UNEMPLOYMENT NEWS TYPE IN MALAYSIA BY TOPIC MODELLING AND CLUSTER ANALYSIS APPROACH

Faiza Rusrianti Tajul Arus[1]

ABSTRAK

Evolusi berita dalam talian menunjukkan pesona yang hebat dalam landskap sosial dan politik di Malaysia. Oleh itu, peluang ini dimanfaatkan untuk memahami isu-isu utama yang terbit dari kandungan berita dalam talian, yang turut mempengaruhi status sosioekonomi semasa di Malaysia. Oleh itu, kajian ini bertujuan untuk menjelaskan bagaimana data tidak terstruktur dalam jenis teks dapat digunakan untuk mengetahui kelompok dan topik yang dihasilkan dari koleksi dokumen ini melalui analisis pengelompokan dan Latent Dirichlet Allocation (LDA). Perlombongan teks adalah proses meneroka koleksi teks dan menukar format dari data tidak berstruktur menjadi data berstruktur untuk analisis selanjutnya manakala analisis kluster bertujuan untuk mengelompokkan teks berdasarkan persamaan dengan menggunakan konsep jarak. Kajian ini mendedahkan bahawa pendekatan LDA adalah praktikal dan dapat digunakan untuk pembaca, penyelidik dan pembuat dasar untuk lebih memahami situasi semasa berdasarkan berita ekonomi dan sosial dalam talian di Malaysia.

Kata kunci: Analisis Pengelompokan, Latent Dirichlet Allocation, Perlombongan Teks

ABSTRACT

Evolution of online news captivates the social and political landscape in Malaysia. Thus, it is an opportunity to understand the main issues arising from the online news content, which affects the current socio-economy status in Malaysia. This paper therefore aims to explain how unstructured data in text type can be used to discover the clusters and topics that occur from a collection of documents through cluster analysis and Latent Dirichlet Allocation (LDA). Text mining is the process of exploring the collections of text and converting the format from unstructured text data into structured data for further analysis while text cluster analysis aims to group text based on their similarity by using the distance concept. This study revealed that the LDA approach is practical and can be used for public readers, researcher and policy makers to have better understanding on current situation created by economic and social online news in Malaysia.

Keywords: Cluster Analysis, Latent Dirichlet Allocation (LDA), Text Mining

[1] Faiza Rusrianti Tajul Arus is currently Senior Assistant Director of Core Team Big Data Analytics (CTADR), Department of Statistics Malaysia.

1. INTRODUCTION

News are information broadcast or written in media to deliver latest updates of happenings around the world. Public interest has moved from reading news on printed newspaper to scrolling for news online (Makaruddin, 2018). Delivering accountable and responsible news is the main role for media organisations. The news is crucial for a country's development by covering and reporting economics, political and social issues. Digital media and its availability are borderless. A new environment is created by communication is able to disperse widely. The news content that is published online shows a big consequence in changing the minds of people based on the influence of the content that they read online which is finally translated into the way people think and thus changing the socioeconomic and political landscape (Alivi et al., 2018). It is important for the media to maintain their credibility as they play a vital role in disseminating information.

Online news plays a significant role in attracting the younger generation in Malaysia by its characteristics which are flexible, accessible and easy to access for search information (Tiung et al., 2016). Media practitioners play a tremendous part in forming the minds of the people through media power. It is undeniable that news plays a vital role in delivering information to Malaysians. Nevertheless, priority should be given to quality of information to increase the people's trust. Nurwidyantoro (2016) suggest that online news has the ability to influence public's readers on the opinion of an article, especially economic news that have direct impact in society such as forecast, current economic situation and simple analysis, such as stock market trends for future investments, banking sector development and effectiveness of the national economic policy implementation.

Text analytics is the way to cover patterns and meaningful themes from unstructured text. This analysis refers to a discipline of computer science that combines machine learning and natural language processing (NLP) to draw meaning from unstructured text documents. Ordenes et al. (2014) revealed that text mining is the process of analyzing textual information in an attempt to discover structure and implicit meanings ‘‘hidden’’ within texts.

Text cluster analysis aims to group text based on their similarity by using the distance concept. Allahyari et al. (2017) described that the clustering is the task of finding groups of similar documents in a collection of documents. The similarity is computed by using a similarity function. Text clustering can be in different levels of granularities where clusters can be documents, paragraphs, sentences or terms.

Generally, this study aims to identify and define the type on unemployment for the year of 2019 based on online news reporting in Malaysia. The main objective of this study is to identify the most key word used by online media practitioner for unemployment news in Malaysia. There are two specific objectives in this study. The first objective is to identify the best method should be used for analysis the number of clusters for unemployment news and the second objective is to identify type of unemployment in Malaysia based on group of words from online news on unemployment by using Latent Dirichlet Allocation.

2. LITERATURE REVIEW AND RESEARCH METHODS

This study used data from online news on unemployment in English language in between January 2019 to December 2019. The news was randomly taken by key word jobless and employment and the keywords were updated according to the frequencies result from 50 articles or documents. The list of new keywords is shown in Table 1:

Table 1: List of Built-In Key Word for Topic Unemployment

No.	Words	No.	Words
1.	jobless	9.	lose jobs
2.	unemployment	10.	laid off
3.	jobs	11.	lay off
4.	job skill	12.	retrenchment
5.	lost jobs	13.	students
6.	critical jobs	14.	talent
7.	graduated	15.	experience
8.	skill job		

The news was extracted manually with full text versions from unstructured to structured format includes the topic, date of news and the statement. A total of 350 news from 20 Media Organisations were used for this analysis namely The Edge Market, The Malay Mail Online, MSN Malaysia, The Malaysian Reserve, The Sun Daily, The Star Online, Free Malaysia Today, The Borneo Post, Yahoo News Malaysia, Daily Express, Channel News Asia, Yahoo News Singapore, New Sabah Times, The Straits Times, Malaysia Kini, The Malaysian Times, Business Insider, The True Net, Business Today dan The Mole.

In this study, several required packages for text mining are loaded in the R environment. Tseng et al. (2007) points out that text mining process involve text mining tools for user to explore the repository to find patterns. A folder contains the selected media was converted into a collection of text documents or corpus. The common pre-processing text was done and then converted into a matrix or known as document term matrix. The matrix is transpose to term document matrix for language analysis purpose. Kadhim et al. (2014) described that matrix as a vector space whose components are that features and their weights which are obtained by the frequency of each feature in that text document.

Text clustering is used to identify similar word groups, based on frequency distance. The *k*-means is also used to group similar data points together and discover underlying patterns or similarities, i.e. grouping the documents into related clusters. Hartigan & Wong (1979) highlights the aim of the K-Means algorithm is to split *M* point in *N* dimensions into *K* clusters so that the within-cluster sum of squares is minimized. *K*-Means clustering is an unsupervised learning method that used the Euclidean distance as measurement of *K* group of similarity. Žalik (2008) found that *k*-Means is attractive in practice, because it is simple, and it is generally very fast. Likas et al. (2003) claims that *k*-Means algorithm is a popular clustering method that lessens the clustering error. The clustering process were divided into five stages as illustrated in figure 1.

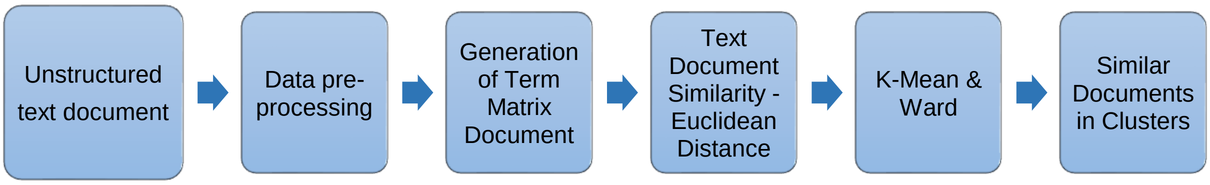
Unstructured text document
Data pre-processing
Generation of Term Matrix Document
Text Document Similarity - Euclidean Distance
K-Mean & Ward
Similar Documents in Clusters

Figure 1: The Five Stages of Clustering Process

The clustering process transform the raw text by data pre-processing such as removal comma, stop words, numbers and other; the creation of Term Document Matrix; TF-IDF (Term Frequency-Inverse Document Frequency); normalization; using Euclidean distance and clustering algorithm based on Cluster Centers. TF-IDF is used as a weight for terms frequently (tf). The weight is a statistical measure used to calculate how important a word is to a document in a corpus. The term's inverse document frequency (idf) are based on weight for frequently used words by decreased the calculation. The combination of tf-idf (the two quantities multiplied together) calculate the frequency of a term adjusted by how rarely it is used. The tf-idf is projected measurement on significant of each word to a document in a collection (or corpus) of documents. The inverse document frequency for any given term is defined as the logarithm of the number of the document in the corpus divided by the number of documents where the specific term appears.

Topic Models is a very useful resolution for document clustering, organizing large blocks of textual data, information retrieval from unstructured text and feature selection. Latent Dirichlet Allocation (LDA) is the most common used topic modelling technique for finding suitable topics from a text data. LDA is a method to automatically discovering topics for a set of statements. Blei et al. (2003) has defined that LDA is a generative probabilistic model for collections of discrete data such as text corpora. LDA is part of Bayesian model with three-level hierarchical, in which each item of a collection is modelled as a finite mixture over an underlying set of topics. LDA is used as a statistical model to discover the topics that occur in a collection of documents.

Liu (2013) expressed that LDA model is one of the most popular probability topic models, and it has more comprehensive assumptions of text generation than other models. LDA is a matrix factorization technique with assumption of each document are forms by mixture of topics. Those topics then create words probability based on their distribution. The composition of documents is based on the amount of topic. In the same vein, Krestel et al. (2009) expresses LDA helps to explain the similarity of data by grouping features of this data into unobserved sets.

As noted by Wang et al. (2012) LDA is more effective methods for classifying, clustering and retrieving textual data because LDA were developed for huge document. This method is particularly useful in studying this online news since the news are lengthy as compared to social media such as Twitter or Facebook, the textual comes in short fragments. Considering all of this evidence, it seems that LDA is the most competent method to be used in this study.

3. RESULT

Table 2 shows the findings on number of clusters by few methods of clustering analysis for unemployment news. All the methods were resulting in three (3) cluster excluding the Silhoutte methods. The details are as follows:

Table 2: The Method of Clustering Analysis and the Optimal Number of Cluster

No.	Method of Analysis	No. of Cluster
1.	Elbow	3
2.	Silhouette	4
3.	Dendogram	3
4.	Clusplot	3
5.	Latern Dirichlet Allocation	3

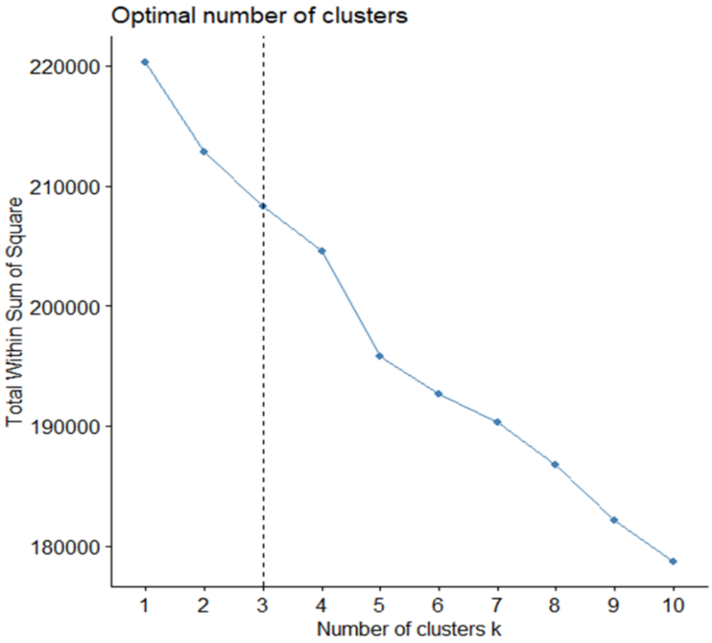

Figure 2: The Optimal Number of Cluster by Using Elbow Method

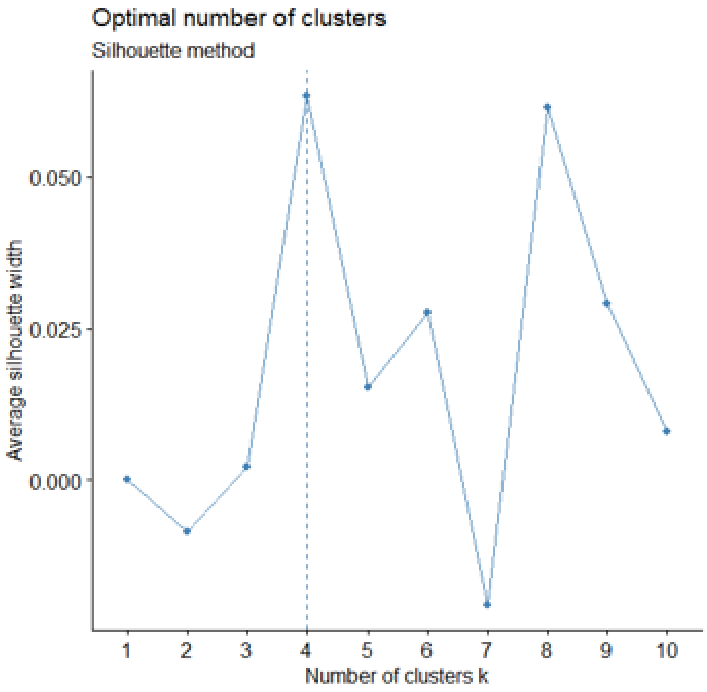

Figure 3: The Optimal Number of Cluster by Using Silhouette Method

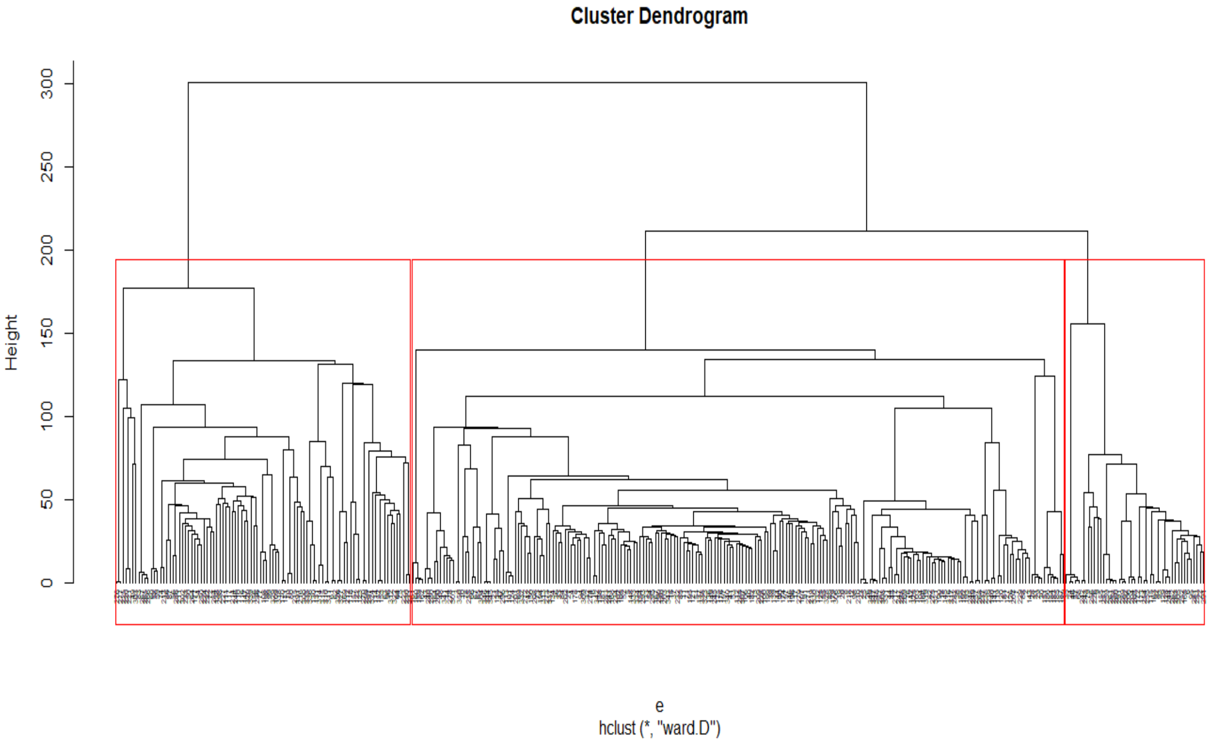

Figure 4: The Optimal Number of Cluster by Using Dendrogram Method

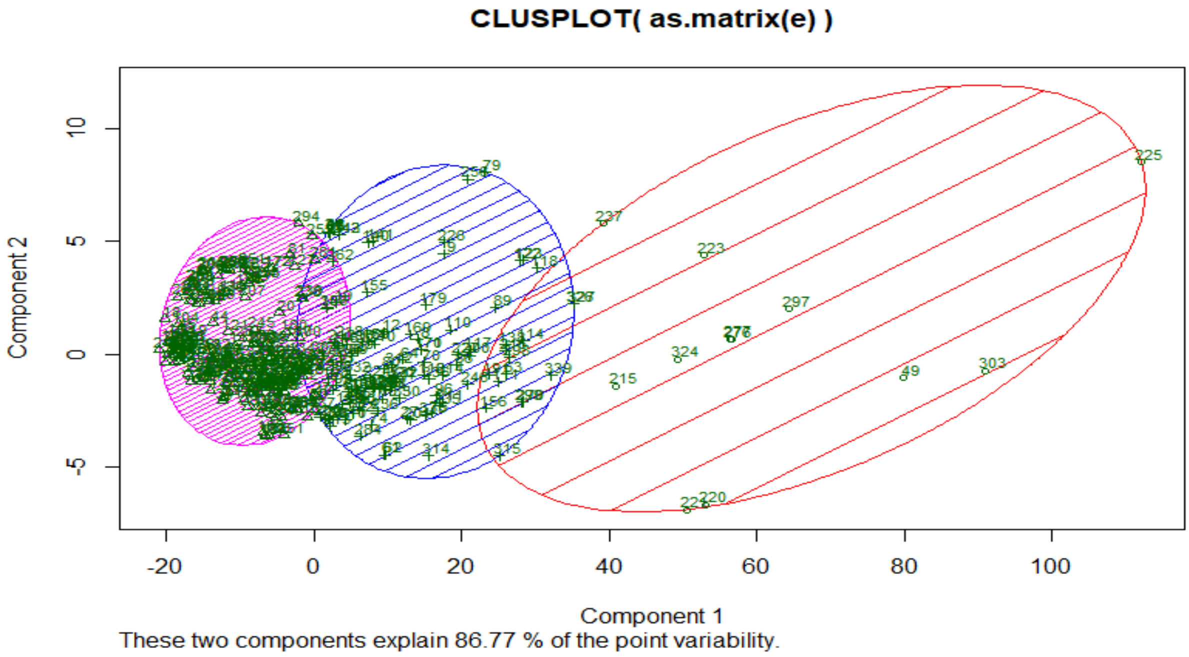

Figure 5: The Optimal Number of Cluster by Using Clusplot Method

Method 1 to 4 did not show clear cluster result while LDA states that each news or document in a corpus is a group of a fixed number of cluster topics. Each document has a probability of producing numerous words based on the observed words in the corpus. The result are as follows:

Table 3: The Clustering Probability for Each Document

No.	Cluster 1	Cluster 2	Cluster 3	No.	Cluster 1	Cluster 2	Cluster 3
1	0.53	0.22	0.25	40	0.43	0.30	0.26
2	0.15	0.74	0.11	41	0.32	0.21	0.46
3	0.73	0.11	0.16	42	0.57	0.28	0.15
4	0.56	0.20	0.24	43	0.75	0.10	0.15
5	0.52	0.19	0.29	44	0.20	0.62	0.19
6	0.20	0.64	0.16	45	0.22	0.58	0.19
7	0.15	0.69	0.16	46	0.08	0.84	0.08
8	0.41	0.18	0.40	47	0.29	0.27	0.44
9	0.24	0.54	0.22	48	0.63	0.17	0.20
10	0.74	0.18	0.08	49	0.56	0.12	0.32
11	0.27	0.17	0.56	50	0.28	0.16	0.56
12	0.18	0.33	0.49	51	0.29	0.22	0.49
13	0.52	0.27	0.21	52	0.42	0.20	0.38
14	0.63	0.16	0.21	53	0.47	0.23	0.29
15	0.52	0.16	0.32	54	0.33	0.14	0.52
16	0.42	0.22	0.36	55	0.55	0.22	0.23
17	0.40	0.22	0.39	56	0.19	0.15	0.66
18	0.24	0.51	0.25	57	0.08	0.83	0.09
19	0.16	0.69	0.15	58	0.17	0.06	0.77

No.	Cluster 1	Cluster 2	Cluster 3	No.	Cluster 1	Cluster 2	Cluster 3
20	0.14	0.68	0.18	59	0.15	0.09	0.76
21	0.19	0.28	0.52	60	0.15	0.08	0.77
22	0.18	0.66	0.16	61	0.68	0.14	0.19
23	0.58	0.22	0.19	62	0.68	0.12	0.19
24	0.08	0.82	0.10	63	0.53	0.15	0.31
25	0.60	0.22	0.18	64	0.32	0.18	0.50
26	0.59	0.22	0.19	65	0.74	0.13	0.13
27	0.42	0.17	0.42	66	0.78	0.14	0.08
28	0.81	0.08	0.11	67	0.78	0.14	0.08
29	0.82	0.08	0.09	68	0.09	0.83	0.09
30	0.81	0.09	0.10	69	0.35	0.34	0.31
31	0.84	0.08	0.08	70	0.11	0.12	0.77
32	0.57	0.13	0.31	71	0.17	0.15	0.68
33	0.20	0.21	0.59	72	0.16	0.19	0.65
34	0.22	0.20	0.58	73	0.57	0.19	0.24
35	0.09	0.82	0.09	74	0.33	0.10	0.57
36	0.29	0.17	0.54	75	0.41	0.18	0.41
37	0.42	0.27	0.31	76	0.46	0.28	0.26
38	0.47	0.26	0.26	77	0.09	0.16	0.75
39	0.52	0.27	0.20	78	0.72	0.16	0.12

Cont.

No.	Cluster 1	Cluster 2	Cluster 3	No.	Cluster 1	Cluster 2	Cluster 3
79	0.10	0.78	0.12	124	0.71	0.15	0.14
80	0.11	0.11	0.78	125	0.44	0.30	0.26
81	0.14	0.74	0.12	126	0.11	0.78	0.11
82	0.33	0.20	0.47	127	0.15	0.72	0.13
83	0.19	0.08	0.73	128	0.16	0.68	0.16
84	0.29	0.17	0.54	129	0.17	0.65	0.18
85	0.23	0.22	0.55	130	0.79	0.12	0.08
86	0.16	0.14	0.70	131	0.81	0.11	0.08
87	0.22	0.16	0.61	132	0.55	0.19	0.26
88	0.14	0.72	0.14	133	0.52	0.17	0.31
89	0.53	0.34	0.14	134	0.50	0.30	0.20
90	0.40	0.16	0.44	135	0.19	0.22	0.59
91	0.36	0.17	0.48	136	0.27	0.25	0.48
92	0.18	0.66	0.16	137	0.26	0.15	0.59
93	0.19	0.63	0.18	138	0.18	0.22	0.59
94	0.57	0.12	0.30	139	0.34	0.48	0.17
95	0.21	0.68	0.12	140	0.07	0.81	0.11
96	0.34	0.17	0.49	141	0.06	0.83	0.10
97	0.48	0.27	0.25	142	0.53	0.24	0.22
98	0.23	0.31	0.46	143	0.14	0.71	0.14
99	0.16	0.14	0.71	144	0.40	0.46	0.14
100	0.56	0.22	0.22	145	0.31	0.33	0.36
101	0.69	0.18	0.13	146	0.36	0.22	0.42

No.	Cluster 1	Cluster 2	Cluster 3	No.	Cluster 1	Cluster 2	Cluster 3
102	0.44	0.18	0.39	147	0.26	0.11	0.63
103	0.45	0.32	0.23	148	0.44	0.22	0.35
104	0.36	0.42	0.21	149	0.40	0.26	0.34
105	0.40	0.14	0.46	150	0.53	0.25	0.23
106	0.43	0.24	0.33	151	0.83	0.06	0.11
107	0.29	0.17	0.55	152	0.54	0.32	0.14
108	0.10	0.13	0.77	153	0.58	0.28	0.13
109	0.18	0.13	0.69	154	0.38	0.20	0.42
110	0.23	0.48	0.28	155	0.21	0.49	0.30
111	0.24	0.16	0.60	156	0.48	0.14	0.39
112	0.37	0.43	0.20	157	0.09	0.76	0.14
113	0.70	0.16	0.14	158	0.27	0.26	0.47
114	0.15	0.17	0.67	159	0.16	0.14	0.69
115	0.10	0.78	0.13	160	0.16	0.14	0.70
116	0.59	0.18	0.23	161	0.19	0.13	0.68
117	0.18	0.22	0.60	162	0.18	0.12	0.70
118	0.19	0.67	0.14	163	0.10	0.79	0.11
119	0.09	0.77	0.13	164	0.63	0.10	0.27
120	0.10	0.76	0.14	165	0.17	0.71	0.13
121	0.12	0.53	0.35	166	0.27	0.25	0.47
122	0.07	0.81	0.12	167	0.18	0.33	0.49
123	0.07	0.79	0.14	168	0.49	0.16	0.36

Cont.

No.	Cluster 1	Cluster 2	Cluster 3	No.	Cluster 1	Cluster 2	Cluster 3
169	0.29	0.53	0.18	214	0.21	0.21	0.58
170	0.10	0.13	0.77	215	0.12	0.11	0.77
171	0.08	0.12	0.81	216	0.13	0.17	0.70
172	0.10	0.75	0.15	217	0.36	0.20	0.45
173	0.12	0.72	0.16	218	0.31	0.32	0.36
174	0.46	0.16	0.38	219	0.11	0.13	0.76
175	0.23	0.33	0.44	220	0.39	0.11	0.50
176	0.27	0.42	0.31	221	0.40	0.08	0.52
177	0.26	0.18	0.57	222	0.24	0.11	0.65
178	0.31	0.43	0.27	223	0.16	0.41	0.43
179	0.13	0.65	0.22	224	0.19	0.20	0.61
180	0.27	0.46	0.27	225	0.13	0.11	0.76
181	0.31	0.13	0.55	226	0.13	0.64	0.23
182	0.42	0.34	0.25	227	0.23	0.16	0.61
183	0.84	0.06	0.10	228	0.37	0.15	0.48
184	0.80	0.07	0.14	229	0.19	0.66	0.15
185	0.83	0.07	0.10	230	0.60	0.15	0.25
186	0.85	0.06	0.10	231	0.59	0.13	0.27
187	0.71	0.12	0.18	232	0.43	0.32	0.25
188	0.63	0.17	0.20	233	0.34	0.13	0.53
189	0.73	0.09	0.18	234	0.55	0.16	0.29

No.	Cluster 1	Cluster 2	Cluster 3	No.	Cluster 1	Cluster 2	Cluster 3
190	0.38	0.22	0.40	235	0.71	0.14	0.15
191	0.21	0.16	0.64	236	0.51	0.10	0.39
192	0.20	0.18	0.63	237	0.11	0.79	0.10
193	0.24	0.18	0.59	238	0.11	0.75	0.13
194	0.57	0.17	0.25	239	0.10	0.76	0.15
195	0.19	0.15	0.66	240	0.74	0.09	0.17
196	0.35	0.35	0.30	241	0.47	0.16	0.36
197	0.49	0.26	0.25	242	0.10	0.80	0.10
198	0.12	0.14	0.74	243	0.11	0.80	0.10
199	0.55	0.16	0.29	244	0.14	0.73	0.13
200	0.66	0.09	0.25	245	0.15	0.53	0.32
201	0.51	0.13	0.36	246	0.20	0.12	0.68
202	0.33	0.13	0.54	247	0.32	0.21	0.47
203	0.43	0.23	0.34	248	0.36	0.24	0.40
204	0.15	0.75	0.10	249	0.38	0.22	0.40
205	0.18	0.74	0.09	250	0.19	0.13	0.67
206	0.18	0.72	0.10	251	0.11	0.79	0.09
207	0.10	0.78	0.12	252	0.10	0.78	0.12
208	0.47	0.17	0.36	253	0.22	0.66	0.12
209	0.49	0.21	0.30	254	0.09	0.78	0.12
210	0.24	0.38	0.38	255	0.10	0.79	0.11
211	0.18	0.08	0.73	256	0.08	0.83	0.08
212	0.25	0.56	0.19	257	0.33	0.27	0.40
213	0.36	0.10	0.54	258	0.13	0.74	0.13

Cont.

No.	Cluster 1	Cluster 2	Cluster 3	No.	Cluster 1	Cluster 2	Cluster 3
259	0.11	0.75	0.14	304	0.67	0.16	0.17
260	0.13	0.75	0.12	305	0.10	0.73	0.17
261	0.13	0.75	0.13	306	0.14	0.69	0.17
262	0.15	0.69	0.17	307	0.38	0.24	0.38
263	0.15	0.64	0.20	308	0.49	0.23	0.28
264	0.26	0.18	0.56	309	0.53	0.19	0.29
265	0.18	0.08	0.74	310	0.52	0.24	0.24
266	0.18	0.08	0.74	311	0.54	0.24	0.22
267	0.20	0.08	0.72	312	0.55	0.19	0.27
268	0.37	0.22	0.41	313	0.62	0.15	0.24
269	0.29	0.21	0.50	314	0.63	0.20	0.17
270	0.58	0.14	0.29	315	0.64	0.19	0.17
271	0.28	0.16	0.56	316	0.25	0.24	0.50
272	0.57	0.25	0.17	317	0.29	0.18	0.54
273	0.30	0.21	0.49	318	0.16	0.15	0.69
274	0.71	0.11	0.19	319	0.20	0.28	0.52
275	0.37	0.29	0.34	320	0.19	0.11	0.70
276	0.62	0.17	0.21	321	0.44	0.25	0.31
277	0.61	0.17	0.23	322	0.38	0.21	0.41

No.	Cluster 1	Cluster 2	Cluster 3	No.	Cluster 1	Cluster 2	Cluster 3
278	0.62	0.16	0.22	323	0.50	0.24	0.26
279	0.21	0.24	0.55	324	0.51	0.18	0.31
280	0.74	0.08	0.18	325	0.23	0.53	0.24
281	0.12	0.80	0.09	326	0.24	0.58	0.18
282	0.38	0.32	0.30	327	0.23	0.59	0.18
283	0.30	0.15	0.55	328	0.35	0.33	0.33
284	0.26	0.13	0.61	329	0.17	0.12	0.71
285	0.34	0.25	0.41	330	0.18	0.15	0.67
286	0.19	0.13	0.68	331	0.44	0.28	0.28
287	0.43	0.30	0.27	332	0.39	0.35	0.26
288	0.49	0.27	0.23	333	0.62	0.15	0.23
289	0.16	0.71	0.13	334	0.34	0.18	0.49
290	0.46	0.24	0.30	335	0.65	0.17	0.18
291	0.15	0.74	0.12	336	0.73	0.13	0.14
292	0.14	0.73	0.13	337	0.54	0.21	0.25
293	0.41	0.20	0.39	338	0.21	0.18	0.61
294	0.11	0.77	0.11	339	0.13	0.07	0.79
295	0.69	0.14	0.18	340	0.55	0.26	0.18
296	0.25	0.25	0.50	341	0.63	0.12	0.25
297	0.60	0.12	0.28	342	0.49	0.17	0.34
298	0.18	0.14	0.67	343	0.55	0.20	0.25
299	0.58	0.14	0.28	344	0.20	0.22	0.58
300	0.08	0.12	0.79	345	0.21	0.21	0.58
301	0.60	0.17	0.23	346	0.21	0.22	0.57
302	0.27	0.10	0.64	347	0.62	0.23	0.15
303	0.35	0.09	0.56	348	0.60	0.24	0.15
349	0.58	0.25	0.17	350	0.71	0.11	0.18

Cont.

A total of 355 probability result were shown in Table 3. Among the result, five (5) documents were sharing equal result namely document 27, 75, 196, 210 and 307. Most documents were in Cluster 1(140) follow by Cluster 3 (126) and Cluster 2 (89). Hence, the results showed LDA is the best method for finding the best topic in a group of words from a collection of documents.

Levine (2013) highlighted that Economist often break down unemployment into three types: frictional, structural, and cyclical. The details are as follows:

- Frictional unemployment arises from the ever-present movement of people into and out of jobs. Examples the unemployment of college graduates while searching for their first jobs;
- Structural unemployment arises from obstacles to the worker-to-job-matching process that lengthen unemployment spells; and

- Cyclical unemployment arises when the economy experiences a decrease in the demand for goods and services. Employers adjust to a downturn in the business cycle by temporarily laying off workers and cutting the hours of employees retained to fill reduced product demand.

The key words frequently used by reporters for the topic of unemployment are jobs, workers job, new, graduates, year, skills, employees, government, utusan, can, work, industry, employers an economy. Table 4 and Figure 6 showed the details frequency of key words that had been used by 350 news in this topic.

Table 4: The Frequency of Key Words Used

No.	Key Words	Frequency
1.	Jobs	775
2.	Workers	548
3.	Job	497
4.	New	444
5.	Graduates	439
6.	Year	430
7.	Skills	427
8.	Employees	410
9.	Government	410
10.	Utusan	371
11.	Can	360
12.	Work	345
13.	Industry	336
14.	Employers	333
15.	Economy	324

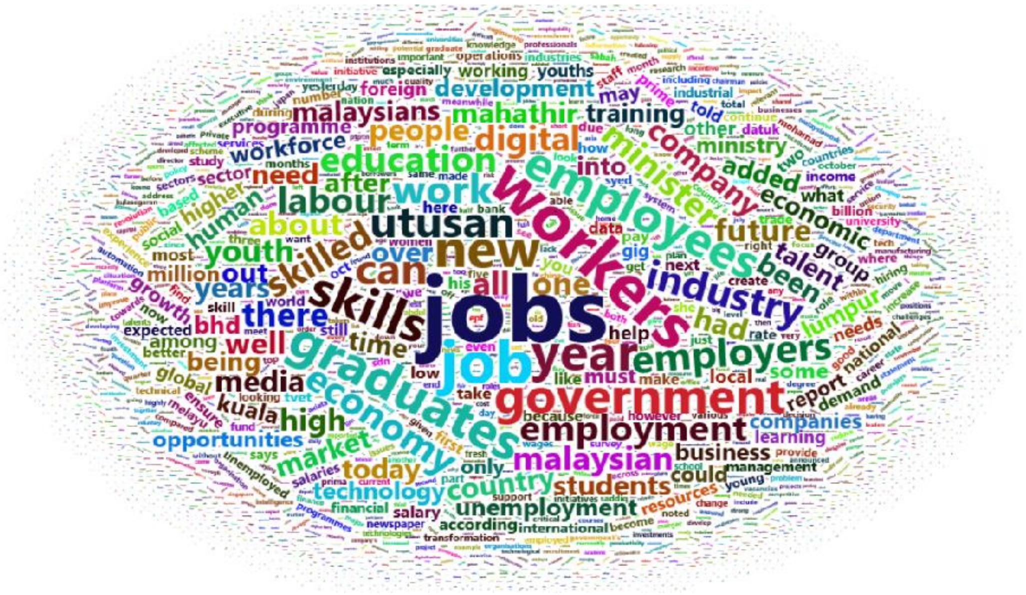

Figure 6: The Word Cloud on Unemployment

This study concluded that most of the media reporting in 2019 was related to Frictional news with total of 140 documents (39 per cent) and the main issue was the need of relevant parties to study the relevant of subjects offered by university rather than the skills required by industry players. Structural news (126 documents or 36 per cent) with major issues was the latest skill required for the preparation of Industrial Revolution 4.0 and Big Data Analytics while Cyclical news at 25 per cent or 89 documents with the major issues reported were the lay-off from Utusan Staff and SOCSO's initiative in developing the System EIS (Employment Insurance System) in assisting workers seeking a new job. Table 5 showed the details key words for each unemployment type in this study.

Table 5: The Key Words by Unemployment Type

Cluster 1 Frictional	Cluster 2 Cyclical	Cluster 3 Structural
graduates	utusan	skills
skilled	employees	digital
youth	media	future
education	workers	students
gig	pay	talent
youths	trade	development
tvet	newspaper	learning
	salary	demand
	off	higher
	union	skill
		high
		industries
		xxperience
		career

4. DISCUSSION AND CONCLUSION

The aim of this study is to formalise the use of news online as source of analysis in understanding the current scenario in Malaysia. In this study, the removal of the unnecessary words was done by using text mining technique. Meanwhile, cluster analysis successfully grouped the key words into the three types of unemployment. This finding suggests that the accurate results can be obtained by using statistical analysis such as data cleaning in text mining, cluster analysis for grouping the documents based on distance and similarity by Euclidean and the used of LDA for topic models.

The current study found that LDA is the best statistical method in identifying key words from a collection of documents as compared to others method in finding details for clustering. Based on the analysis, appropriate news content can be drawn in order to discover the current situation in the country.

Comparing to four others method, the LDA will give clear distance for each document by using probability measurement. This method also can illustrate clear words frequently appear in each cluster and this word can be used for further analysis in determining the theme of the topic. The Dendogram and Clusplot can be used if the documents are small in size since the result are also showing the documents contains in each cluster.

Based on the news reports in Malaysia, the most of unemployment for 2019 is Frictional. The type of unemployment that arises from the ever-present movement of people into and out of jobs. This assumption is based on keywords identified for this cluster namely graduates, skilled, youth, education, gig, youths and tvet. The type of unemployment in Malaysia is then followed by Structural and Cyclical respectively with keywords as stated in Table 5.

However, the limitations of this study are the analysis did not include other mainstream media such as in Bahasa Malaysia or Mandarin. Hence, the result obtained in this study may not be appropriate to the overall news on unemployment in Malaysia.

In future investigations, it might be possible to use a different topic in which online news, can be used to examine different dimensionalities of economic and social news and its result for analysing and better understanding on the current situation in Malaysia. This also to use as comparisons, the capabilities of text data are as good as numerical data in decision and policy making.

REFERENCES

Alivi, M. A., Ghazali, A. H. A., Tamam, E., & Osman, M. N. (2018). A review of new media in Malaysia: Issues affecting society. *International Journal of Academic Research in Business and Social Sciences, 8*(2), 12-29.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. arXiv preprint arXiv:1707.02919.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3*(Jan), 993-1022.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics), 28*(1), 100-108.

Kadhim, A. I., Cheah, Y. N., & Ahamed, N. H. (2014, December). Text document preprocessing and dimension reduction techniques for text document clustering. In 2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology, 69-73. *IEEE.*

Krestel, R., Fankhauser, P., & Nejdl, W. (2009, October). Latent dirichlet allocation for tag recommendation. *In Proceedings of the third ACM conference on Recommender systems*, 61-68.

Levine, L. (2013). The increase in unemployment since 2007: Is it cyclical or structural?

Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition, 36*(2), 451-461.

Liu, Z. (2013). High performance latent dirichlet allocation for text mining (Doctoral dissertation, Brunel University School of Engineering and Design PhD Theses).

Makaruddin, M. H. A. (2018). The future of the newspaper industry in Malaysia in the era of global media and global culture. *Jurnal Komunikasi: Malaysian Journal of Communication, 22*.

Nurwidyantoro, A. (2016, October). Sentiment analysis of economic news in Bahasa Indonesia using majority vote classifier. In 2016 International Conference on Data and Software Engineering (ICoDSE), 1-6. *IEEE*.

Ordenes, F. V., Theodoulidis, B., Burton, J., Gruber, T., & Zaki, M. (2014). Analyzing customer experience feedback using text mining: A linguistics-based approach. *Journal of Service Research, 17*(3), 278-295.

Tiung, L. K., Meri, A., Nayan, L. M., & Othman, S. S. (2016). Kegunaan dan kepuasan portal berita dalam kalangan belia Malaysia. *Jurnal Komunikasi: Malaysian Journal of Communication, 32*(2).

Tseng, Y. H., Lin, C. J., & Lin, Y. I. (2007). Text mining techniques for patent analysis. *Information processing & management, 43*(5), 1216-1247.

Wang, Y., Agichtein, E., & Benzi, M. (2012, August). TM-LDA: efficient online modeling of latent topic transitions in social media. *In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 123-131.

Žalik, K. R. (2008). An efficient k′-means clustering algorithm. *Pattern Recognition Letters, 29*(9), 1385-1391.