How to Rewrite Malaysian History[1]

Ang Li-Lian

28th March 2022

Minerva University

[1] This is one section of my three-part senior Capstone project.

Distant Reading: What do the people say about vernacular education?

Unlike in Section 1, where we close read primary and secondary sources, we've seen two examples of how digital tools gives a macro view of data, allowing us to process and analyse large amounts of data. Distant reading works in the same vein on textual data where you can parse through thousands of works in seconds as opposed to the months you would spend poring over text with your own eyes.[2]

This case study will attempt to understand how vernacular education is represented in the media through web scraped media articles in both English and Malay from 2015 to 2020 and 2017 to 2020, respectively, for a final database of 188 English and 125 Malay articles. Then, the articles were analysed on three NLP tasks: sentiment analysis, emotion analysis and topic modelling. Each of these tasks provides a macro-view of the trends of discussion over time as represented by the media. All code and datasets used for the methods below are available on GitHub.

In the following case study, I chose media articles as a proxy for public opinion based on the premise that journalists write catering to their audience's background and opinion, making sampling from many media houses more representative of public opinion than social media. Since conducting more research about the media landscape, I have understood that the premise is flawed because of the rife media censorship in Malaysia, revealing more about Malaysia's censorship laws than public opinion.[3] Nevertheless, although subject to self-imposed censorship through the annual renewal of printing licenses and laws against freedom of expression, several online media outlets like Malaysiakini have sprung up in the Internet era.[4] Purely online media outlets are not immune to self-censorship, but gathering and analysing their articles do provide a rich, new source base about Malaysia.

Short Background on Malaysian Vernacular Education

Vernacular schools are national-type schools whose primary language of instruction is Mandarin or Tamil instead of the national language, Malay. They exist mainly at the primary level of education and are called Sekolah Jenis Kebangsaan Cina, SJK(C), and Sekolah Jenis Kebangsaan Tamil, SJK(T), respectively. Primary schools whose main language of instruction is Malay are called Sekolah Kebangsaan, SK. These vernacular schools were born out of several debates between the Chinese, Malay and Indian communities, Malaysia's three main ethnic communities, between the British colonial period and Malaysia's independence.[5] The issue of the education system became entwined with the expression of ethnic identity and political power, leading to a compromise between a single stream of national education and

[2] Shawn Graham, Ian Milligan, and Scott Weingart, 'Putting Big Data to Good Use: Historical Case Studies', in *The Historian's Macroscope - Working Title*, Open Draft Version (London: Imperial College Press, 2013), http://web.archive.org/web/20160315222926/http://www.themacroscope.org/?page_id=599.

[3] 'Malaysia : Back to Harassment, Intimidation and Censorship', Reporters without borders, accessed 22 October 2021, https://rsf.org/en/malaysia; 'Malaysia: Freedom in the World 2021 Country Report', Freedom House, accessed 22 October 2021, https://freedomhouse.org/country/malaysia/freedom-world/2021.

[4] Abdul Razak bin Dato' Hussein, *The May 13 Tragedy: A Report*, 86; Goh, *The May 13th Incident and Democracy in Malaysia*, 6; Jonathan Head, 'Malaysiakini: The Upstart That Changed Malaysia's Media Landscape', *BBC News*, 19 February 2021, sec. Asia, https://www.bbc.com/news/world-asia-54277437.

[5] Vivien Wong and Ken Wong Tze, 'Chinese Schools in Malaysia – Between Ethnic Aspirations and the Challenges of Forging a National Education', in *The Cultural Legacies of Chinese Schools in Singapore and Malaysia*, ed. Hoe Yow Cheun and Jingyi Qu (Routledge, 2021), 124–28, https://doi.org/10.4324/9781003009610.

multiple streams serving each community's interests.[6] There are no restrictions on which national schools any race can enrol into, but parents prefer to enrol their children into schools based on their mother tongue, leading to the racial composition of schools seen in Figure 1, where there is increased stratification in primary schools based on racial group.

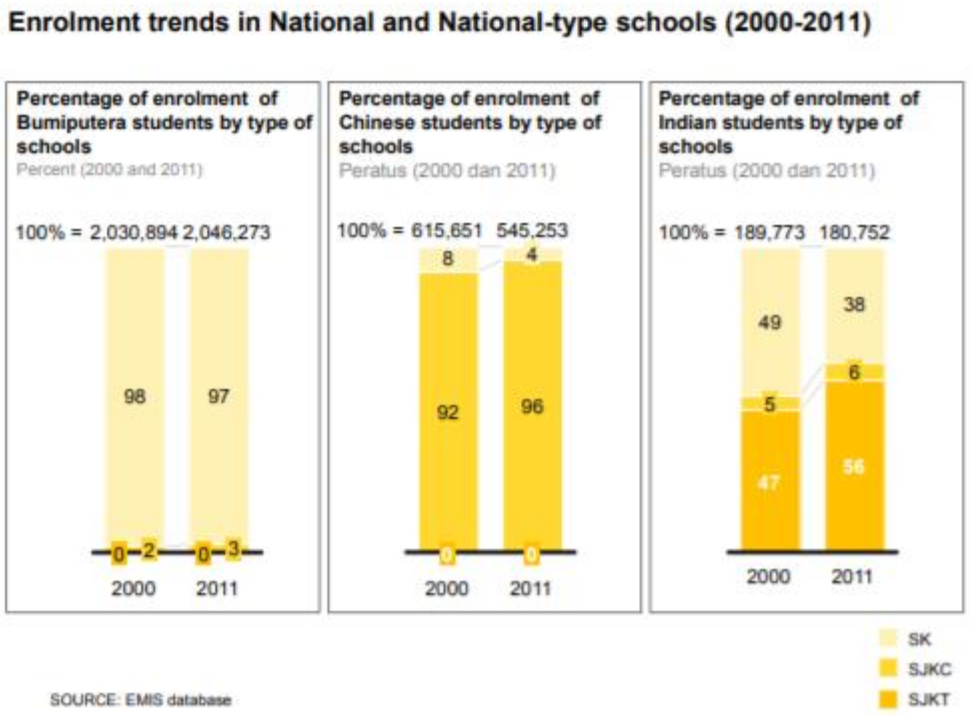


Figure 1 Enrolment trends between 2000 and 2011 by school type between Malay (Bumiputera), Chinese and Indian students in Malaysia.[7]

Building a database of articles

The following methodology follows some of the best practices of a systematic literature review to ensure its reproducibility and reduce bias during screening. I defined the main question, search and filtering criteria in the planning phase before collecting articles to prevent bias from seeing the final database from affecting design choices. I scraped and filtered the articles based on the pre-defined criteria in the execution phase. The scope of articles ultimately scraped were constrained by time and computer processing power. A summary of the results is shown in Figure 2.

[6] Ibid.

[7] Ministry of Education Malaysia, 'Malaysia Education Blueprint 2013-2025 (Preschool to Post-Secondary Education)' (Putrajaya: Ministry of Education, 2013), 324, https://www.moe.gov.my/menumedia/media-cetak/penerbitan/dasar/1207-malaysia-education-blueprint-2013-2025/file.

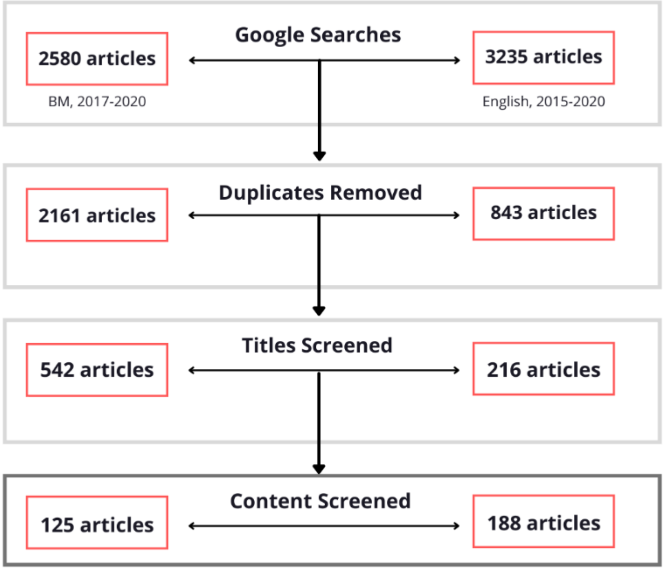


Figure 2 Summary of article filtering process where the left and right column represent Malay and English language articles, respectively.

Phase 1: Planning

The main question for this section is, "how does the media represent vernacular primary education?" I used Google as a search engine because of its easily accessible GoogleNews Python library and the wealth of online resources for web scraping. I defined the search terms as 'vernacular schools malaysia' ('sekolah vernakular malaysia'), 'chinese school malaysia' ('sekolah cina malaysia') and 'tamil school malaysia' ('sekolah tamil malaysia').[8] I scraped all the articles on private browsing to prevent my personal browsing history from skewing the results. Location cookies can also be removed by using a Virtual Private Network (VPN).

After scraping the articles, I defined a specific set of criteria to filter them. Firstly, I removed duplicate articles if they had the same title and media source. Then, based on content, the first screening looks at just the article title and the second screening looks at the article's contents. I filtered the articles if they were about:

- Secondary level education
- Private education
- Education as a whole
- Non-Malaysian education
- Events happening in vernacular schools, but not about vernacular education

The final criteria can be difficult to discern. For example, one of the articles in the dataset, 'Challenge against khat in vernacular schools thrown out' by *Free Malaysia Today*,

[8] Originally, the only search term used was 'vernacular schools malaysia' ('sekolah vernakular malaysia'), but the data set left out many articles which didn't use the specific term. Instead, articles also referred to the vernacular schools as Chinese or Tamil schools, so the search was expanded to capture these articles.

was rejected even though the discusses the article is related to vernacular schools.[9] However, the content of the articles does not express any views on vernacular schools, only on the debate of making Jawi calligraphy a compulsory subject for all students. Nevertheless, I have made all scraped data available on GitHub for others to make discernments of their own, especially for subjective criteria.

Phase 2: Execution (Web scraping)

I scraped the articles to extract their title, media source, date, link and textual content. However, the English and Malay articles had to be scraped using different methods because the Python libraries used for web scraping English articles do not support the Malay language.

The alternative method of scraping Malay articles is slower and more computationally expensive than English ones. It takes upwards of five hours to scrape each year's worth of Malay articles compared to English articles, which take less than an hour and has far fewer bugs. Due to time constraints, Malay articles were only scraped from 2017 to 2020, whereas the English articles were scraped from 2015 to 2020. Ultimately, the different periods did not drastically impact the amount of English to Malay articles.

Dates were extracted twice because some of the dates given by Google are not specific enough, like '1 month ago' or '2 weeks ago'. For these cases, the date from the article's webpage was extracted based on the page metadata (Malay) or the algorithm from the Python library (English).

English Articles

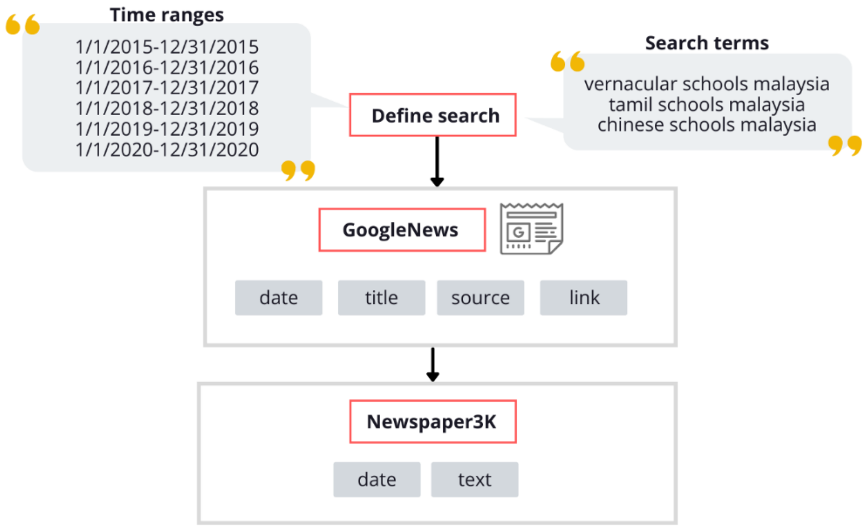


Figure 3 Visual representation of web scraping English articles.

The GoogleNews Python library was developed to search through GoogleNews and return the title, source, date, link and description for all search results per page. However, bombarding a website with requests is typically bad practice because it slows down website

[9] FMT Reporters, 'Challenge against Khat in Vernacular Schools Thrown Out', *Free Malaysia Today (FMT)* (blog), 20 April 2021, https://www.freemalaysiatoday.com/category/nation/2021/04/20/challenge-against-khat-in-vernacular-schools-thrown-out/.

traffic and is usually a flag for abuse. This could result in your IP address being blocked, so there must be a waiting time between each request, which I set as twenty seconds.[10]

Next, I used the Newspaper3k Python library to extract the full articles text and the article's publication date. The Newspaper3k library was developed specifically to scrape articles.[11] It has been trained to detect the article text, author, publishing date and performs NLP tasks like summarising and generating keywords.[12] Even though the library is available in multiple languages, Malay is not one of them. I only used Newspaper3k for extraction for more control over the NLP tasks.

Malay Articles

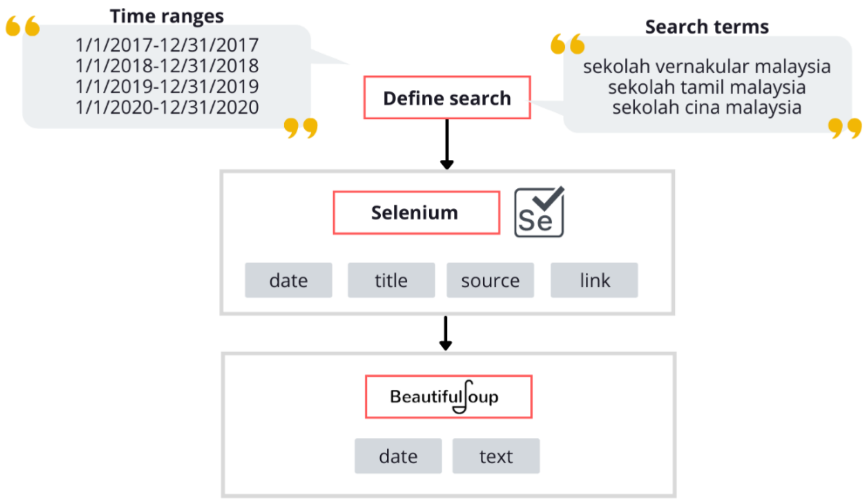


Figure 4 Visual representation of web scraping Malay articles

Since GoogleNews only returns English language articles, even with Malay search terms, I've opted to use Selenium. Selenium is a Python library that automates repetitive tasks, which in this use case is copying the title, source, date and link for all the search results on the page and repeating the process for all search results pages.[13] Then, Selenium goes to each link and waits for BeautifulSoup4, a Python library that parses through the HTML of each web page to extract the article content and publishing date.[14] BeautifulSoup4 does this by looking for all the 'p' (paragraph) tags of the web page and extracting text and searching for date-like objects in the 'script' tags.

[10] Hurin Hu, *GoogleNews: Google News Search for Python*, version 1.5.9, OS Independent, Python, accessed 4 December 2021, https://github.com/Iceloof/GoogleNews.

[11] 'Newspaper3k: Article Scraping & Curation — Newspaper 0.0.2 Documentation', accessed 4 December 2021, https://newspaper.readthedocs.io/en/latest/.

[12] Ibid.

[13] '1. Installation — Selenium Python Bindings 2 Documentation', accessed 4 December 2021, https://selenium-python.readthedocs.io/installation.html#introduction.

[14] 'Beautiful Soup Documentation — Beautiful Soup 4.9.0 Documentation', accessed 4 December 2021, https://www.crummy.com/software/BeautifulSoup/bs4/doc/.

This process is much slower than the one for scraping English articles because Selenium was built to test websites, so it goes through each task as a human would, whereas GoogleNews and Newspaper3k were built for speed. Nevertheless, both methods returned the relevant information.

Phase 3: Execution (Filtering)

Based on the predefined criteria, I filtered through all the articles. The first task was to remove all duplicates. I considered two entries duplicates if they shared the same title and media source. Then, I read the title of each article and removed those which were obviously unrelated, such as those about foreign affairs or reporting school closures due to flooding. Finally, I skimmed through the contents of all remaining articles and removed those that did not fit the criteria. I manually populated articles with missing content, media sources or publication dates and corrected those with miscategorised language tags.

Although there was an imbalance in articles in the initial database, it was reversed after the duplicates were removed and ultimately became balanced. With the extra two years of articles for English, media articles are proportionally the same for both English and Malay.

Natural Language Processing (NLP)

NLP is a form of distant reading, pulling out patterns from the text that you would not normally notice when close reading. NLP leverages machine learning models, which are trained to identify patterns. Sentiment analysis and emotion analysis are supervised learning tasks meaning that they are trained on a data set that was tagged based on specific categories. It outputs the probability that each article holds a specific sentiment (positive, negative and neutral) or emotion (happy, angry, sad, surprise and fear). The output for both analyses is a list of numbers between 0 and 1 which sum to 1, whereas topic modelling is an unsupervised learning task that finds patterns without predefined categories. The output is a set of topics that each have ten words. Each word is associated with a probability to show how much they represent the topic. These topics are defined subjectively, based on one's perception of what the group of words mean.

BERT

In this paper, I use BERT (Bidirectional Encoder Representations from Transformers), one of the breakthrough models in NLP, for most tasks. Developed in 2018 by Google AI, the BERT model was designed for transfer learning, serving as a foundation for more specific and complicated NLP processes.[15] In this section, I aim to provide a high-level, intuitive understanding of the model because to interpret the results from the model fully, we must understand how to distil hundreds of articles down to a matrix of probabilities. I will not dive into the mathematics or introduce any equations to make this section accessible to those outside the field of computational sciences.

BERT is a training architecture for NLP rather than a canned model. That is, it isn't a fully formed model that you can just apply to your data but a method for training your data to create a model that can explain your data.[16] Let's draw on the analogy of teaching babies to

[15] Jacob Devlin et al., 'BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding', *ArXiv:1810.04805 [Cs]*, 24 May 2019, http://arxiv.org/abs/1810.04805.

[16] Jonathan Hui, 'NLP — BERT & Transformer.', Medium, 5 November 2019, https://jonathan-hui.medium.com/nlp-bert-transformer-7f0ac397f524.

speak to understand what BERT is at the highest level of abstraction. Parents usually exaggerate each syllable while speaking and expose the baby to language with books read aloud, conversations (unidirectional to start), videos or songs. Babies eventually associate certain mouth movements to specific sounds, sequences of sounds to meanings and patterns of word ordering to grammar and sentence structure. Essentially, we teach babies how to recognise and replicate patterns, and the same is true when you pick up a new language. With BERT, researchers teach the model how language works, so it can apply those learnings when given any other language or when it encounters new vocabulary. Developers can utilise BERT's highly sophisticated language-processing architecture and refine it to suit their needs. This saves developers and researchers time because they don’t have to train the model from scratch.

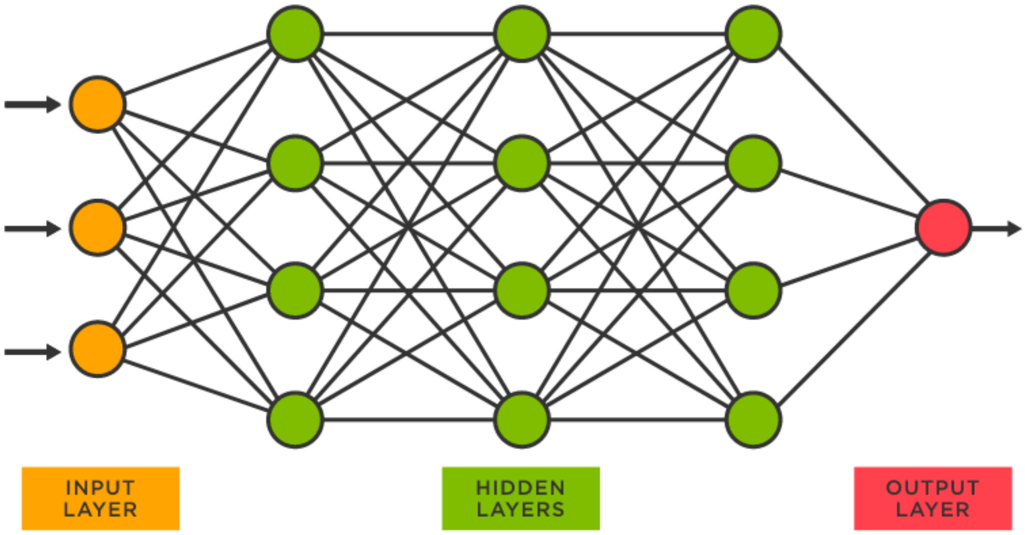


Figure 5 Simplified graphical representation of a neural network.[17]

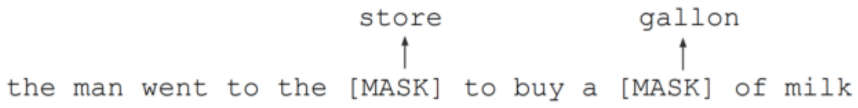
BERT is built on neural networks, which you can think of as a dense connection of ‘neurons’ to help you visualise the more technical aspects of how it works. There are broadly three parts to the neural network, as shown in Figure 5. We first transform our data into a form that the model can take and feed it into the input layer. For example, we might transform the words in the article into a numerical representation based on some rules. Then, the data is passed into the hidden layers where complicated processing happens and finally distilled in the output layer as a set of probabilities, a tag or any objective of our task. Each hidden layer is connected to all nodes from the previous layer and represents a specific feature of the data. The differences between layers are how many nodes there are in the layer, the weight or strength of the signal they pass on and how the node is activated. Much like how neurons in our brain are connected, each neuron fires at different strengths and takes different stimulus levels to activate it. Typically, the researcher fixes the number of nodes and the activation type for each layer, and the neural network is trained to find the weights for each layer to yield optimal results. We train a neural network through unsupervised learning where the objective is to find the set of weights which will minimise the loss function – our metric of success. The loss function helps score each set of weights and decides which set of weights to try next. An example of a loss

[17] ‘What Is a Neural Network?’, TIBCO Software, accessed 4 December 2021, https://www.tibco.com/reference-center/what-is-a-neural-network.

function is gradient descent, where the model always chooses the weights, which minimise the loss.

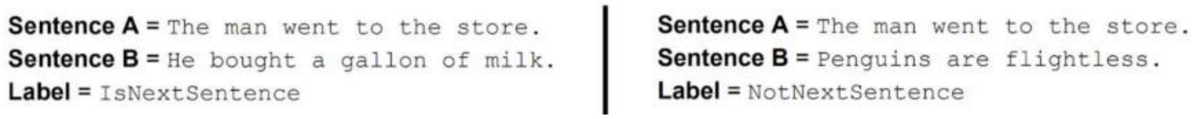
With BERT, we want to train our neural network to create a rich but abstract representation of language. Our first step is to transform the data via input embedding. Input embedding encodes the context of a word into a single numerical representation with three key pieces of information derived from simplifying words to reduce vocabulary size (word piece embedding), identifying which sequential context of words each token belongs to (segment embedding) and determining the absolute and relative position of a word (position embedding).[18]

Then, the data is inputted into the model, where it is pre-trained on two tasks to set the weights within the hidden layer – masked language model (Masked LM) and next sentence prediction (NSP).[19] Masked LM aims to teach the model how one word could be used in different contexts within a sentence by replacing 15% of the words in the corpus with an empty placeholder (Figure 6).[20] The model then predicts what the masked word is based on the context of the sentence. In this case, the model aims to maximise the number of correctly predicted words. NSP, on the other hand, trains the model to understand how different sentences are related to each other by picking two sentences (A & B) and getting the model to classify whether Sentence B follows Sentence A or not (Figure 7).[21] The loss function, in this case, is the number of correctly classified pairs of sentences. After pre-training the model, we end up with weights on the hidden layers that represent the linguistic properties of language.



```
the man went to the [MASK] to buy a [MASK] of milk
```

Figure 6 Example of how Masked LM masks words within a sentence.[22]



Sentence A = The man went to the store.
Sentence B = He bought a gallon of milk.
Label = IsNextSentence

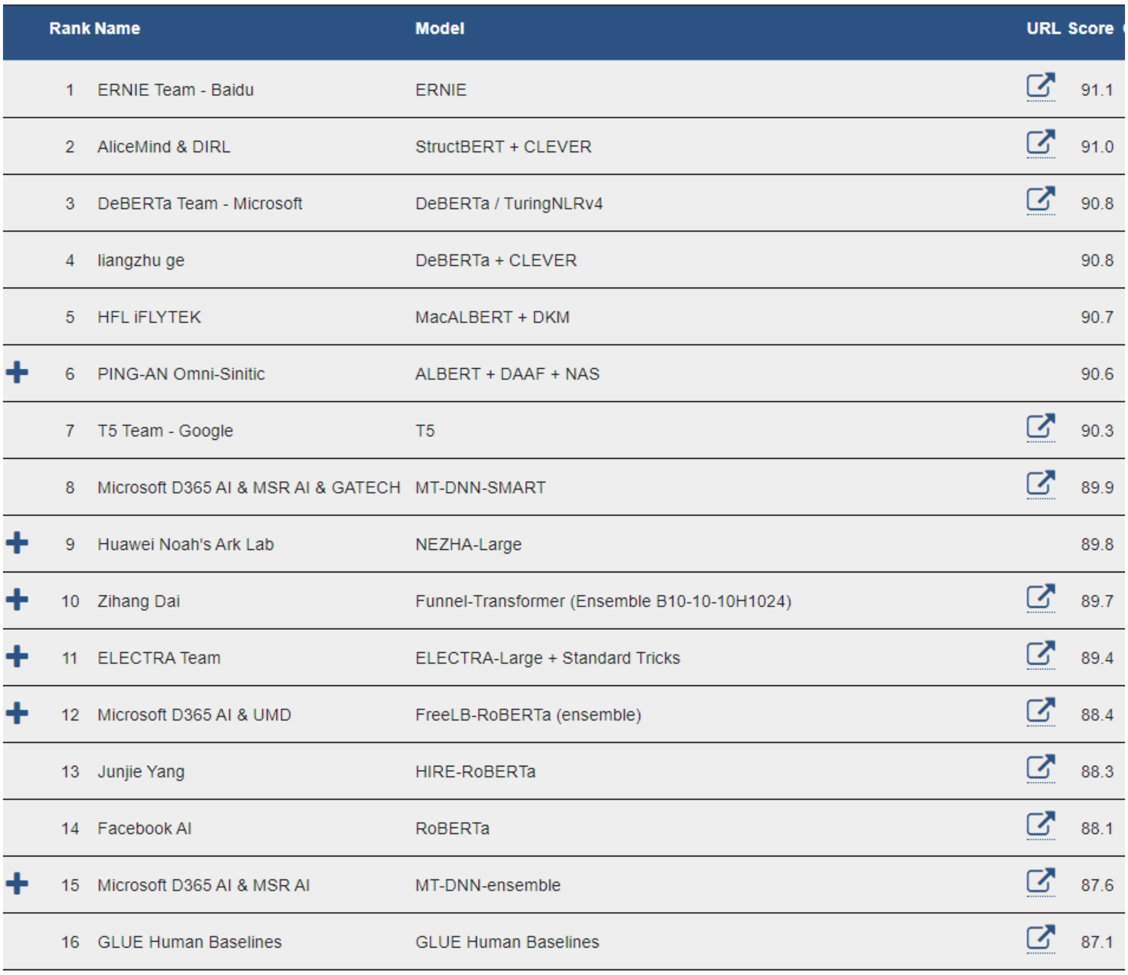
Sentence A = The man went to the store.
Sentence B = Penguins are flightless.
Label = NotNextSentence

Figure 7 Example of how NSP is used to categorise whether Sentence B does or does not come after Sentence A.[23]

Finally, BERT is fine-tuned to a specific NLP task like sentiment analysis, emotion analysis or topic modelling (there are a wide variety of NLP tasks other than these available!). Bear in mind that BERT is not being trained to apply to a specific language or dataset just yet. We are simply building the architecture on which data will be trained. This step defines the output layer and refines the weights in the hidden layer to output our desired objective, whether that is a sentiment or emotion classification or a list of topics.

[18] Rani Horev, 'BERT Explained: State of the Art Language Model for NLP', Medium, 17 November 2018, https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270.
[19] Horev.
[20] Ibid.
[21] Ibid.
[22] Hui, 'NLP — BERT & Transformer.'
[23] Ming-Wei Chang, 'BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding', n.d., 42.

To show how effective BERT is, we use GLUE (General Language Understanding Evaluation) to evaluate a model's natural language understanding (NLU) based on nine tasks benchmarked against human performance.[24] Unlike other tests, which only give the model one task, GLUE evaluates a model's ability to perform transfer learning which makes it more flexible.[25] An example of a complicated task is coreference resolution, which identifies what a pronoun is referring to in a complex sentence. For example, given the sentence "Lily spoke to Donna, breaking *her* concentration", what noun would best replace 'her'. Another example task more relevant to this paper is sentiment analysis, where the model determines the degree of positive, negative or neutral sentiment based on the given text. Each model (which can be a mix of multiple transformers) performs all nine tasks without tweaking the main model each time to make it specific to a new task.



	Rank	Name	Model	URL	Score
	1	ERNIE Team - Baidu	ERNIE	↗	91.1
	2	AliceMind & DIRL	StructBERT + CLEVER	↗	91.0
	3	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4	↗	90.8
	4	liangzhu ge	DeBERTa + CLEVER		90.8
	5	HFL iFLYTEK	MacALBERT + DKM		90.7
+	6	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6
	7	T5 Team - Google	T5	↗	90.3
	8	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART	↗	89.9
+	9	Huawei Noah's Ark Lab	NEZHA-Large		89.8
+	10	Zihang Dai	Funnel-Transformer (Ensemble B10-10-10H1024)	↗	89.7
+	11	ELECTRA Team	ELECTRA-Large + Standard Tricks	↗	89.4
+	12	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)	↗	88.4
	13	Junjie Yang	HIRE-RoBERTa	↗	88.3
	14	Facebook AI	RoBERTa	↗	88.1
+	15	Microsoft D365 AI & MSR AI	MT-DNN-ensemble	↗	87.6
	16	GLUE Human Baselines	GLUE Human Baselines	↗	87.1

Figure 8 GLUE leader board taken on 26 October 2021 with human performance highlighted placed at 16. The score column is the average, while the subsequent columns show the scores for each of the nine tasks.[26]

As you can see, BERT isn't the only model out there, and these models are often combined to improve NLU. In Figure 8, fifteen NLP models outperform the human benchmark based on their average over all nine tasks. This does not mean that BERT or any of these models

[24] Chris McCormick and Nick Ryan, 'GLUE Explained: Understanding BERT Through Benchmarks', Chris McCormick, 5 November 2019, https://mccormickml.com/2019/11/05/GLUE/.
[25] Ibid.
[26] Chris McCormick and Nick Ryan, 'GLUE Benchmark', GLUE, accessed 4 December 2021, https://gluebenchmark.com/.

are better in *all* tasks, but they do exhibit exceptional performance while saving the time and effort it takes to manually label each article and reducing personal bias during the process. However, machine learning models do not completely eliminate bias since bias can also be encoded during the model's training and development based on the assumptions and datasets used. Therefore, a combination of careful interpretation and sampling data to check the model's classification is necessary.

BERT can be trained with a labelled dataset with the desired language and corpus with this training architecture in hand. For the Malay language, I used Husein Zolkepli's Malaya[27] BERT transformer, which was trained on scraped Malay text from various sources.[28] While for the English language, we will use BERTopic for topic modelling, a combination of BERT and TF-IDF (term frequency-inverse document frequency) developed by Maarten Grootendorst.[29]

NLP Tasks: Sentiment Analysis, Emotion Analysis & Topic Modelling

Sentiment analysis and emotion analysis return a probability of each article holding a particular sentiment or emotion. This is known as a discriminative model that determines how information should be classified based on pre-defined labels. It uses the tags from pre-training and fine-tuning to determine how the articles should be classified. Outputting a probability instead of a single category gives us a better sense of how confident the model is in its prediction. For example, the article is given a 55% negative, 30% neutral and 15% positive probability instead of classifying an article as having a negative sentiment. This gives a richer representation of the sentiment scale rather than an arbitrary boundary deciding which sentiment an article holds.

On the other hand, topic modelling is a generative model which does not rely on a labelled dataset. Instead, the model determines how articles and words should be grouped. It outputs each grouping and probabilities representing the degree to which each word or article represents the group.

Model Metrics

The models need to be evaluated on their performance. Sentiment analysis and emotion analysis are evaluated based on three different metrics – precision, recall and F-score. Precision shows how many articles were correctly categorised for all the labelled articles. It shows how sensitive the model is in discerning which articles are of a certain category. Whereas precision shows, based on the actual labels, how many did we correctly categorise. It shows how good the model is at picking which articles are relevant to the category. However, having an extremely high score on one measure but not the other is a bad sign. For example, you could achieve 100% recall if you labelled all the data as one category since there are no false negatives – you have selected all the relevant articles. But you would score poorly on precision since the model failed to distinguish between classes. We balance these metrics using the F-score, given by the following formula in the case of having only two categories.

[27] Malaya also works with several other transformers like XL-Net, ELECTRA and ALBERT, but we will be using BERT here because for the three NLP tasks we are interested in, BERT gives the best scores overall. The difference in scores are only within a few percentage points of each other.

[28] Husein Zolkepli, 'Malaya', GitHub, 27 July 2021, https://github.com/huseinzol05/malaya.

[29] Maarten Grootendorst, 'Interactive Topic Modeling with BERTopic', Medium, 12 January 2021, https://towardsdatascience.com/interactive-topic-modeling-with-bertopic-1ea55e7d73d8.

$$F_1 = 2\frac{precision \cdot recall}{precision + recall}$$

With multiclass classification, the score weights each class by its size and averages it over all classes. The lowest score is 0 when precision and recall are zero, meaning everything is mislabelled. The highest score is 1 when precision and recall are perfect, meaning everything is labelled correctly.

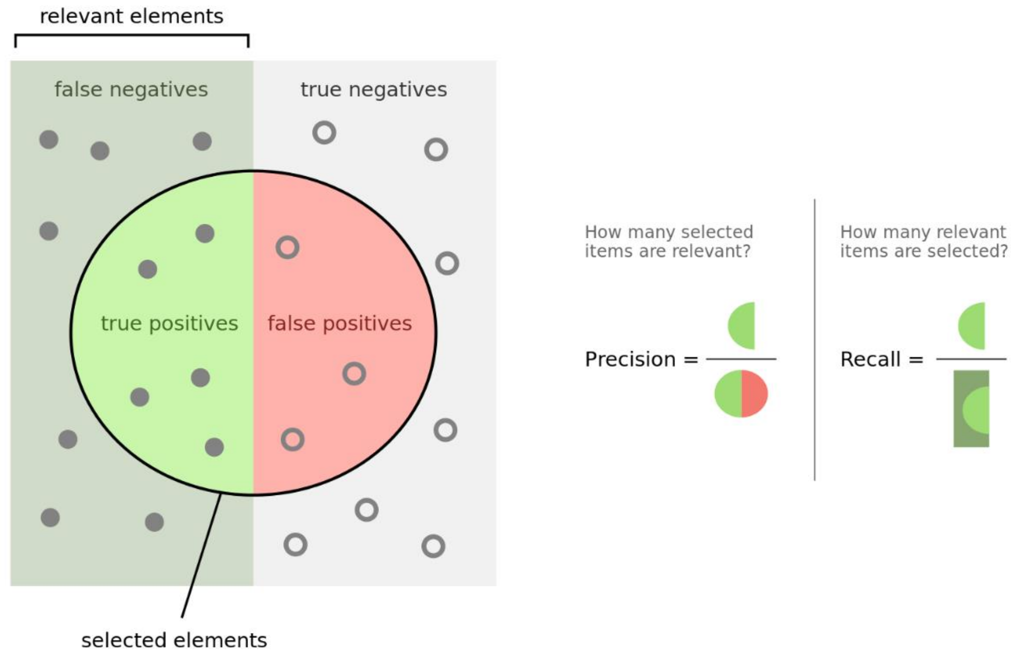


Figure 9 Visual representation of how precision and recall are calculated.[30]

For topic modelling, there are no labels to compare the outputs since the model itself generates the labels. Instead, one could use topic coherence, which measures how similar the words in each topic are. If the words support each other and have a similar theme or meaning, one can say that the topic is coherent. Note that topic coherence is used to evaluate a model's *general* ability to group words. For this case study, I will simply look at the output to determine if the topics outputted were coherent or not.

The following sub-sections will investigate the specific models applied to the English and Malay articles and critically evaluate their training methods.

Malaya (Analysing Malay Articles)

Malaya's training dataset differs for each NLP task, with sentiment analysis being trained on predominantly Twitter data and a significantly smaller set of news articles, Amazon, Yelp, and IMDB comments. Emotion analysis is also predominantly trained on translated Twitter data and Reddit comments from *Goemotion*'s tagged dataset. The transformers were trained on a larger variety of datasets like Wikipedia, social media, school essays, news articles,

[30] Walber, *Precision and Recall*, 22 November 2014, 440 x 800 pixels, 22 November 2014, https://en.wikipedia.org/w/index.php?title=Precision_and_recall&oldid=1050491609.

academic articles, and parliament proceedings. Their datasets are also open-source and available on GitHub.[31]

After gathering the raw data, the text was tagged through a four-step semi-supervised process to fine-tune the classification models and assess the model.[32] The main principle behind the process was to reduce the human effort and error of manually tagging millions of words, sentences and paragraphs. Firstly, Malaya generated a lexicon of words classed based on sentiment and emotion. To put it simply, they generated the lexicon from a small corpus of labelled words and attempted to find other words in the database that have a similar sentiment and emotion based on context to build an even larger lexicon.[33] Secondly, they used this lexicon and translated the text of English tagged datasets for weak learning. Weak learning is a fast way to label a large corpus with imprecise labels to overcome the constraints of resources required for precise tagging, often involving paying several trained, qualified individuals to manually tag and check a dataset.[34] Thirdly, the unsupervised tagged dataset goes through confident learning, which identifies label errors.[35] Finally, the labels on the dataset were checked through '5 iterations from humans'.[36] It is unclear how exactly this final step occurs, but the idea appears to inject some form of supervised learning into building the dataset.

Understanding how the dataset was tagged is crucial because it is the same dataset used to validate the model's accuracy. Malaya sets aside 20% of the data as testing data.[37] Therefore, receiving a high score on the model metrics does not necessarily mean that the model is accurate, but rather that it is accurate to the labels of the dataset. There are two points of entry for label errors. The first is from translated tagged English datasets. The labels were preserved with only the text translated using Google Translate, assuming that its translations were accurate enough.[38] The second is from the vagueness of the supervised portion of data tagging. It was alluded that linguists were paid, but it is unclear how much of the data they tagged, the linguists' background and how standardised the process was.[39] Therefore, I will also randomly sample the tags given by the model to assess the model's accuracy.

Of the transformers available, I chose to use BERT because it is the model I understand best, and the difference between the models is marginal given the close scores across precision, recall and F-score (Table 1 & Table 2).

[31] Husein Zolkepli, 'Malay Dataset', GitHub, 26 November 2021, https://github.com/huseinzol05/malay-dataset.

[32] Ibid.

[33] Read more from Hamilton, Clark, Leskovec & Jurafsky's paper "Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora". Retrieved from https://arxiv.org/pdf/1606.02820.pdf

[34] Alex Ratner et al., 'Weak Supervision: The New Programming Paradigm for Machine Learning · Stanford DAWN', Stanford Dawn, 16 July 2017, https://dawn.cs.stanford.edu/2017/07/16/weak-supervision/.

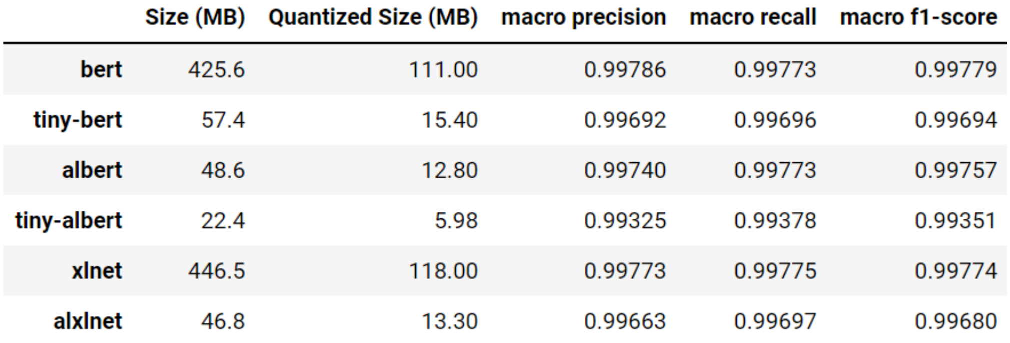
[35] Read more from Northcutt, Jiang and Chuang's paper "Confident Learning: Estimating Uncertainty in Dataset Labels". Retrieved from https://arxiv.org/abs/1911.00068

[36] Zolkepli, 'Malay Dataset', sec. How we gather the dataset.

[37] 'Emotion Analysis — Malaya Documentation', accessed 4 December 2021, https://malaya.readthedocs.io/en/latest/load-emotion.html; Husein Zolkepli, 'Sentiment Analysis — Malaya Documentation', Malaya, accessed 4 December 2021, https://malaya.readthedocs.io/en/latest/load-sentiment.html.

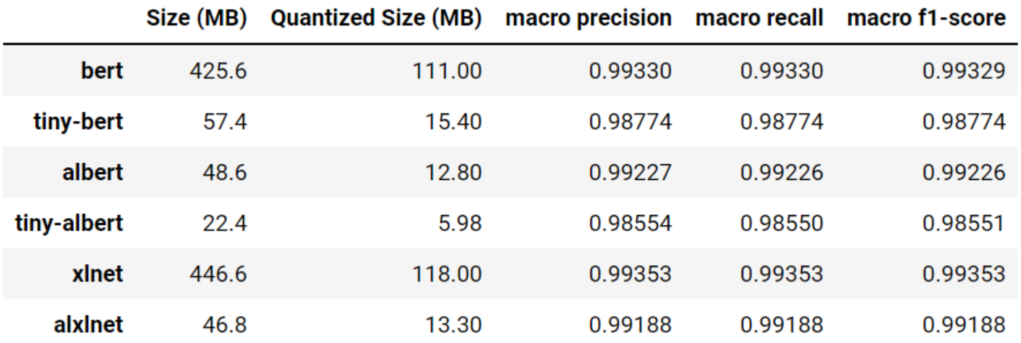
[38] Zolkepli, 'Malay Dataset'.

[39] Ibid.



	Size (MB)	Quantized Size (MB)	macro precision	macro recall	macro f1-score
bert	425.6	111.00	0.99786	0.99773	0.99779
tiny-bert	57.4	15.40	0.99692	0.99696	0.99694
albert	48.6	12.80	0.99740	0.99773	0.99757
tiny-albert	22.4	5.98	0.99325	0.99378	0.99351
xlnet	446.5	118.00	0.99773	0.99775	0.99774
alxlnet	46.8	13.30	0.99663	0.99697	0.99680

Table 1 Metrics for transformers in emotion classification. The prefix macro simply means an aggregate of the score across each emotion category.[40]



	Size (MB)	Quantized Size (MB)	macro precision	macro recall	macro f1-score
bert	425.6	111.00	0.99330	0.99330	0.99329
tiny-bert	57.4	15.40	0.98774	0.98774	0.98774
albert	48.6	12.80	0.99227	0.99226	0.99226
tiny-albert	22.4	5.98	0.98554	0.98550	0.98551
xlnet	446.6	118.00	0.99353	0.99353	0.99353
alxlnet	46.8	13.30	0.99188	0.99188	0.99188

Table 2 Metrics for transformers in emotion classification. The prefix macro simply means an aggregate of the score across each sentiment.[41]

Implementing the code itself is simple. After converting the article content into a list of strings with one article as each list element, I inputted the list into each model. Each article will be given two sets of probabilities for sentiments and emotions and a list of ten topics with ten words. Malaya outputs one additional emotion (love), which I will omit from the results for consistency with text2emotion.

BERTopic – English Topic Modelling

BERTopic is an algorithm that extracts the context of the words based on the text, clusters words into topics and creates topic representations.[42] By default, it extracts embeddings using BERT, but it can also be swapped with any other embedding technique. BERTopic differentiates between each cluster using c-TF-IDF[43], the class-based variant of the algorithm that defines the importance of words within a set of documents but defines each set as a topic with all related documents compiled. With this, c-TF-IDF will output the most important words within each topic.

BERTopic also supports multiple languages, so I will compare the topic model results for the Malay articles with that of Malaya. The only change is in defining the language to be ‘multilingual’. Since the dataset is relatively small, I can determine topic coherence by looking at the output of topics BERTopic by eye.

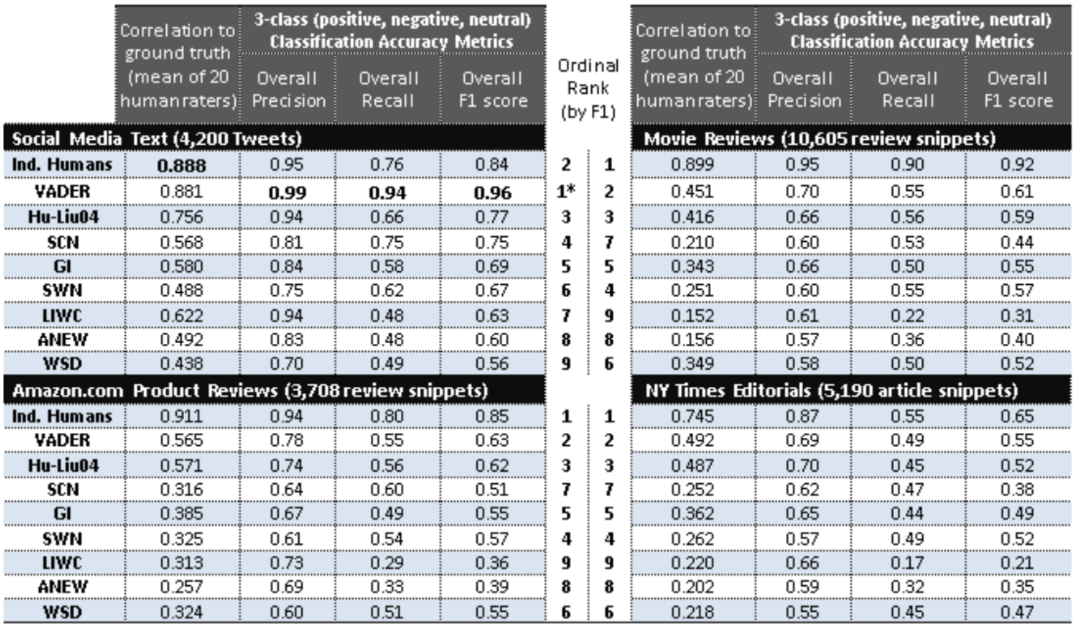
[40] ‘Emotion Analysis — Malaya Documentation’.
[41] Zolkepli, ‘Sentiment Analysis — Malaya Documentation’.
[42] Maarten Grootendorst, ‘The Algorithm’, BERTopic, accessed 4 December 2021, https://maartengr.github.io/BERTopic/tutorial/algorithm/algorithm.html.
[43] term frequency-inverse document frequency

VADER and text2emotion – English Sentiment and Emotion Analysis[44]

VADER (Valence Aware Dictionary for sEntiment Reasoning) is a Python library trained on social media content to determine the sentiment of text by a naïve summation of scores of each word based on a lexicon.[45] Similarly, text2emotion determines emotion based on a naïve summation of emotions based on a lexicon.[46] Both libraries remove stop-words and clean up text to calculate the proportion of words of a particular sentiment or emotion that make up the text.

In terms of NLP models, VADER and text2emotion are simplistic since they do not consider the word's context. Moreover, VADER is old, released in 2014. However, I could not find a pre-trained BERT model for English sentiment and emotion analysis, so these two libraries are the most convenient options.

VADER performs best with the domain of text it was originally trained on (social media) and performs the worst with article snippets (Table 3). In contrast, a similar model assessment is not available for text2emotion. Nevertheless, we can judge VADER and text2emotion's performance on our specific set of articles by randomly sampling some articles to compare our sentiment and emotion scores to the models' before evaluating their usefulness for the analysis.



	Correlation to ground truth (mean of 20 human raters)	3-class (positive, negative, neutral) Classification Accuracy Metrics			Ordinal Rank (by F1)		Correlation to ground truth (mean of 20 human raters)	3-class (positive, negative, neutral) Classification Accuracy Metrics		
		Overall Precision	Overall Recall	Overall F1 score				Overall Precision	Overall Recall	Overall F1 score
Social Media Text (4,200 Tweets)							Movie Reviews (10,605 review snippets)			
Ind. Humans	**0.888**	0.95	0.76	0.84	2	1	0.899	0.95	0.90	0.92
VADER	0.881	**0.99**	**0.94**	**0.96**	1*	2	0.451	0.70	0.55	0.61
Hu-Liu04	0.756	0.94	0.66	0.77	3	3	0.416	0.66	0.56	0.59
SCN	0.568	0.81	0.75	0.75	4	7	0.210	0.60	0.53	0.44
GI	0.580	0.84	0.58	0.69	5	5	0.343	0.66	0.50	0.55
SWN	0.488	0.75	0.62	0.67	6	4	0.251	0.60	0.55	0.57
LIWC	0.622	0.94	0.48	0.63	7	9	0.152	0.61	0.22	0.31
ANEW	0.492	0.83	0.48	0.60	8	8	0.156	0.57	0.36	0.40
WSD	0.438	0.70	0.49	0.56	9	6	0.349	0.58	0.50	0.52
Amazon.com Product Reviews (3,708 review snippets)							NY Times Editorials (5,190 article snippets)			
Ind. Humans	0.911	0.94	0.80	0.85	1	1	0.745	0.87	0.55	0.65
VADER	0.565	0.78	0.55	0.63	2	2	0.492	0.69	0.49	0.55
Hu-Liu04	0.571	0.74	0.56	0.62	3	3	0.487	0.70	0.45	0.52
SCN	0.316	0.64	0.60	0.51	7	7	0.252	0.62	0.47	0.38
GI	0.385	0.67	0.49	0.55	5	5	0.362	0.65	0.44	0.49
SWN	0.325	0.61	0.54	0.57	4	4	0.262	0.57	0.49	0.52
LIWC	0.313	0.73	0.29	0.36	9	9	0.220	0.66	0.17	0.21
ANEW	0.257	0.69	0.33	0.39	8	8	0.202	0.59	0.32	0.35
WSD	0.324	0.60	0.51	0.55	6	6	0.218	0.55	0.45	0.47

Table 3 VADER 3-class classification performance compared to individual human raters and seven established lexicon baselines across four distinct domain contexts (clockwise from upper left: tweets, movie reviews, product reviews, opinion news articles).[47]

[44] https://ojs.aaai.org/index.php/ICWSM/article/view/14550/14399

[45] C. Hutto and Eric Gilbert, 'VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text', *Proceedings of the International AAAI Conference on Web and Social Media* 8, no. 1 (16 May 2014): 216–25.

[46] Karan Bilakhiya, 'How I Created A Python Package - Text2emotion', *Analytics India Magazine* (blog), 10 September 2020, 2, https://analyticsindiamag.com/how-i-created-a-python-package-text2emotion/.

[47] Hutto and Gilbert, 'VADER', 224.

Results

Sentiment and Emotion Analysis

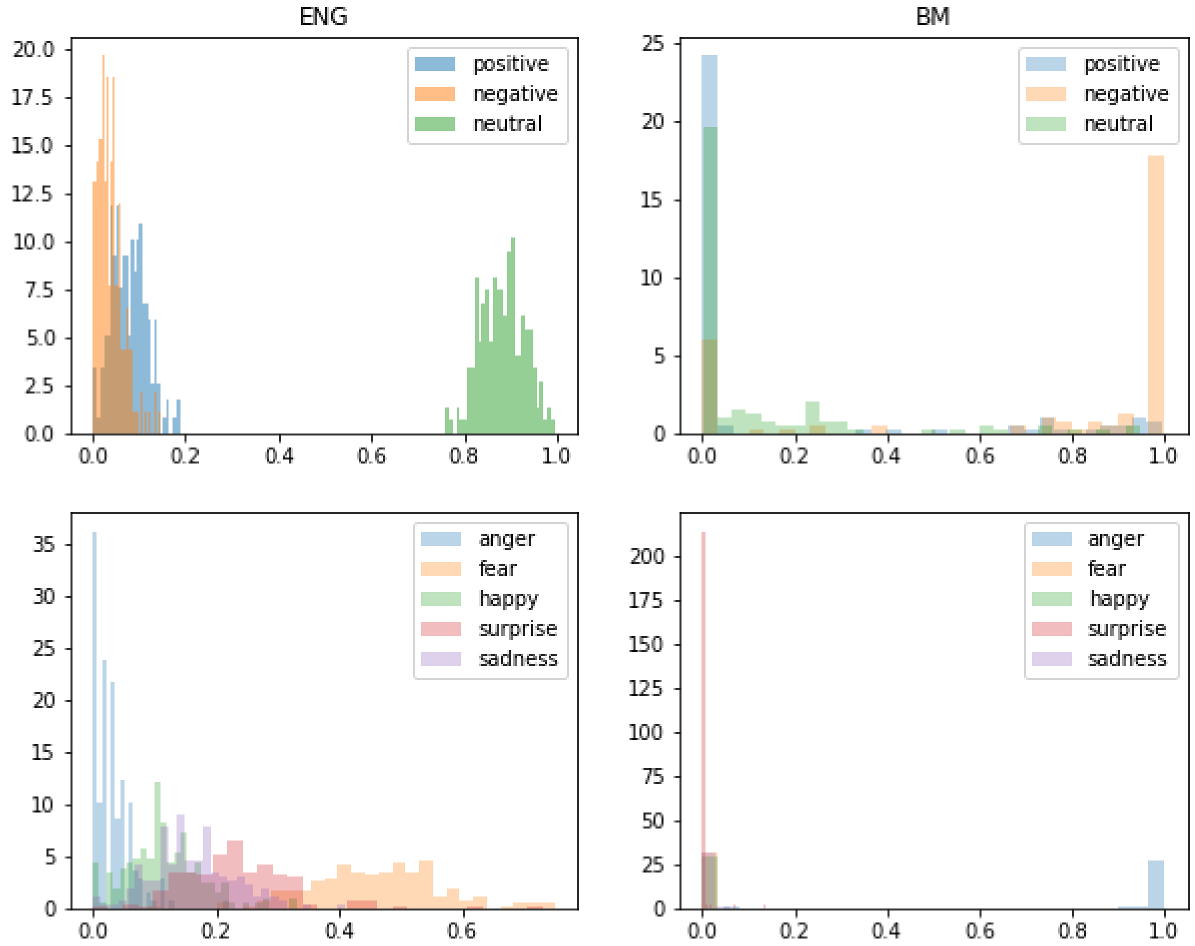


Figure 10 Overall sentiment and emotions of English and Malay articles.

Intuitively, I would expect the media articles' overall sentiment to be skewed primarily to neutral, given that most articles tend to be written objectively. The English articles follow this intuition, with most articles exhibiting neutral sentiments and slight skew towards positive sentiments, whereas the Malay articles exhibit extreme negative sentiments and emotions.

Let's compare a randomly selected English and Malay article to illustrate the flaws I discovered within the model (Table 4). Reading the sampled English article from *The Star* critiquing the education system in Malaysia, the naïve scoring method of the English model (calculating the proportion of words representing each sentiment or emotion) is congruent to how I might have scored the article independently.[48] Even with the advertisements dominating the first portion and tail end of the article, the overall sentiment and emotion summarised by the final score remain unaffected. However, the sampled Malay article from *Sinar Harian* reporting on when the High Court will rule on a lawyer's lawsuit against vernacular schools'

[48] Thiam Hock Tan, 'The Education Dilemma', The Star, 11 January 2020, https://www.thestar.com.my/business/business-news/2020/01/11/the-education-dilemma.

constitutionality is suspect.[49] I would have given a significantly higher neutral score and a more distributed score for emotions since I would not attribute a particular emotion to the article. The Malay model is more complex, so it is not apparent why the model scored the article as such. Still, it is most likely a deficiency within the training data that skewed the sentiments and emotions attached to specific words. The limitation of a transformer model is that it is difficult to peer into each layer or an individual parameter to determine where exactly the model went wrong. However, these results give us an indication of how the model might be improved. In this case, I can investigate articles in the training set with high anger and negative scores to look for sample imbalances or glaring mistakes in tagging.

Category		**English Score (%)**	**Malay Score (%)**
Sentiment	Positive	8.6	0.3
	Negative	3.8	69.7
	Neutral	87.7	30
Emotion	Anger	3	99.98
	Fear	53	0
	Happy	10	0
	Surprise	18	0
	Sadness	16	0.02

*Table 4 Sentiment and emotion scores for one English (**Error! Reference source not found.**) and Malay (**Error! Reference source not found.**) article.*

When looking at the proportion of emotion over time, it is stable for Malay articles and more volatile for English articles. Even though I filtered articles between 2015 and 2020 for English articles, some earlier articles were still returned by Google. The high volatility in emotions for English articles can be attributed to the small sample size of articles before 2015 (Figure 12).

[49] BERNAMA, 'Keputusan cabar tubuh sekolah vernakular diketahui', Sinar Harian, 4 November 2019, https://www.sinarharian.com.my/article/55626/BERITA/Mahkamah/Keputusan-cabar-tubuh-sekolah-vernakular-diketahui-11-November.

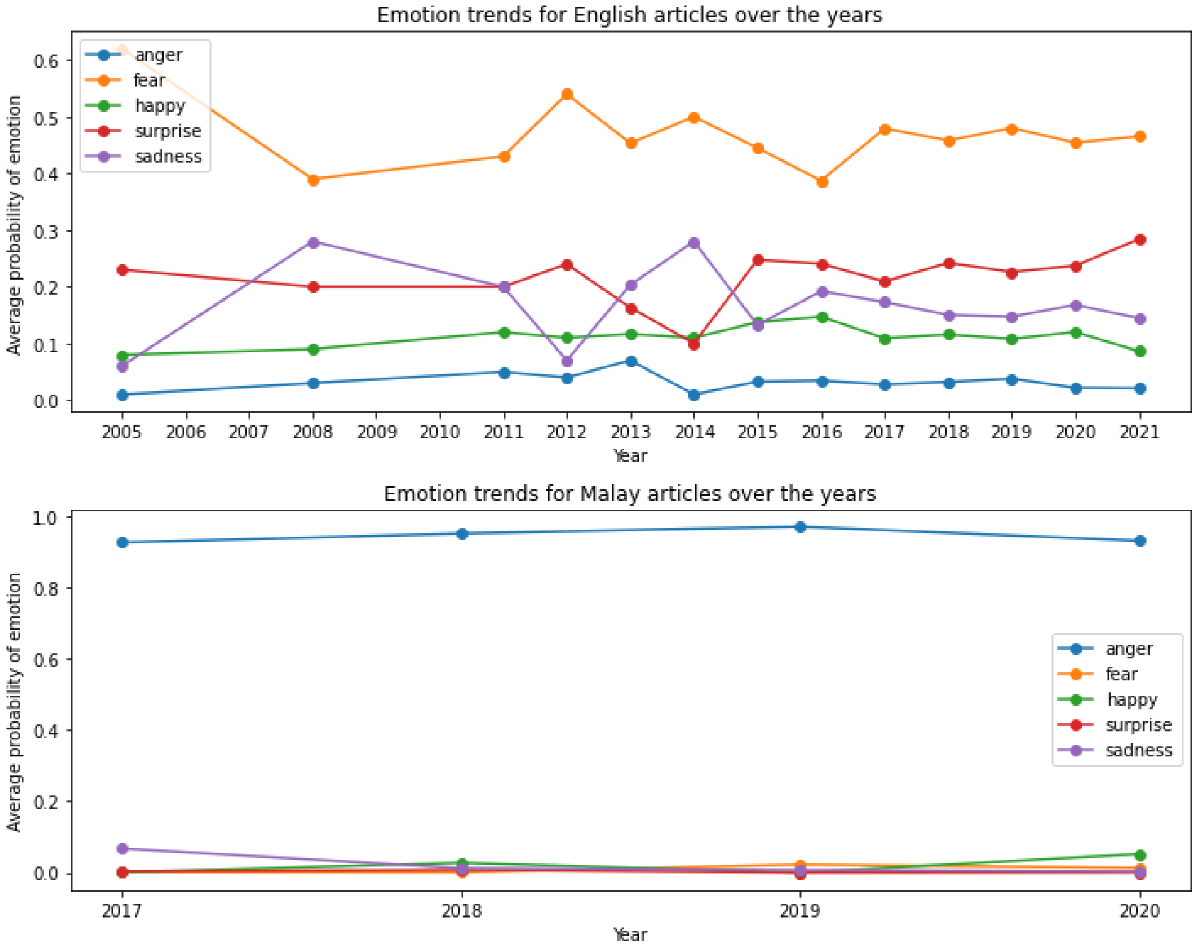


Figure 11 Emotion trends for English (top) and Malay (bottom) articles over time.

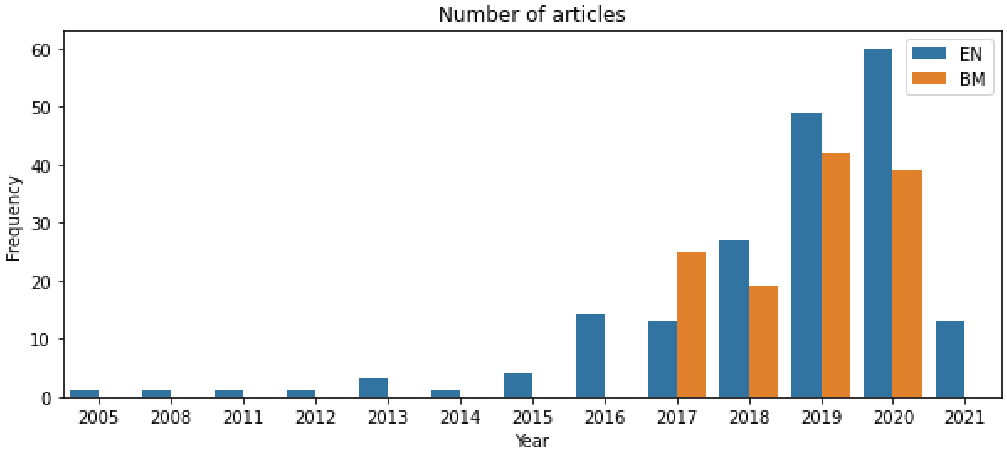


Figure 12 Number of English and Malay articles each year.

The interesting feature to note from Figure 11 is the flip in the most represented emotion. English articles are consistently most characterised by fear and least by anger and vice versa for Malay articles. Both emotions indicate dissatisfaction with vernacular education but from different positions of power. Fear implies a vulnerable and defensive position, whereas anger is offensive. The emotional power imbalance is congruent with Malaysian social dynamics, given that the English articles are directed toward the Chinese and Indian communities who constitute the minority.

With sentiments, Malay articles grow in negativity, jumping up by 20% within four years, while English articles remain relatively neutral over the fifteen-year span captured (Figure 13).

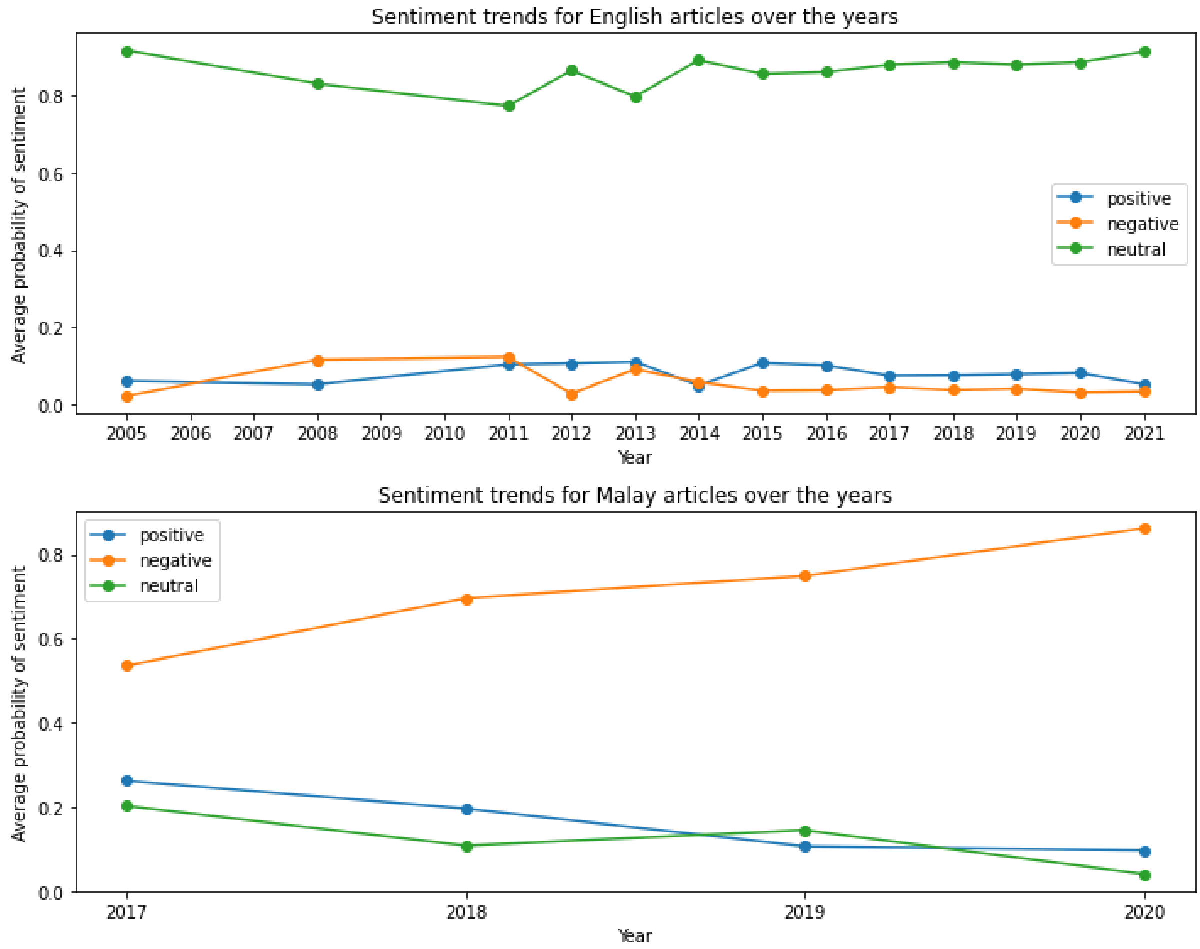


Figure 13 Sentiment trend of English (top) and Malay (bottom) articles over time.

The jump in negative sentiment for Malay articles from 2017 to 2018 coincides with Pakatan Harapan's (PH) victory in the 2018 General Elections, the first change in the ruling party in Malaysian history.[50] The former ruling party, Barisan Nasional (BN) (originally known as the Alliance Party), was a coalition predominantly controlled by the United Malays National Organisation (UMNO) who advocate for the maintenance of special privileges for Malays, including maintaining quotas for Malays in higher education institutions and upkeeping Malay as the only official and national language in Malaysia. The change in leadership may have sparked negative emotions among the Malays (not limited to anger) in the uncertainty of what might happen to their special privileges. These emotions continued to escalate in 2019 with announcements by the Ministry of Education in August to introduce Islamic calligraphy as a compulsory subject in vernacular schools, sparking debates between Islamic societies and

[50] Ding, 'In Malaysia, Young People Find Their Voice amid a Pandemic'.

Dong Jiao Zong (two Chinese educationist groups).[51] Two months later, in October, a Malay lawyer and member of the Malay nationalist party, Parti Bumiputera Perkasa Malaysia (PUTRA), filed a lawsuit against vernacular schools as unconstitutional, which escalated to the Federal court in 2021 and, at the time of writing, is still ongoing.[52]

These events only appear to impact the sentiment and emotions in Malay language articles with no jump or apparent trend with English articles.

Topic Modelling

Figure 14 and Figure 15 show the topics outputted by BERTopic for English and Malay articles, respectively. I plotted the top ten words associated with each topic against their percentage representation of the topic. Topic -1 contains all the miscellaneous words which do not fit into a topic.[53]

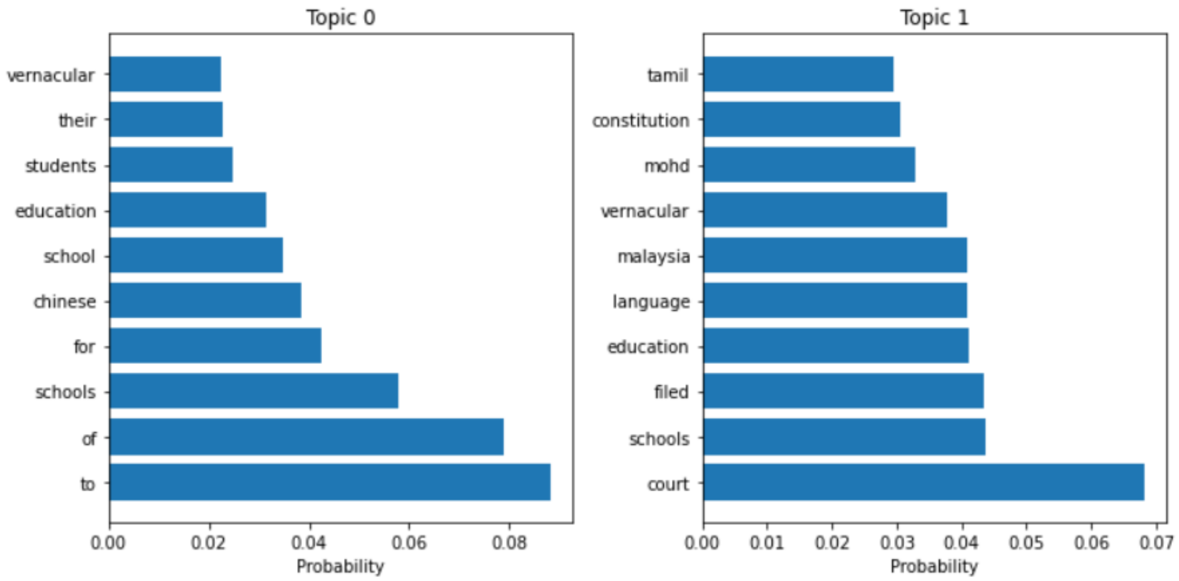


Figure 14 BERTopic English topics.

[51] Radzi Razak, 'May 13 May Recur as Long as Dong Zong Still around, Muslim Students' Group Warns', Malay Mail, 24 December 2019, https://www.malaymail.com/news/malaysia/2019/12/24/may-13-may-recur-as-long-as-dong-zong-still-around-muslim-students-group-wa/1822037.

[52] 'In Court Papers, Govt Says Vernacular Schools Part of Education System', Free Malaysia Today (FMT), 10 May 2021, https://www.freemalaysiatoday.com/category/nation/2021/05/10/in-court-papers-govt-says-vernacular-schools-part-of-education-system/; Kenneth Tee, 'Lawyer Makes Second Legal Jab to Declare Vernacular Schools Unconstitutional', Malay Mail, 17 December 2019, https://www.malaymail.com/news/malaysia/2019/12/17/lawyer-makes-second-legal-jab-to-declare-vernacular-schools-unconstitutiona/1819959.

[53] 'Getting Started - BERTopic', accessed 4 December 2021, https://maartengr.github.io/BERTopic/tutorial/quickstart/quickstart.html.

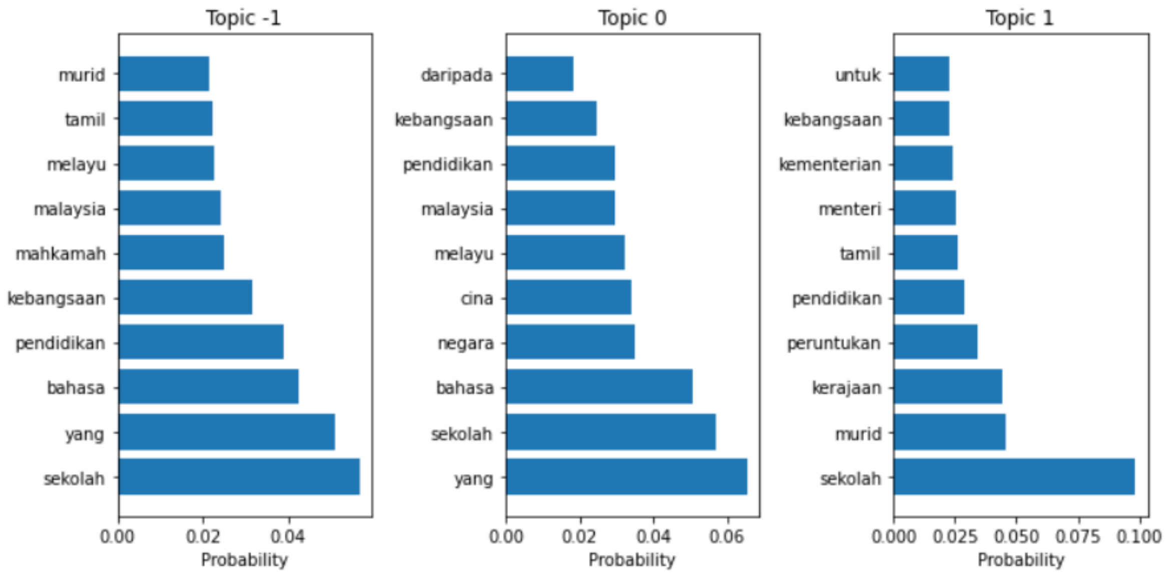


Figure 15 BERTopic Malay topics.

The topics lose some of their informative powers because stop words and word lemmatisations (different forms of a word with the same meaning) were not removed in the post-processing of the model.[54] Topic 0 for the English articles is mainly represented by stop words like ‘of’ and ‘to’. Similarly, for the Malay articles, Malay stop words like ‘daripada’ and ‘yang’ remained. Additionally, the model failed to omit ‘mohd’, a common part of Malay names, showing a limitation in the model’s training set, which most likely lacks representation of Muslim names. An example of words that should have been lemmatised is ‘school’ and ‘schools’ in Topic 0 for English. Both words give the same meaning, so having two versions of the word to represent a topic is redundant and splits the percentage representation of the topic between both words.

With sufficient knowledge of Python, one could use the Natural Language Toolkit (NLTK) library, which is equipped to remove stop words and lemmatise. It would require getting a list of all words for each topic and their associated probabilities, combining lemmatised words, removing the stop words, and recalculating the probabilities, so the sum adds up to one. Although canned models like BERTopic are easy to use, they are also limiting in their inflexibility, especially without good programming knowledge.

Based on prior understanding of significant events related to vernacular schools between 2015 to 2020, we can infer that Topic 1 of the English articles refers to the lawsuit filed by a lawyer to rule vernacular schools unconstitutional based on the words ‘court’, ‘vernacular’, ‘constitution’ and ‘filed’ (Figure 14).

By contrast, it is difficult to find meaning in the topics given from the Malay articles, most likely because BERTopic was not fine-tuned for the Malay language. Based on the words given, we can infer that Topic 0 and 1 are related to vernacular education based on the words ‘kebangsaan’ (national), ‘sekolah’ (school), ‘bahasa’ (language), ‘pendidikan’ (education) and

[54] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, ‘Stemming and Lemmatization’, in *Introduction to Information Retrieval* (Cambridge University Press, 2008), https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html.

‘melayu’, ‘cina’ and ‘tamil’ which correspond to the three languages in question. The topic model does not tell us more about the articles than our keyword inputs.

topic 0	topic 1	topic 2	topic 3	topic 4
Sekolah	negara	sekolah	masjid	sekolah
penuh	pendidikan	mahkamah	masjid surau	telus
media	bersatu	malaysia	bersebelahan masjid	baharu
penyelenggaraan	memikirkan	pendidikan	masjid surau hubungan	persekutuan
sk	sekolah	sekolah vernakular	bersebelahan masjid surau	pakatan
hadir	ibunda	perlembagaan	terletak bersebelahan masjid	melayu
membabitkan	negara melaksanakannya	parlimen	sekolah	perlembagaan
sekolah jenis	kemajuan negara	peguam	malaysia	aidilfitri
penuh komited	bersatu padu	sekolah sekolah	melayu	sekolah vernakular
membimbing penuh	peranan pendidikan	persekutuan	bahasa	perlembagaan persekutuan

Table 5 Malaya’s output for the top 5 Malay topics and the top 10 words for each topic.

Using a model built for the Malay language, we get significantly different results that are more informative than the topics given by BERTopic because it understands the structure of the Malay language from its training.[55] Unlike BERTopic, we cannot access each word’s percentage representation, but the words are arranged from top to bottom by largest representation. Topic 0 is vaguely related to the commitment to education, inferred from ‘penyelenggaraan’ (maintenance) and ‘membimbing penuh’ (fully take care of), but the subject is not apparent. Topic 1 is related to building national unity through education, inferred from the words ‘peranan pendidikan’ (education plan), ‘bersatu padu’ (united) and ‘kemajuan negara’ (national progress). Topic 2 is related to a lawsuit against vernacular education from Topic 1 from the English BERTopic model, inferred from the words ‘mahkamah’ (court), ‘peguam’ (lawyer) and ‘parliamen’ (parliament) and ‘perlembagaan’ and ‘persekutuan’ (together meaning Federal Constitution). Topic 3 references the placement of a mosque or *surau* beside a school. Topic 4 does not appear to form a coherent topic.

Malaya does a better job at removing stop words than BERTopic but could improve its tokenisation (splitting a sequence of characters into meaningful words or terms).[56] For example, ‘sekolah vernakular’ (vernacular school) and ‘bersatu padu’ (unity) are valid tokens, whereas ‘negara melaksanakannya’ (the country carries it out) and ‘penuh komited’ (full of commitment) should have been split into individual words. It also inconsistently identifies ‘perlembagaan persekutuan’ as a term or two separate words across the topics.

[55] I also ran the model on the English articles for completeness, but the topics were almost entirely comprised of stop words and generic words related to vernacular education in Malaysia. These results are to be expected given that Malaya’s post-processing would not account for English stop words.

[56] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, ‘Tokenization’, in *Introduction to Information Retrieval* (Cambridge University Press, 2008), https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html.

The model could have failed because of the advertisements captured within the article text, which may have created erroneous associations between tokens. It was difficult to remove these advertisements because there were no clear markers I could code to identify where they start and end. These advertisements usually prompt readers to read another article from the media source, so they include the title and snippets of the advertised article. Alternatively, the model might not have been well trained, given that it is not entirely clear how exactly Malaya was fine-tuned for topic modelling

What can we say about public opinion?

Our examination of public opinion on vernacular education through media articles has clarified the benefits and limitations of distant readings. It does not remove the question of accessibility because applying these methods properly requires familiarity with the methodology and, for greater flexibility, coding experience. Canned models are convenient but severely limited for specific use cases and dangerous when applied in ignorance.

The conclusions we have drawn on public opinion are cautious at best, with wide confidence intervals, showing the possibility for a considerable variation of actual public opinion. Some might doubt the usefulness of the methodology in comparison to regular polls, but neither methodology should be discounted. Instead, these methods should be combined with a close reading of the articles to bring out nuances that NLP methods cannot yet capture.

Our exploration points us in several new directions towards understanding race relations in Malaysia, which we would not have conceived of otherwise. Firstly, NLP methods for the main languages of Malaysia are underdeveloped. We have shown the spaces for improvement with Malay (and even English) articles without even considering a lesser-used language like Tamil and a more syntactically complicated language like Mandarin Chinese. However, unlike Tamil and Malay, Mandarin Chinese has a large and strong community of researchers developing NLP methods for the language.

Secondly, our analysis shows growing anger within the Malay community concerning vernacular education. Are these emotions fairly represented? And if so, are they being fuelled or characterised by the media? Who are the leading voices in these conversations, and what are their aims? We might extend the analysis to map out the actors involved for a topography of power in vernacular education discussions.