International Journal on Advanced Science Engineering Information Technology

Vol.9 (2019) No. 6
ISSN: 2088-5334

Investigating the Relevant Agro Food Keyword in Malaysian Online Newspapers

Mohamad Farhan Mohamad Mohsin[#], Siti Sakira Kamaruddin[#], Fadzilah Siraj[#], Hamirul Aini Hambali[#], Mohammed Ahmed Taiye[#]

[#]*School of Computing, College of Arts & Sciences, Universiti Utara Malaysia, Kedah, Malaysia*
E-mail:farhan@uum.edu.my; sakira@uum.edu.my; fad173@uum.edu.my; hamirul@uum.edu.my; tfeatslekan@gmail.com

Abstract— **Online newspaper is a valuable resource of information for decision making. To extract relevant information from them is a challenging process when their volume is massive, and its knowledge is in an unstructured form that is scattered on every page. This situation becomes more complicated when different news providers have different styles of journalism when reporting a similar event and use different concepts and terms. In this study, we examined the three Malaysian English online newspapers in order to identify knowledge in terms of the most relevant keywords used in daily online news. The news articles related to Agro-food industries were taken from online news websites - The Star Online, The Sun Daily, and The News Straits Times. During the extraction, about 458 Agro-food industries news articles were scrapped from the website within the time frame of 2014-2017. The keywords were extracted using the RAKE algorithm and were classified into 4 groups i.e. agriculture, livestock, fishery and miscellaneous. The agriculture keywords group was found as the most frequent keywords in all newspapers (58%) and it was followed by the livestock (23%), fishery (12%), and miscellaneous (7%). Through the analysis, there were 146 Agro-related keywords found in all newspapers, repeated 720 times, and the highest Agro terms were found in The Star Online (35.13%), followed by The Sun Daily (33.78%), and The News Straits Times (31.08%). There were 12 Agro keywords0 which considered as the most relevant when they appear in all newspapers- palm oil, rice, fruits, fish, vegetable, livestock, paddy, crop, chicken, animal, meat, and beef. The 'palm oil' is the most popular keyword among the three newspapers and it was found 37 times (38.9%) in The Star Online, 26 times (37.9%) in News Straits Time, and repeated 22 times (23.2%) in the Sun. The identified keywords can be recommended as input to form a future Agro inventory.**

Keywords— **agro-food keywords; news mining; RAKE algorithm; text mining; online newspaper.**

I. Introduction

Across the world, the internet has enabled various tasks to be performed through a computer or smartphone. With fast internet and diverse computer networking technologies, people are using the internet platform to promote business, communicate, search for information and keep abreast of the latest news. Journalism through the online newspaper publication is one of the areas that has overgrown due to internet explosion. With the widespread practice of online newspapers since 1970 [1], it has enabled the society to receive environmentally friendly, free, and instant interactive news updates that can be produced within a short time. Although the online news does not provide a detail description of an event, it offers a quick synopsis about what happened [2]. The online newspaper also benefits the news providers when it has fewer barriers to entry, more extensive distribution coverage, and lower distribution costs.

The online newspaper is a valuable resource for information not only to update the reader about what has happened but also to provide input for decision making through the news mining approach. Since the daily news, their volume is massive; thus, the processing and analysis of that news become more challenging. In line with volume, each page in the newspaper often contains many unrelated topics and unstructured knowledge that are scattered on every page that causes difficulty in extracting them [3]. The information in newspaper comprises of the cultural, social, and historical facts of specific regions that bring value to readers and interested parties but the knowledge value is difficult to be extracted easily from the newspapers [4]. Another challenge during the process is to overcome the issues of different words and concepts because of the different journalism practices. The news providers tend to use different words and concepts when reporting a similar event, although they are referring to similar contexts.

As countries are moving towards open data initiatives, the public inventories and databases need to be frequently updated to benefit the stakeholders to make the decision. Currently, the existing Malaysian databases shared in open data repositories are less informative and prone to slow updating rates when it highly relies on data providers to update the newest information into repositories. Since the benefit of the ever-growing mountains of digital newspapers, it is an opportunity to have an updated data repository through the recent news.

In this study, we identified the most relevant keywords used in daily online news that can be recommended as input for data repositories based on the mined keyword and concept. To extract the keyword, three Malaysian English online newspapers experimented – The Star Online, The Sun Daily, and The News Strait Time. This study chose the news about Malaysian Agro-food industries as a case study. Based on the text mining approach, specifically news mining, the hidden concept relevant to the agri-food industry that is hidden in the newspapers was extracted. These news articles from the three newspapers were scrapped from their websites with a predefined description of the time frame 2014-2017. Using the new mining algorithm called Rapid Automatic Keyword Extraction (RAKE) [18], the Agro keywords were identified based on the frequency of the highest terms, and their pattern among the three online newspapers was further analyzed. Thus, an adaptable dataset that can be periodically updated from various recourses related to Argo such as daily news can be recommended.

This paper is organized as follows. Section II is the related works that involve three discussions; the online newspaper, news mining, the Malaysian online newspaper, and the Agro-food industries. Then, the methodology used to conduct this study is discussed in Section III. In Section IV, the findings and discussion of the study are presented. The final sections conclude this work.

A. Recent Work

This section discusses the recent work on the online newspaper, news mining, and the information about three newspapers, and the back-ground of Agro-food industries

B. Online News Paper

News is a report that contains information about past, current, or future events, which is essential for the public to know. News often answers to the 5Ws questions- who, what, when, where and which or how. It can be delivered via many channels such as word of mouth, broadcasting in television and radio, print in the paper, or publish it over the internet. Article in the news generally has one or more of the following requirements - power elite, celebrity, entertainment, surprise, bad news, good news, magnitude, relevance, follow-up, and newspaper agenda [5]. News has two types of indexing that are time-based and news providers.

Online news or e-newspaper is remote access news that is published over the internet. It is like the hardcopy version; however, it contains additional content in other formats such as video, audio, hyperlinks, and an additional user can response to the published news. At the beginning years of its implementation, the reader's reactions were discouraging because they were accustomed to read printed newspapers [6]. However, this situation has changed when all ages are using smartphones, such the daily print newspaper in the UK was declined because of the widespread use of smartphones among its people [7]. [8] list out the top ten well-known e-newspaper websites in 2018 that lead by Yahoo! News, Google News, Huffington Post, CNN, New York Times, Fox News, NBC News, Mail Online, Washington Post, and The Guardian. Until May 2018, the Yahoo! News is the most popular online news websites with 175,000,000 readers and followed by Google News with 150,000,000 readers and Huffington Post with 110,000,000 readers. In Malaysia, many news providers provide online newspaper to Malaysian [9]. The Malaysian newspapers can be classified according to languages (Malay, English, Cantonese, Tamil, Multi Languages) and coverage (national or regional). Among the Malaysian online newspaper (based on language) are:

Malay

Berita Harian	- www.bharian.com.my/
Harian Metro	- www.hmetro.com.my/
Harakah Daily	- www.harakahdaily.net/
Utusan Malaysia	- www.utusan.com.my/
Kosmo Onlline	- www.kosmo.com.my/

English

Borneo Post	- www.theborneopost.com/
Daily Express Sabah	- www.dailyexpress.com.my/
Digital News Asia	- www.digitalnewsasia.com/
Free Malaysia Today	- www.freemalaysiatoday.com/
Ipoh Echo Ipoh	- www.ipohecho.com.my/
Malay Mail Online	- www.themalaymailonline.com/
Malaysia Chronicle	- www.malaysia-chronicle.com/
My Cen News	- www.mycen.com.my/news/
New Sarawak Tribune	- www.newsarawaktribune.com/
The New Straits Times	- www.nst.com.my/
The Edge Markets	- www.theedgemarkets.com/
The Heat Online	- www.theheatonline.asia/
The Rakyat Post	- www.therakyatpost.com/
The Star Online	- www.thestar.com.my/
The Sun Daily	- www.thesundaily.my/
The Malaysian Insider	-wwwthemalaysianinsider.com/

Cantonese

China Press	- www.chinapress.com.my/
Guang Ming Daily	- www.guangming.com.my/
Nanyang Siang Pau	- www.nanyang.com/
Oriental Daily	- www.orientaldaily.com.my/
Sin Chew	- www.sinchew.com.my
Overseas Chinese Daily News	- www.ocdn.com.my/

Tamil

Tamil Nesan	-www.tamilnesan.com.my/

Multiple Languages

Bernama	- www.bernama.com/ (Malay, English, Cantonese, Spanish)
Malaysiakini	- www.malaysiakini.com/ (Malay, English, Cantonese, Tamil).

C. *News Mining*

Web mining is the process of mining the web using data mining or text mining techniques. It combines techniques from different disciplines including Natural Language Processing, Machine Learning, Probabilistic and Statistical techniques. Web mining can generally be categorized into 3 categories i.e. Web usage mining, web structure mining, and web content mining. Web usage mining is the process of investigating the web user's browsing behavior on the internet [21], while in the web structure mining analyses the website's hyperlinks to obtain the structure of the website [20]. Web content mining acts as a tool to discover useful information from web content [19]. Web content mining is the focus of this study particularly on the extraction of valuable content from the online news articles.

News mining is under the text mining area that explores and analyses large amounts of unstructured text to discover patterns, concepts and interesting keywords hidden in the data [10]. The news document such as articles in newspapers, magazines, and blogs are among the input of news mining. It is a challenging process to extract knowledge from a newspaper because news document is often vague, inconsistent and contradictory, however, the information inside them are accumulated from various parties which are valuable for decision making.

New mining technology has been practiced in several areas. For example, in stock price forecasting, the information form mass media is used as a basis to forecast inter-day stock prices. The behaviors identified from the media is used to determine the stock's future market condition whether it will rise or fall [11] and utilized news mining in analyzing the 2007 Kenyan elections news report [12]. In the studies, the political information from the local newspapers was analyzed and compared with the news published in Western newspapers (British and US). In line with that, a system to determine the polarity of Saudi public opinion called Aara' was developed using a new mining approach. In the system, the comments left by readers in e-newspaper were taken for mining to extract the public opinion polarity [13]. Furthermore, the online newspaper also can be utilized to inspect sentiment on certain occasions or events. Such as in [14], about 900 sentences from Malay newspapers were mined to seek valuable behavior among its readers using sentiment analysis.

Furthermore, news mining also has been explored to improve business performance. For example, [15] utilized the news mining technology to reveal the relationship interestingness among its competitor based on the occurrence of the company being cited in online news articles. In [16], the web news report related to solar cells were used to identify weak signal topics by exploiting keyword-based text mining. In their studies, weak signals are early indicators of critical events or trends to formulate new potential business ideas.

D. *Online News in Malaysia*

There are many online newspapers in Malaysia as explained in the previous section. In this study, three online English newspapers were chosen for experiments which are The Star Online, The Sun Daily, The NST Online.

1) *The Star Online*: The Star Online (http://www.thestar.com.my) is Malaysia's first news website, which was launched on June 23, 1995. With the aim to provide readers with up to date breaking news and comprehensive information, the Star Online covers current news, business, sports, community, technology, property, job, world news, and lifestyle. In 2014, this website has been awarded as one of the best in Asia by the World Association of Newspapers and News Publishers (WAN-IFRA). As to build a strong loyal relationship with its online community readers, the Star connects them through Twitter and Facebook. Evolving with times, The Star Online also offers its contents through The Star ePaper and mobile app instead of the conventional printed paper. Besides that, readers' on-the-go who want breaking news and business updates delivered to them can opt for their SMS services available via Maxis and DiGi.

2) *The Sun Daily*: The Sundaily (http://www.thesundaily.my/) is an online newspaper published by Sun Media Corporation Sdn. Bhd, which is part of the Berjaya Media Group. The printed version is called The Sun was launched on 1 June 1993. The Sun is the first national free daily newspaper in tabloid form. Which is available from Mondays to Fridays except on public holidays with a target audience of white-collar workers and urban youth. It covers national and international news as well as sports, property, media & marketing and lifestyle & entertainment. The Sun has won several awards for advertising such as The Media Partner of the Year Award (2004/2005) and Excellence in Public Service Journalism and Opinion Writing (2007 and 2008).

3) *The NST Online*: The NST Online (https://www.nst.com.my/) is an online newspaper of the printed version New Straits Times. It is the oldest newspaper in Malaysia, which has been founded as The Straits Times in 1845 and was re-established as the "New Straits Times" in 1974. This newspaper belongs to Media Prima Berhad and printed by The New Straits Times Press (M) Bhd. Since its birth, this newspaper previously followed the example of British Newspapers, The Times and The Independent and later it has been revolutionized into tabloid size in April 2005. The section in the New Straits Times consists of Business Times, Property Times, Tech n U, Travel, and Premier League Plus. While the weekly section includes Sunday People, Focus, Learning Curve, and Cars, Bikes & Trucks.

E. *Agro Food Industries.*

The Agro-food industry refers to the business of producing agriculturally based food that covers a wide range of activities that utilizing farm, animal and forestry-based products as raw materials. The industry can be categorized into two; on the land that refers to agriculture in nature where its harvest is the final product. Secondly, it is on the table where they are processed where the harvest becomes foods [17]. There is a broad range of Agro-food commodities such as grains and pulses; oilseeds and nuts; fruits and vegetables; roots and tubers; meat and dairy products; honey and sugar; spices and stimulants. In general, the Agro sector can be classified into three major agriculture tasks that are fishery, livestock, and agriculture. The fishery is an activity that has a relation with water, ocean, catching fish, or other sea animals. The livestock is any Agro activity related to animal farming, while agriculture involves the cultivation of the soil for the growing of crops.

In Malaysia, the Agro-food sector is one of the national key result areas. It is monitored by the Ministry of Agriculture and Agro-Based Industry, which responsible for designing, coordinating and ensuring the implementation of the agricultural development Agro-food program. Under the Malaysian National Agro-Food Policy (2011-2020), the government is keen to ensure the availability, affordability and accessibility of food security and safety as well as the competitiveness and sustainability of the Agro-food industry. One of the critical areas in the Eleventh Malaysia Plan (EMP) (2016-2020) is modernizing agriculture which aims to accelerate growth and promote the agriculture sector to be a modern and dynamic sector as well as increase the competitiveness of agriculture produce. In the EMP, the agriculture sector, namely the Agro-food and industrial commodity subsectors will be transformed and modernized into a high-income and sustainable sector. This sector is expected to grow at 3.5% per annum, contributing 7.8% to GDP in 2020.

II. Materials and Method

This study applied the data mining methodology precisely to the text mining approach. The methodology is divided into 4 phases that are news extraction, news pre-processing, news mining and news analysis. The methodology is depicted in Fig. 1.

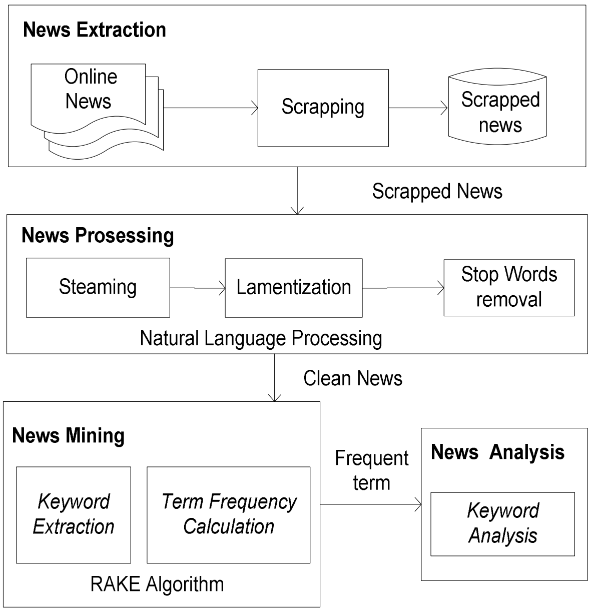


Fig. 1: The methodology

A. News Extraction

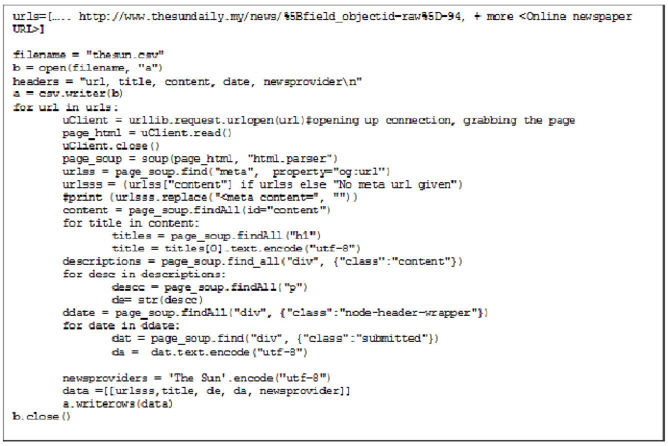
This phase involved scrapping news articles from online newspaper websites. Three Malaysian online English newspaper providers were chosen – The Star Online (http://www.thestar.com.my), The Sun Daily (http://www.thesundaily.my/), and The News Strait Time. These news articles were related to the agricultural events, policies and development made by the Malaysian Government raging from various sectors and products of agricultural implementation of the country. The searching process focused on Agro-Food sector news whereby the keyword ‘Agro-food’ and ‘’agriculture food’ were taken as a filter key. The timestamp to sample the news was set from 2014-2017. Table I depicts the sampling period for each online news website.

TABLE I
Sampling Period for Malaysian Online Newspaper Website

Online Newspaper	Period	#Days
The Star Online	6/03/2014- 24/06/2017	1206
The Sun Daily	12/06/2014-23/07/2017	1137
News Strait Time	12/06/2014-23/07/2017	1137

A scrapping algorithm written in Python has applied to the online newspaper websites as shown in Fig. 2. During the process, five important news components were recorded that are

- URL - URL link of the news
- Title - Title of the news
- Description - Full news text
- Date – Published date (Day, DD Month//YYYY)
- Newspaper provider name- The Sun, The Star, The NST



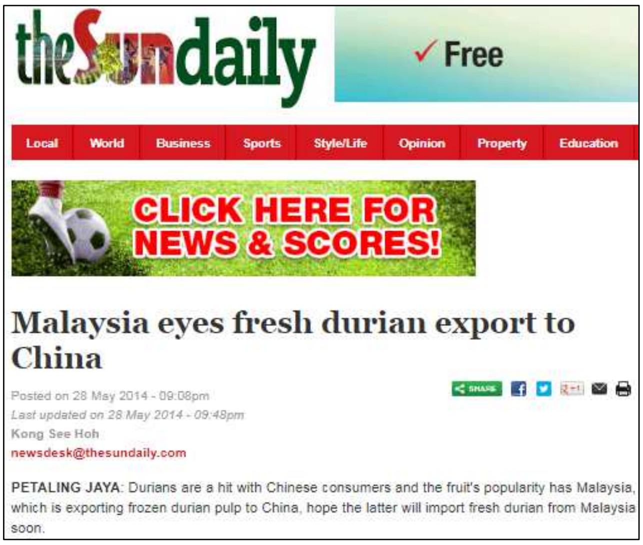
```
urls=[.... http://www.thesundaily.my/news/%5Bfield_objectid-raw%5D-94, + more <Online newspaper
URL>]

filename = "thesun.csv"
b = open(filename, "a")
headers = "url, title, content, date, newsprovider\n"
a = csv.writer(b)
for url in urls:
        uClient = urllib.request.urlopen(url)#opening up connection, grabbing the page
        page_html = uClient.read()
        uClient.close()
        page_soup = soup(page_html, "html.parser")
        urlss = page_soup.find("meta",  property="og:url")
        urlsss = (urlss["content"] if urlss else "No meta url given")
        #print (urlsss.replace("<meta content=", ""))
        content = page_soup.findAll(id="content")
        for title in content:
                titles = page_soup.findAll("h1")
                title = titles[0].text.encode("utf-8")
        descriptions = page_soup.find_all("div", {"class":"content"})
        for desc in descriptions:
                descc = page_soup.findAll("p")
                de= str(descc)
        ddate = page_soup.findAll("div", {"class":"node-header-wrapper"})
        for date in ddate:
                dat = page_soup.find("div", {"class":"submitted"})
                da =  dat.text.encode("utf-8")

        newsproviders = 'The Sun'.encode("utf-8")
        data =[[urlsss,title, de, da, newsprovider]]
        a.writerows(data)
b.close()
```

Fig. 2: The methodology

The information displayed in Fig. 3 is the sample of the original online news from The Sun Daily newspaper on 28th April 2014 (http://www.thesundaily.my/news/1061922). The news was scrapped as shown in Fig. 4.



Malaysia eyes fresh durian export to China

Posted on 28 May 2014 - 09:08pm

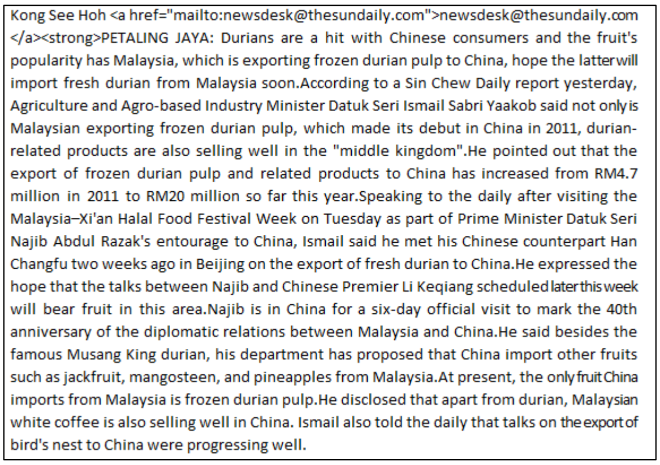
Last updated on 28 May 2014 - 09:48pm

Kong See Hoh

newsdesk@thesundaily.com

PETALING JAYA: Durians are a hit with Chinese consumers and the fruit's popularity has Malaysia, which is exporting frozen durian pulp to China, hope the latter will import fresh durian from Malaysia soon.

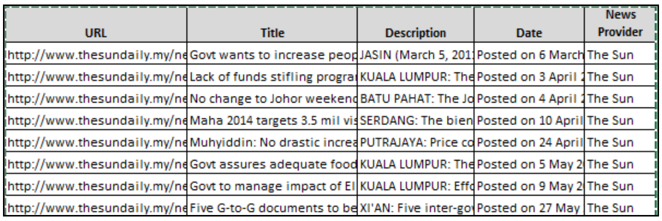
Fig. 3: The Sun Daily newspaper on 28th April 2014



Kong See Hoh <a href="mailto:newsdesk@thesundaily.com">newsdesk@thesundaily.com </a><strong>PETALING JAYA: Durians are a hit with Chinese consumers and the fruit's popularity has Malaysia, which is exporting frozen durian pulp to China, hope the latter will import fresh durian from Malaysia soon.According to a Sin Chew Daily report yesterday, Agriculture and Agro-based Industry Minister Datuk Seri Ismail Sabri Yaakob said not only is Malaysian exporting frozen durian pulp, which made its debut in China in 2011, durian-related products are also selling well in the "middle kingdom".He pointed out that the export of frozen durian pulp and related products to China has increased from RM4.7 million in 2011 to RM20 million so far this year.Speaking to the daily after visiting the Malaysia–Xi'an Halal Food Festival Week on Tuesday as part of Prime Minister Datuk Seri Najib Abdul Razak's entourage to China, Ismail said he met his Chinese counterpart Han Changfu two weeks ago in Beijing on the export of fresh durian to China.He expressed the hope that the talks between Najib and Chinese Premier Li Keqiang scheduled later this week will bear fruit in this area.Najib is in China for a six-day official visit to mark the 40th anniversary of the diplomatic relations between Malaysia and China.He said besides the famous Musang King durian, his department has proposed that China import other fruits such as jackfruit, mangosteen, and pineapples from Malaysia.At present, the only fruit China imports from Malaysia is frozen durian pulp.He disclosed that apart from durian, Malaysian white coffee is also selling well in China. Ismail also told the daily that talks on the export of bird's nest to China were progressing well.

Fig. 4: The scrapped news

The deliverables of this phase are a collection of Malaysian News regarding Agro-Food Sector in RSS format and stored in .csv format. Fig 5 shows the list of the Agro related news which was scrapped from The Sun Daily Newspaper.



URL	Title	Description	Date	News Provider
http://www.thesundaily.my/ne	Govt wants to increase peop	JASIN (March 5, 201	Posted on 6 March	The Sun
http://www.thesundaily.my/ne	Lack of funds stifling progra	KUALA LUMPUR: The	Posted on 3 April 2	The Sun
http://www.thesundaily.my/ne	No change to Johor weekend	BATU PAHAT: The Jo	Posted on 4 April 2	The Sun
http://www.thesundaily.my/ne	Maha 2014 targets 3.5 mil vis	SERDANG: The bien	Posted on 10 April	The Sun
http://www.thesundaily.my/ne	Muhyiddin: No drastic increa	PUTRAJAYA: Price co	Posted on 24 April	The Sun
http://www.thesundaily.my/ne	Govt assures adequate food	KUALA LUMPUR: The	Posted on 5 May 2	The Sun
http://www.thesundaily.my/ne	Govt to manage impact of El	KUALA LUMPUR: Effo	Posted on 9 May 2	The Sun
http://www.thesundaily.my/ne	Five G-to-G documents to be	XI'AN: Five inter-go	Posted on 27 May	The Sun

Fig. 5: The scrapped news

B. *News Pre-processing*

The aim of this phase is to prepare data for mining in the next step. In this phase, the scrapped news ware pre-processed. It involved removing irrelevant item and symbol. To do that, the natural language processing (NLP) approach was applied to the raw input file. The general NLP steps were stemming i.e. removing suffixes and prefixes from words, lemmatization i.e. a process of finding root word by just not removing suffixes and prefixes but to refer to dictionaries to find word’s lemma (roots), and removing stop word i.e. removing words such as prepositions that do not contribute to the knowledge.

C. *News Mining*

In this phase, the relevant Agro-food keyword that is hidden inside the scrapped news was mined. There were two processes involved in keyword extraction and term frequency calculation.

1) Keyword extraction: Extracting keywords is one of the most important tasks when working with text. Keywords describe the main topics expressed in a document. In this study, we focused on two specific tasks and their evaluation. It is extracting the essential words and phrases that appear in each text. There are three main components in keyword extraction algorithms, which are candidate selection, properties collection, and scoring and selecting keywords. The candidate selection process involves extracting all possible words, phrases, terms or concepts that can potentially be keywords. After that, it is followed by properties calculations that count prominent candidate text occurrence. It means that, for each possible candidate, the algorithm counts properties that indicate that it may be a keyword for Agro-food. Lastly is the scoring and selecting keywords process where each candidate is given a score. All candidates can be scored by either combining the properties into a formula or using a machine learning technique to determine the probability of a candidate being a keyword. A score on the number of keywords is then used to select the final set of keywords.

2) Term Frequency Calculation: Term frequency is often used in information retrieval and text mining. It is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Typically, the tf-idf weight is composed by two terms: the first computes the normalized Term Frequency (TF), aka. The number of times a word appears in a document, divided by the total number of words in that document. The second term is the Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.

- TF: Term Frequency, which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization: TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).
- IDF: Inverse Document Frequency, which measures how important a term is. While computing TF, all terms are considered equally important. However, it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus, we need to weigh down the standard terms while scaling up the rare ones by computing the following: IDF(t) = loge (Total

number of documents / Number of documents with term t in it).

To mine the news, we adopted the Rapid Automatic Keyword Extraction (RAKE) [18] which includes extracting the keyword and calculating the term frequency, as shown in Algorithm 1. RAKE is domain independence where it can operate independently on documents without referring to a corpus and has good performance in terms of precision, simplicity and computational efficiency. It tries to determine key phrases in a body of text by analyzing the frequency of word appearance and its co-occurrence with other words in the text. The results from the algorithm were sorted descending.

```
Input: Scrapped news from the website
Output: A list of the most occurrence keywords
1.  Start
2.  Scrap news from websites
3.  Pre-process the news to filter out unwanted text
      -  Split the document into an array of words, breaking it at word
         delimiters (like spaces and punctuation).
4.  Split the words into sequences of contiguous words, breaking each
    sequence at a stop word. Each sequence is now a "candidate
    keyword".
5.  Calculate the "score" of each individual word in the list of candidate
    keywords. This is calculated using the metric
                    degree(word)/frequency(word)

6.  For each candidate keyword, add the word scores of its constituent
    words to find the candidate keyword score.
7.  Take the first one-third highest scoring candidates from the list of
    candidates as the final list of extracted keywords.
8.  Sort the extracted keywords
9.  End
```

Algorithm 1: Rapid Automatic Keyword Extraction (RAKE) [18]

D. News Analysis

News analysis is the final stage where the identified keywords in news mining phase are analyzed using the descriptive analysis. This involves the identification of the most frequent keywords and classification of keywords based on Agro food term used in all newspapers.

III. RESULT AND DISCUSSION

This section reports the finding of the study. It covers the outcomes of the scrapped Agro news from three online news providers and the extraction of relevant Agro keyword using news-mining approach. The first activity was to scrap the news articles that contain information related to Agro food. Using the Python scrapper, the Agro food related news were scrapped into a .csv file format. From this process, 458 news articles were extracted for further analysis as shown in Table II.

TABLE II
THE NUMBER OF EXTRACTED ONLINE NEWS RELATED TO AGRO FOOD FROM THREE NEWSPAPER PROVIDERS

Newspaper	# Agro Related News
The Star Online	143 (31.2%)
The Sun	173 (37.8%)
News Strait Time	142 (31.0%)
Total	458

The extracted news from the three online newspapers then presented to the RAKE algorithm to discover the important Agro keywords and calculate the term frequency. The result is displayed in Table III. The mining result discovered that there were 146 Agro-related keywords in The Star, The Sun, and The News Strait Time and the distribution of the figures between newspapers was not significantly different. The highest Agro terms can be seen in The Star Online (35.13%), followed by The Sun (33.78%), and The News Straits Times (31.08%). Out of the total identified keyword's, they were repetitively found 720 times in all newspapers – The Sun (35.14%), News Straits Times (29.03%), and The Star Online (35.83%).

TABLE III
TOTAL NUMBER OF IMPORTANT AGRO KEYWORDS EXTRACTED FROM THE STAR, THE SUN, AND THE NEWS STRAITS TIMES.

<table>
<tr><th colspan="3">The Sun #253</th></tr>
<tr><td colspan="3">fish (24), palm oil (22), rice(19), fruits(18), livestock (16), vegetables (16), chicken (14), crops (12), beef (8), durian (8), farmers (8), paddy (8), padi (7), seafood (7), breeders (5), coffee (5), poultry (5), cattle (4), farms (4), meat (4), seeds (4), food (3), herbs (3), plants (3), animals (1), beverages (1), biodiesel (1), bird (1), citrus (1), cockle (1), crude (1), elephant (1), fields (1), frozen (1), lamb (1), mangos teen (1), mussels (1), mutton (1), onion (1), orange (1), orchard (1), park (1), pork's (1), prawn (1), roselle (1), sea (1), seedlings (1), shellfish (1), turtle (1), wildlife (1)</td></tr>
<tr><th colspan="3">News Straits Times #209</th></tr>
<tr><td colspan="3">palm oil (36), fruits (16), fish (15), rice (14), vegetable (14), livestock's (12), plants (8), durian (7), padi (7), animals (7), meat (4), crop (4), animal (4), honey (4), rubber (4), coconut (4), spices (4), sea (4), chicken (4), plantations (3), beef (3), milk (3), farming (3), poultry (3), pudding (1), oyster (1), pelagic (1), sharks (1), dairy (1), grass (1), juice (1), crude (1), pure (1), deer (1), goats (1), seedlings (1), boneless (1), tropical (1), stingless (1), green (1), dorsata (1), super fruits (1), lemon (1), fauna (1), drink (1), salmon (1)</td></tr>
<tr><th colspan="3">The Star Online #258</th></tr>
<tr><td colspan="3">Palm oil (37), Rice (23), Vegetables (21), Fruits (20), Livestock (16), Fish (13), Crops (9), Meat (9), Farming (8), padi (8), corn (7), poultry (7), animal (6), food (6), beef (5), breeders (4), cattle (4, chicken (4), eggs (4), seafood (4), seeds (4), fishmeal (3), organic (3), plantation (3), planting (3), arowana (1, beverage (1), birds (1), black (1), buffalo (1), coconut (1), cooking (1), cow (1), cultivation (1), fishing (1), gliders (1), granaries (1), grass (1), lizard (1), peelers (1), pepper (1), prawn fishing (1), programmes (1), rear (1), seaweed (1), spices (1), staple (1), stockpiles (1, sweet (1), tea (1), tropical (1), wild (1)</td></tr>
<tr><td></td><td>#Occurrence</td><td>#Agro keywords</td></tr>
<tr><td>The Sun</td><td>253 (35.14%)</td><td>50 (33.78%)</td></tr>
<tr><td>News Straits Times</td><td>209 (29.03%)</td><td>46 (31.08%)</td></tr>
<tr><td>The Star Online</td><td>258 (35.83%)</td><td>52 (35.14%)</td></tr>
<tr><td>Total</td><td>720 (100%</td><td>146 (100%)</td></tr>
</table>

From the observation in Table III, there were certain similar keywords used in each newspaper when to report news about Agro such the palm oil, rice, fruits, and fish terms. However, there were also some keywords that only found once in every newspaper such as pork, prawn, and tea. Besides that, there were also different keywords but referring to similar concept which cause duplication. One of the issues is because of the grammar and the inconsistency used of Malay and English keywords such 'Padi' and 'Paddy'., the similar keywords were merge as one keyword. After resolved this conflict by merging the similar keywords, the total number of important Agro keywords was reduced to 102 from 146 as displayed in in Table IV.

Based on the identified keyword list in Table III, the next analysis was to identify the keyword similarities between newspapers. For that, a frequency matric among the newspapers was constructed as depicted in Table IV. The table sorts the Agro relevant keywords based on the descending total frequencies.

TABLE IV
Frequency matric BETWEEN The Star Online, News Straits Time, and The Sun

No	Agro Food Code	Agro Terms	The Star Online	News Straits Times	The Sun	Total
1	3	palm oil	37	36	22	95
2	3	rice	23	14	19	56
3	3	fruits	20	16	18	54
4	1	fish	13	15	24	52
5	3	vegetable	21	14	16	51
6	2	livestock	16	12	16	44
7	3	paddy	8	7	15	30
8	3	crop	9	4	12	25
9	2	chicken	4	4	14	22
10	2	animal	6	11	1	18
11	2	meat	9	4	4	17
12	2	beef	5	3	8	16
13	3	durian		7	8	15
14	2	poultry	7	3	5	15
15	4	farming	8	3		11
16	3	plants		8	3	11
17	1	seafood	4		7	11
18	2	breeders	4		5	9
19	4	food	6		3	9
20	2	cattle	4		4	8
21	4	farmers			8	8
22	3	seeds	4		4	8
23	3	corn	7			7
24	3	plantation	3	3		6
25	3	coconut	1	4		5
26	3	coffee			5	5
27	1	sea		4	1	5
28	3	spices	1	4		5
29	2	eggs	4			4
30	4	farms			4	4
31	4	honey		4		4
32	3	rubber		4		4
33	1	fishmeal	3			3
34	3	herbs			3	3
35	2	milk		3		3
36	3	organic	3			3
37	3	planting	3			3
38	3	beverage	1		1	2
39	2	bird	1		1	2
40	3	crude		1	1	2
41	3	grass	1	1		2
42	3	seedlings		1	1	2
43	3	tropical	1	1		2
44	1	arowana	1			1
45	3	biodiesel			1	1
46	4	black	1			1
47	2	boneless		1		1
48	2	buffalo	1			1
49	3	citrus			1	1
50	1	cockle			1	1
51	4	cooking	1			1
52	2	cow	1			1
53	4	cultivation	1			1
54	2	dairy		1		1
55	2	deer		1		1
56	4	dorsa		1		1
57	4	drink		1		1
58	4	elephant			1	1
59	2	fauna		1		1
60	3	fields			1	1
61	1	fishing	1			1
62	4	frozen			1	1
63	3	gliders	1			1
64	2	goats		1		1
65	3	granaries	1			1
66	3	green		1		1
67	3	juice		1		1
68	2	lamb			1	1
69	3	lemon		1		1
70	4	lizard	1			1
71	3	mangos teen			1	1
72	1	mussels			1	1
73	2	mutton			1	1
74	3	onion			1	1
75	3	orange			1	1
76	3	orchard			1	1
77	1	oyster		1		1
78	4	park			1	1
79	3	peelers	1			1
80	1	pelagic		1		1
81	3	pepper	1			1
82	2	pork’s			1	1
83	1	prawn			1	1
84	1	prawn fishing	1			1
85	4	programmes	1			1
86	3	pudding		1		1
87	4	pure		1		1
88	4	rear	1			1
89	3	roselle			1	1
90	1	salmon		1		1
91	4	seaweed	1			1
92	1	sharks		1		1
93	1	shellfish			1	1
94	4	staple	1			1
95	1	stingless		1		1
96	4	stockpiles	1			1
97	3	super fruits		1		1

98	3	sweet	1			1
99	3	tea	1			1
100	1	turtle			1	1
101	4	wild	1			1
102	4	wildlife			1	1
	Total		258	209	253	720
Agro Food Code: 1- Fishery, 2- Livestock, 3- Agriculture, 4-Others						

The information in Table IV shows that there are 12 Agro keywords (item no 1-12) which are considered as the most important as their appearance in all news providers- palm oil, rice, fruits, fish, vegetable, livestock, paddy, crop, chicken, animal, meat, and beef. The bar chart in Fig.6 shows the frequency (%) of top 12 Agro keywords found in The Star Online, The Sun Daily, and News Straits Time. The keyword 'palm oil' is the most popular keywords when it has the highest frequency- 95 keywords (13.1%) appeared in three newspapers. Palm oil was found 37 times (38.9%) in The Star Online, 26 times (37.9%) in New Straits Time, and repeated 22 times (23.2%) in the Sun Daily. The important keywords were followed by rice, fruits, fish, and vegetable with the total counts recorded between 56-51 counts (7.1-7.1%). The other important keywords found in all newspapers are livestock, paddy, crop, chicken, animal, meat, and beef. Moreover, the keywords that rank between 44 and 102 in Table IV are considered as less important because they were appeared only once in one of the newspapers which total about 58 keywords

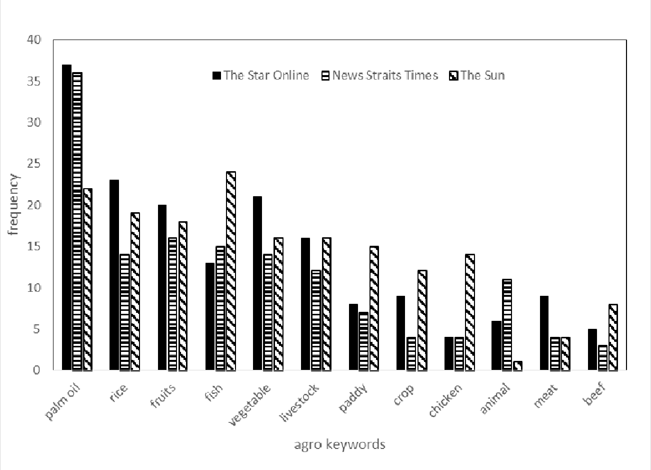


Fig. 6: The frequency of 12 Agro keywords that found in The Star Online, News Straits Time, and the Sun

We also classified the identified keywords in Table IV into four Agro classification codes that are 1 - 'Fishery', 2- 'Livestock', 3- 'Agriculture', 4- 'Miscellaneous' in order to identify which newspapers are more likely to what areas (the second column). The classification is based on the Agro food definition, which are focusing on the four major Agro food tasks that are fishery, livestock, agriculture and miscellaneous. The miscellaneous is a group to represent any keyword that can be categorized into more than one Agro food classification codes. The classification result is displayed in Fig. 7 where the agriculture was found as the most frequent Agro keywords in all newspaper (58%) and it were followed by the livestock (23%), fishery (12%), and miscellaneous (7%). This shows that the agriculture is the most prominent issues which frequently reported in Malaysian newspapers than the other three Agro sectors. This is in line with the agriculture sector as one of the main contributors to the national economy. As indicated in Table IV, the first, second, and third important Agro keywords are from the agriculture sector.

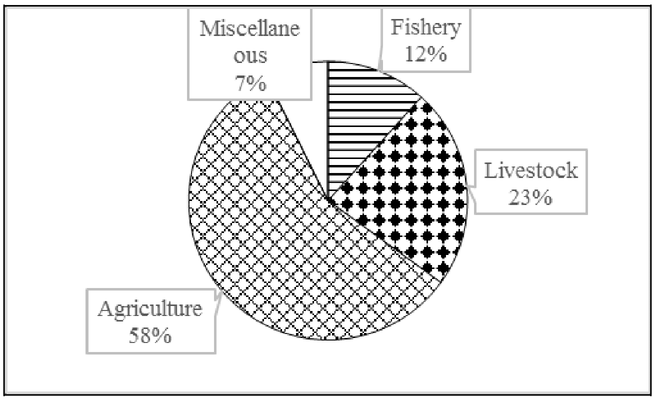


Fig. 7: The classification of newspaper according to Agro food classification code in all newspaper

In Table V, we then filtered and sorted the top ten important keywords based on classification code and highest ranking – agriculture, livestock, fishery, and miscellaneous. The information in Table V indicates the most important keywords in each classification code.

TABLE V
TOP TEN KEYWORDS BASED ON CLASSIFICATION CODE

Rank	Classification Code	Keywords	The Star Online	NST	The Sun	Total
1	Agriculture	palm oil	37	36	22	95
		rice	23	14	19	56
		fruits	20	16	18	54
		vegetable	21	14	16	51
		paddy	8	7	15	30
		crop	9	4	12	25
		durian		7	8	15
		plants		8	3	11
		seeds	4		4	8
		corn	7			7
		Total	129	106	117	352
2	Livestock	livestock	16	12	16	44
		chicken	4	4	14	22
		animal	6	11	1	18
		meat	9	4	4	17
		beef	5	3	8	16
		poultry	7	3	5	15
		breeders	4		5	9
		cattle	4		4	8
		eggs	4			4
		milk		3		3

		Total	59	40	57	156
3	**Fishery**	seafood	4		7	11
		sea		4	1	5
		fishmeal	3			3
		arowana	1			1
		cockle			1	1
		fishing	1			1
		mussels			1	1
		oyster		1		1
		pelagic		1		1
		prawn			1	1
		Total	9	6	11	26
4	**Miscellaneous**	farming	8	3		11
		food	6		3	9
		farmers			8	8
		farms			4	4
		honey		4		4
		black	1			1
		cooking	1			1
		cultivation	1			1
		dorsata		1		1
		drink		1		1
		Total	17	9	15	41

We further investigated the Agro food classification code in each newspaper and the results are different when we emphasize at each single newspaper. From the analysis, the Sun Daily newspaper published more news related to fishery than the others when it has the highest score (44%). Interestingly, The Star Online has the highest score for the others three areas–livestock (36.9%), agriculture (35.9%), and miscellaneous (45.3) keywords. This is summarized in Figure 8.

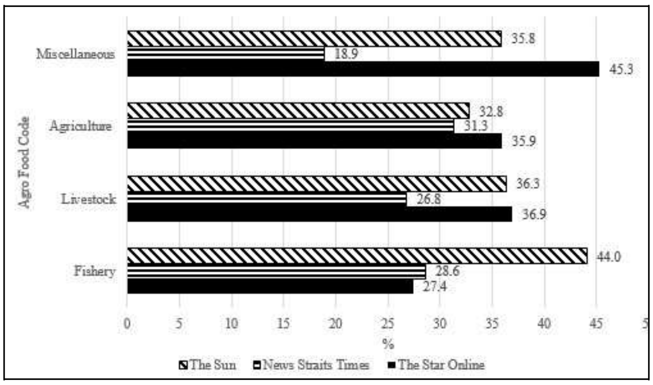


Fig. 8: The classification of newspaper according to Agro food classification code in each online newspaper

To aid the visualization of the important keywords, we constructed word clouds based on the frequencies of the keywords. Fig. 9 depicts the word clouds of the important Agro keywords found in this study. As can be seen in Fig. 9, the 12 words which has bigger fonts are palm oil, rice, fruits, fish, vegetable, livestock, paddy, crop, chicken, animal, meat, and beef.

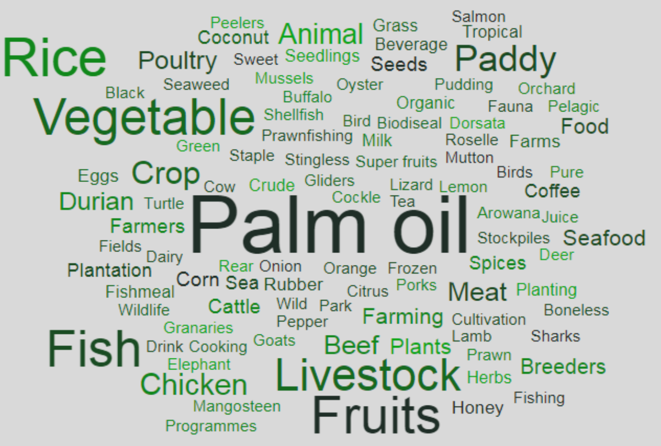


Fig. 9: The classification of newspaper according to Agro food code

IV. Conclusions

In this study, we identified the most relevant Agro food keywords that had been frequently used in The Star Online, The Sun Daily, and The Straits Times News. Using the RAKE algorithms, the agriculture-based keywords was found as the most frequent keywords in all newspaper (58%) and it was followed by the livestock (23%), fishery (12%), and miscellaneous (7%). Through the analysis, 146 keywords related to Agro food sector were found, being repeated 720 times in all newspapers and the highest Agro terms was found in The Star Online (35.13%), followed by The Sun (33.78%), and The News Straits Times (31.08%). Moreover, there were 12 keywords considered as the most relevant keywords in Agro food sector when they had appeared in all three newspapers. The 'palm oil' was the most popular keywords and it was found 37 times (38.9%) in The Star Online, 26 times (37.9%) in News Straits Time, and repeated 22 times (23.2%) in the Sun Daily. The other keywords after the palm oil were rice, fruits, fish, vegetable, livestock, paddy, crop, chicken, animal, meat, and beef. The identified keywords can be recommended as input to form future Agro inventory. In future, this study will further examine the relevant Agro food keywords in several layers where we believe that the keyword relationship found in each layer can improve the reliability of the finding.

Acknowledgment

The authors gratefully acknowledge the UUM Generation Research Grant Scheme (SO Code 13436) for supporting this research project.

References

[1] S. Steve, “"Plato People” Reunite, Honor Founder,” Culture, 1997. [Online]. Available: https://www.wired.com/1997/03/platopeople-reunite-honor-founder/. [Accessed: 28-Jun-2018].

[2] A. Taylor, The People’s Platform: Taking Back Power and Culture in the Digital Age. 2014.

[3] R. Ø. Nørv˚ag, Kjetil, “News Item Extraction for Text Mining inWeb Newspapers,” in International Workshop on Challenges in Web Information Retrieval and Integration, 2005.

[4] M. S. and R. W. A Yzaguirre, “Newspaper archives + text mining = rich sources of historical geo-spatial data,” IOP Conf. Ser. Earth Environ. Sci. 34, vol. 34, pp. 1–8, 2016.

[5] D. O. Tony Harcup, "WHAT IS NEWS?" Journal. Stud., vol. 18,no. 12, pp. 1470–1488, 2017.

[6] S. Carina, Ihlström Eriksson, Åkesson, Maria Nordqvist, "From Print to Web to e-paper - the challenge of designing the e- newspaper," in International Council for Computer Communication (ICCC), 2004, pp. 249–260.

[7] J. Edwards, "For every £154 newspapers lose in print revenue, they gain only £5 on the digital side," Business Insider UK, 2017. [Online]. Available: http://uk.businessinsider.com/statistics-smartphones-print-newspaper-revenues-2017-2/?IR=T. [Accessed: 09-Jul-2018].

[8] EBizMBA, "Top 15 Most Popular News Websites," eBizMBAInc, 2018. [Online]. Available: http://www.ebizmba.com/articles/ newswebsites.

[9] Malaysia Central, "Malaysian News: List Of Online Media, Newspapers, Dailies, Print Versions, News Portals, Independent Media, Alternative Press & News Agencies, News Sources & Publications," MALAYSIA CENTRAL: The Leading Malaysia-Centric Info Portal, 2016. [Online]. Available: http://www.mycen.com.my/malaysia/news.html. [Accessed: 18- Jul-2018].

[10] G. S. L. Vishal Gupta, "A Survey of Text Mining Techniques and Applications," J. Emerg. Technol. Web Intell., vol. 1, no. 1, pp.60–79, 2009.

[11] J. Z. Xiangyu Tang, Chunyu Yang, "Stock Price Forecasting by Combining News Mining and Time Series Analysis," in ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, 2009, pp. 1–5.

[12] S. Pollak, R. Coesemans, W. Daelemans, and N. Lavrač, "Detecting contrast patterns in newspaper articles by combining discourse analysis and text mining," *Pragmatics. Q. Publ. Int. Pragmat. Assoc.*, vol. 21, no. 4, pp. 647–683, 2011.

[13] S. M. A. Aqil M. Azmi, "Aara'– a system for mining the polarity of Saudi public opinion through e-newspaper comments," J. Inf. Sci., vol. 40, no. 3, 2014.

[14] S. P. and N. A. R. Mazidah Puteh, Norulhidayah Isa, "Sentiment Mining of Malay Newspaper (SAMNews) Using Artificial Immune System," in Proceedings of the World Congress on Engineering, 2013, pp. 1–6.

[15] O. R. L. S. ZhongmingMa, GautamPant, "Mining competitor relationships from online news: A network-based approach," Electron. Commer. Res. Appl., vol. 10, no. 4, pp. 418–427, 2011.

[16] J. Yoon, "Detecting weak signals for long-term business opportunities using text mining of Web news," Expert Syst. Appl., vol. 39, no. 16, pp. 12543–12550, 2012.

[17] E. D. Goodman, "Agro- Food Studies in the 'Age of Ecology': Nature, Corporeality, Bio- Politics," J. Eur. Soc. Rural Sociol., vol. 39, no. 1, pp. 17–38, 1999.

[18] S. R. D. E. N. C. W. Cowley, "Automatic Keyword Extraction from Individual Documents," in Text Mining: Applications and Theory, M. W. B. Kogan, Ed. John Wiley & Sons, Ltd, 2010, pp. 1–120.

[19] Vidya, S., & Banumathy, K. (2015). Web Mining-Concepts and Applications. International Journal of Computer Science and Information Technologies, 6(4), 3266-3268.

[20] Mughal, M. J. H. (2018). Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview. International Journal of Advanced Computer Science and Applications, 9(6).

[21] Mebrahtu, A., & Srinivasulu, B. (2017). Web Content Mining Techniques and Tools. International Journal of Computer Science and Mobile Computing, 6(4).