Development of Multilingual Social Media Data Corpus: Development and Evaluation

*Fitrah Rumaisa[a], Halizah Basiron[b], Zurina Saaya[c], Noorli Khamis[d],
[a]Information Technology Department, Universitas Widyatama, Bandung, Indonesia, [b,c,d]Faculty of Information and Communications Technology, Universiti Teknikal Malaysia Melaka (UTeM), Melaka, Malaysia,
 *Corresponding Author E-mail: [a]fitrah.rumaisa@widyatama.ac.id, [b]halizah@utem.edu.my, [c]zurina@utem.edu.my, [d]noorli@utem.edu.my

The purpose of this study is manual annotating, a corpus for Bahasa Indonesia and Bahasa Melayu. Corpus for both languages has been made by many researchers before, but the focus of this research is only on words with the same vocabulary but which have very different meanings. The data were obtained from social media, so informal words were found. As many as 2100 words for each language were identified which were then randomly selected so that 300 words with the same vocabulary but with different meanings were used. The objective of this study was to confirm that this condition can influence the results of polarity sentiment. At the end of this paper, we will show the results of the influence of the conditions of the two languages on the polarity of sentiments. From the manual annotation, an annotation agreement test was made by three Bahasa Indonesia annotators and three Bahasa Melayu annotators. The results of the annotation found that there were 63 out of 300 words that experience different polarity. Results of score agreement among annotations for each language show that there is good agreement among the annotators during annotation process.

Key words: *Corpus; Bahasa Indonesia; Bahasa Melayu; Annotation; Social Media.*

Introduction

A corpus is now understood as a collection of texts that are presented electronically, which are able to be analyzed automatically or semi-automatically rather than manually, including written or oral artefacts from various sources on different topics and by different authors/

speakers. According to Kruger, there are two types of the corpus that can be used as learning materials instead of a dictionary. The first is the parallel corpus which compares the original text with the translation. The second is a bilingual comparable corpus which compares the different text languages within the same topic (Kruger, 2004).

In general, textual information consists of two categories: facts and sentiments. Facts are an objective expression of people, actions, and properties. For example, "I come from Bandung". The sentence is a fact, which states that the person is from the city of Bandung. Sentimental text shows more subjective expressions that indicate feelings, opinions, or judgments about a person, event, organization or its properties (Liu, 2010), as in the phrase "Bandung is very beautiful". The meaning of the phrase implies a subjective sentiment, it could be positive, negative or neutral. The general feeling of documents, sentences or words are called polarity.

Research in the area of sentiment analysis is, however, diverse. Researchers aim to explore various models, datasets, languages, and methods. Popular datasets used in this area of research are often taken from social media. One of the most popular applications of sentiment analysis is Natural Language Processing (NLP). NLP can be integrated in Machine Learning (ML) approaches that enable a computer to understand, analyze, manipulate, and potentially generate human language.

NLP and ML are derivatives of Artificial Intelligence (AI) which work together to solve many data problems. NLP concentrates on computer interactions with human language. ML allows computers to learn and develop themselves without needing to be programmed again when meeting new data. In ML processing, a dataset that is usually called a corpus, is needed (Dorothy and Rajini, 2016).

Annotation activities for the formation of corpus are urgently needed. This is because corpus are formed not only in English. In the current era of microblogging, people can express their opinions using a variety of languages and informal languages. For example, opinions that are posted on Twitter; all people from different countries can post their opinions on Twitter using their own language.

In this study, an annotation manual was constructed using 6 annotators on a corpus that had been developed in previous studies. There are 300 words with the same vocabulary but have different meanings from Bahasa Indonesia and Bahasa Melayu (Rumaisa et al., 2019).

Related Work

This section will discuss the background of the basic concepts, datasets, and methodologies

that have been used by previous researchers.

A. Previous Manual Annotation Research

Annotated data are very important for researchers in the field of linguistics. The data must be reliable. Therefore, the annotation is not enough if it is not accompanied by reliability tests of the annotation data. Reliability is the extent to which various methods, results of research, or groups of people, arrive at the same conclusions or facts (Krippendorff, 2011).

In this section, the methodologies that have been used by previous researchers will be explained. The dataset highlighted in this session are Bahasa Indonesia and Bahasa Melayu in the construction of the corpus. Furthermore, a number of corpus studies will be described using manual annotation techniques. The next paragraph will discuss Bahasa Indonesia and Bahasa Melayu corpus development techniques that have been used in studies previously.

In addition to this, research also uses a lexical-based for corpus formation in Bahasa Melayu. Darwich et.al used human-coded corpus annotation methods for sentiment analysis. The researcher used the WordNet Language and English WordNet. The steps to this process are to map WordNet Language to English WordNet, set the word "baik" instead of "good" and "buruk" instead of "bad", then the set of words is used to detect synonyms and antonyms in WordNet. Words that have been labelled by the propagation algorithm serve as a basis for training the classification of unseen words not seen in polarity (Darwich and Noah, 2016).

While Noah et.al uses the semantic pre-process similarity method for machine-readable dictionaries in Malay to calculate semantic words using two possible and normalized intersection methods. Then they use the equivalence of semantic words per word to identify the similarity of semantic sentences, finally evaluating the effects of lexical components. These results are then annotated manually using human-rating. Several groups of sentences are compared to find similarities in meaning. The annotators will rank the sentences against the target sentence (Noah et al., 2015). In contrast to the two researchers discussed earlier, the next study evaluated the accuracy of the Part of Speech (POS) markers in Corpus Malay found from police reports. The methodology used is only two stages of tokenization and POS tagging. This POS tagging process is done manually through annotation using human tagging, and automatically through annotation using computer tagging. The value obtained from the two annotation processes determines the accuracy of POS tagging (Hamzah and Na'imah, 2014).

The next language that has also been studied is the Bahasa Indonesia. So far there are two researchers who discussed the Bahasa Indonesia corpus of Aliandu, 2013, and Franky, 2015.

The first Bahasa Indonesia corpus study discusses Naive Bayes as a method used in manually annotating the results of corpus formation that can determine the classification of sentences in tweets in Indonesian, with Support Virtual Machines used as a comparison method in the validation phase. Steps taken include collecting corpus using the Twitter API, Feature Extraction, Learning-Naive Machine Method, Accuracy Measurement, and Testing Data (Aliandu, 2019).

In the case of the second Bahasa Indonesia corpus, research discusses the analysis of lexicon subjectivity from positive and negative revelation in the Bahasa Indonesia by automatically translating lexical English. Resources used consist of two types, first, a small collection of user opinions across multiple domains; second, a collection of sentiments created by translating existing lexicon using several translation methods. The annotation process is done by two native speakers of Bahasa Indonesia, and then an agreement using the Kappa statistic method is done (Franky et al., 2015).

Based on several studies using Bahasa Indonesia and Bahasa Melayu as the above dataset, there are several research gaps:
- There is lack of researchers having discussed Bahasa Melayu and Bahasa Indonesia at the same time, either the similarities or the differences between the two languages.
- There is lack of researchers having discussed the same vocabulary, yet a different meaning, between the two languages.
- Manual annotation methodologies have not been widely used for the construction of both language corpus.

The gaps explained above becomes the basis of the research conducted by the author and becomes a significant contribution to the formation of the corpus.
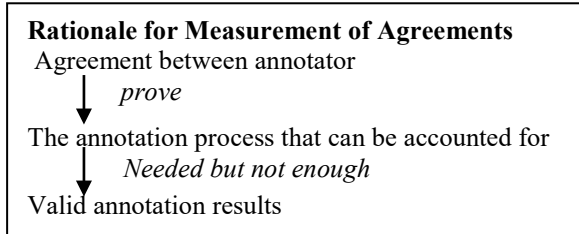
B. Inter-Annotator Agreement (IAA)

To do manual annotation, human annotators are needed. If more than one human annotator is doing the manual annotations, a measurement to find a good agreement on annotations among the annotators is required. The reason for this, is to get a good measure of validity among the annotations. The results of reliable annotation values will be able to provide consistent and reproducible annotation data in the future. Inter-annotator agreement has become the standard for testing the accuracy of manual annotations. The more annotators that are involved and agree on the same material, the more the annotation results are free from deviant variations (Artstein and Poesio, 2008).

To check for consistency, it is necessary to apply multiple annotation processes to the same source, and also necessary to use different annotators or more than one person

(Artstein, 2009). These steps are shown in Figure 1.

Fig. 1. Rationale for Measurement of Agreements (Artstein, 2009)

Rationale for Measurement of Agreements
Agreement between annotator
    ↓   *prove*
The annotation process that can be accounted for
    ↓   *Needed but not enough*
Valid annotation results

To test the reliability value of the annotation results, Cohen Kappa (McHugh, 2012) determines the range of values that signify the quality of the annotation results between annotators. This value can be seen in Table 1. The Kappa values will be used when conducting an inter-annotator agreement process in the last section of this study.
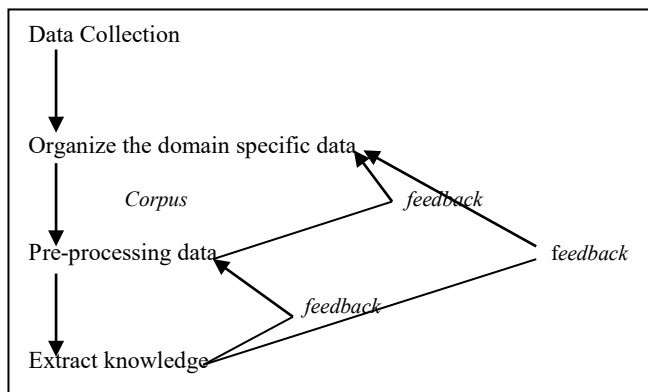
TABLE 1: Level of Kappa Agreement (McHugh, 2012)

Value of Kappa	Agreement Level
Above 0.90	Almost Perfect agreement
0.80 – 0.90	Strong agreement
0.60 – 0.79	Moderate agreement
0.40 – 0.59	Weak agreement
0.21 – 0.39	Minimal agreement
0 – 0.20	None agreement

Design and Development

This section will explain in detail the steps that are followed in the formation of the corpus, starting from data collection, preprocessing (stopwords, tokens, stemming), and knowledge extraction. Each stage of this process will be evaluated, so as to get results accordingly. These steps are shown in Figure 2. Explanation of Figure 2 will be detailed in the next subsection.

Fig. 2. Preparation steps of corpus formation

Data Collection
↓
Organize the domain specific data
  Corpus   *feedback*
↓
Pre-processing data   *feedback*
    feedback
↓
Extract knowledge

A. Data Collection

The data used was collected from social media, namely Twitter and Facebook. Data from Twitter was taken by using specified language and domain filters. In accordance with its purpose, the language used is Bahasa Melayu (code = msa) and Bahasa Indonesia (code = id). While the selected domains are political, Samsung and I-Phone. As many as 1000 tweets were taken from each language. Also, using raw data in the form of tweets taken from Malay Chat-Style-Text Corpus (MCC) (Saloot et al., 2016).

B. Data Preprocessing

The collected data were preprocessed which included tasks such as data cleaning, tokenizing, stemming, and stopwords.

a. Cleaning

This stage cleans up punctuation and unnecessary symbols. Figure 3 and Figure 4 respectively show tweets before the cleaning process and after the cleaning process.

Fig. 3. Sample tweets

```
Saya baru create apple id baru. Dan baru download iTunes. Kenapa nak log in iTunes
tak boleh? Saya 5S user
salam..tumpang tnya klau nak buat sebelah line celcom tu mcm mana ea ? tolong ajar
kan
```

Fig. 4. Cleaning Result

```
saya baru create apple id baru dan baru download itunes kenapa nak log in itunes tak
boleh saya 5s user
salam tumpang tnya klau nak buat sebelah line celcom tu mcm mana ea tolong ajar kan
```

As seen in Figure 3, there are uppercase letters, symbols and punctuation marks. After the cleaning process, Figure 4 shows all tweets have changed to all lowercase letters, while the symbols and punctuation have been deleted.

b. Stop Words

Stop words are words that are ignored and will not be taken into account in the corpus. The stopwords of Bahasa Melayu are obtained from Saloot et al., but some informal languages are not included in MCC, such as "x" which stands for "tak" or "tidak". Bahasa Indonesian stopwords are obtained from website pages   https://github.com/masdevid/ID-Stopwords. Some examples of stopwords can be seen in Table 2. Table 2 consists of a combination of Bahasa Indonesia and Bahasa Melayu stopwords.

TABLE 2: Sample of stopwords list

Stopwords list		
jangan	tidak	sampai
kamu	aku	i
tak	x	apa
asal	jadi	amat
agar	akan	sangat
jangan	sebab	jika
…	…	…

c. Tokenizing

This process breaks the sentence into a word, so that the collected words can be seen. Table 3 shows some samples of words that have been produced.

TABLE 3: Sample of tokenizing result

Tokenizing result		
modem	pelik	kena
wifi	salam	salam
elok	samsung	beli
lampu	tukar	duit
jokowi	prabowo	seronok
tolak	tolak	butuh
…	…	…

After going through preprocessing, 7000 words for each language were found. However, only 30% taken for inclusion in the corpus (2100 words), in accordance with the results of Asmah's research which indicated that 30% of Bahasa Indonesia and Bahasa Melayu are considered inexplicable, incomprehensible and unusual (Asmah, 2001). Of these 2100 words, as many as 300 words were randomly selected as words with the same vocabulary but a different meaning. Examples of words obtained as in the Table 4.

TABLE 4: Sample of words that have the same vocabulary but have a different meaning

Sample words	
bandar	bual
setor	wayang
dewan	patung
kekal	seronok
acar	baja
saman	ceroboh

kacak	kacak
perasan	bekas
mesra	Jemput
…	…

C. *Manual Annotation*

About six (6) annotators worked on this corpus for 3 weeks; three (3) of them worked on the Bahasa Indonesia corpus, and three (3) worked on the Bahasa Melayu corpus. All annotators were native-speakers of each language, namely 3 native-speakers Bahasa Indonesia and 3 native-speakers of Bahasa Melayu. The annotators were required to have a minimum educational background of a Master in the field of IT or Linguistic. The 300 words that are the focus of this corpus, and the 618 sentences formed, were reviewed by each annotator. Each sentence has 6 components that must be reviewed. The draft annotation form given for each annotation is shown in Table 5.

TABLE 5: The draft of annotation form

Words	Sentence	Lang	Tagset	Meaning	Polarity/words	Polarity/sentence	Spelling
Bandar	Saya duduk bandar. Boleh tak nak request						
	Gempa bumi yang akan melanda bandar aceh sebentar lagi						

The annotators were given annotation guidelines that had been prepared previously. More detailed information about each category in the table above will be discussed in the next sub section.

D. *Annotation Scheme*

To do manual annotation, an annotation scheme is needed that can help to get the desired results. In this study the same scheme was used for manual annotations. There are five categories that will be determined by the annotator:

a. Language classification
b. Tag set corrections
c. Meaning corrections
d. Polarity
e. Spelling Corrections

The sentence given for one kind of vocabulary may exceed one sentence, since it may have another meaning. In addition, the sentence is used based on concordance formed from the corpus that had been compiled. Focus tagging on the word being bolded and its use in a sentence. The description of each category will be explained in the next section.

a. Language Classification

The category BM will be inputted for Bahasa Melayu and BI for Bahasa Indonesia. Annotators will determine if the available sentence is a Bahasa Melayu or Bahasa Indonesia sentence. If the category has no argument, annotators just write the code BM or BI as exampled in Table 6.

TABLE 6: Language Classification

Sentence	Lang
Pimpin tangan dia ke syurga. Bukan **tolak**, tendang, terajang, sepak dia	BM
Mahasiswa mulai angkat isu **tolak** Pemilu Mahasiswa di Yogyakarta	BI

b. Tag set Corrections

Bahasa Indonesia and Bahasa Melayu have similarities in the form of Tag set. Bahasa Indonesia has five (5) main parts of speech that include verb, adjective, adverb, noun, and function words (Pisceldo et al., 2009). While Bahasa Melayu has four (4) main parts of speech, that include noun, verb, adjective and adverb (Hamzah and Na'imah, 2014). This makes each language unique, but both have the same post tagger form as shown in Table 7.

TABLE 7: Tag set Corrections

SYM Symbols	**VBT** Transitive Verb	**IN** Prepositions
NNC Countable common nouns	**VBI** Intransitive Verbs	**CC** Coordinate conjunction
NNU Uncountable common nouns	**MD** Modal or auxiliaries verbs	**SC** Subordinate conjunction
NNG Genetive common nouns	**ADJ** Adjectives	**RB** Adverbs
NNP Proper nouns	**CDP** Primary cardinal numerals	**UH** Interjections
PRP Personal pronouns	**CDO** Ordinal cardinal numerals	**DT** Determinations
PRN Number pronouns	**CDI** Irregular cardinal numerals	**WDT** WH-determinations
PRL Locative pronouns	**CDC** Collective cardinal numerals	**RP** Particles
WP WH-pronouns	**NEG** Negations	**FW** Foreign Word

Tag set formed from each language are as shown in Table 8.

TABLE 8: Tag set Corrections

Words	Sentence	Lang	Tagset
Acar	Awak bawa acar lemon?	BM	NNU
	Acar timun atau mentimun merupakan salah satu jenis makanan yang sering kita jumpai pada masakan soto	BI	NN
Saman	Mahkamah tolak saman Eskay RM20 juta terhadap pemaju Jambatan bengkok.	BM	IN
	Tari Saman adalah sebuah tarian Suku Gayo yang biasa ditampilkan untuk merayakan peristiwa-peristiwa penting dalam adat	BI	NN

As shown in Table 8, there is the same tag set and there is also a different one for the same word.

c. Meaning Corrections

This category contains the actual meaning for the word that is bolded. Before the Meaning categories column, there is a column of meaning derived from a machine translator such as Google translate between Bahasa Indonesia → Bahasa Melayu or vice versa. Annotators tag the true meaning according to the language they know; therefore the annotator must be the

person who understands the language. One word can have more than one meaning. This process will be explained in Table 9.

TABLE 9: Meaning Corrections

Words	Sentence	Lang	Tagset	Meaning
Acar	Awak bawa acar lemon?	BM	NNU	Jeruk
	Acar timun atau mentimun merupakan salah satu jenis makanan yang sering kita jumpai pada masakan soto	BI	NN	Pickle
Saman	Mahkamah tolak saman Eskay RM20 juta terhadap pemaju Jambatan bengkok.	BM	IN	Sesuai
	Tari Saman adalah sebuah tarian Suku Gayo yang biasa ditampilkan untuk merayakan peristiwa-peristiwa penting dalam adat	BI	NN	Saman

d. Sentiment Polarity

In this category, the annotator will determine the polarity of the formatted bold word. Either negative, positive or neutral. This process is shown in Table 10.

TABLE 10: Sentiment Polarity

Words	Sentence	Lang	Tagset	Meaning	Polarity
Acar	Awak bawa acar lemon?	BM	NNU	Jeruk	Neutral
	Acar timun atau mentimun merupakan salah satu jenis makanan yang sering kita jumpai pada masakan soto	BI	NN	Pickle	Neutral
Saman	Mahkamah tolak saman Eskay RM20 juta terhadap pemaju Jambatan bengkok.	BM	NNC	Sesuai	Negative
	Tari Saman adalah sebuah tarian Suku Gayo yang biasa ditampilkan untuk merayakan peristiwa-peristiwa penting dalam adat	BI	NN	Saman	Neutral

e. Spelling Corrections

This category improves the spelling of the bold words. This category aims to improve the word spelling that can result in different meanings of the word. This is because the sentences used are taken from social media so that sometimes informal words are used (Sinaga et al., 2019; Saudi et al., 2019). This process is shown in Table 11.

TABLE 11: Spelling Corrections

Sentence	Spelling Correction
*Sila bagi mahal sikit, kain sutera atau **kapan***	Kafan

E. Inter-Annotator Agreement Result

For the purposes of this paper, our focus is only on sentiment analysis. The statistics of the annotation results are calculated as two categories, namely Polarity per words and Polarity per sentence.

Then the inter-annotator agreement was calculated for the two categories using Fleiss 'kappa. The scores are presented in the Table 12.

TABLE 12: Fleiss Kappa scores

Corpus	Fleiss Kappa (Polarity/words)	Fleiss Kappa (Polarity/ sentence)
Bahasa Indonesia	0.9345	0.8142
Bahasa Melayu	0.7251	0.6855

When referring to Table 1 regarding the level of Kappa agreement, it can be seen that results of the Fleiss Kappa score for Bahasa Indonesia show a very good result, "Almost Perfect" for polarity per words and "Strong" for polarity per sentence. On the other hand, Bahasa Melayu Fleiss Kappa score showed "Moderate" results for both types of polarity.

When viewed from these two languages, both experienced a decrease in Kappa values at Polarity per sentence. This is because each annotator has a different understanding of the sentence.

Comparison of the two results of the agreement between the two languages above shows unsatisfactory results in Bahasa Melayu. This might happen because Bahasa Melayu rarely has a more expressive word to show a word is negative, positive or neutral. For example, in the sentence " Ketika suaminya pulang, disambutlah mesra," the word that is addressed is "mesra". Two of the 3 annotators consider this to be neutral, while one considered the word "mesra" in the sentence to have positive sentiment. In language that is used daily, the word "mesra" is used for something that is friendly. For example, the name of a company that wants to show that they are a friendly company (Icha Mesra Tours and Travel), so the results of sentiment are considered neutral.

On the other hand, in the use of Bahasa Indonesia, the "mesra" word is used only for an

intimate relationship, like the relationship of husband and wife, male and female. "Mesra" is rarely used to show a close relationship between parents and their children, or to name a company. So the three respective annotators agreed that the above sentence had a positive sentiment. Although the results of Bahasa Melayu annotations do not show "Almost Perfect" results, they can still be accepted as a result of valid annotations. With these agreement results, the annotated corpus can be used for further research.

Conclusion

This study develops a corpus of Bahasa Indonesia and Bahasa Melayu where the vocabulary is the same but has different meanings. Of the 2100 words obtained from the preprocessing process for each language, there were 300 words that have the same vocabulary but different meanings. The words were then annotated manually by 6 annotators and the results of the agreement test were good for both languages. Given this, the results can be considered to have validity. In addition, from the results of the annotation it is known that there are 63 words (21%) that have different polarity for the same word. This will likely affect the validity of the next research sentiment analysis. Therefore, future research will aim to make an annotation scheme model for both languages that are similar; so as to minimize the invalid results of polarity in future studies.

Acknowledgement

REFERENCES

A. Kruger, "Corpus-based translation research: its development and implications for general, literary and Bible translation," Acta Theol., vol. 22, no. 1, pp. 70–106, 2004.

B. Liu, "Sentiment Analysis and Subjectivity," Handb. Nat. Lang. Process., no. 1, pp. 1–38, 2010.

M. Dorothy and S. Rajini, "The Various Approaches for Sentiment Analysis : A Survey," vol. 5, no. 1, pp. 2014–2016, 2016.

F. Rumaisa, H. Basiron, and Z. Saaya, "Development of Multilingual Social Media Data Corpus for Sentiment Classification," J. Adv. Res. Dyn. Control Syst., vol. 11, pp. 286–293, 2019.

K. Krippendorff, "Communication Methods and Measures Agreement and Information in the Reliability of Coding Agreement and Information in the Reliability of Coding," vol. 5, no. March 2013, pp. 37–41, 2011.

M. Darwich and S. A. Noah, "AUTOMATICALLY GENERATING A SENTIMENT," Asia-Pacific J. Inf. Technol. Multimed. J. Teknol. Mklm. dan Multimed. Asia-Pasifik, vol. Vol. 5 No., no. December, pp. 49–59, 2016.

S. A. Noah, N. Omar, and A. Y. Amruddin, "Evaluation of Lexical-Based Approaches to the Semantic Similarity of Malay Sentences," J. Quant. Linguist., vol. 22, no. 2, pp. 135–156, 2015.

M. P. Hamzah and S. F. Na'imah, "Part of Speech Tagger for Malay Language Based," vol. 2014, no. October, pp. 1499–1502, 2014.

P. Aliandu, "Sentiment Analysis on Indonesian Tweet," Proc. Int. Conf. Information, Commun. Technol. Syst., pp. 203–208, 2013.

Franky, O. Bojar, and K. Veselovská, "Resources for Indonesian Sentiment Analysis," Prague Bull. Math. Linguist., vol. 103, no. 1, pp. 21–41, 2015.

R. Artstein and M. Poesio, "Inter-Coder Agreement for Computational Linguistics," Comput. Linguist., vol. 34, no. 4, pp. 1–42, 2008.

R. Artstein, "Inter-Annotator Agreement," no. July, pp. 297–313, 2009.

M. L. McHugh, "Interrater reliability: the kappa statistic," Biochem. Medica, vol. 22, no. 3, pp. 276–282, 2012.

M. A. Saloot, N. Idris, A. T. Aw, and D. Thorleuchter, "Twitter corpus creation: The case of a Malay Chat-style-text corpus (MCC)," Digit. Scholarsh. Humanit., vol. 31, no. 2, pp. 227–243, 2016.

H. O. Asmah, "The Malay Language In Malaysia And Indonesia: From Lingua Franca To National Language," Asianists' ASIA, vol. 2, pp. 1–21, 2001.

F. Pisceldo, M. Adriani, and R. Manurung, "Probabilistic Part of Speech Tagging for Bahasa Indonesia," Proc. 3rd Int. MALINDO Work. Coloca. event ACL-IJCNLP, 2009.

Sinaga, O., Saudi, M. H. M., Roespinoedji, D., & Razimi, M. S. A. (2019). The Dynamic Relationship between Natural Gas and Economic Growth: Evidence from Indonesia. International Journal of Energy Economics and Policy, 9(3), 388-394.

Saudi, M. H. M., Sinaga, O., Roespinoedji, D., & Razimi, M. S. A. (2019). The role of renewable, non-renewable electricity consumption and carbon emission in development in Indonesia: Evidence from Distributed Lag Tests. International Journal of Energy Economics and Policy, 9(3), 46-52.