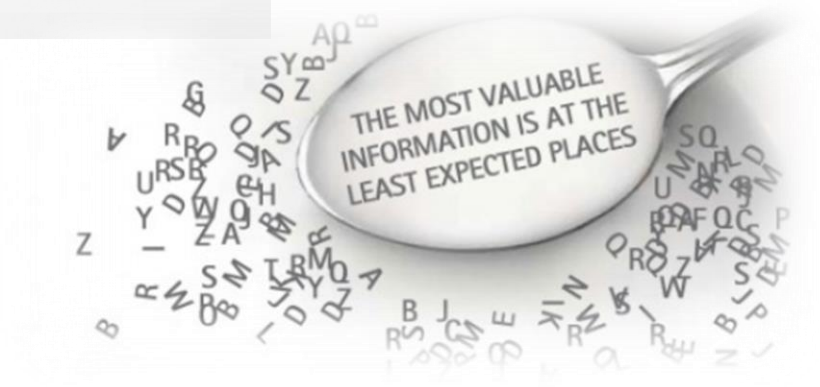
Identifying The Key Words Of Online Economic News In Malaysia Through Text Analytics

UNIVERSITI KEBANGSAAN MALAYSIA
The National University of Malaysia

Jabatan Perangkaan MALAYSIA

STATSBDA BIG DATA ANALYTICS



By:

Faiza Rusrianti Bt Tajul Arus

Capstone Project

School of Mathematical Sciences, UKM

Presentation Outline

1. Introduction
2. Literature Review
3. Methodology
4. Data Collection
5. Statistical Analysis
6. Result
7. Conclusion

1. Introduction

01

Social media and electronic media play a vital role as a main reference for publics to gather information.

02

The rapid development of electronic media provides organizations with the opportunity to measure perceptions more quickly and transparently by implementing mediums such as a dedicated portal for monitoring online media reporting.

03

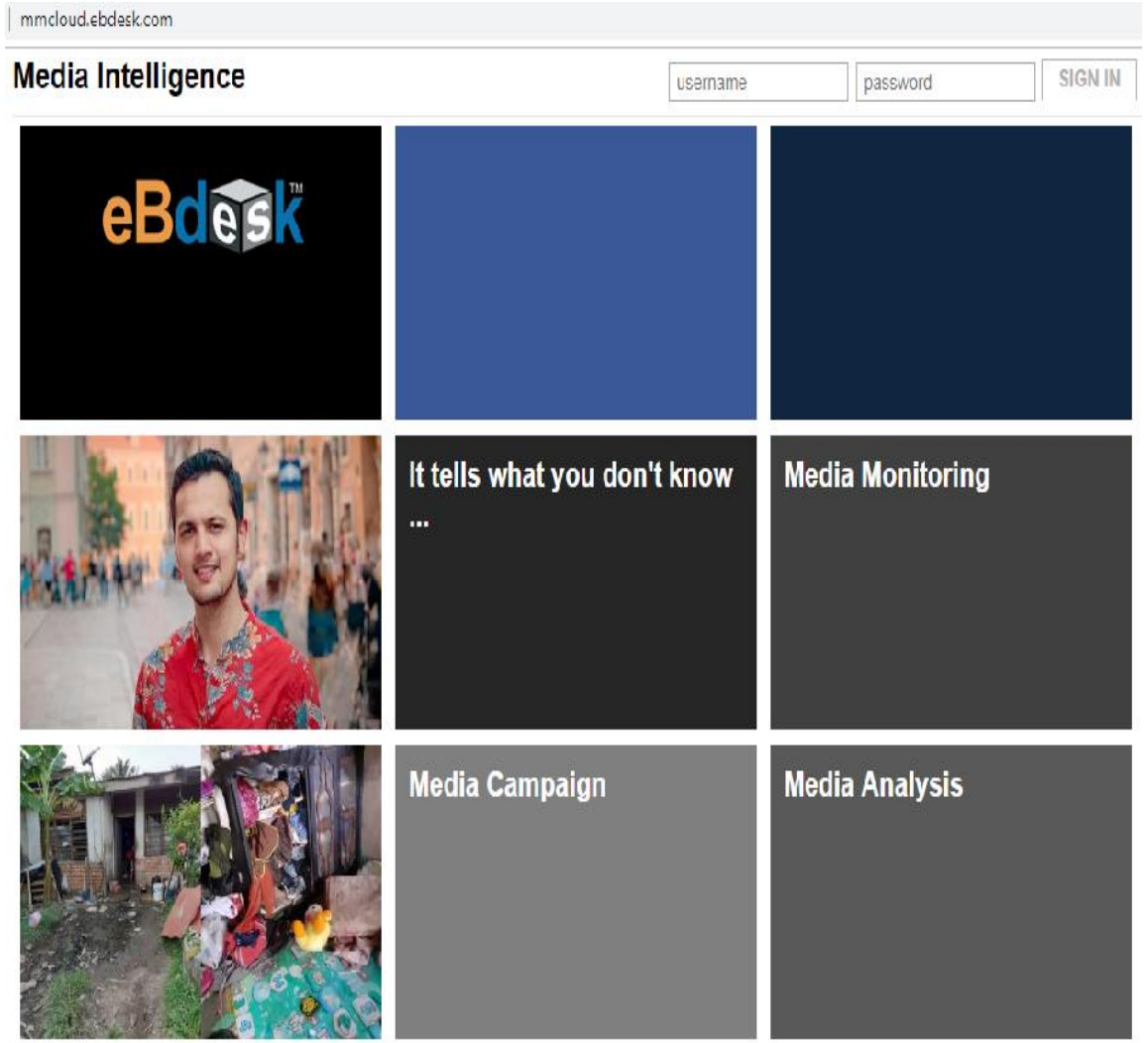
Agarwal et al. (2016) identified that news mining based on sentiment analysis could be analysed and alkhwarizmi test can be use to classified news statements with positive or negative sentiments that could help the government in decision making process.

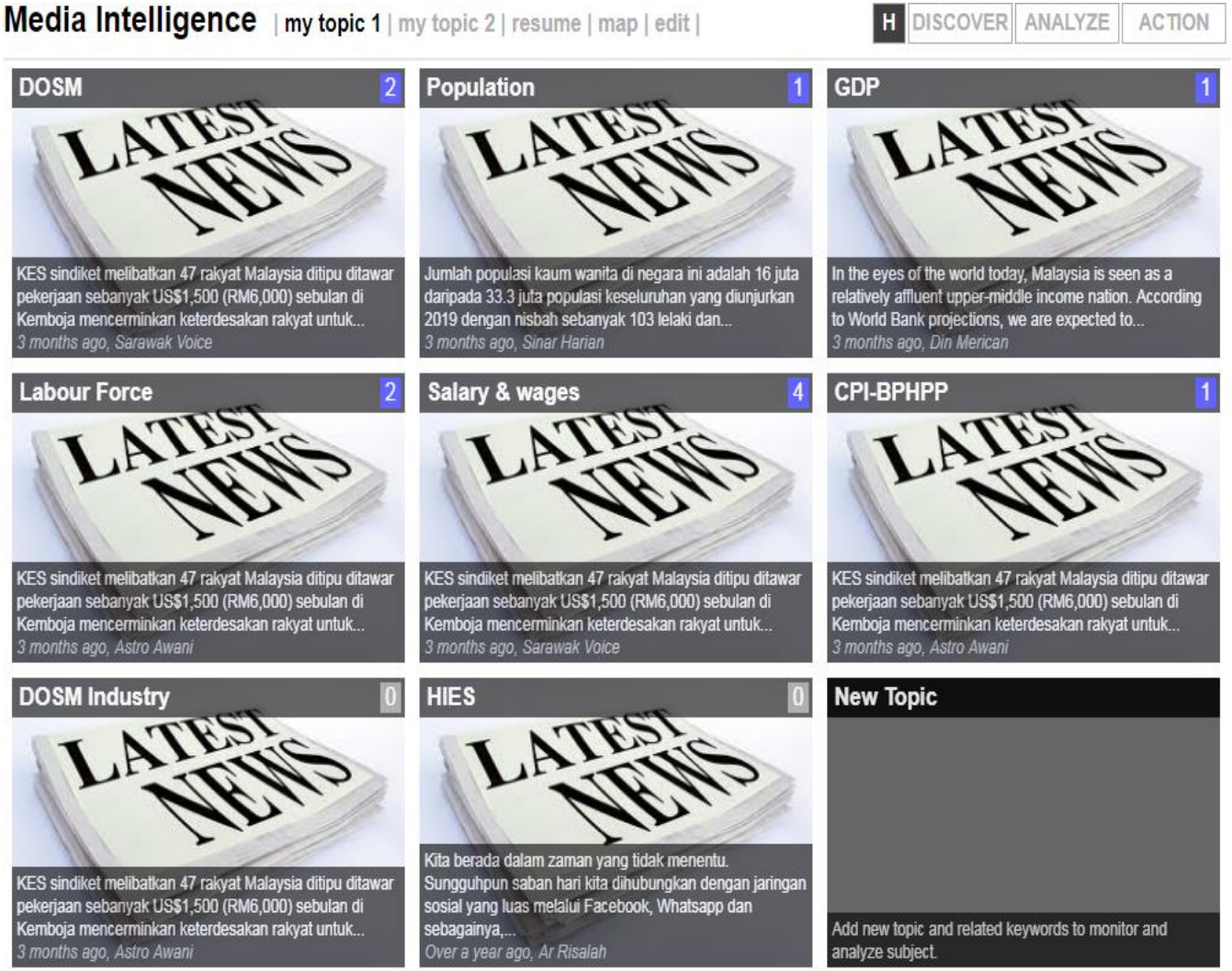
04

In order to monitoring online economic news is in line with official statistical released and at par with national policy, DOSM has embarked on a data project initiative known as Statistics Big Data Analytics or STATSBDA.

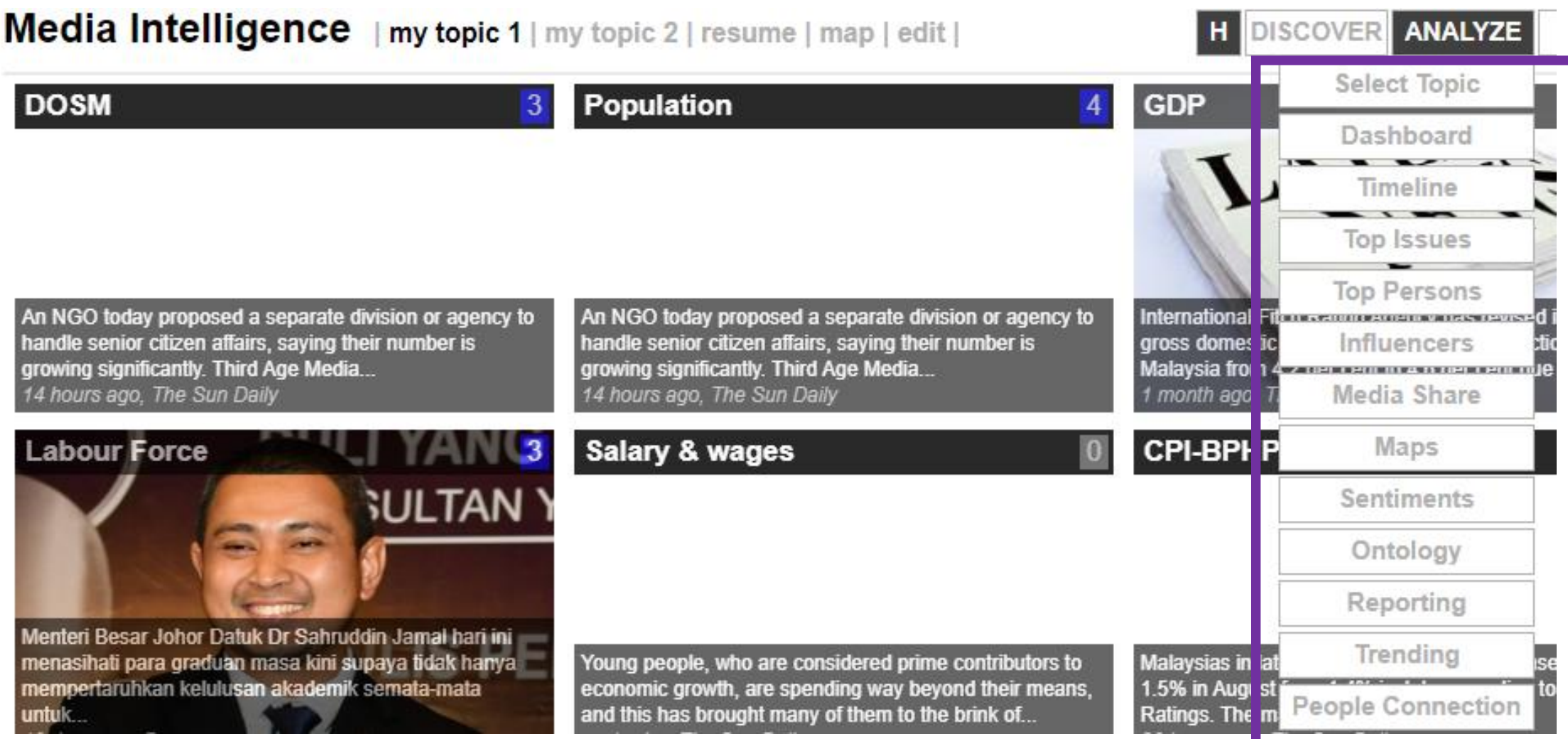
Public Maturity Assessment on Official Statistics (PMAOS) and Real Time News on Official Statistics (RTOS) were developed to provide DOSM with holistic and comprehensive insight of perception developed by media.

At a glance of PMAOS & RTOS Analysis

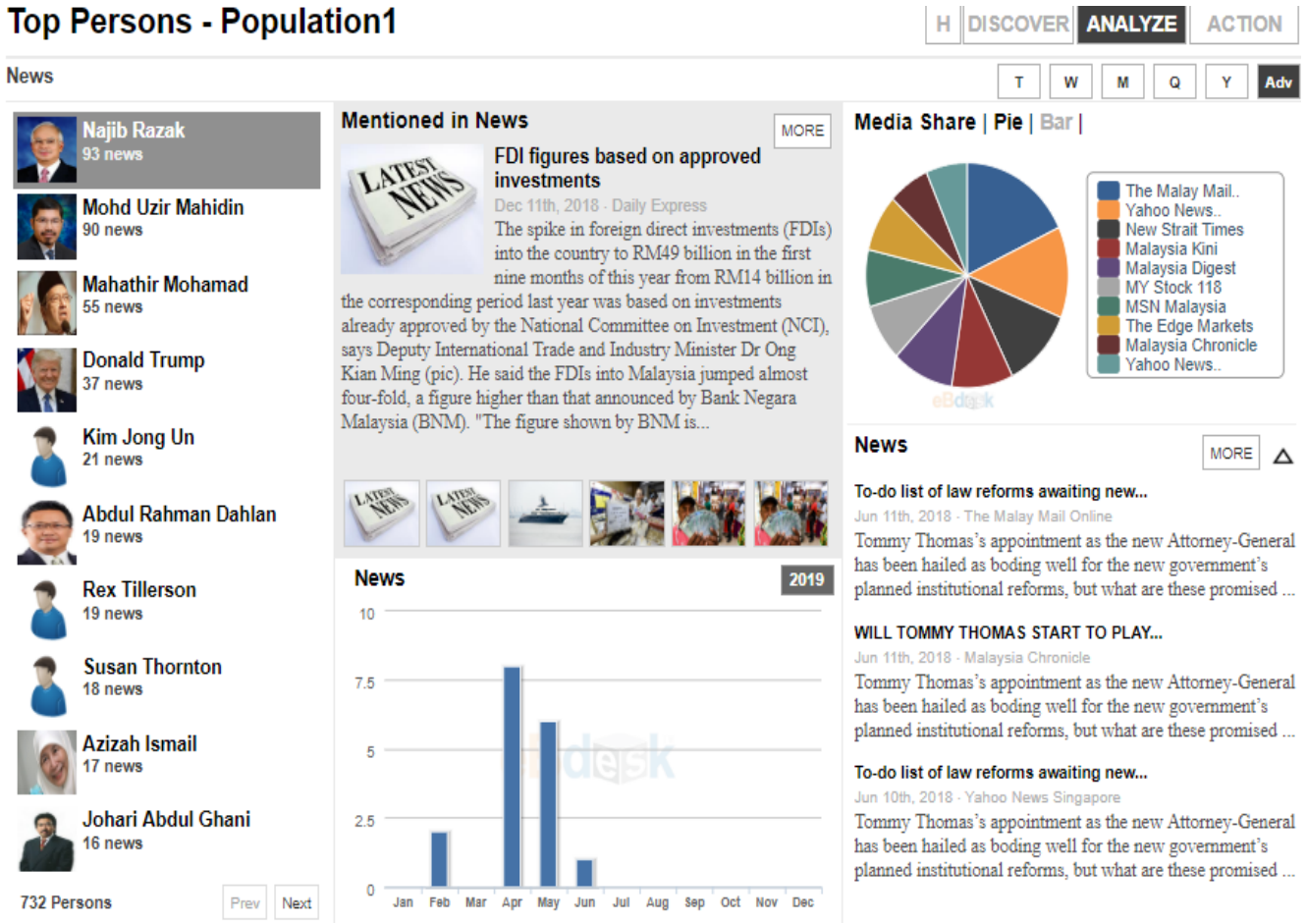


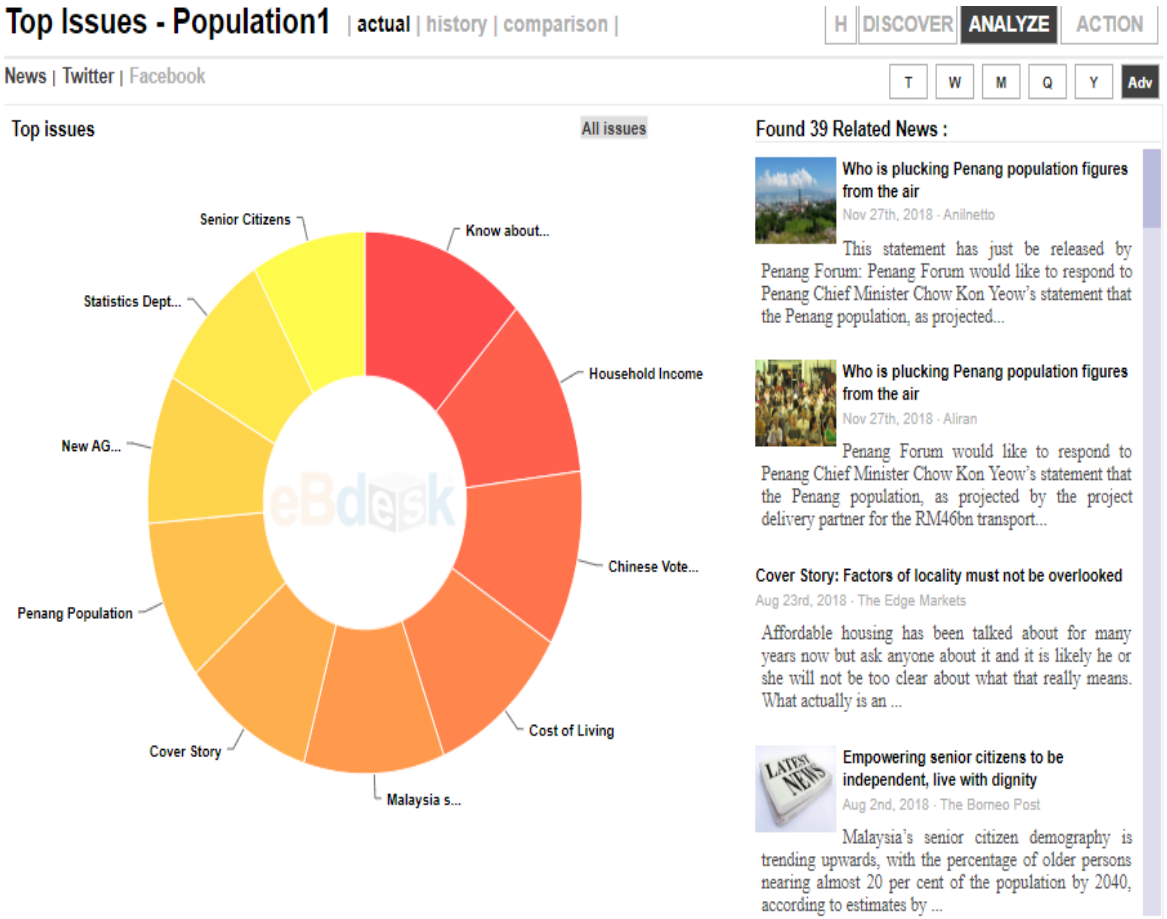


Media Intelligence Analysis

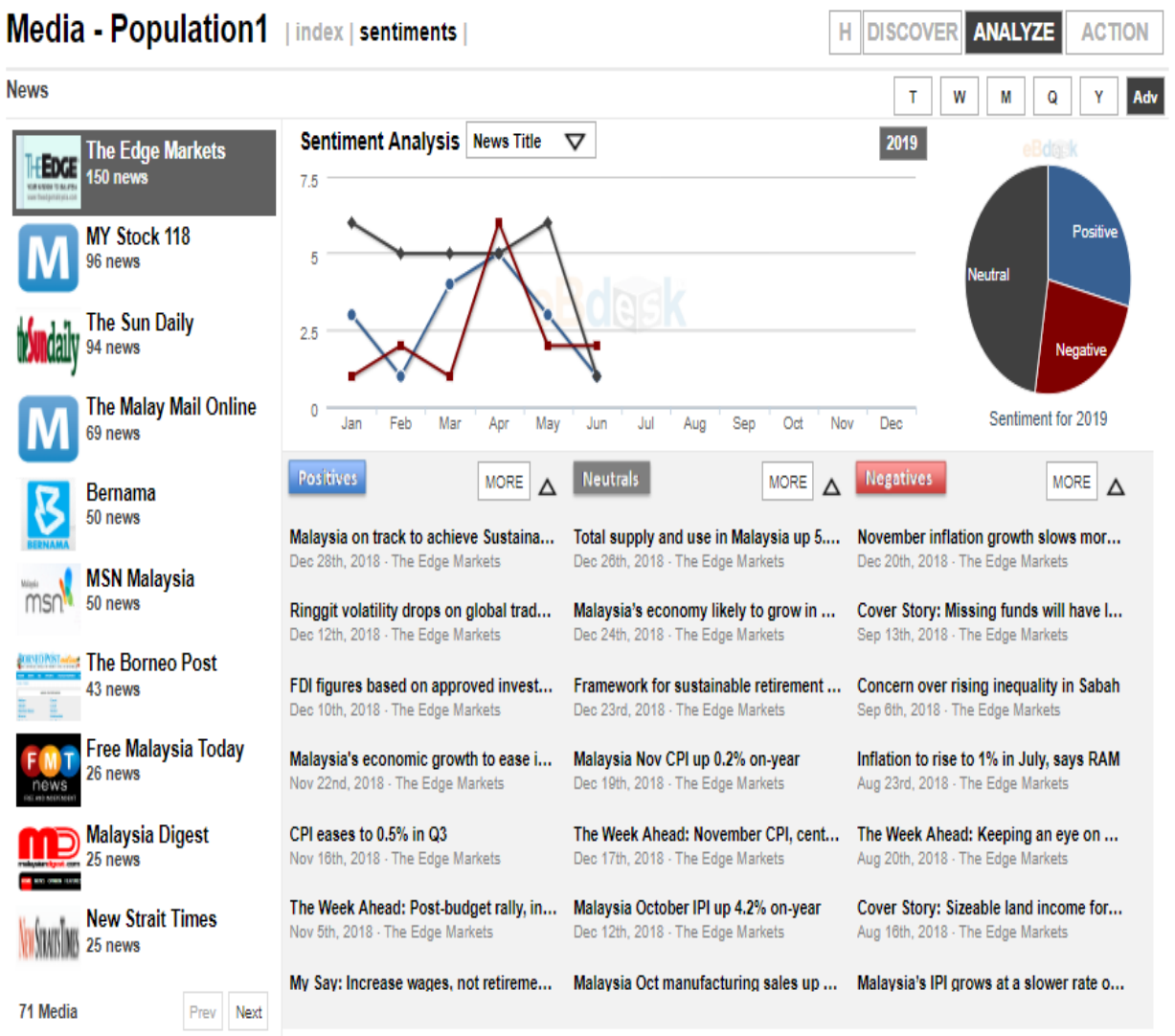


Top Persons & Top Issues

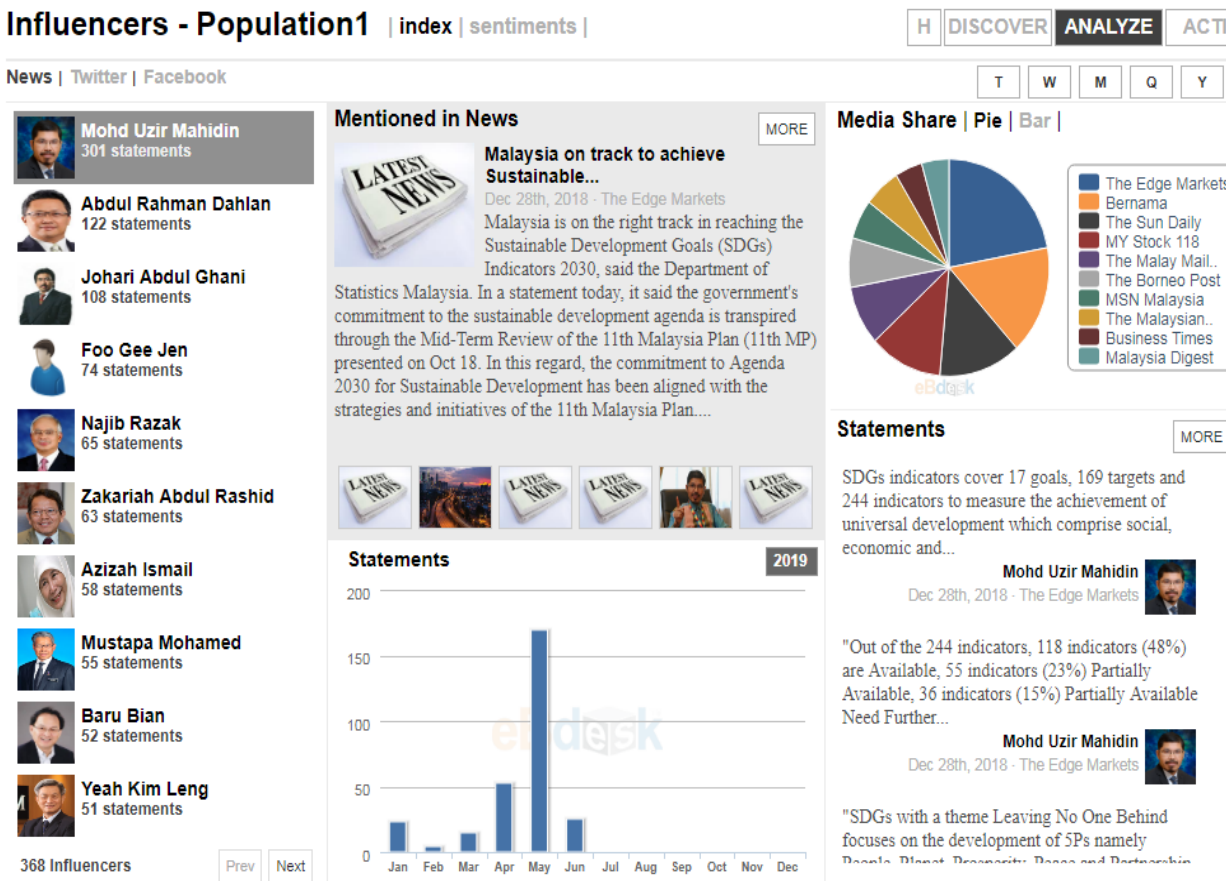


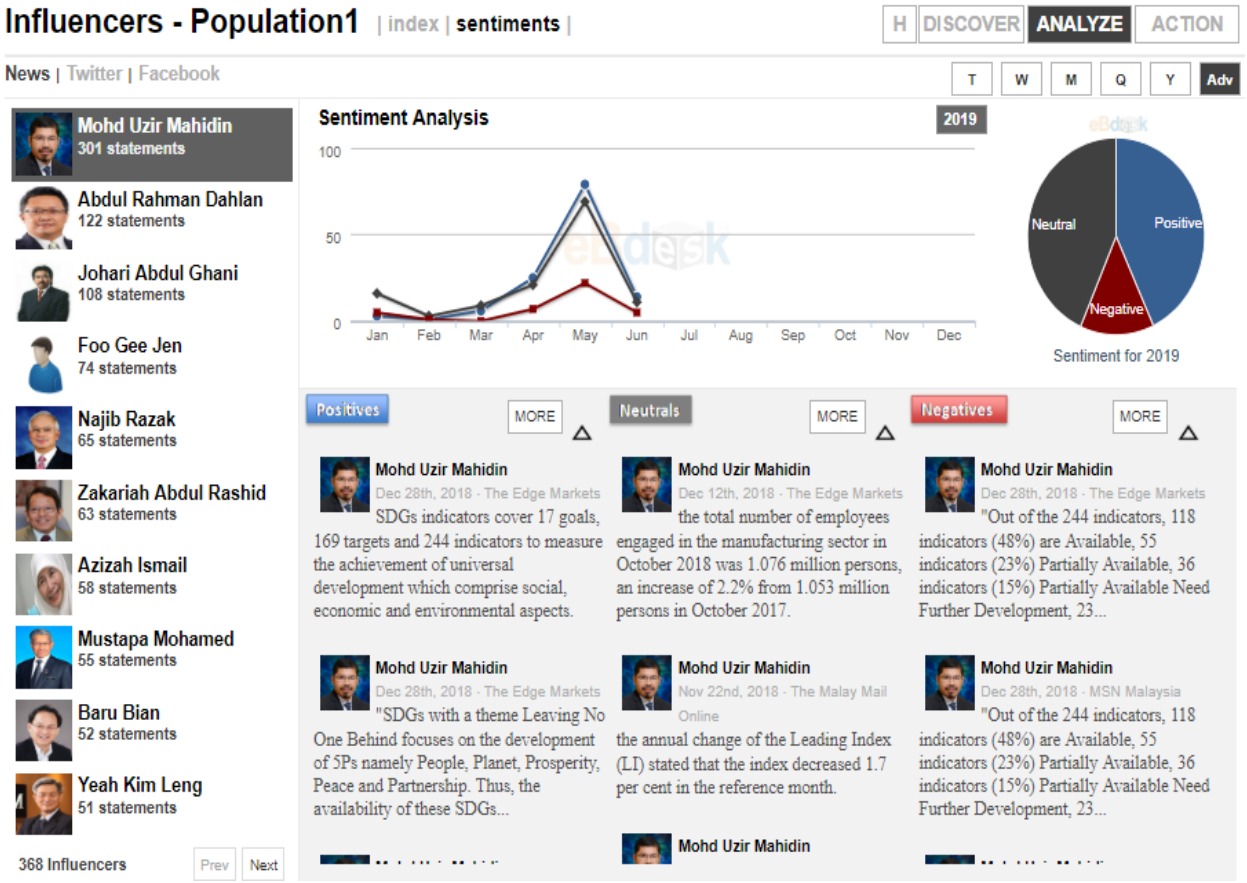


Media Share & Sentiment Analysis (News Title)

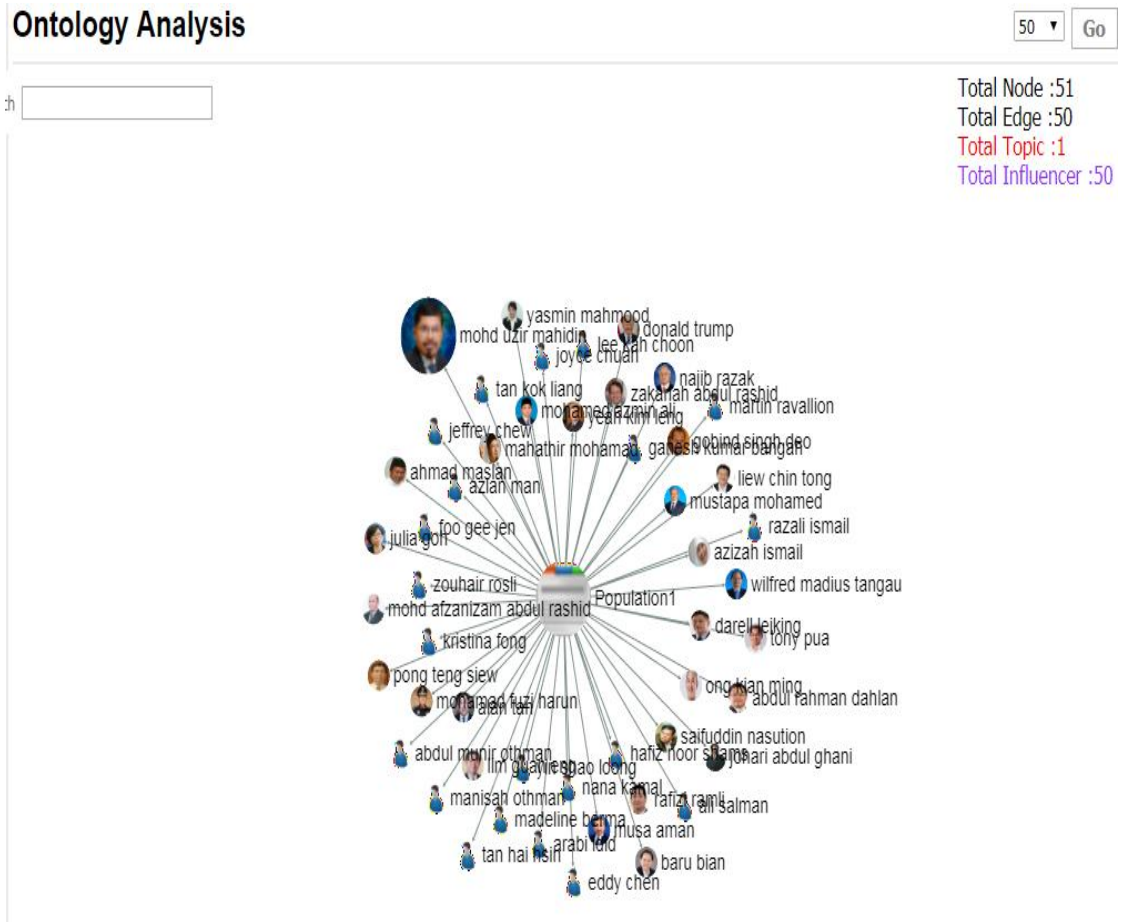


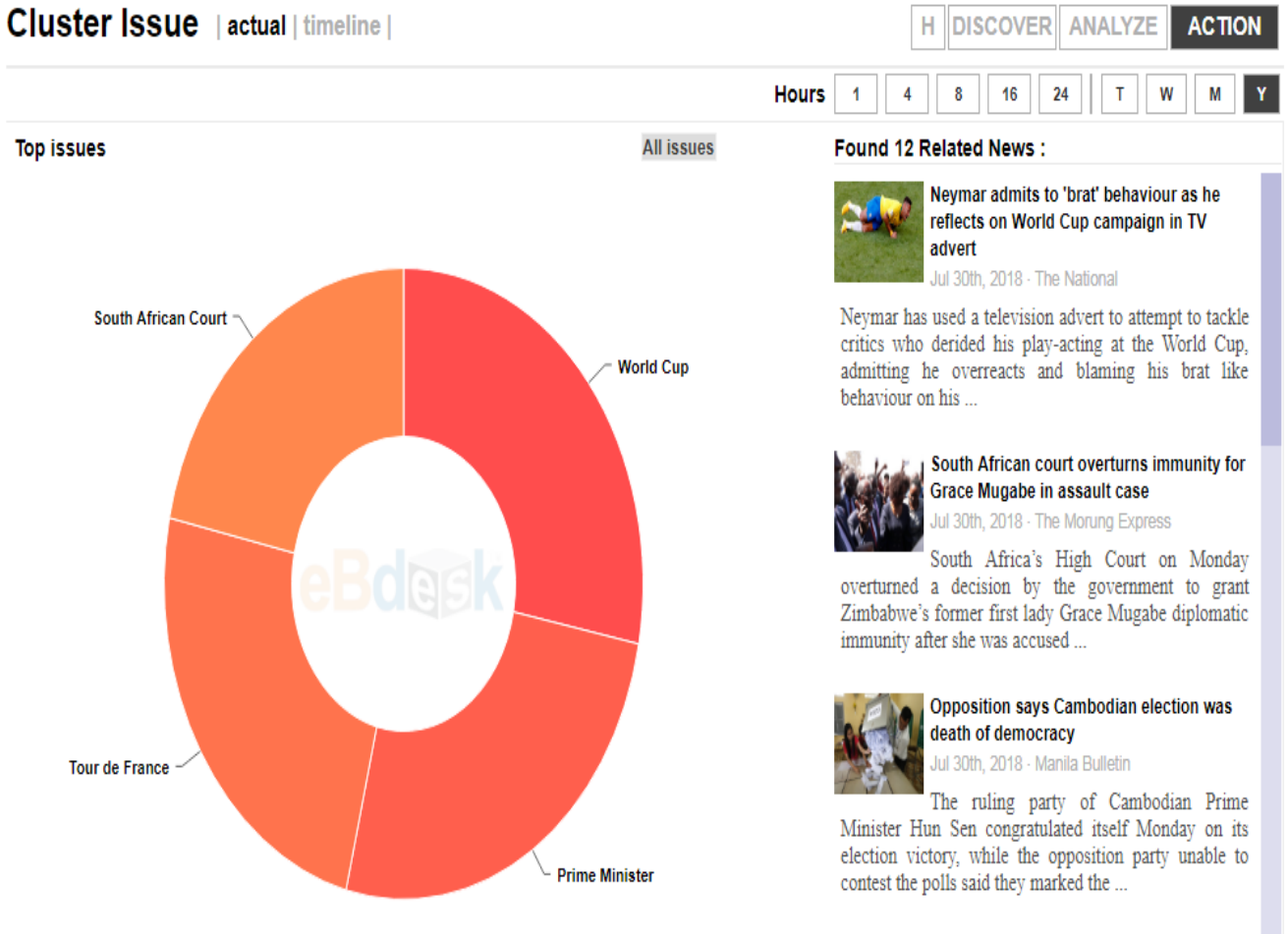
Media Share & Sentiment Analysis - Influencers



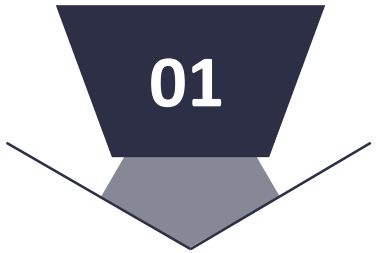


Ontology Analysis & Cluster Issue

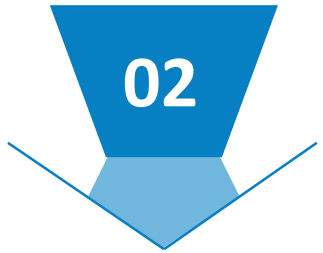




Problem Statement



Requirement to monitor online economic news is in line with official statistics released as well as current national policy implementation;



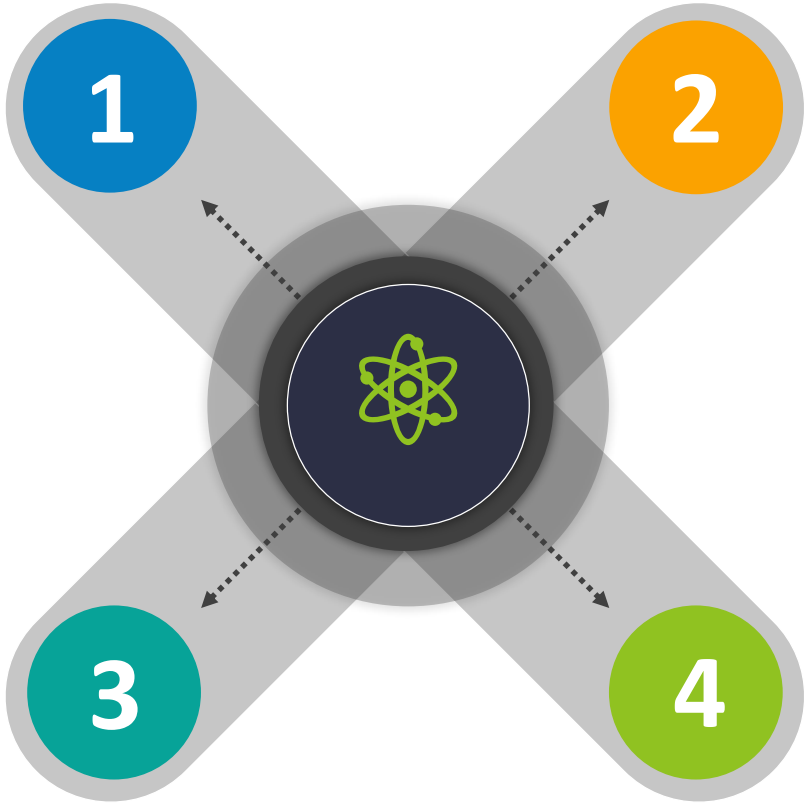
Increasing trend in using the catchy topics as news headlines for capturing the reader's attention; and

Efforts to attract readers attention is needed for enhance credibility of online news at par with written or printed news.

Generally, this study aims to identify and define the important of key words in layman for public attractions to read the entire economic report.

Objectives

Identify the key words used by media practitioners for online statistical and economic news on **population** issues in Malaysia;



measure perception developed by media to publics readers towards statistical information released by DOSM through news online (sentiment analysis);

conduct cluster analysis by using lexicon and corpus approaches and develop topics model by using Latent Dirichlet Allocation; and

develop list of layman key words as guidelines for publication write up and editorial in Department of Statistics Malaysia.

Scope Of The Research

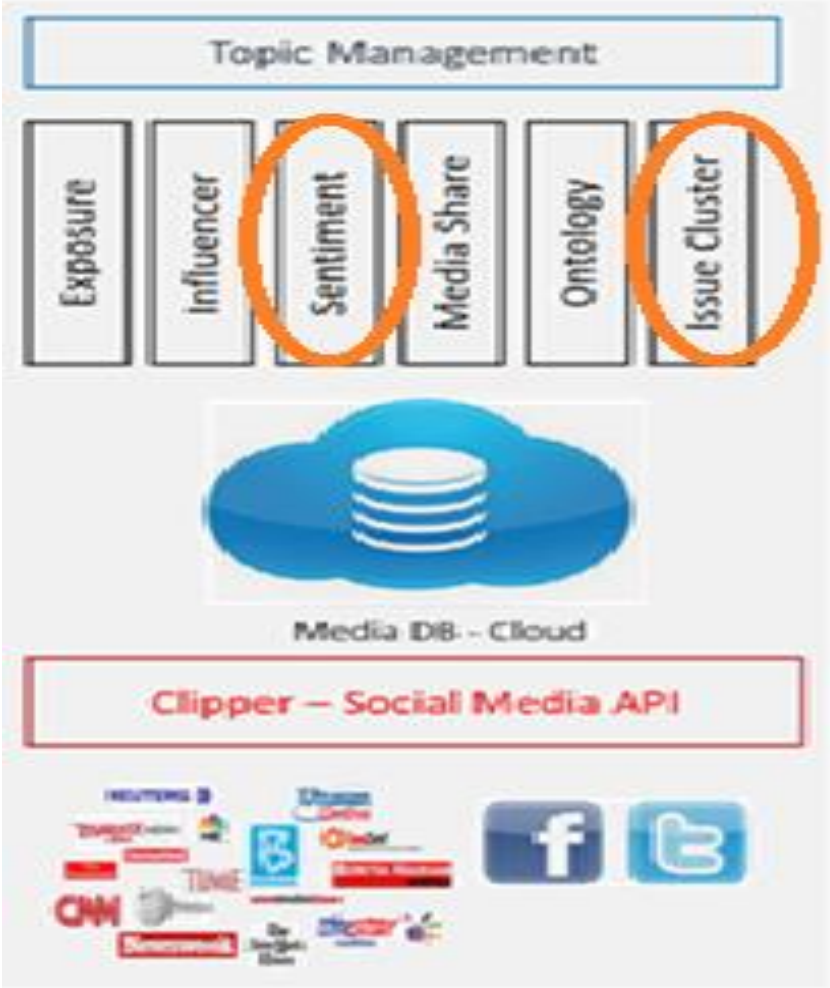
Data

- Media Intelligence Portal
- Topic: Population1

Time

- 1 Jan 2017 – 31 Dec 2018
- News Online

Scope Of The Research vs PMAOS RTOS



The Significant Of The Research

DEPARTMENT OF STATISTICS MALAYSIA
TRANSFORMATION PLAN
2015 - 2020

THRUST 4 STRATEGIC
Strengthening the Role as a Producer of Statistics

STRATEGY 1
Strengthening the Statistical Awareness, Understanding and Usage in Evidence-Based Decision Making

Programme 2
Strengthening Statistical Communication

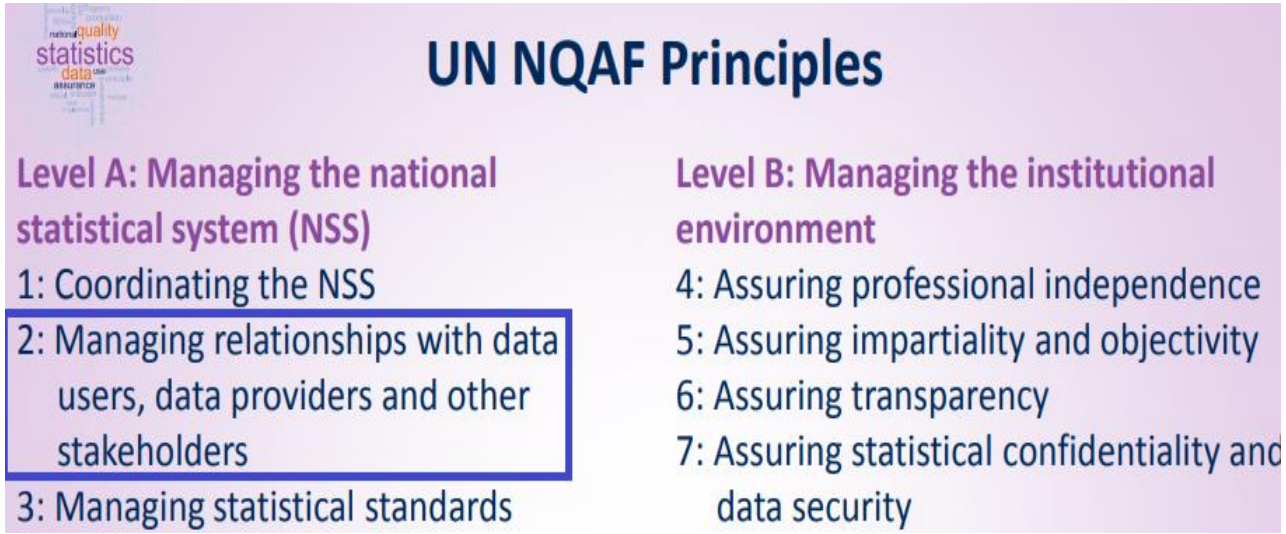
1. Strengthening media communications plan through:
   a) expanding media partnerships;
   b) creating special statistical column in the media;
   c) increasing participation in journalistic writing;
   d) highlighting the statistical products in the premier media;
   e) developing a team with expertise in statistical writing in the field of economy, social, demographic and environment; and
   f) creating a talk show on current issues "Statistics and You".
2. Using portal as the main medium of dissemination of statistical information with a user-friendly concept

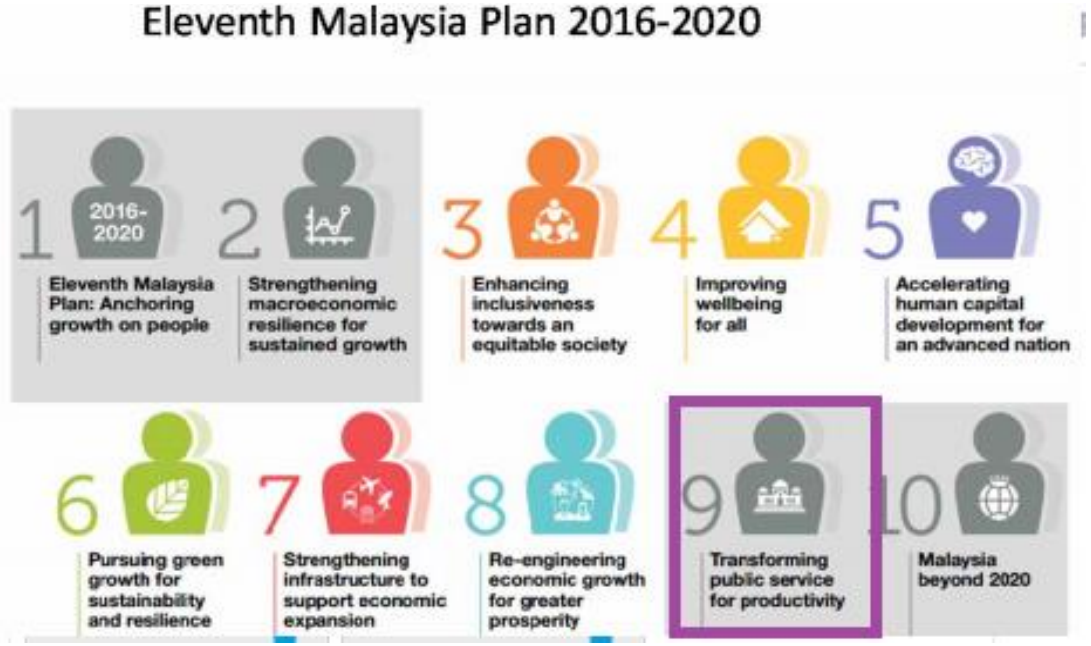
1. to produce a list of layman words that can be used for readers better understanding;

2. to assist the Public Relations Officer in DOSM to share the list with media practitioners to avoid the use of unclear, indescribable for technical terms or negative tone which will cause confusion among news online readers; and

3. to assist respective officer or editorial team as guidelines in any write-up or report to be published by DOSM.

The Significant Of The Research





2. Literature Review

Author: Ordenes el al. (2014)

Title: Analyzing Customer Experience Feedback Using Text Mining: A Linguistics-based Approach.

Findings: Text mining is the process of analyzing textual information in an attempt to discover structure and implicit meanings “hidden” within text.

Author: Nanli et al. (2012)

Title: Sentiment Analysis: A Literature Review

Findings: An analysis of sentiments or opinion aiming to investigate, analyse, extract subjective texts that involve user opinion, preferences and sentiments.

Author: Zamahsyari and Nurwidyantoro (2016)

Title: Sentiment Analysis of Economic News in Bahasa Indonesia Using Majority Vote Classifier

Findings: Online news has the ability to influence public readers on the opinion of an article.

Author: Allahyari et al. (2017)

Title: A Brief Survey pf Text Mining: Classification, Clustering and Extraction Techniques

Findings: Clustering is the task of finding groups of similar documents in a collection of documents.

Literature Review

Author: Yatim el al. (2016)

Title: A Corpus-Based Lexicon Building in Indonesian Political Context Through Indonesian Online News Media.

Findings: The sentimental lexicon is a vocabulary that has a polarity whether the test is positive or negative

Author: Kamps et al. (2004)

Title: Using WordNet to Measure Semantic Orientations of Adjectives.

Findings: Lexicon-based analysis approach is a semantic-oriented that is a non-supervisory approach in sentiment analysis.

Author: Blei et al. (2003)

Title: Latent Dirichlet Allocation

Findings: LDA is a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model.

Author: Liu (2013)

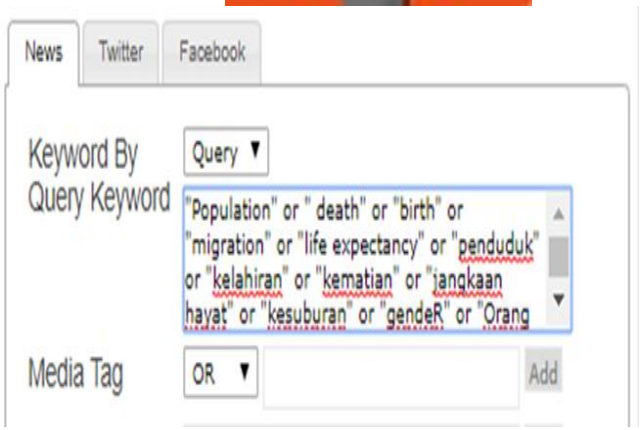
Title: High Performance Latent Dirichlet Allocation for Text Mining

Findings: LDA model is one of the most popular probability topic models and it has more comprehensive assumptions of text generation than other models.

3. Methodology

Tools	Statistical Analysis	Result
• R • Package: • tm, NLP, wordcloud, wordcloud2, ggplot2, tidyverse, tidytext, topicmodels, cluster,NbClust, RColorBrewer	• Text Mining • Cluster Analysis • Topic Model – LDA • Sentiment Analysis	• Structured data • Frequency of words • Group of similar contents (distance) • Suitable topic • Positive and negative words

4. Data Collection



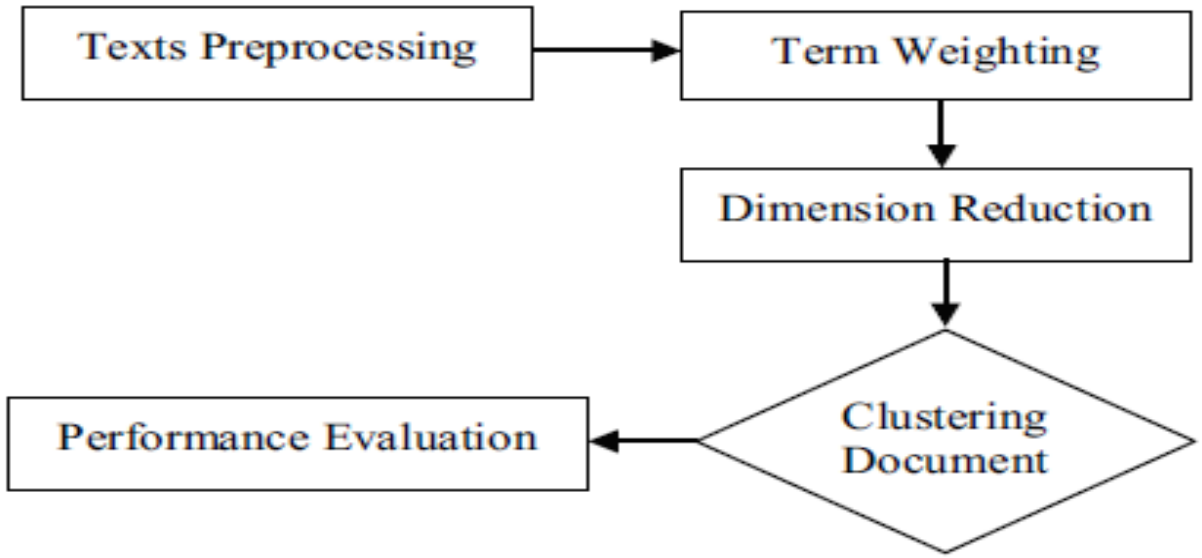
News on Malay Language are excluded (6 news from blog i.e. Blog Kamaluddins, The Patriots, Generasi Pejuang Bangsa

No.	Folder	Media	Total News	Grand Total
1	Bernama	Bernama	50	130
		Malay Mail	69	
		Mtimes	11	
2	Borneo	Borneo	43	51
		Sabahtime	8	
3	Business Time	Business Time	18	18
4	Chronicle	Chronicle	19	19
5	Daily n Digital	Daily Express	9	14
		Digital News Asia	5	
6	Focus	Focus	18	18
7	Free M	Free M	26	26
8	M Digest	M Digest	25	25
9	M Insight	M Insight	23	23
10	MSN	MSN	50	50
11	MyStock 118	MyStock 118	96	96
12	NST	NST	25	25
13	Others Mainstream	Cilisos	6	38
		Dap	6	
		Din Merican	5	
		kini	14	
		M MustKnow	1	
		M Today	4	
		People's Parliament	1	
		Umno	1	
14	Singapore	Btimes SG	6	32
		S BizReview	4	
		Yahoo S	22	
15	The Edge	The Edge	149	149
16	The Star	The Star	12	12
17	The Sun Daily	The Sun Daily	92	92
18	Vietnam	V Plus	5	8
		VietNam Net	3	
19	Yahoo M	Yahoo M	20	20

No.	Folder	Media	Total News	Grand Total
20	Others	ABC	1	86
		ABS CBN	1	
		Aliran	3	
		Aninetto	1	
		Asia One	3	
		Astro Awani	1	
		Bangkok Post	1	
		Biz Insider	1	
		Biz Today	1	
		Blogs Wsj	1	
		Çhannel	6	
		CoverBureau	4	
		Express	1	
		Ipoh Echo	2	
		Jakarta Post	1	
		Khmertimeskh	1	
		KTemocKonsiders	1	
		Lim Sian See	2	
		LKS Blog	1	
		M Reserve	9	
		Mole	10	
		My Sinchew	6	
		News Xin	6	
		Reuters	3	
		Says M	4	
		Srwk Voice	1	
		Stop	1	
		Strait Times	6	
		Taiping	1	
		Tech In Asia	1	
		Today Online	4	
		Xinmsn	1	
TOTAL NEWS				**932**

5. Statistical Analysis – Text Mining

Kadhim et al. (2014) proposed system approach that can be used for this study. The proposed system is divided into four main stages such as: preprocessing stage, term weighting stage, dimensional reduction and performance evaluation as shown in Figure



01 upload data as corpus

02 start pre-processing

03 remove stopword by use standard stopword list and build-in

04 format vector

05 create document-term matrix

06 convert dtm to matrix

07 collapse matrix by summing over columns

08 inspect most frequently occurring terms

09 change to data frame format

5. Statistical Analysis – Cluster Analysis

Aims to grouping the text based on distance between document vectors, evaluated on the similarity by using the distance concept

5. Statistical Analysis - Sentiment Analysis

The main two methods of sentiment analysis are lexicon-based method and machine learning based approach. The lexicon-based approach, uses sentiment lexicons - dictionaries of words with labels specifying their sentiments - to identify the sentiment of text

This research is based on lexicon-based method with aims to classify the term either positive or negative and determined the sentiment polarity score at word level analysis based on the formula used by Kolchyna et al. (2015) as stated below:

Type of sentiment lexicon

```
> get_sentiments("afinn")
# A tibble: 2,476 x 2
   word       score
   <chr>      <int>
 1 abandon       -2
 2 abandoned     -2
 3 abandons      -2
 4 abducted      -2
 5 abduction     -2
 6 abductions    -2
 7 abhor         -3
 8 abhorred      -3
 9 abhorrent     -3
10 abhors        -3
# ... with 2,466 more rows
```

```
> get_sentiments("bing")
# A tibble: 6,788 x 2
   word        sentiment
   <chr>       <chr>
 1 2-faced     negative
 2 2-faces     negative
 3 a+          positive
 4 abnormal    negative
 5 abolish     negative
 6 abominable  negative
 7 abominably  negative
 8 abominate   negative
 9 abomination negative
10 abort       negative
# ... with 6,778 more rows
```

```
> get_sentiments("nrc")
# A tibble: 13,901 x 2
   word        sentiment
   <chr>       <chr>
 1 abacus      trust
 2 abandon     fear
 3 abandon     negative
 4 abandon     sadness
 5 abandoned   anger
 6 abandoned   fear
 7 abandoned   negative
 8 abandoned   sadness
 9 abandonment anger
10 abandonment fear
# ... with 13,891 more rows
```

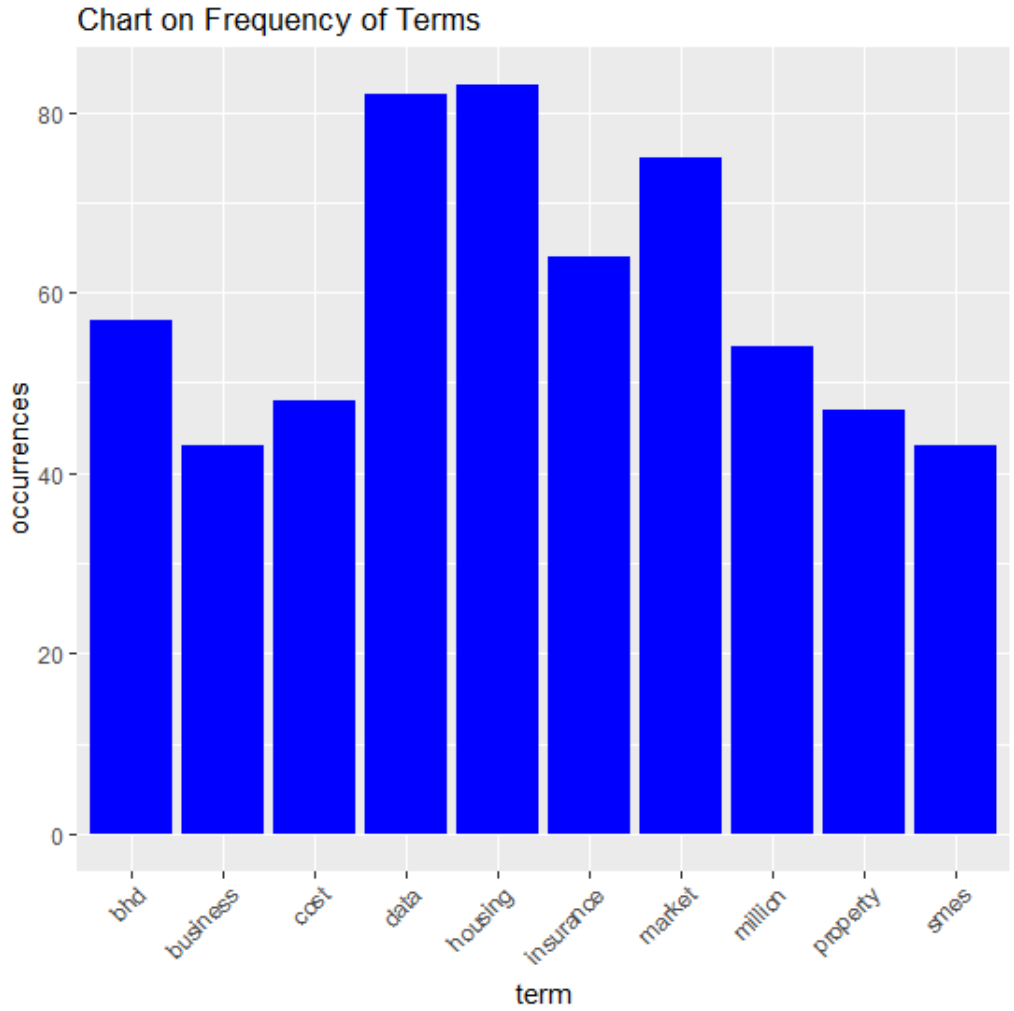
01

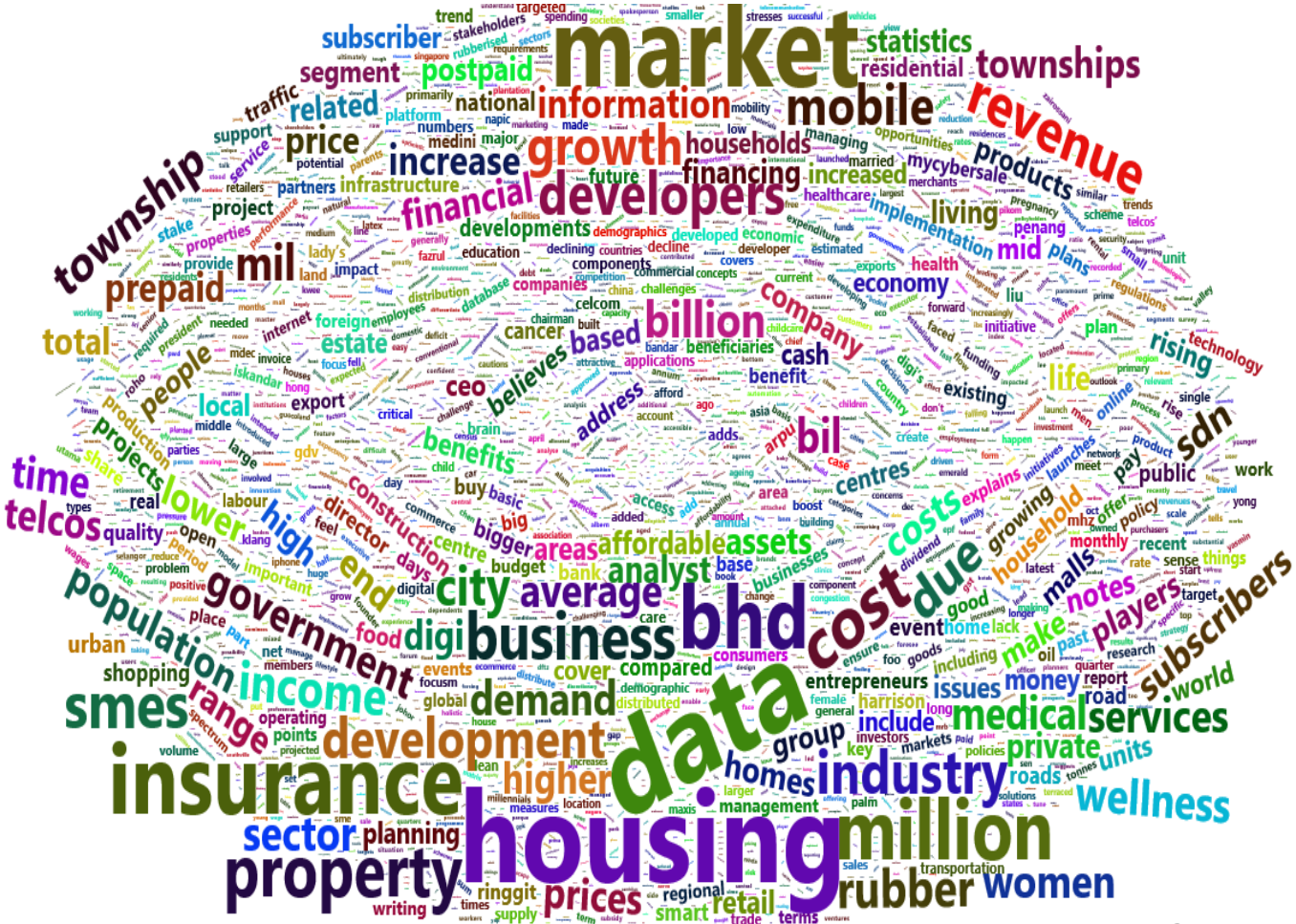
The AFINN lexicon categorizes words using a score, with negative scores indicating a negative sentiment. The bing lexicon categorizes words into a positive or negative flag ; the nrc lexicon categorizes words into emotions like anger, disgust, fear, joy and others.

02

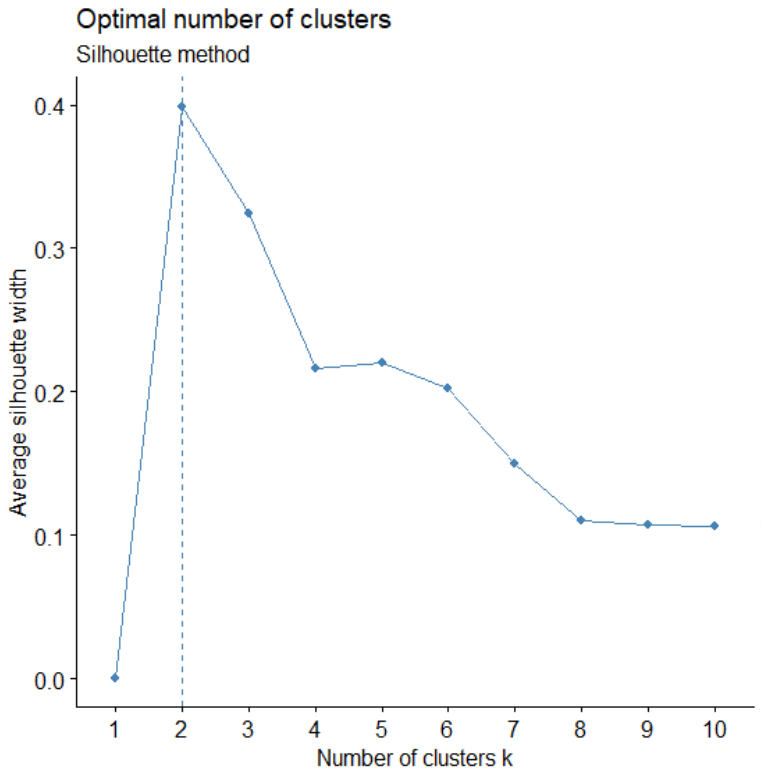
Bing and NRC have more negative than positive words, but the ratio of negative to positive words is higher in the Bing lexicon than the NRC lexicon. This will contribute to the effect in the plot.

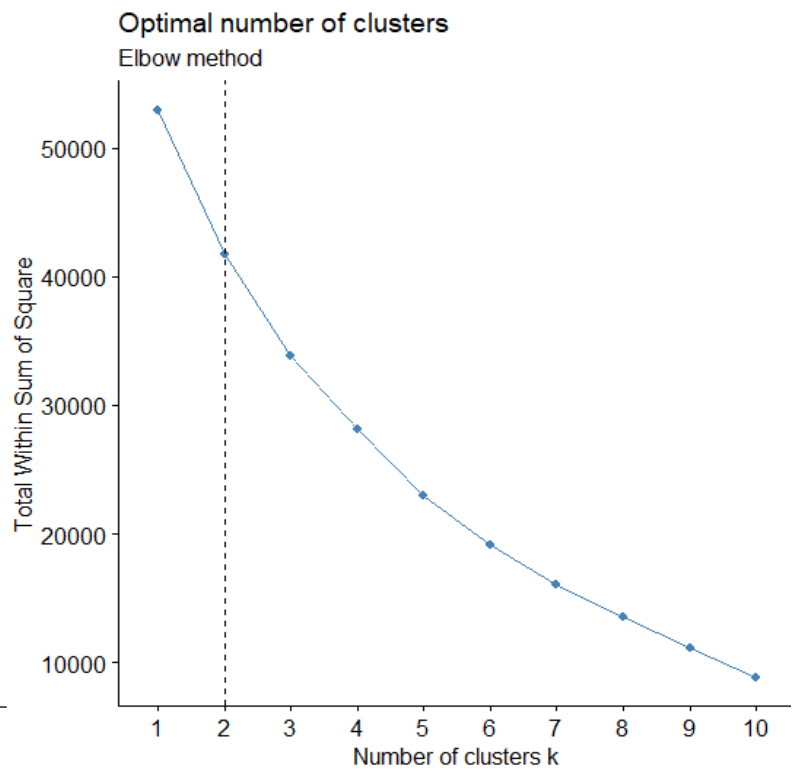
6. Result - Frequency of Words

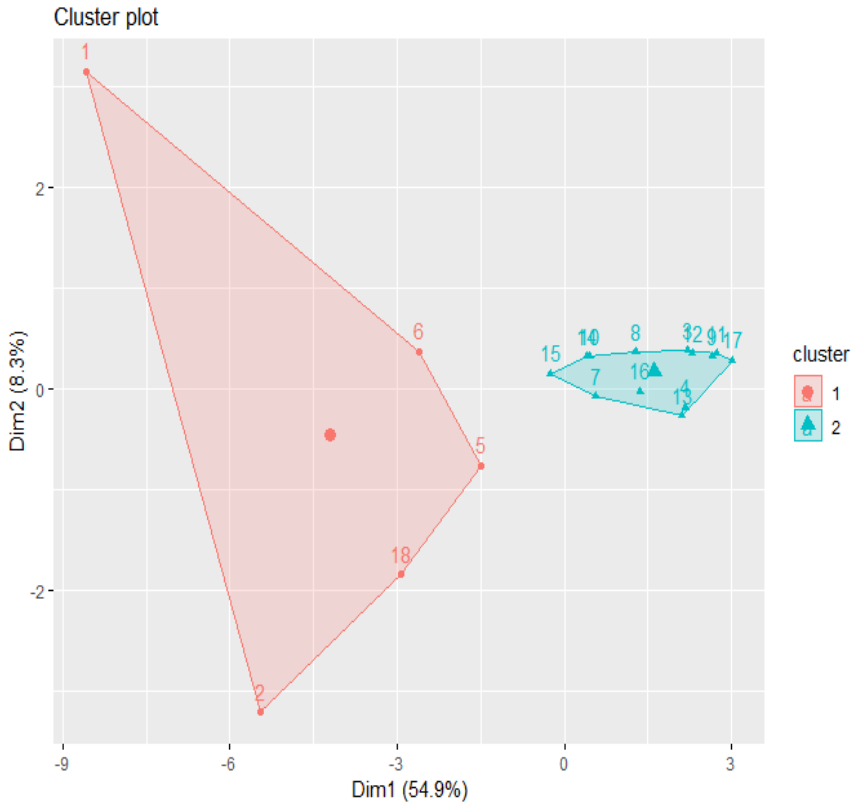




6. Result – Cluster Analysis







$$\text{Idf (term)} = \ln \frac{ndocuments}{\text{ndocuments containing term}}$$

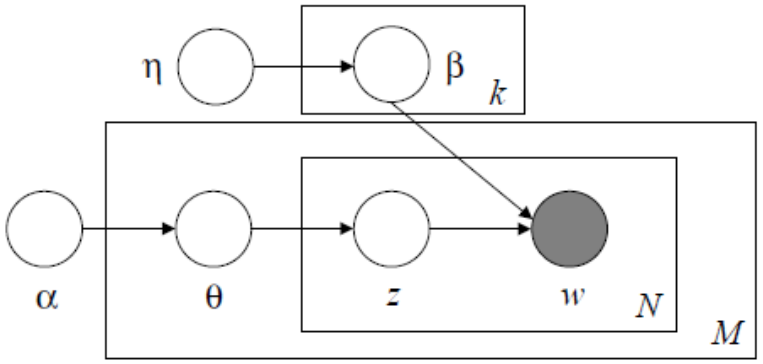
6. Result - Latent Dirichlet Allocation (LDA)

LDA

LDA is a method for finding the best topic in a group of words from a collection of documents.

LDA

LDA is a generative probabilistic model for collections of discrete data such as text corpora and a three-level hierarchical Bayesian model.



Graphical Model Representation of the Smoothed LDA Model.

Document	Cluster	Probabilities	Document	Cluster	Probabilities
1	1	0.866	10	1	0.690
2	2	0.865	11	1	0.696
3	1	0.639	12	1	0.669
4	2	0.790	13	2	0.777
5	2	0.857	14	2	0.722
6	1	0.885	15	1	0.850
7	2	0.852	16	2	0.676
8	1	0.747	17	1	0.695
9	1	0.737	18	2	0.885

6. Result - Topic Model (The Edge)

No.	Topic 1 (Economy)	Topic 2 (Social)	Topic 3 (General)
1	growth	market	income
2	bhd	property	tax
3	billion	retirement	labour
4	prices	housing	workers
5	exports	price	government
6	statistics	financial	higher
7	manufacturing	affordable	economic
8	data	time	low
9	trade	population	foreign
10	expected	development	high
11	index	high	cost
12	products	house	household
13	inflation	age	wage
14	million	klang	living
15	global	land	median
16	oil	work	average
17	compared	developers	economy
18	sector	government	report
19	week	public	people
20	gdp	term	wages
21	due	long	lower
22	quarter	units	monthly
23	bank	people	state
24	rate	adds	rate
25	sales	good	jobs
26	increase	working	gst
27	research	home	revenue
28	export	prices	growth
29	increased	million	households
30	production	demand	country

#top 30 terms in each topic

6. Result -Sentiment Analysis

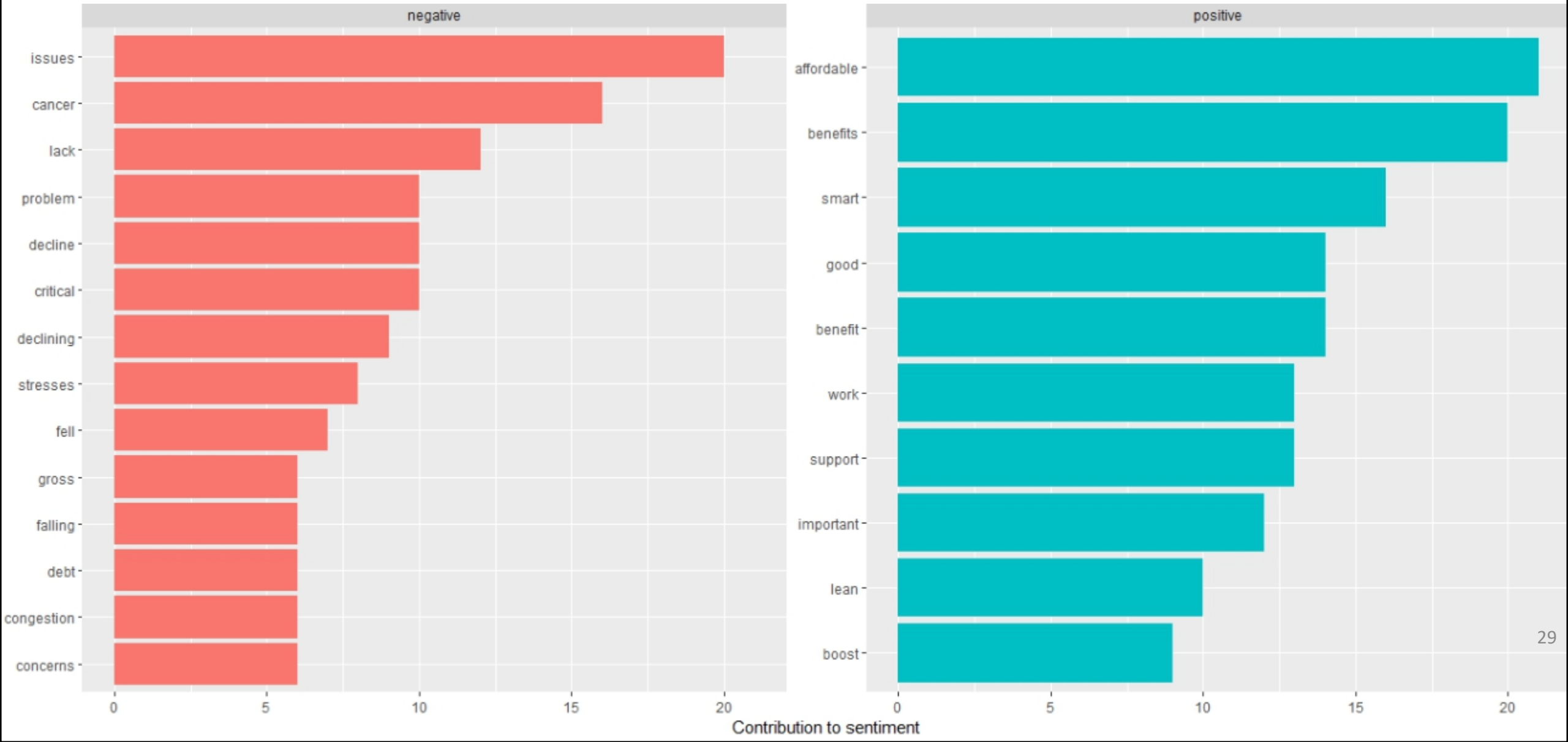
Folder	Positive	Negative
Bernama	leading	gross
	work	decline
	improved	opposition
	good	crude
	support	issues
	strong	issue
	savings	unemployed
	top	negative
	free	tell
	worth	concerns

Folder	Positive	Negative
Focus	affordable	issues
	benefits	cancer
	smart	lack
	good	problem
	benefit	decline
	work	critical
	support	declining
	important	stresses
	lean	fell
	boost	gross

Folder	Positive	Negative
Borneo	positive	poverty
	top	negative
	tender	poor
	strong	lack
	support	issue
	stable	decline
	work	risk
	free	gross
	benefit	debt
	improve	crude

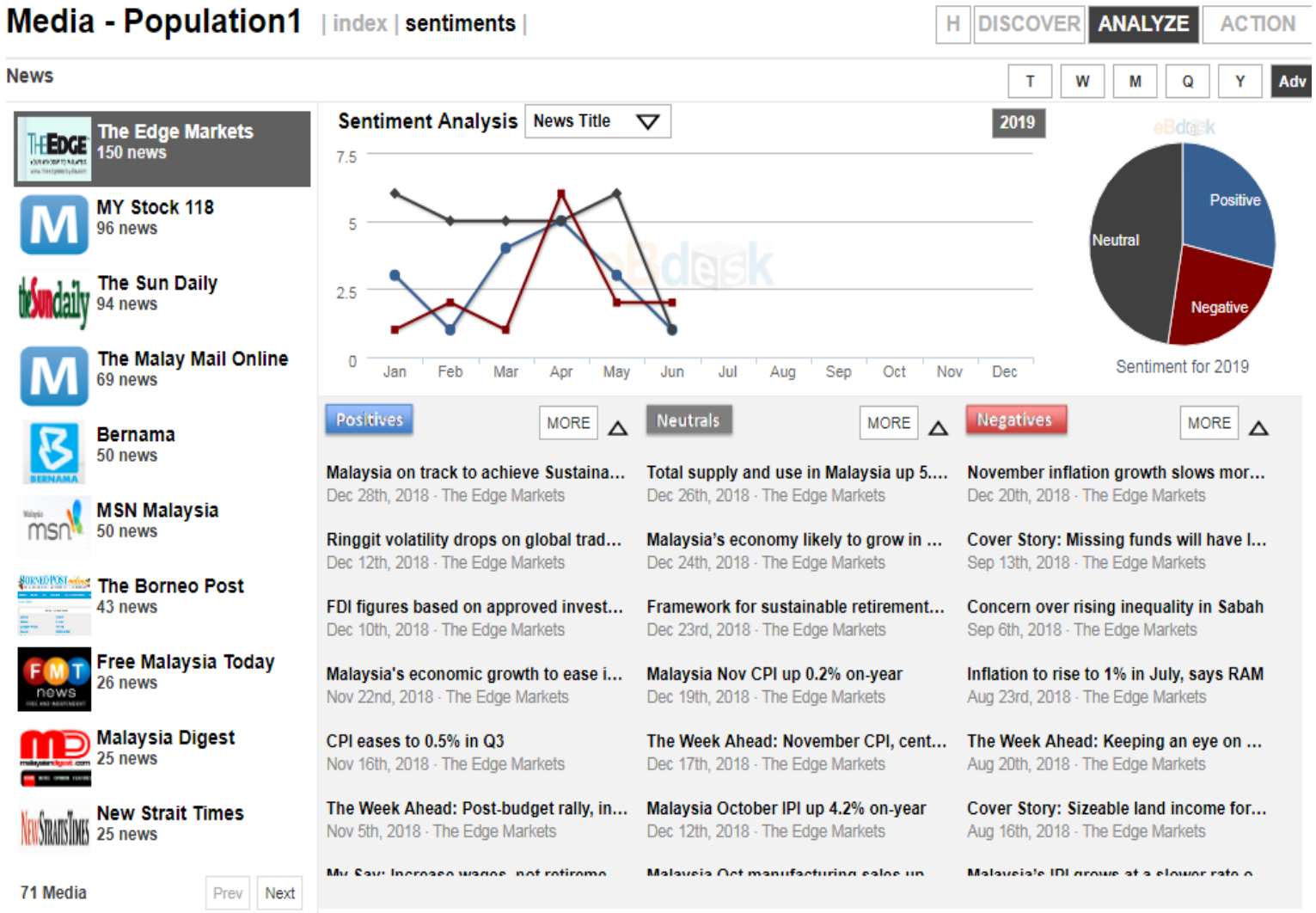
Folder	Positive	Negative
The Edge	affordable	gross
	top	inequality
	good	crude
	work	debt
	strong	decline
	savings	issue
	support	risk
	skilled	problem
	important	issues
	supported	slower

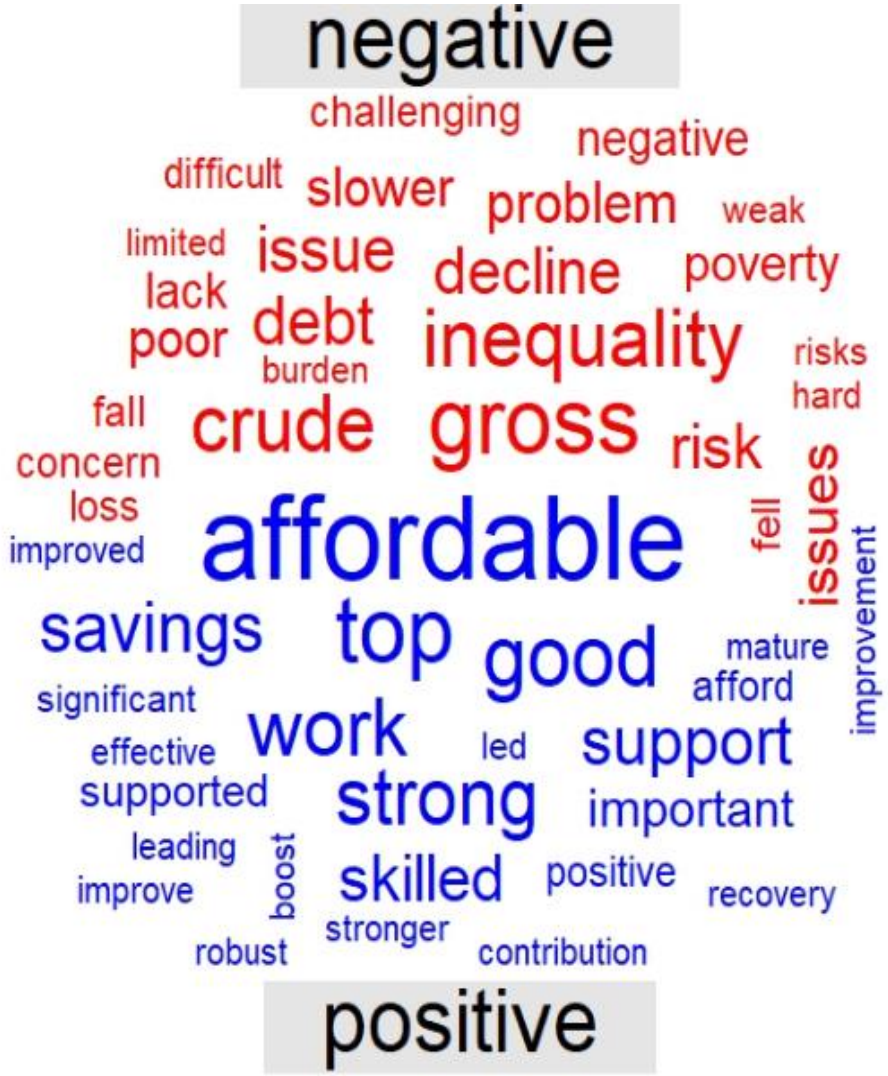
6. Result -Sentiment Analysis:
most common positive & negative words (Focus)



7. Conclusion

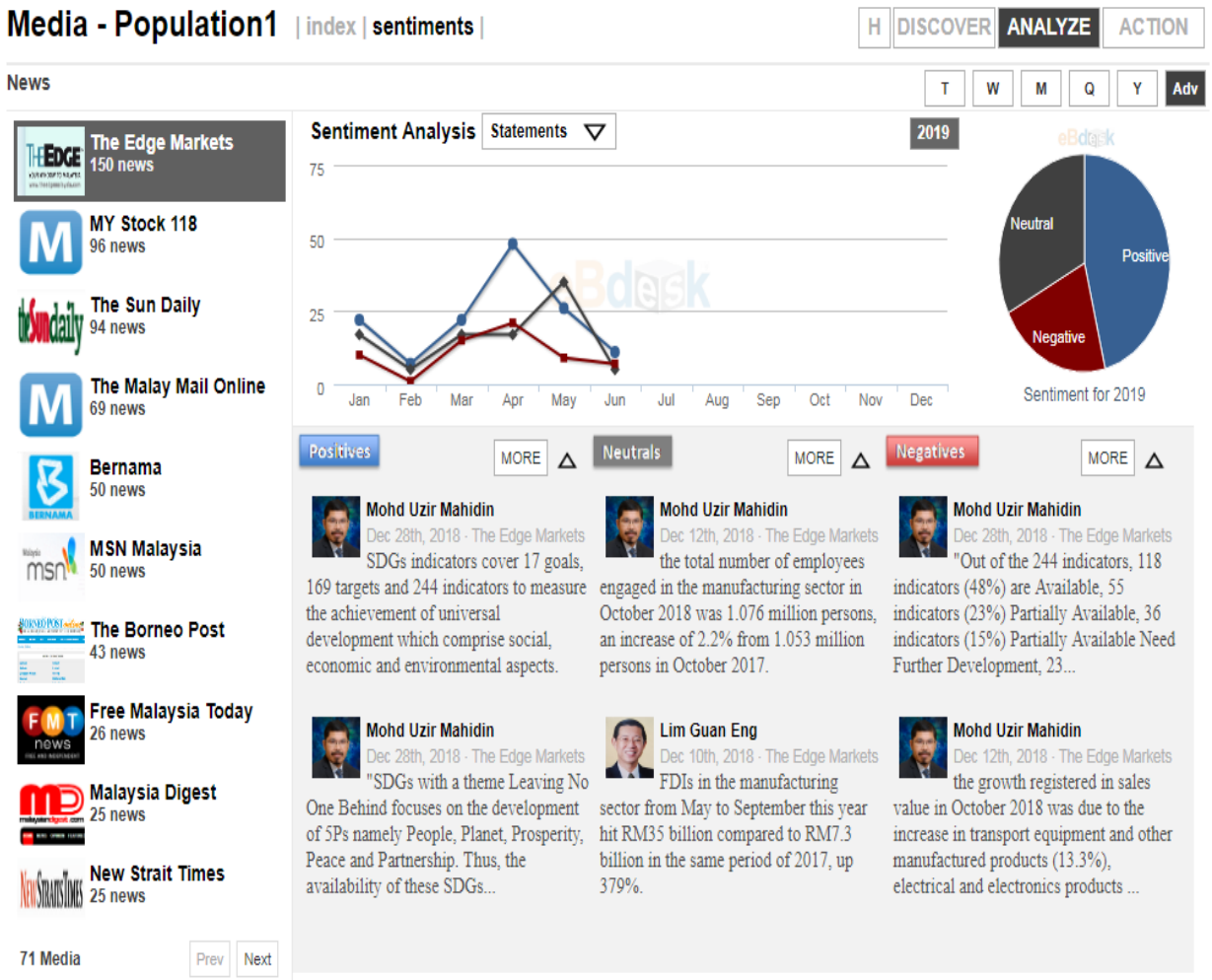
Analysis Comparison with STATSBDA (PMAOS and RTOS)

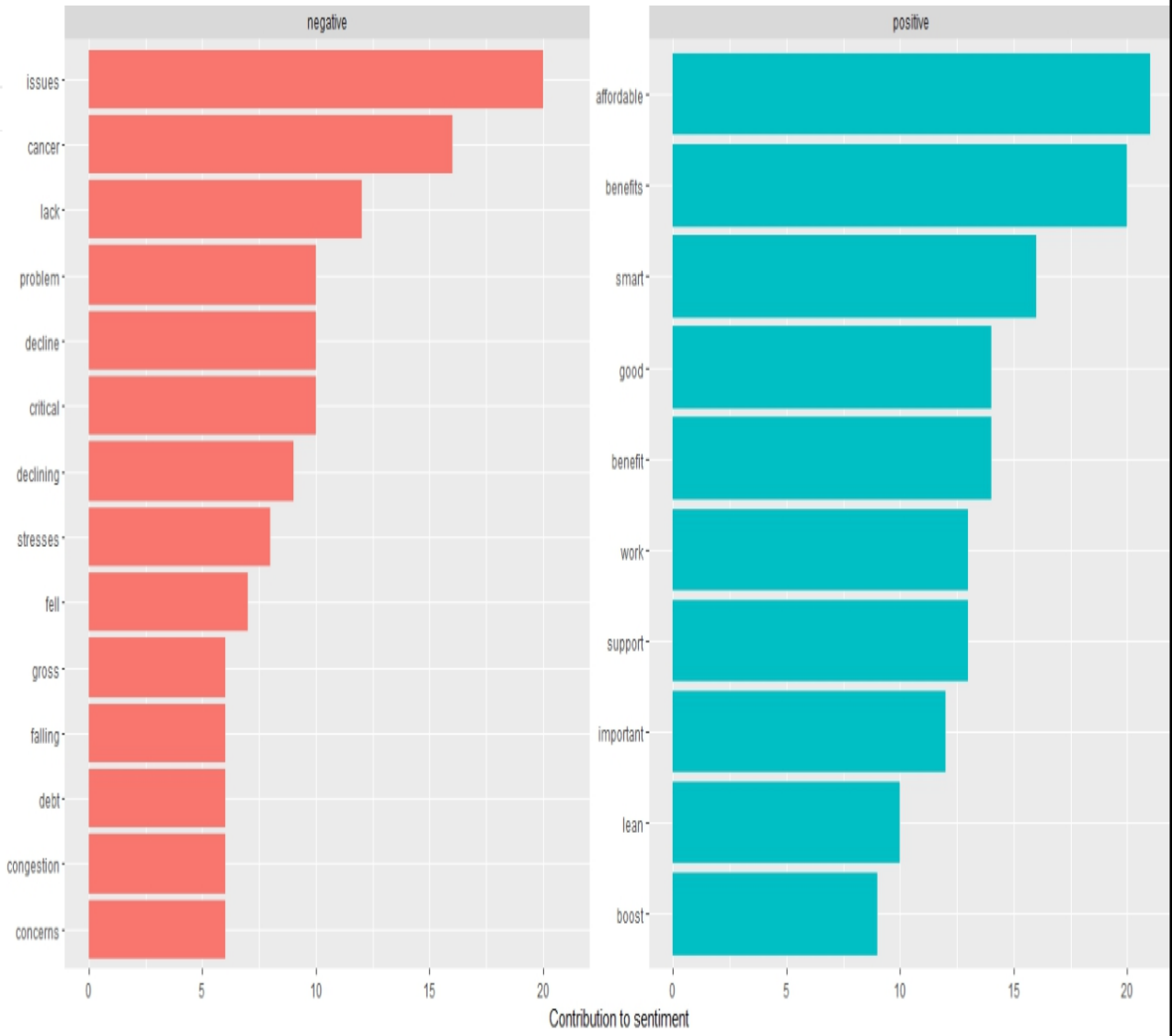




7. Conclusion

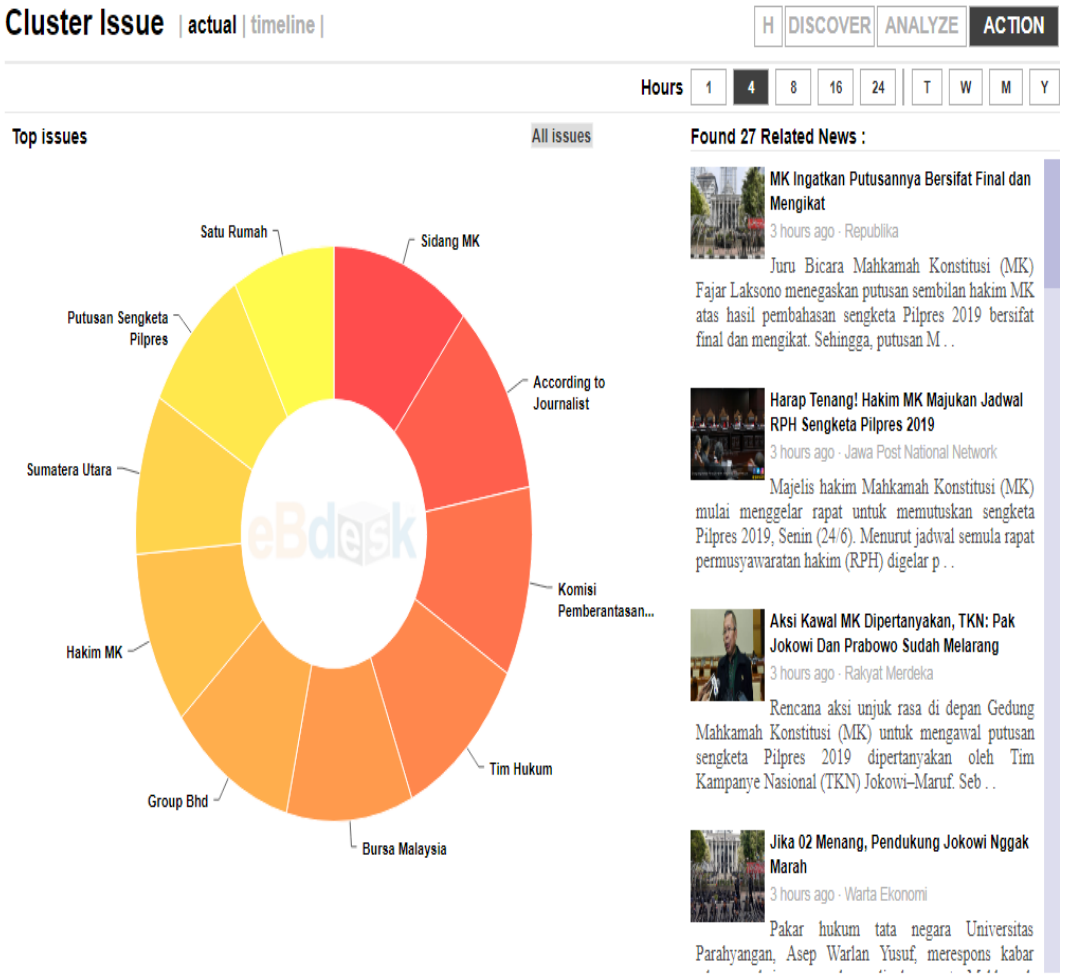
Analysis Comparison with STATSBDA (PMAOS and RTOS)

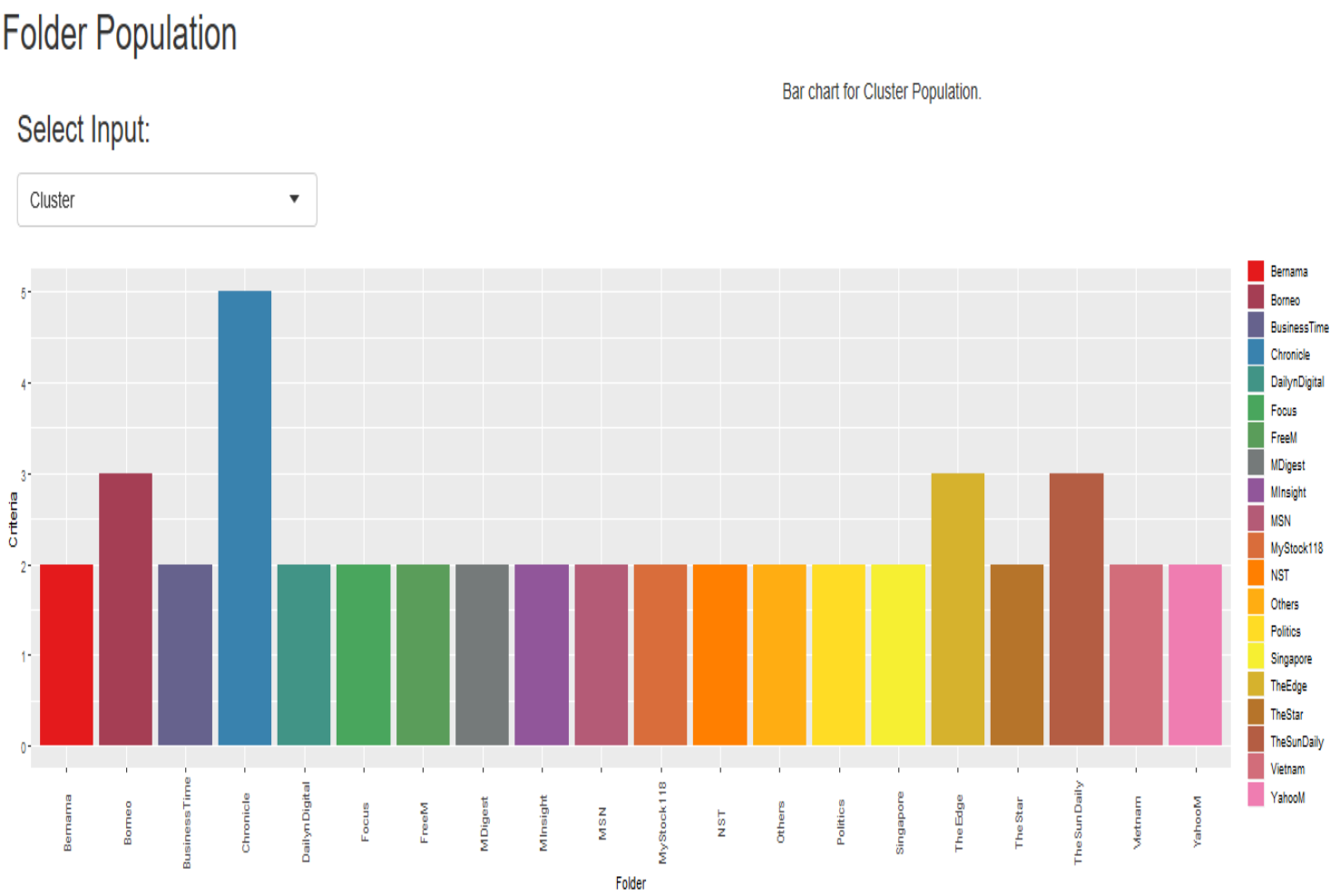




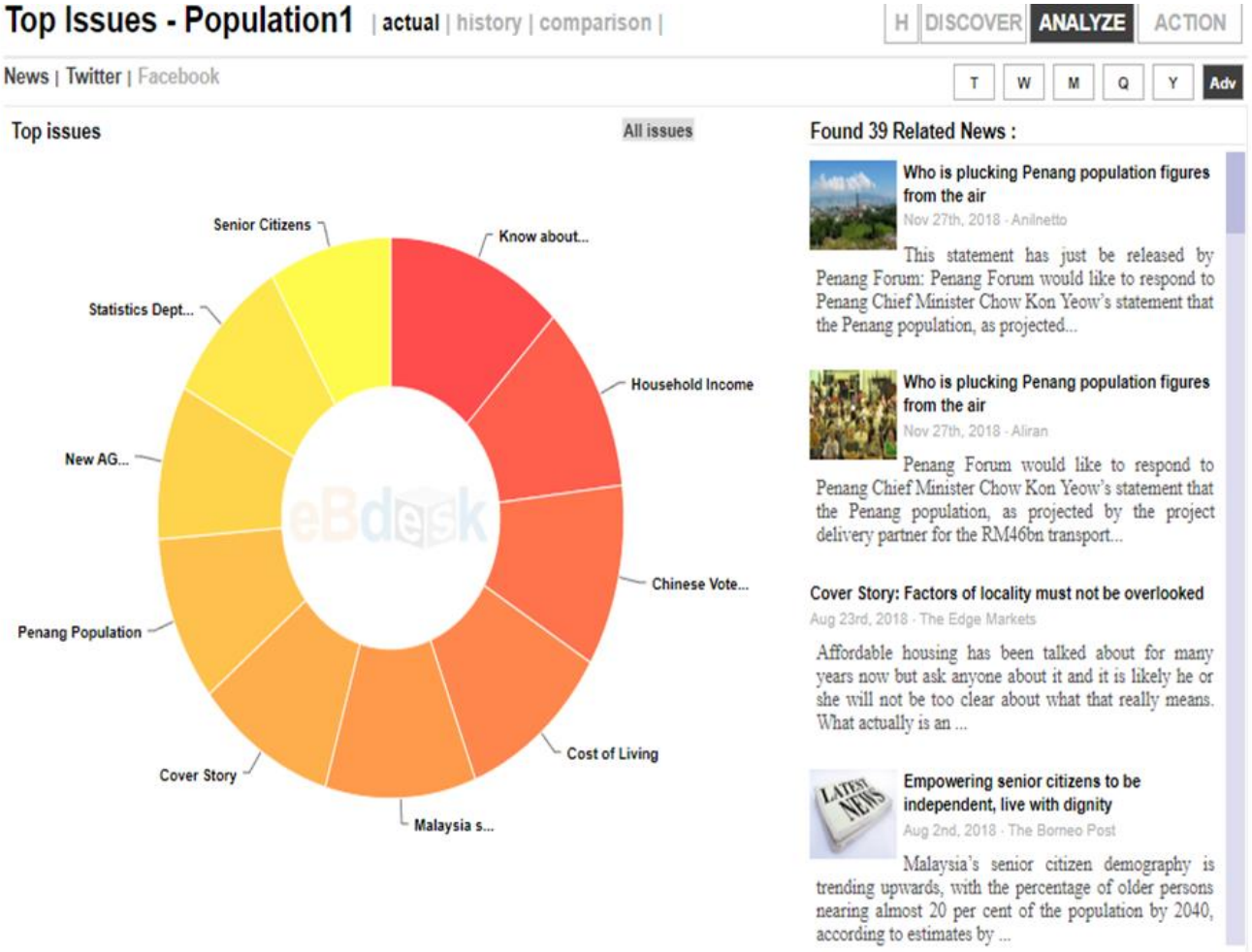
7. Conclusion

Analysis Comparison with STATSBDA (PMAOS and RTOS)





7. Conclusion
Analysis Comparison with STATSBDA (PMAOS and RTOS)



No.	Key Word	Frequency
1	growth	19
2	statistics	15
3	billion	14
4	economic	14
5	million	14
6	government	13
7	economy	12
8	income	12
9	people	11
10	data	10

7. Conclusion

- Current project were based on machine learning method (maximum entropy);
- Based on this capstone project result, the sentiment analysis can be used for determine the degree of “happiness” of the community towards the official statistics media released and the key words are related to statistics theme in DOSM; and
- Text analytics can be used as tools for planning any improvements in producing statistical products and services that meet user’s expectations.

Positive	Negative
affordable	issues
benefits	cancer
smart	lack
benefit	critical
good	decline
support	problem
work	declining
important	stresses
lean	fell
boost	concerns
afford	congestion
positive	debt
attractive	falling
free	gross
easier	challenging
easy	difficult
fast	poor
integrated	risk
lead	
successful	
top	