” is highly associated with everything about Norway in that they co-occur frequently. We see it is correlated to the terms “aftenposten” (Norway's largest printed newspaper), “karlsoy” (Karlsøy is an island municipality in Norway), “nordreisa” (municipality in Norway), “norway”, “norwegian” and such, which are all very Norway themed. The word association analysis has indicated an underlying theme or context within the news articles gathered, which should be investigated further to gain more insights.

1.6 Conclusion

In a nutshell, implementing LDA topic modelling and simple text analysis on news articles regarding natural disasters has brought us many valuable insights that can be further explored. We were able to extract prevalent themes, identify patterns and analyse the relationship between different terms in the text data, all aiding to enhance our understanding of how natural disasters are portrayed by the media. Firstly, the LDA model has identified four different topics, all made up of distinct natural disasters. This could help abundantly in the categorisation and organisation of text data, dividing natural disasters into the right groups, especially if a bigger data set is utilised. The beta spread confirms the distinction between themes, but nevertheless reminds us that certain terms tend to overlap between topics as it is not a hard clustering algorithm. Apparently, 34 out of the 35 news articles gathered were generated almost solely from one specific topic, showing a good separation of topics by the LDA model.

The analysis carried out also sheds some light on frequently discussed issues regarding natural disasters by the media. It seems to be that there is a comprehensive coverage of the victims involved, scale of the impact, areas affected, rescue efforts and so on in the text. Interestingly, certain natural disasters such as earthquakes and storms are more discussed in the news articles gathered as compared to other disasters. The word associations have also discovered certain underlying themes that might need to be looked into further. With further exploration of the results obtained, they could go as far as to aid emergency management agencies, policymakers and enthusiastic researchers in the field. The results certainly provide a deeper comprehension of the terminology and central issues associated with natural disasters, allowing for more informed decision-making and even targeted interventions by relevant parties.

TASK 2

By utilising the data from Task 1, this task aims to construct data clustering using three different clustering algorithms, namely k-means, hierarchical and HDBScan.

2.1 K-Means Clustering Algorithm

Firstly, for the k-means clustering algorithm, the number of clusters needed must be determined beforehand. As such, the optimal number of clusters can be defined using Silhouette analysis, which is a similarity index of data points within a cluster and between clusters (Banerji 2021). Silhouette scores can range from -1 to +1, with higher scores indicating better clustering performance.

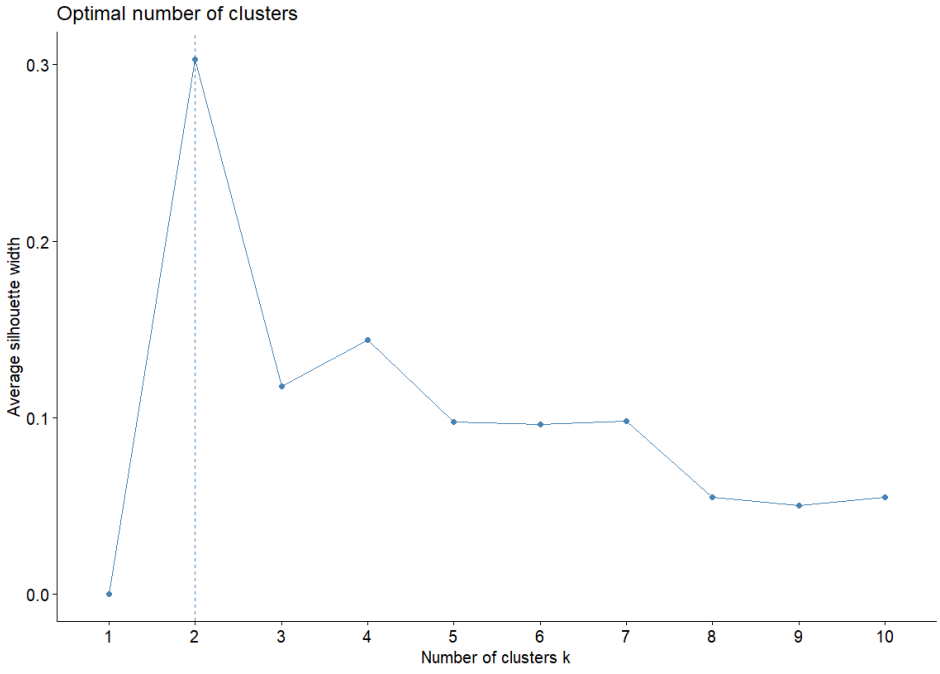

Figure 8 Silhouette analysis to determine the optimal number of clusters for k-means clustering algorithm

From the Silhoutte analysis carried out, 2 clusters have the highest score, indicating that this method has chosen 2 as the optimal number of clusters. However, based on our specific objective of clustering news articles on natural disasters, 2 clusters might not be such a good

choice. This is because the news articles encompass various types of natural disasters, caused by different hazards and also possess different characteristics. Increasing the number of clusters to the next highest silhouette score, 4 clusters, would be a much better choice. A more comprehensive analysis can be carried out with the groups capturing the variety and unique characteristics of the natural disasters.

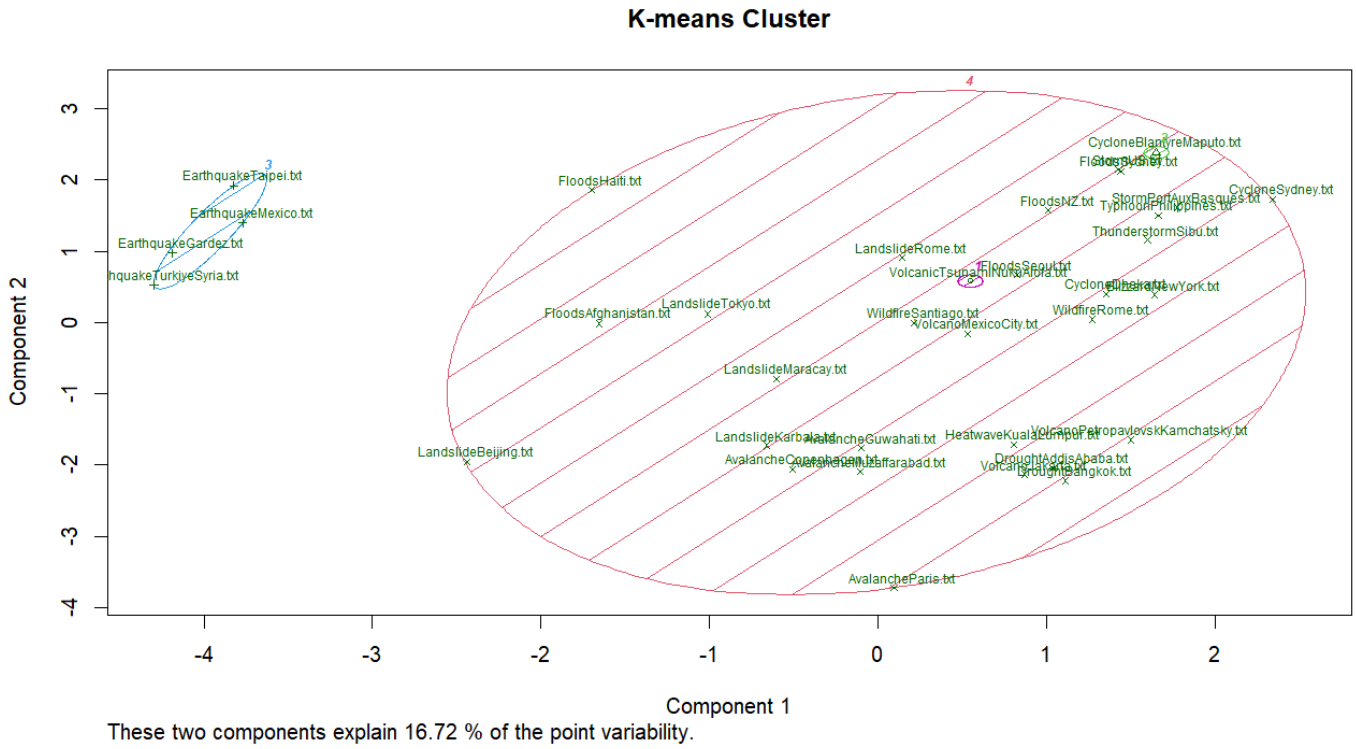

Figure 9 Plot for the k-means clustering algorithm results

From the k-means cluster plot shown in Figure 9 above, we notice that the news articles are being clustered based on their specific types of naturals disasters. First of all, notice how *VolcanicTsunamiNukuAlofa.txt* is all on its own in one cluster. This article discusses how an underwater volcano has triggered a tsunami in Tonga. Maybe due to its distinctive nature, the algorithm has moved it to a cluster of its own to highlight its uniqueness. Observe how *CycloneBlantyreMaputo.txt* is also clustered on its own in one group. The article is about Cyclone Freddy striking Mozambique and Malawi, and as such the cluster displays a singular representation of the cyclone event in those countries. Moving on to the next cluster, there are four news articles, namely *EarthquakeTaipei.txt*, *EarthquakeMexico.txt*, *EarthquakeGardez.txt* and

EarthquakeTurkiyeSyria.txt. Notably, all of them are specifically related to earthquakes, taking place in different countries and areas. It can be said that this particular cluster is dedicated to natural disasters related to earthquake contents. The last cluster is rather large, consisting of the remaining 29 news articles covering all kinds of natural disasters from storms, to landslides, wildfires and so on. Despite the news articles representing different types of disasters, the algorithm seems to assume that they have common characteristics that can be grouped together.

2.2 Hierarchical Clustering

Next, hierarchical clustering is carried out and visualised using a dendrogram constructed from the *ward.D2* method in R, so that we can identify a more logical cluster combination ourselves. Essentially, the number of clusters for hierarchical clustering can be determined based on self-judgement by looking at the structure of the dendrogram. The vertical lines represent the distance between clusters in which the distance between clusters is inversely correlated with the length of the vertical lines.

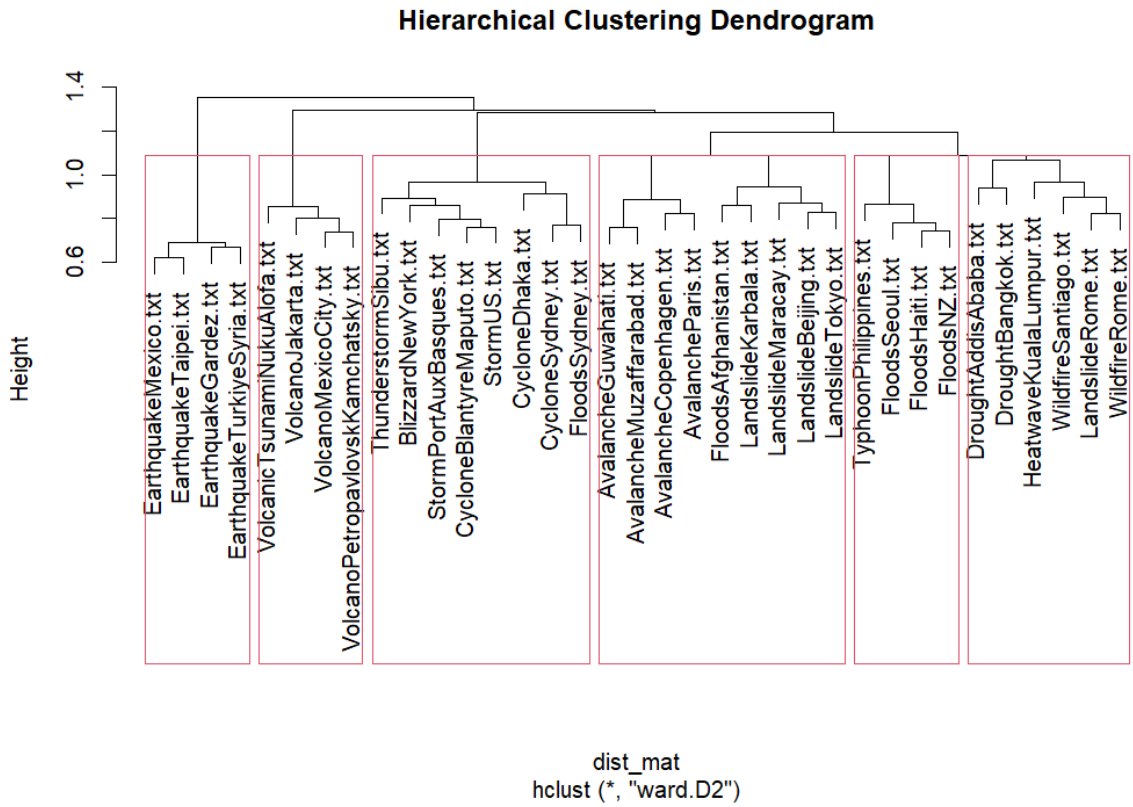

Figure 10 Dendrogram for the hierarchical clustering algorithm results

Judging from the dendrogram produced, six clusters seem to be the best choice for clustering the news articles. In the first cluster, all the four articles related to earthquakes that were extracted are put in one group, similar to the k-means algorithm from before. The next cluster on the other hand consists of all four articles related to volcanoes. Cluster 3 portrays news articles of natural disasters such as storms, thunderstorms, cyclones, blizzards and floods. Most of these are due to extreme weather conditions such as strong winds or cold weathers, except for floods which is actually a water hazard. *FloodsSyney.txt* might have been grouped in this cluster due to the fact that it has many similarities with *CycloneSydney.txt*, since both the events took place in the same area.

Moving on, cluster 4 consists of all the news articles regarding avalanches and landslides, both are geological hazards that involve the fall and motion of ice and ground respectively. However, *FloodsAfghanistan.txt* was also grouped in this cluster as the algorithm might have found some similar characteristics it has with the other articles. Three floods and one typhoon news articles were grouped into cluster 5. *TyphoonPhilippines.txt* might be a little out of place but looking at the text, it is quite logical that it was grouped with articles on floods since it discusses rain, large waves and even flash floods, all water-related as well. The last cluster has 6 news articles, regarding droughts, heatwaves and wildfires, which are essentially due to hot and dry weather conditions. *LandslideRome.txt* was also grouped in this cluster, most probably due to it sharing the same area of disaster as *WildfireRome.txt*.

2.3 Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBScan)

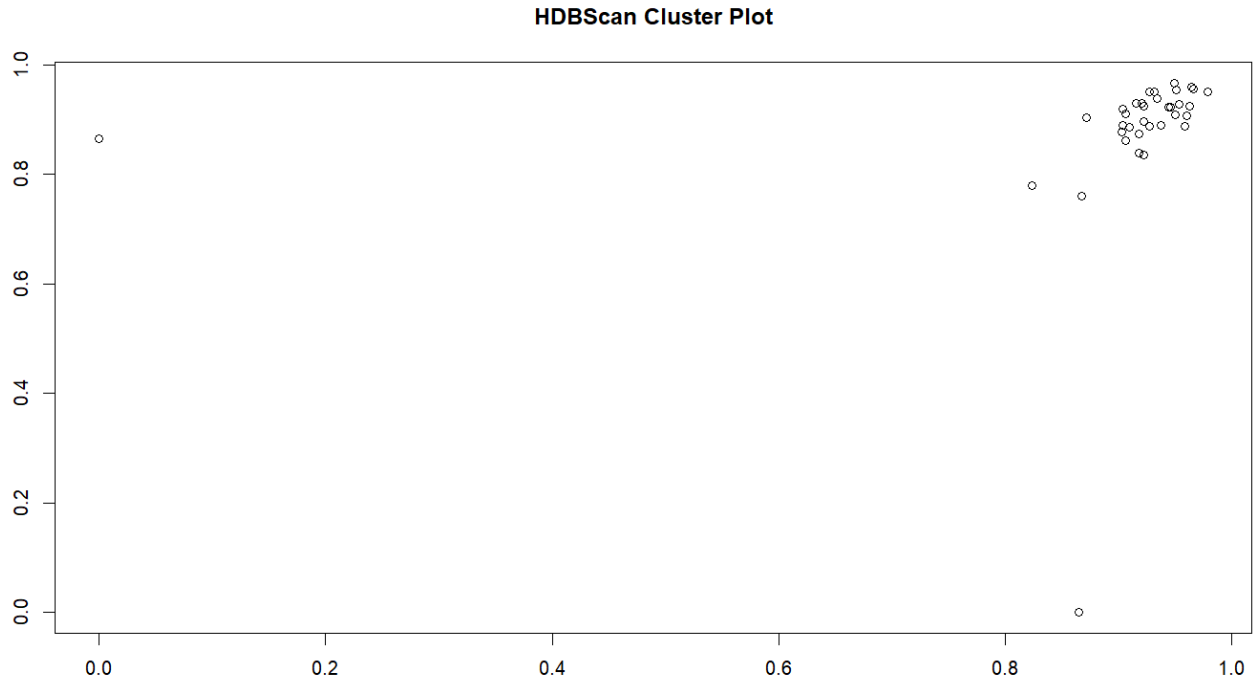

Figure 11 Scatter plot to portray the results of the HDBScan algorithm

HDBScan is a density-based approach that clusters data based on their densities. It is important to take note that in R, noise points are classified as having a cluster ID of 0. From the results obtained (also by referring to Table 3 for confirmation), it seems that there is no clustering and all 35 news articles have been labelled as noise. This means that the HDBScan algorithm was unable to find any meaningful and distinct clusters from the news articles. This may be due to the fact that the news articles come from diverse types of natural disasters with different writing styles, since it was from three different news sources. Seeing as how the algorithm is density-based, it could struggle to find clear density-based structures in the data set. As such, this algorithm is not suitable for this particular data set even though it may be a powerful algorithm in other cases.

2.4 Overall Clustering Results

Table 3 Number of news articles in each cluster for all three clustering algorithms

Clustering Algorithm	No Cluster	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
K-means		1	1	4	29		
Hierarchical		9	8	6	4	4	4
HDBScan	35						

To provide an overall view of the results obtained, Table 3 displays the number of articles in each cluster for all three clustering algorithms. As mentioned earlier, the HDBScan algorithm is not suited for this data set as it is unable to locate any discerning clusters due to lack of density structure. The k-means algorithm was executed with 4 clusters as determined by the Silhouette analysis and self-judgement, whereas the dendrogram structure portrayed 6 clusters as most appropriate. As can be seen in Table 3, the k-means algorithm has two clusters with only one news article each, and another two clusters with four and 29 news articles respectively. It seems that the distribution is a little unequal as cluster 4 comprises of a much bigger portion of the news articles than the other 3 clusters. The hierarchical clustering algorithm on the other hand, generated a much more diverse set of clusters in that the clusters vary in size. The distribution of the news articles is much more even with 9, 8, 6, 4, 4 and 4 news articles in each cluster.

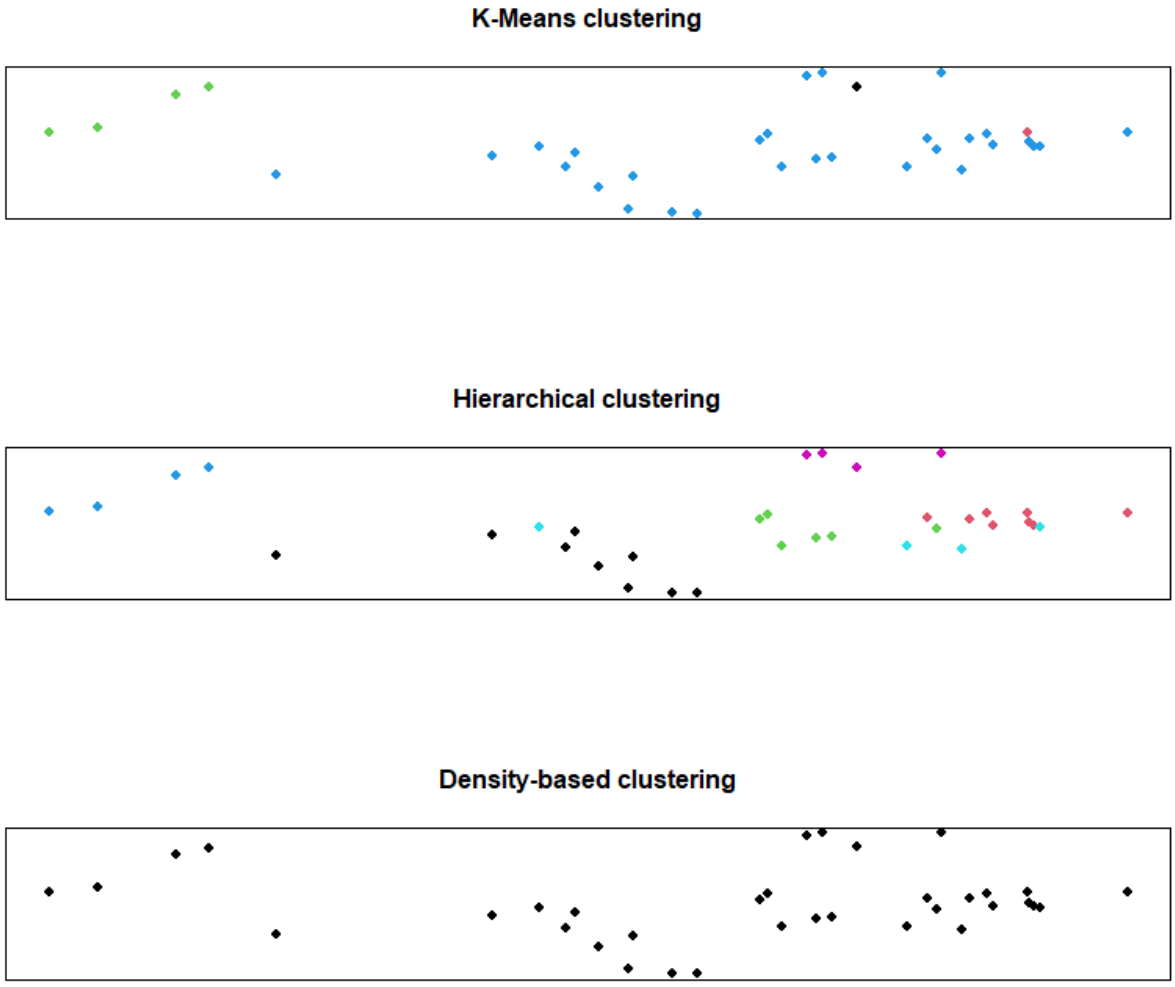

Figure 12 Scatter plots to visualise the clusters obtained from k-means, hierarchical and HDBScan clustering algorithms respectively

The scatter plots in Figure 12 represent points where the distances between the points are roughly proportional to their dissimilarities. From the scatter plots, only the k-means and hierarchical clustering algorithms were able to distinguish some groupings within the news articles. HDBScan on the other hand could not find any clusters (as is shown by only one single colour of points). We notice that most of the points in the k-means clustering belong to one cluster (blue), whereas the others are in very much smaller clusters. This visualisation makes it evident that hierarchical clustering does the best job in grouping the news articles into clusters as points closer to each other belong in the same cluster with varying sizes across clusters. Even though one of the light blue points is a little further away from its group, it might still have the closest

similarities to this cluster instead of any other based on the hierarchical algorithm. In fact, this is most probably *TyphoonPhilippines.txt* being grouped with the floods articles as explained earlier.

2.5 Conclusion – Most Appropriate Clustering Algorithm is Hierarchical Clustering

All in all, the hierarchical clustering algorithm does the best job in clustering the 35 news articles regarding natural disasters that were extracted. This is due to the fact that the news articles are distributed more evenly across clusters as compared to the other two algorithms. There are at least four articles in one cluster. Besides that, as was seen through the dendrogram in Figure 10, even with a diverse range of natural disaster types, the algorithm successfully managed to group the articles according to their types, such as caused by geological hazards, water hazards, extreme weather conditions and so on. Although some articles seemed to be out of place, the algorithm still managed to capture similarities in that the articles were grouped with another article discussing the same area impacted by the disaster. The algorithm may not be perfect, but it is much better in segregating the news articles regarding natural disasters based on similar characteristics than k-means or HDBScan algorithms.

The findings demonstrate the variety and diversity of natural disasters, which encompass many sources of hazards. By diving deeper into analysing the clusters, they may serve as a good guide in identifying underlying themes and patterns displayed by news articles in each cluster. The underlying characteristics of each natural event are taken into account by the clusters, allowing us to carry out a more focused analysis on the contents of each cluster to understand them deeper.

TASK 3

3.1 Introduction

"I scream, you scream, we all scream for ice cream", is a popular song from the 1920s by the popular American band, Waring's Pennsylvanians (Nielsen 2009). Even from back then, ice cream, a lovely frozen treat has delighted the taste buds of millions of people worldwide. Ice cream is so indulged by many that there is even research stating that its ability to quench thirst through mouth cooling, is one of the reasons ice creams are considered to be satisfying (Eccles et al. 2013). The question is, to what extent do people actually love ice cream? Do certain ice cream flavours from one's favourite brand triumph over other flavours? Simple questions like these can be answered through sentiment analysis. In order to do so, the ice cream data set with 241 ice cream flavours across four brands, namely Ben & Jerry's, Häagen-Dazs, Breyers and Talenti is chosen to be analysed.

In the sentiment analysis carried out, only the Ben & Jerry's ice cream data set is chosen so that more emphasis can be put on one particular brand with a variety of different ice cream flavours. In fact, Ben & Jerry's is one of the biggest giants in the ice cream manufacturing industry, with 2022 sales of $910.68 million (Kolmar 2023). For the purposes of this analysis, only the *reviews.csv* and *products.csv* data sets (from the Ben & Jerry's directory) are downloaded, and then merged together to obtain both the flavours of ice creams and their many reviews from customers. Essentially, there are 57 different ice cream flavours in the data set with 7943 reviews. By utilising the reviews, many valuable insights into the opinions of customers regarding their products can be mined, to understand customer preferences and gauge satisfaction besides fixing any areas in which they are lacking. Following the immense popularity of Ben & Jerry's ice cream, sentiment analysis is able to unlock a wealth of knowledge for the manufacturers in order for them to act wisely and further boost their sales.

3.2 Sentiment Scores Using Different Lexicons

In conducting sentiment analysis, it is important to look at the sentiment scores of reviews as they represent the degree of sentiment expressed by the customers and not just whether they are positive or negative. As such, the sentiment scores for the overall Ben & Jerry's reviews are obtained using the Syuzhet, Bing and AFINN lexicons to provide a more comprehensive analysis as the amount

of linguistic coverage that different lexicons offer varies. Table 4 below provides the sentiment score summaries for the Syuzhet, Bing and AFINN lexicons respectively. It provides us with an overview of the overall reviews' sentiment along with its distribution.

Table 4 Summary statistics of sentiment scores according to the Syuzhet, Bing and AFINN lexicons

Lexicon	**Sentiment Score Summary**					
	Minimum	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Maximum**
Syuzhet	-5.200	0.750	1.600	1.705	2.550	10.700
Bing	-8.000	1.000	2.000	1.751	3.000	12.000
AFINN	-15.000	3.000	5.000	5.570	8.000	37.000

From a glance at the sentiment score summary, it is noticeable that the mean sentiment score for the overall reviews of Ben & Jerry's ice cream is positive across all three lexicons even though their magnitudes vary. Even if there were to be any outlier reviews, the median scores provided (1.600, 2.000, 5.000) all lean towards positive reviews. In comparison to the other lexicons, the AFINN lexicon has the greatest median and mean sentiment scores, indicating a higher positive sentiment. Although, this can be attributed to the difference in their scoring ranges, where AFINN scores range from -5 to +5 (integers), Syuzhet from -1 to +1 (float) and Bing only -1 and +1 (integers). The bigger sentiment scoring for AFFIN explains why its values are bigger, whereas the similar ranges for Syuzhet and Bing explain their similarities in mean sentiment score. Between the 1st quartile and 3rd quartile is the interquartile range (IQR) representing the central distribution, where a significant number of reviews lie. Observe that the IQR is positive, with values of 1.800, 2.000 and 5.000 for each lexicon respectively, indicating that a big proportion of reviews are positive. The range of reviews on the other hand is the narrowest for the Syuzhet lexicon with a value of 15.900, where the most extremely negative review has a score of -5.200 and the most positive review has a score of 10.700. This suggests a more compressed range of sentiment scores for Syuzhet as compared to Bing and AFINN.

3.3 Most Common Positive and Negative Words

For the next component of sentiment analysis, we look at the most common positive and negative words for the overall reviews. This will help to shed some light on customer preferences, providing insights into what aspects of the ice cream products that customers seem to like and dislike.

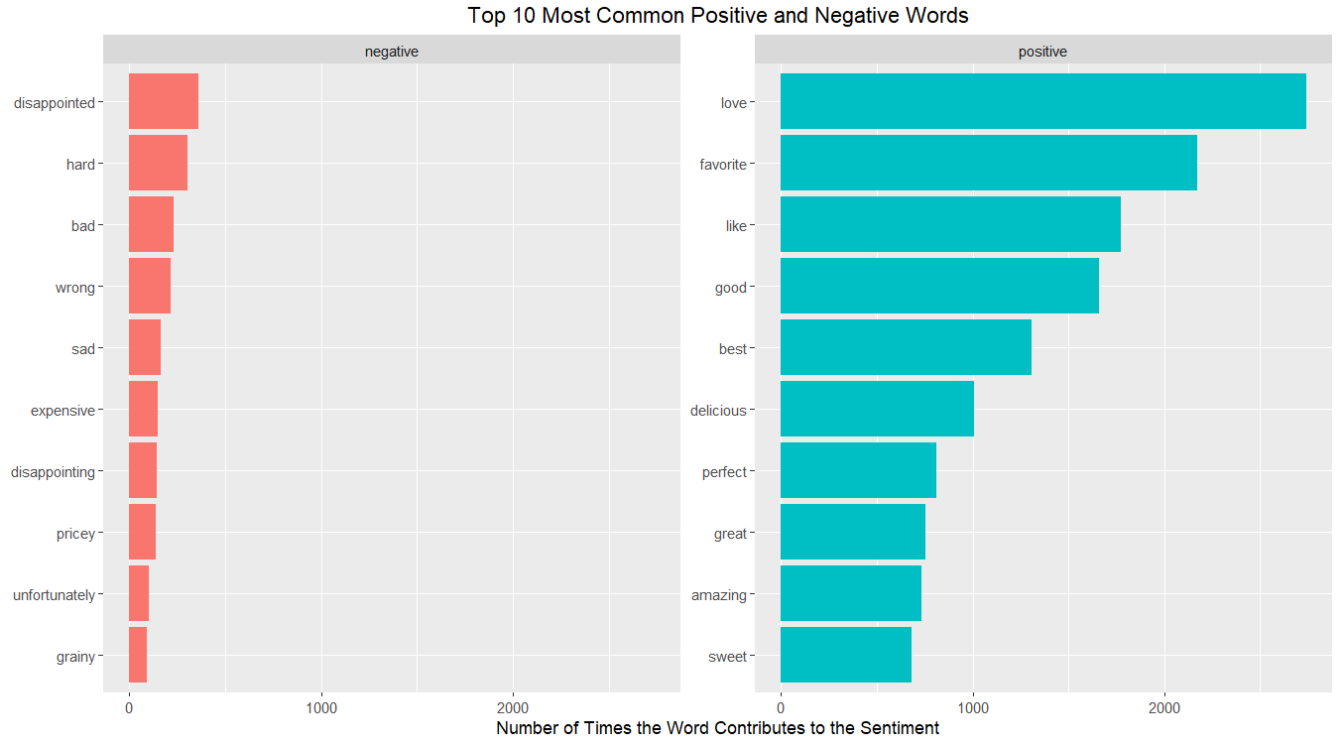

Figure 13 Top 10 most common positive and negative words in the ice cream reviews data set

Evidently from Figure 13, it is very obvious that the number of times positive words appear in the reviews is much higher than negative words. The most common positive words convey a sense of satisfaction and high regard for the Ben & Jerry's products, where some even mention that it is the "best". The words "love" and "favorite" appear the most often, that is 2740 and 2172 times respectively. This represents the immense fondness and enthusiasm of customers towards the ice cream, which may be their number one choice of ice cream. The word "perfect" on the other hand, might be an indication that the ice cream meets the customers' needs in every aspect, where this is true for 813 customers, provided that there are no repetitions of the word "perfect" in a review.

Alternatively, the most common negative words convey a sense of disappointment, dissatisfaction and even criticism on their negative experiences with the ice cream. The words "expensive" (148) and "pricey" (137) indicate that these customers find the ice cream products to be relatively more costly than their counterparts, leading to their overall dissatisfaction and impacting future purchase decisions. Notice however that only a very small fraction of customers feel this way, which may mean that they are not the right target audience. As opposed to the word "perfect" that appears 813 times, the word "unfortunately" appears 103 times, where these customers feel that the ice cream does not meet their expectations. The word "grainy" (89) shows criticism on the texture of the ice cream in which it could be too coarse. Ben & Jerry's can look into this, gather more feedback on the texture of their ice cream and further refine their production process.

Figure 14 Comparison cloud portraying the negative and positive terms in the ice cream reviews data set

Besides that, a comparison cloud is plotted in Figure 14 above. The comparison cloud will help us in getting a quick understanding of the overall sentiment aspects. Note that the size of the

words cannot be compared across the negative and positive sentiments. Besides the 10 most common words from earlier, we see that customers even use the word “pretty”, which could be compliments about the packaging aspect. They find the ice cream to be “heaven” like, and would even “recommend” this “worth” ice cream to others. As expected, even if there are customers who mentioned about the ice cream being “grainy”, there are also reviews with the word “smooth”, affirming that every person has different reflections. Some customers go as far as saying that the ice cream is “gross” and “disgusting”. The Ben & Jerry’s team should look into extreme comments like these to ensure that these are genuine comments and not fake ones, in which they would need to handle accordingly.

Most customers seem to have experienced a positive sensory sensation with positive outlooks on the quality and superiority of the ice cream products. Conversely, only a handful of customers find the ice cream to have flaws. Of course, it is not possible to satisfy every single customer since everyone has different tastes in food. However, Ben & Jerry’s can work closely to improve in certain areas that are commonly commented on by dissatisfied customers and work towards boosting their sales with a more targeted audience.

3.4 Emotion Classification

Instead of just knowing whether the reviews are generally positive or negative, emotion classification can be done to gain deeper insights into the emotional feedbacks expressed by customers. Figure 15 below portrays a bar chart with 8 different emotion sentiments and their respective contributions to the overall review sentiment.

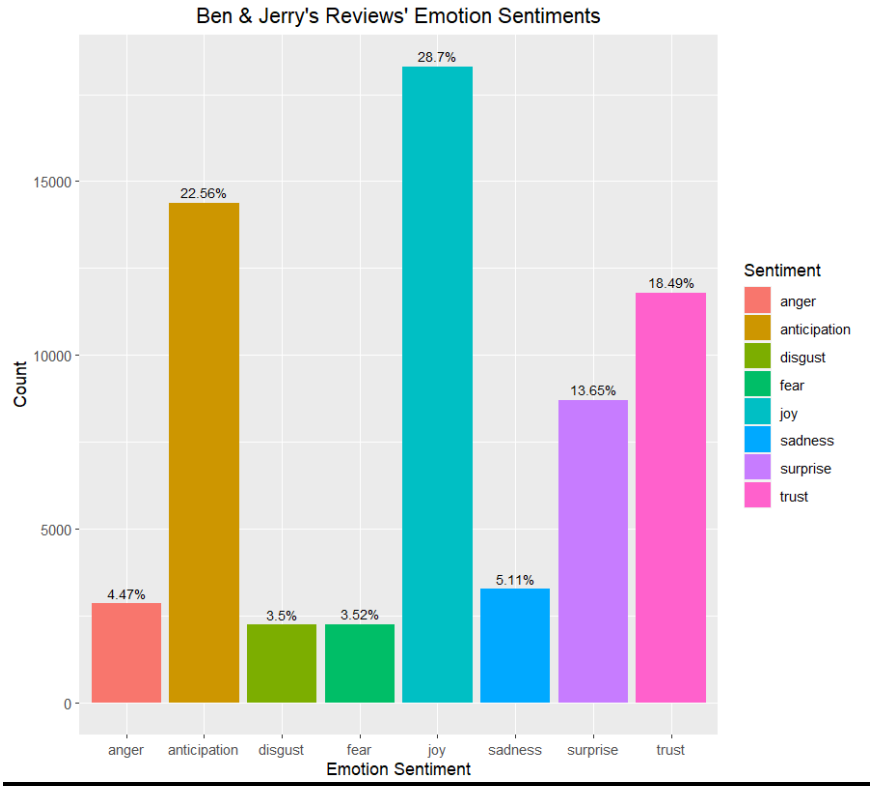

Figure 15 Emotion sentiment classification for the ice cream reviews data set

Based on customer experiences, the most prevalent emotions were all positive, namely joy, anticipation and trust, making up 28.7%, 22.56% and 18.49% of the emotion sentiment respectively. With a sentiment count of 18,302, the joy sentiment shows that customers are delighted and generally satisfied with the ice cream. Anticipation (14,384) on the other hand is a positive sentiment in that the customers might have been eager or excited to indulge in the ice cream, whereas trust (11,789) shows that customers can rely on the reputation of Ben & Jerry's ice cream, emphasising on the consistent quality and continuous positive experiences with the products. Besides that, the surprise (8,703) sentiment portrays unexpected reviews from customers that could be pleasant or bad, in which would need to be figured out further. Nevertheless, looking at the amount of positive reviews, they are most probably pleasant surprises about the taste, flavour combinations, quality and so on. The four negative sentiments, namely sadness (3,261), anger (2,850), fear (2,246) and disgust (2,233) make up only 16.6% of the overall sentiment count. These customers were most probably frustrated with the unpleasant flavours, bad texture or any other issues encountered during their ice cream consumption. Additionally, they could feel disappointed

that the ice cream did not live up to their expectations or even fear about the food safety such as allergens and so forth. Generally, most reviews have positive sentiments in nature.

3.5 Top 5 Most and Least Reviewed Flavour Ice Creams

Now that we know most reviews are positive in nature, let us look at which ice cream flavours are the most and least popular, based on their number of reviews. Ice cream flavours with more reviews are considered to be more popular in the sense that it is highly visible and customers are aware about it. They are even eager enough to leave reviews (good or bad ones) of their experiences with the flavour, increasing brand recognition and therefore perceived to be popular. Figure 16 and Figure 17 display the percentage of positive and negative sentiments for the top 5 most popular and least popular ice cream flavours respectively.

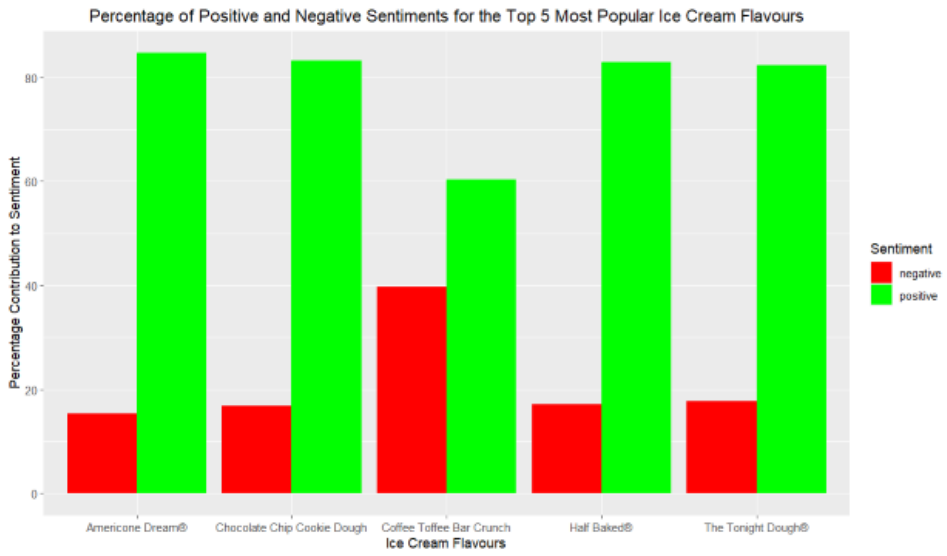

Figure 16 Bar chart of the percentage of positive and negative sentiments for the top 5 most popular ice cream flavours

According to analysis, the top 5 most popular ice cream flavours are American Dream, Chocolate Chip Cookie Dough, Coffee Toffee Bar Crunch, Half Baked and The Tonight Dough. Essentially, four out of the five flavours have high positive sentiments, all above 80%. This suggests that the flavours are well-received and enjoyed by customers, despite several negative comments. Unlike these four flavours, Coffee Toffee Bar Crunch has a higher percentage of negative sentiments at around 39.6% out of the 1269 reviews it received. This implies that while

there are still a lot of good comments about the flavour, a relatively higher percentage of customers have voiced discontent or had negative encounters with it. It would do good for Ben & Jerry's to look deeper into the concerns regarding this particular flavour and further improve their quality since it is such a popular flavour among its customers.

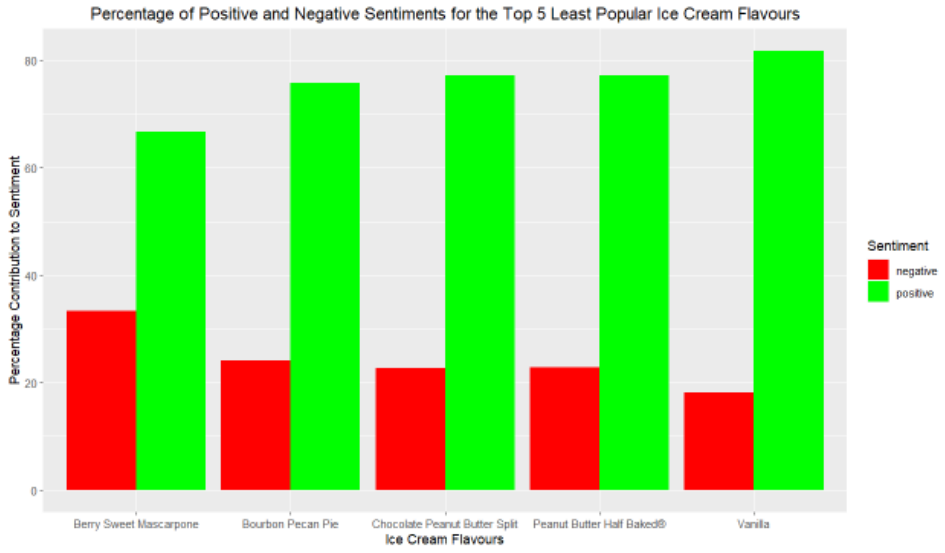

Figure 17 Bar chart of the percentage of positive and negative sentiments for the top 5 least popular ice cream flavours

Let us then look at the top 5 least popular ice cream flavours, namely Berry Sweet Mascarpone, Bourbon Pecan Pie, Chocolate Peanut Butter Split, Peanut Butter Half Baked and Vanilla. Keep in mind that these flavours have only received 45, 29, 22, 57 and 33 reviews respectively. Therefore, the percentage of positive and negative sentiments for these ice cream flavours must be interpretated with caution since the sample size is small. For all the flavours except Berry Sweet Mascarpone, the percentage of positive sentiments from customers is quite high, above 75%. Even though there aren't many reviews, most of the consumers who did share their opinions appreciated these flavours. Berry Sweet Mascarpone has a higher percentage of negative sentiments at 33.3%. The team should look into these five flavours and figure out ways to potentially increase their customer count or just scrap the flavour all at once if it does not bring in profit in the long run. Although these five flavours are not as popular as compared to other flavours, there is still a portion of loyal customers who love and enjoy them. This further highlights the diversity in their customers' tastes and preferences. Ben & Jerry's could then curate better new and improved flavours based on these feedbacks.

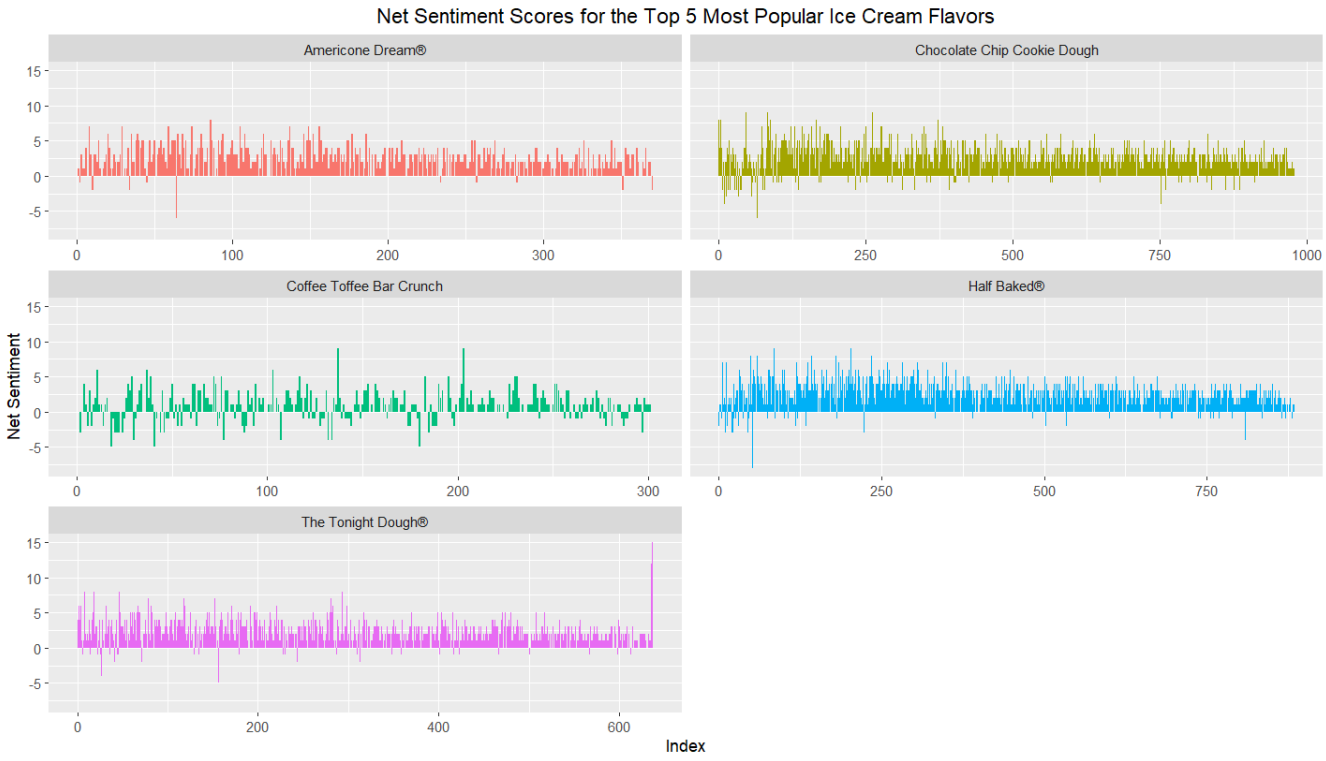

Figure 18 Net sentiment scores for the top 5 most popular ice cream flavours

Figure 18 above portrays the net sentiment scores for the top 5 most popular ice cream flavours to get a closer look at the sentiment distribution across index of reviews. This gives us a sense of whether the reviews are predominantly positive, negative or neutral while making comparisons between flavours. In a general sense, all five flavours show a mixture of positive and negative net sentiments, of course leaning more towards positive reviews. As was also discovered earlier, Coffee Toffee Bar Crunch shows the highest amount of negative net sentiment scores, with a higher magnitude as well. This shows that this flavour not only receives more backlash from overall reviews, but it is also apparent when looking at smaller portions of reviews. This indicates that it is not outlier reviews causing the higher count of negative sentiments. Americone Dream has the least negative net sentiments, meaning that most of its negative sentiments come from only a few reviews that show pure hatred for this flavour. These reviews should be looked into to avoid fake comments that are intended to bring down the brand's reputation. Certain reviews with extremely high positive net sentiment scores for The Tonight Dough flavour should also observed closely to prevent fake comments intending to boost positive reviews.

3.6 Conclusion

Concludingly, the sentiment analysis carried out on Ben & Jerry's ice cream products have brought about many interesting insights that could be used to make data-driven decisions. By utilising a large number of ice cream reviews, the general overall sentiment of customers was able to be gauged, besides learning about the distribution of sentiment across different emotions, identifying the aspects appreciated and disliked by customers, and also looking at the sentiment distribution across the most and least popular ice cream flavours. All in all, a very big portion of the overall reviews had predominantly positive sentiments in which customers are satisfied with their consumption. This was observed through the positive sentiment scores and also the frequent appearance of words such as "love", "favourite", "amazing" and so on in the reviews. There are however, customers that complain about the costly price of the ice cream and also the unpleasant texture in that it is grainy. Negative reviews, especially extreme ones should be taken close care of by the relevant team in order to improve their customers' trust and boost brand reputation. The mostly positive emotion classifications can help in the customisation of the ice cream products and marketing tactics to elicit the necessary emotional reactions of customers.

Furthermore, finding out how people feel about popular flavours can help determine customer preferences and inform product development strategies. Out of the five most popular ice cream flavours, four of them had over 80% of positive sentiments. Ben & Jerry's could further investigate the aspects of these flavours that customers enjoy to improve their existing flavours or even come up with new ones based on customer preferences. They could even have different target markets in which they promote certain flavours to certain segments of customers, which could then boost their sales further. Through the frequent monitoring and analysis of current sentiment patterns, Ben & Jerry's can actively cater to their customers' needs and remain as the big guns in the highly competitive ice cream manufacturing industry. It is of course important to interpret all the results with a grain of salt since sentiment analysis has its limitations in that it cannot capture the intentions or true context of the reviews such as sarcasm, fake reviews and irony among others. Nevertheless, the results provided a general overview of customer sentiment, which is majorly positive for Ben & Jerry's in this case.

REFERENCES

Banerji, A. 2021. K-mean: Getting the optimal number of clusters. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/ [16 June 2023].

CRED. 2023. *2022 Disasters in Numbers*. Brussels: Centre for Research on the Epidemiology of Disasters.

Eccles, R., Du-Plessis, L., Dommels, Y. & Wilkinson, J.E. 2013. Cold pleasure. Why we like ice drinks, ice-lollies and ice cream. *Appetite* 71: 357–360.

Kolmar, C. 2023. The 10 largest ice cream companies in the United States. *Zippia*. https://www.zippia.com/advice/largest-ice-cream-companies/ [14 June 2023].

Nielsen, M. 2009. I scream, you scream, we all scream for ice cream. 1925. https://www.youtube.com/watch?v=-0pfP_MD6xA [14 June 2023].

WSL Institute for Snow and Avalanche Research SLF. n.d. Avalanche types. https://www.slf.ch/en/avalanches/avalanche-science-and-prevention/avalanche-types.html [14 June 2023].