ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse179942020.pdf
https://doi.org/10.30534/ijatcse/2020/179942020

BMBI: A Development of a Special Corpus on Homonyms for Multi-Lingual Sentiment Analysis

Fitrah Rumaisa [1], Halizah Basiron[2], Zurina Saaya[3], and Yoki Muchsam[4]
[1]Department of Information Engineering, Widyatama University, Bandung, INDONESIA, fitrah.rumaisa@widyatama.ac.id
[2] Centre for Advanced Computing Technologies (C-ACT), Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka (UTeM), Hang Tuah Jaya, 76100, Durian Tunggal, Melaka, MALAYSIA, halizah@utem.edu.my
[3] Centre for Advanced Computing Technologies (C-ACT), Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka (UTeM), Hang Tuah Jaya, 76100, Durian Tunggal, Melaka, MALAYSIA, zurina@utem.edu.my
[4] Department of Medical Record, Akademi Perekam Medis dan Informatika Kesehatan (APIKES), Bandung, West Java, INDONESIA, yoki.muchsam@apikesbandung.ac.id

ABSTRACT

Research in the area of sentiment analysis is growing rapidly. Along with this the need for a corpus that can help in increasing the validity of the sentiment results is very much needed. But there are special cases against languages where data sources are very rare. One of them is a homonym word which means the word which has the same vocabulary but has a different meaning. In this study, an annotation scheme model named BMBI annotation scheme model will be designed to meet the needs of the corpus. This model has several elements namely <KATA>, <KALIMAT>, <HOLDER>, <APPRAISAL GROUP>, <TARGET>, and <MODIFIER>. The annotation process of the scheme model was done by six (6) annotators with certain criterion. An agreement evaluation of the annotation process was performed using the Fleiss. The calculation of the agreement among annotators is focused on 7 tasks namely language identifiers, BI polarity on the <KATA> element, BM polarity on the <KATA> element, BI polarity on the <KALIMAT> element, BM polarity on the <KALIMAT> element, BI tagset, and BM tagset; produce a Moderate value of the agreement which indicates that the agreement results are feasible to be used as a basis for further research.

Key words: BMBI, homonym, annotation, Fleiss Kappa, Inter-annotation, model

1. INTRODUCTION

Sentiment analysis, or commonly referred to opinion mining is one part of the text mining. Sentiment analysis is an intersection of information retrieval, natural language processing, and artificial intelligence [1]. This field of study is to discuss the people's opinion, sentiment, evaluation, behaviour, and emotions to an entity such as products, services, organizations, individuals, issues and topics, events, and attributes [2]. This research requires data sources from various sources that can reflect opinions on the entity.

The need for corpus to help sentiment analysis research is increasingly high, especially for Multilanguage corpus. But for special cases in some languages that lack a lot of data sources, it is very difficult to find a suitable corpus. One of them is for homonym languages which mean have the same vocabulary but different meanings. The meanings here focus on the polarity of sentiment. The purpose of this study is to establish a special corpus to explain the difference of a word that has two or more meanings of a multilingual. This corpus will be used for the classification of sentiment analysis on Social Media.

Bahasa Melayu (BM) that is used in Indonesia and Malaysia likewise varies in the comprehension and the general observation by people in general of the two nations. This distinction can trigger a false impression [3]. This is because the two languages have the same vocabulary but have different meanings, so that there are differences in Part of speech-tags and sentiment, thus allowing the results of sentiment analysis research to be contradict. Based on research from [4], explains that the existence of Bahasa Indonesia (BI) words and phrases known by people of Malaysia but has a different meaning. Also explained also by [5] that the same words in different languages can be interpreted with different or even contradictory semantics and this can lead to other contextual ambiguities

For example, refer to the following Bahasa Indonesia tweet:

"*Sidang gugatan Rp 13 milyar Bupati tolak hadiri sidang Serambi-Lhokseumawe*" (Rp13 billion suit The Regent refused to attend the Serambi-Lhokseumawe hearing)

Compare this tweet with the following Bahasa Melayu tweets:

"*Selina Kyle jadi Catwoman sebab Bruce Wayne tolak dia jatuh dari tempat tinggi lepas tu dia...*" (Kyle becomes

Catwoman because Bruce Wayne pushes her down from the heights, and then she...)

If noted there is a word "*tolak*" on both tweets above. In Bahasa Indonesia, the word is used, as well as in Bahasa Melayu. However, if considered based on the sentence, the word "*tolak*" in the Bahasa Indonesia means reject/refused while in the Bahasa Melayu, "*tolak*" sometimes means the push.

Another example in Bahasa Indonesia:

"*Kamu kenapa sih comel banget?*" (Why are you so nagging?)

Compare with the use of the same word (*comel*) in Bahasa Melayu:

"*Comel sangat budak kecil tu*" (That kid is so cute)

The use of the word "*comel*" in both sentences is different. In addition, it has different sentiments. In Bahasa Indonesia, "*comel*" has a negative sentiment, namely nagging. While in Bahasa Melayu, "*comel*" means cute which has a positive polarity sentiment.

In addition to these conditions, data collection from social media will also experience difficulties. Researchers who will conduct sentiment analysis will collect data using ISO (The International Organization for Standardization) language code. Bahasa Indonesia has a code "**id**", and Bahasa Melayu has a code "**msa**".

So that at certain moments, one word can affect the analysis of sentiment in the future. Sentiment results can't match the intended. This query validation process requires expertise from human annotations.

To overcome this, we need an annotation scheme model that can help the language annotation process. However, there is a lack of annotation scheme models that can be accessed for similar languages such as Bahasa Indonesia and Bahasa Melayu, and the corpus containing vocabulary words is the same but has different meanings.

The annotation scheme model found at this time can only identify one type of language and cannot support similar languages, such as TimeML [6], ISO-TimeML [7] OpinionMining-ML [8], SentiML [9], SentiML++ [10] dan OpinionML [5].

This paper will first describe the methods used to form the scheme model that is known as the BMBI model (section 2). Then in section 3, the process of forming the BMBI Model and its guidelines is described and followed by the annotation process using the BMBI Model in section IV. Section V will explain about the evaluation results of the annotations using the BMBI Model. Section VI concludes this paper with future enhancement.

2. METHODS

This section will explain the improved model, the BMBI Annotation Scheme Model. The methodology used is MATTER (Model, Annotation, Train, Test, Evaluation, and Revise) [11]. Phases in MATTER are depicted in Figure 1. In accordance with Figure 1, the steps of the annotation stage consist of 4 (four) steps, namely the model and guidelines, annotate, evaluate, and revise. The first sub section describes the BMBI scheme model and guidelines. In it will be explained in detail about the structure and usefulness and it will also be explained what the benefits are of using the BMBI Annotation Scheme Model. The next section is an annotation process of models for BMBI corpus that have been built before, and then proceed with evaluation of the agreement calculation of 6 (six) annotators using Fleiss Kappa [12]. The last section explains the revision process of the annotation process using the BMBI scheme model.

In annotation process, manual annotation method is performed to ensure that no data is different from predetermined guidelines. But it is not possible to validate manual annotations directly. Therefore, we need at least two annotators who can evaluate the results of annotations using the Kappa family, one of which is. Annotators must get the same text samples in parallel and compute the results using coefficients [13].

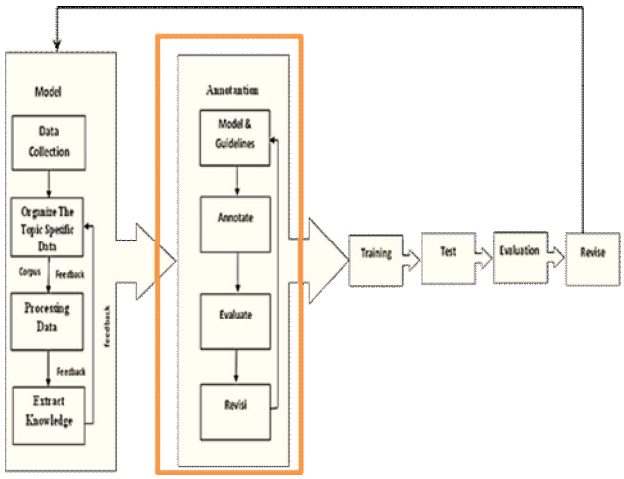


Figure 1: MATTER cycle

About 6 (six) annotators worked for this corpus for 3 weeks. 3 (three) of them worked on the Bahasa Indonesia corpus and 3 (three) worked on the Bahasa Melayu corpus. Below is the specification of the annotator that performs annotations using the BMBI Annotation Scheme Model, including:

1) Consists of 6 (six) annotators, namely 3 native-speakers Bahasa Indonesia and 3 native-speakers of Bahasa Melayu.
2) All annotators are native speakers of each language.
3) The annotators must have a minimum educational background of a Master in the field of IT or Linguistic.

The next section will explain the implementation of the methodology used to improve the model for Indonesian and Malay annotation schemes which have the same vocabulary but have different meanings.

3. BMBI ANNOTATION MODEL AND GUIDELINES

The BMBI Annotation scheme model is a model designed to help annotate several languages that have similarities, in this case a dataset from Bahasa Indonesia and Bahasa Melayu that has the same vocabulary but has a different meaning.

As explained for example in Introduction there is a number of words that have vocabulary similarities but differ in meaning from Bahasa Indonesia and Bahasa Melayu. This has an impact on determining the results of polarity and sentiment. These words can be annotated automatically using a previously available corpus. However, due to the difficulty of identifying which words are in Bahasa Indonesia and which are Bahasa Melayu words. Data collection from social media will also experience difficulties. Researchers who will conduct sentiment analysis will collect data using ISO (The International Organization for Standardization) language code. Indonesian has a code "**id**", and Melayu has a code "**msa**". In practice, when taking data in Indonesian, all data in Indonesian and Melayu is taken. However, when only taking data in Melayu, no data is retrieved. Table 1 displays the results of automatic annotations.

Table 1: Sentiment Polarity Bahasa Melayu And Bahasa Indonesia

Words	Sentiment	
	BM	**BI**
jimat	positive	negative
percuma	positive	neutral
asyik	neutral	positive
bual	positive	negative
gampang	negative	positive
…	…	…

The above results prove that there is a difference in polarity even though it says the same. But if you pay attention, in the word "*percuma*" (useless) the results of sentiment in Indonesian is not very precise. The use of the word "*percuma*" will be negative if placed in a sentence like "*Percuma saja kamu belajar kalau tidak dipahami dengan baik*". Therefore, the proposed model allows annotating words from the word level and sentence level sides. The following paragraph will explain the structure contained in the BMBI Annotation Model.

The BMBI Annotation Scheme model is divided into two structural parts namely the KATA structure group and the KALIMAT structure group. Element KATA becomes the outermost structure in which there is a KALIMAT structure group. In KATA there are *lang*, *tagset*, *polarity* and *spelling* attributes. On the KALIMAT element there is a polarity attribute which indicates the polarity of the sentence to the chosen KATA element. In addition to these 2 (two) elements, there are 4 (four) semantic elements that have been modified from previous models, namely SentiML, SentiML ++ and Opinion ML. The details will be explained one by one as follows:

1. The <KATA> element is the element that indicates the intended word. This <KATA> element has been selected or inputted according to a previously formed corpus. In this case the corpus contains the same vocabulary words but has different meanings from Bahasa Indonesia and Melayu.
2. The <KALIMAT> element is a sub-element of the <KATA> element that contains sentences that are examples of words intended by the <KATA> element. Annotation results from one word can get different annotation results. In the <KALIMAT> element, there are more attributes than the <KATA> element, although the polarity attribute appears also in the <KALIMAT> element. This is to prove that polarity can change if a word is included in a sentence.
3. The <HOLDER> element is a semantic element that describes the holder of the <KALIMAT> element that has been selected.
4. The <APPRAISAL GROUP> element is the link between the <TARGET> element and the <MODIFIER> element.
5. The <TARGET> element is a semantic element that explains the words that are the target of words that contain opinions.
6. The <MODIFIER> element is the next semantic element that search for words that contain negative, positive, neutral or ambiguous sentiment meaning from a sentence.

[1] <KATA> element

The <KATA> element is the element that indicates the intended word. This <KATA> element has been selected or inputted according to a previously formed corpus. In this case the corpus contains the same vocabulary words but has different meanings from Bahasa Indonesia and Bahasa Melayu.

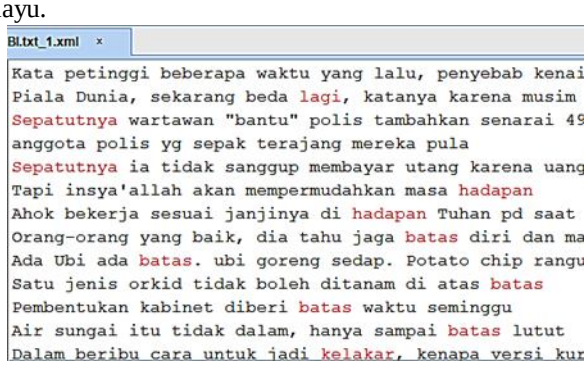
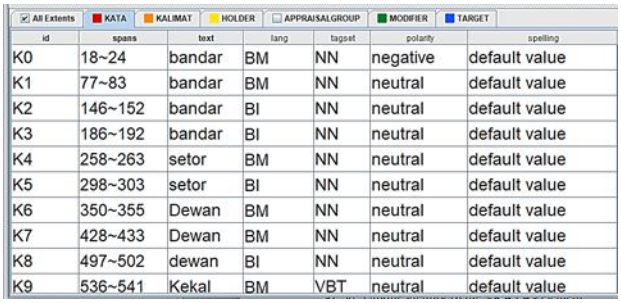


Figure 2: The red word has been selected in the <KATA> element

Figure 2 shows there are several words with red highlights. These words have been chosen using the <KATA> element. In this process, the words were predetermined according to the corpus that had been built previously, namely BMBI Corpus which contained Bahasa Indonesia and Bahasa Melayu words with the same vocabulary but had different meanings.

```
<KATA ELEMENT>
<KATA id=<ID>>
<descriptions text= <string>
spelling="default value" polarity=
<positive/negative/neutral>
tagset="NN" lang=<BM/BI> />
</KATA>
```

Figure 3: KATA Element



id	spans	text	lang	tagset	polarity	spelling
K0	18~24	bandar	BM	NN	negative	default value
K1	77~83	bandar	BM	NN	neutral	default value
K2	146~152	bandar	BI	NN	neutral	default value
K3	186~192	bandar	BI	NN	neutral	default value
K4	258~263	setor	BM	NN	neutral	default value
K5	298~303	setor	BI	NN	neutral	default value
K6	350~355	Dewan	BM	NN	neutral	default value
K7	428~433	Dewan	BM	NN	neutral	default value
K8	497~502	dewan	BI	NN	neutral	default value
K9	536~541	Kekal	BM	VBT	neutral	default value

Figure 4: Display of DTD results from <KATA> element

Figure 3 shows that inside the <KATA> element there are 4 main attributes and Figure 4 shows the display of it. The main attributes of <KATA> element which are:

a. id: Unique identity of the <KATA> element
b. text: The exact text of the word as found
c. lang: i.e. determine the type of language of the word. Figure 5 shows that there are two choices that the annotator must choose, namely BM (Bahasa Melayu) and BI (Bahasa Indonesia).

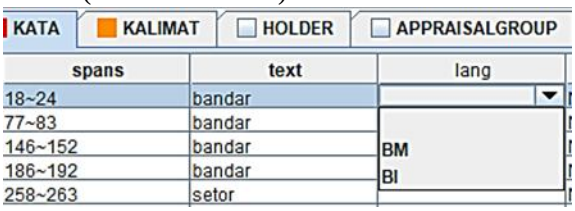


Figure 5: The "lang" attribute in the <KATA> element

d. tagset: i.e. determine the type of tagset of the word in question. The annotator will choose one of the 27 tagset based on the standard reference from each language [14], [15] as seen in Figure 6.

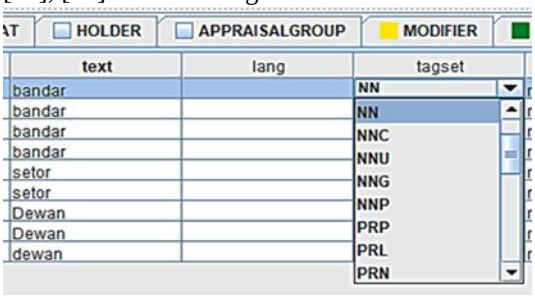


Figure 6: The "tagset" attribute in the <KATA> element

e. polarity: i.e. determine the polarity or sentiment of the word in question. Figure 7 shows that there are 3 (three) choices: positive, negative and neutral.

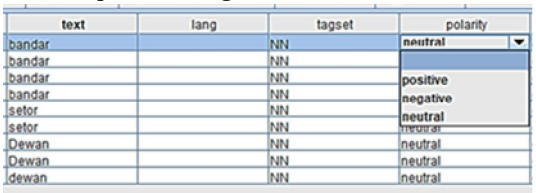


Figure 7: The "polarity" attribute in the <KATA> element

f. spelling: for this attribute only, a description will be provided which will be filled out by the annotator in case of spelling changes to the word. Example of the word "*kapan*" (shroud) in Bahasa Melayu is changed to "*kafan*".

[2] <KALIMAT> element

The <KALIMAT> element is a sub-element of the <KATA> element that contains sentences that are examples of words intended by the <KATA> element. Annotation results from one word can get different annotation results. In the <KALIMAT> element as seen in Figure 8 and the display of DTD in Figure 9, there are less attributes than the <KATA> element, although the polarity attribute appears also in the <KALIMAT> element. This is to prove that polarity can change if a word is included in a sentence.

```
<KALIMAT ELEMENT>
<KALIMAT id=<ID>>
<descriptions text= <string>
polarity=<positive/negative/neutra
l> />
</KALIMAT>
</KALIMAT ELEMENT>
```

Figure 8: <KALIMAT> element

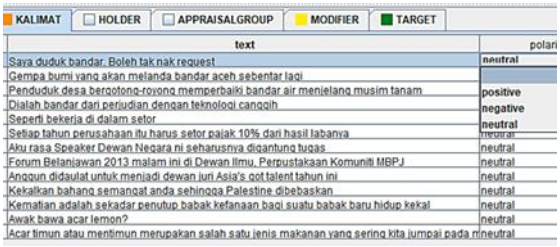


Figure 9: Display of DTD results from <KALIMAT> element

Figure 10 shows that there are several sentences with orange highlights. These sentences have been chosen using the <KALIMAT> element. The sentences are obtained from social media where there are words listed in BMBI corpus.

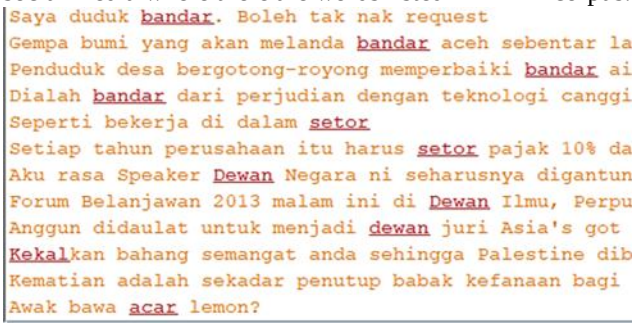


Figure 10: The sentences have been selected in the <KATA> element

The attributes contained in the <KALIMAT> element are as follows:

a. id: Unique identity of the <KALIMAT> element
b. text: The exact text of the sentence as found
c. polarity: same as the use of the polarity attribute on the <KATA> element as seen in Figure 11, this attribute also contains 3 (three) choices, positive, negative, and neutral. But the results can be different from the polarity results in the <KATA> element along with the formation of sentences in the <KALIMAT> element. For example,

for words "*bandar*" (port). In the <KATA> element, the polarity of the word is neutral. But when included in the sentence "*Dialah bandar dari perjudian dengan teknologi canggih*" (He is the bookie of gambling with advanced technology) the polarity of the sentence becomes negative.

id	fromID	fromText	toID	toText	orientation
A0	T0	harga	M0	kenaikan	negative
A1	T1	uang	M1	habis	negative
A2	T2	masa	M2	mempermuda	positive
A3	T3	ubi	M3	sedap	positive
A4	T4	orkid	M4	boleh	ambiguous
A5	T5	waktu	M5	batas	neutral

Figure 11: The "polarity" attribute of the <KALIMAT> element

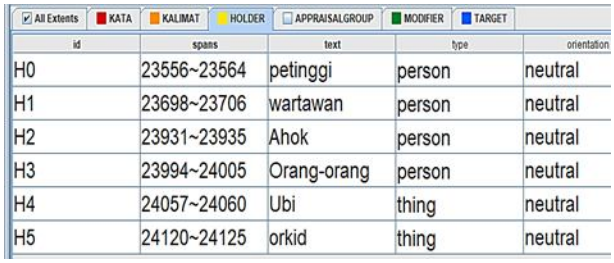
The next section is an explanation of other elements which are semantic elements adapted from several previous models. These elements include the HOLDER, TARGET, MODIFIER and APPRAISAL GROUP as non-consuming elements which are the link tags of the TARGET and MODIFIER elements.

[3] <HOLDER> element

An opinion holder means an entity that has a specific opinion on a particular topic or problem. The <HOLDER> element contains many unique <HOLDER> elements. Each <HOLDER> has a unique identity that can be used as a reference anywhere in the document. Element <HOLDER> is a semantic attribute that indicates who owns the sentence in question as seen in Figure 12 and Figure 13.

```
<HOLDER ELEMENT>
<HOLDER id=<ID>>
<descriptions text= <string> type=
"thing" orientation=
<positive/negative/neutral/ambig
uous> />
</HOLDER>
</HOLDER ELEMENT>
```

Figure 12: < HOLDER> element



id	spans	text	type	orientation
H0	23556~23564	petinggi	person	neutral
H1	23698~23706	wartawan	person	neutral
H2	23931~23935	Ahok	person	neutral
H3	23994~24005	Orang-orang	person	neutral
H4	24057~24060	Ubi	thing	neutral
H5	24120~24125	orkid	thing	neutral

Figure 13: Display of DTD results from <HOLDER> element

The attributes contained in the <HOLDER> element are as follows:

a. id: Unique identity of the <HOLDER> element
b. text: The exact text of the word as found
c. type: Type of the holder entity i.e. person, organization, country, place, concept or thing.
d. orientation: contains positive, negative, neutral and ambiguous values.

[4] <APPRAISAL GROUP> element

The <APPRAISAL GROUP> element is the link between the <TARGET> element and the <MODIFIER> element, so it can be seen the relationship between the two elements, including the orientation of the relationship (positive, negative, neutral, ambiguous). This relationship can be divided into several forms:

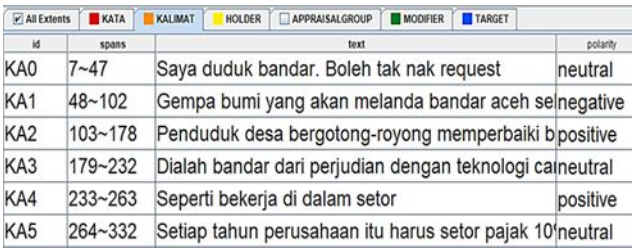
This relationship can be divided into several forms:

1. A noun with adjective. For example, in the word "*Saya suka*" (I like). The word "*Saya*" (I) is a noun from the <TARGET> element while the word "*suka*" (like) is an adjective from the <MODIFIER> element.
2. A verb with noun. As an example in the word "*kacak pinggang*" (akimbo). The word "*Kacak*" (conceited) is a verb of the <MODIFIER> element while the word "*pinggang*" (waist) is a noun from the <TARGET> element.
3. An adjective with verb. As an example in the word "*suka berbual*" (like to brag). The word "*suka*" (like) is the adjective of the <TARGET> element while the word "*berbual*" (boast) is a verb of the <MODIFIER> element.

The <APPRAISAL GROUP> element has attributes namely id, fromID, from Text, ToID, orientation as seen in Figure 14 and Figure 15.

```
 APPRAISAL GROUP ELEMENT>
<APPRAISAL id=<ID>>
<descriptions fromID=<ID>
fromText=<string> toID=<ID>
toText="<string>
orientation=<positive/negative/neu
tral/ambiguous> />
</APPRAISAL GROUP>
</APPRAISAL ELEMENT>
```

Figure 14: <APPRAISAL GROUP> element



id	spans	text	polarity
KA0	7~47	Saya duduk bandar. Boleh tak nak request	neutral
KA1	48~102	Gempa bumi yang akan melanda bandar aceh se	negative
KA2	103~178	Penduduk desa bergotong-royong memperbaiki b	positive
KA3	179~232	Dialah bandar dari perjudian dengan teknologi ca	neutral
KA4	233~263	Seperti bekerja di dalam setor	positive
KA5	264~332	Setiap tahun perusahaan itu harus setor pajak 10	neutral

Figure 15: Display of DTD results from <APPRAISAL GROUP> element

The attributes contained in the <APPRAISAL GROUP> element are as follows:

a. id: Unique identity of the <APPRAISAL GROUP> element.
b. fromID: is the ID of the <TARGET> element selected to

be associated with the <MODIFIER> element.

c. fromText: contains the text from the fromID attribute.
d. toID: is the ID of the <MODIFIER> element that will be associated with the <TARGET> element in the fromID attribute. This attribute selects a modifier that will be linked to the target in the fromID attribute.
e. toText: contains the text of the toID attribute.
f. orientation: contains positive, negative, neutral, and ambiguous values.

[5] <TARGET> element

The <TARGET> element is an entity that is addressed by a sentiment. One target can have more than one sentiment value. The <KATA> element can also be a <TARGET> element so that one word can contain two elements as seen in Figure 16 and Figure 17. For example, "*Kalau dilihat pada rupa, memang semua manusia itu cantik dan kacak*". The word "*manusia*" is an element of <TARGET> which explains the sentiment of "*cantik*" and "*kacak*".

```
 TARGET ELEMENT>
<TARGET id=<ID>>
<descriptions text= <string>
orientation=<positive/negative/neu
tral/ambiguous> />
</TARGET>
</TARGET ELEMENT>
```

Figure 16: <TARGET> element

id	spans	text	type	orientation
T0	23609~23614	harga	thing	neutral
T1	23861~23865	uang	thing	neutral
T2	23918~23922	masa	concept	neutral
T3	24072~24075	ubi	thing	neutral
T4	24120~24125	orkid	thing	neutral
T5	24193~24198	waktu	thing	neutral

Figure 17: Display of DTD results from <TARGET> element

The attributes contained in the <TARGET> element are as follows:

a. id: Unique identity of the <TARGET> element
b. type: Type of the target entity i.e. person, organization, country, place, concept, or thing.
c. orientation: contains positive, negative, neutral, and ambiguous values. This attribute is filled in if the target contains sentiment.

[6] <MODIFIER> element

The <MODIFIER> element is used to search for words that contain negative, positive, neutral, or ambiguous sentiment meaning from a sentence. In one sentence can have more than one element <MODIFIER> as seen in Figure 18 and Figure 19. The <KATA> element can also be a <MODIFIER> element so that one word can contain two elements. For example, "*Kalau dilihat pada rupa, memang semua manusia itu cantik dan kacak*". The word "*kacak*" is selected in the <KATA> element, but also is chosen as the <MODIFIER> element.

```
<MODIFIER ELEMENT>
<MODIFIER id=<ID>>
<descriptions attitude=
<affect/judgement/appreciation>
orientation=<positive/negative/n
eutral/ambiguous>
force=<high/low/normal/reverse>
polarity=<marked/unmarked> />
</MODIFIER>
</MODIFIER ELEMENT>
```

Figure 18: MODIFIER element

id	spans	text	attitude	orientation	force	polarity
M0	23600~	kenaikan	affect	neutral	normal	unmarked
M1	23875~	habis	affect	neutral	normal	unmarked
M2	23903~	mempermudahkan	affect	neutral	normal	unmarked
M3	24083~	sedap	appreciati	neutral	normal	unmarked
M4	24132~	boleh	judgemen	negative	high	marked
M5	24187~	batas	affect	neutral	normal	unmarked

Figure 19: Display of DTD results from <MODIFIER> element

The attributes contained in the <MODIFIER> element are as follows:

a. id: Unique identity of the <MODIFIER> element
b. attitude: Type of the target entity i.e. person, organization, country, place, concept, or thing.
c. orientation: contains positive, negative, neutral, and ambiguous values. This attribute is filled in if the target contains sentiment.
d. force: contains the intensity values of the <MODIFIER> element, which are low and high.
e. polarity: this attribute to give a sign if in the sentence contains the word negation.

4. BMBI ANNOTATION PROCESS

The annotation process uses the BMBI scheme model that has been designed and uses the BMBI corpus that has been formed. As many as 1000 tweets were collected for each language. From the 1000 tweets, 7000 words were generated from the data pre-processing process in each language. Then 2100 words were taken that had the same vocabulary in both languages, and then only 300 words had different meanings which would later be included in the BMBI corpus.

The annotation process requires several annotators who understand the language being annotated both formal and informal. However, the annotators must have guidelines and use language references that are recognized by their respective language rules in working on this process.

About 6 (six) annotators worked for this corpus for 3 weeks. 3 (three) of them worked on the Bahasa Indonesia corpus and 3 (three) worked on the Bahasa Melayu corpus. Below is the specification of the annotator that performs annotations using the BMBI Annotation Scheme Model, including:

1. Consists of 6 (six) annotators, namely 3 native-speakers Bahasa Indonesia and 3 native-speakers of Bahasa

Melayu.

2. All annotators are native speakers of each language.
3. The annotators must have a minimum educational background of a Master in the field of IT or Linguistic.
4. A total of 300 words are the focus of this corpus, and 80% of 618 sentences are formed that will be reviewed by each annotator.

The annotators may not carry out annotations without standard references according to the rules and part-of-speech of each language. In this study, Indonesian and Malay are used; the references used are taken from the sources below:

1. Bahasa Indonesia dictionary (Kamus Besar Bahasa Indonesia)
2. Bahasa Melayu Dictionary
3. Link https://kbbi.web.id/ for Bahasa Indonesia
4. Link http://prpm.dbp.gov.my/ for the Bahasa Melayu (but no tags are provided)
5. Link http://prpmv1.dbp.gov.my/ for the Bahasa Melayu (but no tags are provided)

After all the above conditions are fulfilled, the annotators will be given annotation guidelines and how to use the BMBI Annotation Scheme Model as described in points [1] to [6]. This annotation process has been revised 3 (three) times to get the appropriate model shape and will be explained in section D.

5. BMBI EVALUATION

This section will explain the evaluation results of BMBI training data and BMBI testing data. This evaluation uses a Fleiss Kappa calculation of 6 (six) annotators. This process is carried out 3 (three) times if the agreement does not reach a minimum value of 70%.

A. Inter-Annotator Agreement

The reliability of the model that has been designed will be tested using the inter-annotator agreement technique using Fleiss Kappa. The use of human annotators to check the reliability and validity of the model is expected to get more detailed and thorough results, especially for the determination of POS tags and sentiment based on references that have been determined for each language. This stage processes data of each language namely 80% training data and 20% testing data. It takes each 6 (six) annotator to process Bahasa Melayu and Bahasa Indonesia sentences. All annotators must agree on the results of this evaluation process at least 70% [16]. This process uses the Inter-Annotator Agreement and its statistical calculations using Fleiss Kappa. If the result is below 70%, then revision of the model or data must be made.

Kappa (κ) can be defined as:

$$\kappa = \frac{P_0 - P_e}{1 - P_e}$$

P is the actual agreement and P_e is the expected agreement. The agreement value is based on how many annotators agree on a tag. Assessments are made for each language, i.e. the maximum score is 3, because each language is assessed by 3 (three) annotators such as those shown in Table 2.

Table 2: Diagnose on 10 sentences by three annotators per sentence

n_{ij}	Negative	Positive	Neutral	P_i
Sentence 1	0	3	0	1
Sentence 2	0	0	3	1
Sentence 3	0	0	3	1
Sentence 4	3	0	0	1
Sentence 5	3	0	0	1
Sentence 6	0	0	3	1
Sentence 7	0	3	0	1
Sentence 8	1	0	2	0.333
Sentence 9	0	0	3	1
Sentence 10	0	0	3	1
Total	7	6	17	9.333
p_j	**0.233**	**0.2**	**0.567**	

a. Evaluation of Kappa value in Language Identification (lang)

This evaluation aims to determine the level of agreement between annotators in identifying a word, whether it is Bahasa Indonesia or Bahasa Melayu. Specifically, for this attribute, the annotation results are taken from a combination of the six (6) annotators, meaning that the results are not distinguished between the results of annotations from Bahasa Indonesia annotators and Bahasa Melayu annotators.

Table 3: Annotation data from lang attribute

Words	BI	BM	P_i
W1	0	6	1
W2	5	1	0.666667
W3	5	1	0.666667
W4	5	1	0.666667
W5	0	6	1
W6	6	0	1
W7	0	6	1
…	…	…	…
W618	6	0	1
Total	1617	2091	539
p_j	**0.436084**	**0.563916**	

Based on Table 3 the value is known n=6, N= 618, k=2, sum of Pi = 539 and sum of cells= 618. Then the calculation of $\bar{P}$, $\bar{P}_e$ and k (kappa) as follows:

$$\bar{P} = \frac{1}{618}(539) = 0.872$$
$$\bar{P}_e = 0.436^2 + 0.564^2 = 0.508$$
$$\kappa = \frac{0.872 - 0.508}{1 - 0.508} = 0.739$$

Then the result of kappa value for lang attribute is 0.739 or 73.9%.

b. Evaluation of Kappa value in polarity (on the <KATA> element)

This evaluation aims to determine the level of agreement between annotators in determine the polarity of the <KATA> element, whether it is Bahasa Indonesia or Bahasa Melayu. The annotation results are taken from each language, i.e. 3 (three) from Bahasa Indonesia annotators and 3 (three) from Bahasa Melayu annotators.

Table 4: Annotation Bahasa Indonesia data from polarity attribute on <KATA> element

Words	negative	positive	neutral	P_i
W1	0	0	3	1
W2	0	0	3	1
W3	0	0	3	1
W4	1	0	2	0.333333
W5	0	0	3	1
W6	0	0	3	1
W7	0	0	3	1
…	…	…	…	…
W618	0	0	3	1
Total	366	178	1310	
p_i	**0.197411**	**0.096009**	**0.70658**	

Based on Table 4 the value is known n=3, N= 618, k=3, sum of Pi = 600.67 and sum of cells= 618. Then the calculation of $\bar{P}$, $\bar{P}_e$ and k (kappa) as follows:

$$\bar{P} = \frac{1}{618}(600.67) = 0.972$$
$$\bar{P}_e = 0.197^2 + 0.096^2 + 0.707^2 = 0.548$$
$$k = \frac{0.972 - 0.548}{1 - 0.548} = 0.938$$

Then the result of kappa value for Bahasa Indonesia data from polarity attribute on <KATA> element is 0.938 or 93.8%. Furthermore, the polarity of the <KATA> element in Bahasa Melayu will also be calculated.

Table 5: Annotation Bahasa Melayu data from polarity attribute on <KATA> element

Words	negative	positive	neutral	P_i
W1	0	0	3	1
W2	0	0	3	1
W3	0	1	2	0.33333
W4	0	0	3	1
W5	0	0	3	1
W6	0	0	3	1
W7	0	0	3	1
…	…	…	…	…
W618	0	1	2	0.33333
Total	342	212	1300	
p_i	**0.18447**	**0.11435**	**0.7012**	

Based on Table 5 the value is known n=3, N= 618, k=3, sum of Pi = 542 and sum of cells= 618. Then the calculation of $\bar{P}$, $\bar{P}_e$ and k (kappa) as follows:

$$\bar{P} = \frac{1}{618}(542) = 0.877$$
$$\bar{P}_e = 0.185^2 + 0.114^2 + 0.701^2 = 0.539$$
$$k = \frac{0.877 - 0.539}{1 - 0.539} = 0.733$$

Then the result of kappa value for Bahasa Melayu data from polarity attribute on <KATA> element is 0.733 or 73.3%.

c. Evaluation of Kappa value in polarity (on the <KALIMAT> element)

The purpose of this evaluation is to determine the value of the agreement between annotators regarding differences or changes in polarity between the <KATA> element and the <KALIMAT> element. This annotation process is also carried out separately by each language with 3 (three) annotators per language.

Table 6: Annotation Bahasa Indonesia data from polarity attribute on <KALIMAT> element

Sentences	negative	positive	neutral	P_i
S1	0	0	3	1
S2	1	0	2	0.333
S3	0	1	2	0.333
S4	1	0	2	0.333
S5	0	3	0	1
S6	0	0	3	1
S7	0	0	3	1
…	…	…	…	…
S618	1	0	2	0.333
Total	642	444	768	
p_j	**0.346278**	**0.2395**	**0.41424**	

Based on Table 6 the value is known n=3, N= 618, k=3, sum of Pi = 543.33 and sum of cells= 618. Then the calculation of $\bar{P}$, $\bar{P}_e$ and k (kappa) as follows:

$$\bar{P} = \frac{1}{618}(543.33) = 0.879$$
$$\bar{P}_e = 0.346^2 + 0.240^2 + 0.414^2 = 0.349$$
$$k = \frac{0.879 - 0.349}{1 - 0.349} = 0.814$$

Then the result of kappa value for Bahasa Indonesia data from polarity attribute on <KALIMAT> element is 0.814 or 81.4%. Furthermore, the polarity of the <KALIMAT> element in Bahasa Melayu will also be calculated.

Table 7: Annotation Bahasa Melayu data from polarity attribute on <KALIMAT> element

Sentences	negative	positive	neutral	P_i
S1	0	0	3	1
S2	0	0	3	1
S3	0	0	3	1
S4	0	0	0	1
S5	1	0	2	0.3333
S6	0	0	3	1
S7	0	0	3	1
…	…	…	…	…
S618	0	1	2	0.3333
Total	465	290	1098	
p_i	**0.25081**	**0.15642**	**0.5922**	

Based on Table 7, the value is known n=3, N= 618, k=3, sum of P_i = 519.5 and sum of cells= 618. Then the calculation of $\bar{P}$, $\bar{P}_e$ and k (kappa) as follows:

$$\bar{P} = \frac{1}{618}(519.5) = 0.841$$
$$\bar{P}_e = 0.251^2 + 0.156^2 + 0.592^2 = 0.438$$

$$k = \frac{0.841 - 0.438}{1 - 0.438} = 0.717$$

Then the result of kappa value for Bahasa Melayu data from polarity attribute on <KALIMAT> element is 0.717 or 71.7%.

d. Evaluation of Kappa value in tagset (on the <KATA> element)

The purpose of this tagset annotation is to find out the value of the agreement between annotators in each language. The results of this evaluation will indicate whether there is a difference in tagset for the same word in the two languages. This evaluation still uses 3 annotators for each language.

Table 8: Annotation Bahasa Indonesia data from tagset attribute on <KATA> element

Words	FW	SYM	NN	NNC	…	RP	P_i
W1	0	0	3	0	…	0	1
W2	0	0	3	0	…	0	1
W3	0	0	3	0	…	0	1
W4	0	0	3	0	…	0	1
W5	0	0	3	0	…	0	1
W6	0	0	0	0	…	0	1
W7	0	0	3	0	…	0	1
…	…	…	…	…	…	…	…
W618	0	0	3	0	…	0	1
Total	5	0	1122	98		12	
p_j	**0**	**0**	**0.60517**	**0**		**0**	

Based on Table 8 the value is known n=3, N= 618, k=28, sum of Pi = 519.5 and sum of cells= 618. Then the calculation of $\bar{P}$, $\bar{P}_e$ and k (kappa) as follows:

$$\bar{P} = \frac{1}{618}(524.7) = 0.849$$

$$\bar{P}_e = 0^2 + 0.605^2 + 0^2 + \cdots + 0.076^2 + \cdots + 0.129^2 + \cdots + 0^2 = 0.388$$

$$k = \frac{0.849 - 0.388}{1 - 0.388} = 0.753$$

Then the result of kappa value for Bahasa Indonesia data from tagset attribute on <KATA> element is 0.753 or 75.3%. Furthermore, the tagset of the <KATA> element in Bahasa Melayu will also be calculated.

Table 9: Annotation Bahasa Melayu data from tagset attribute on <KATA> element

Words	FW	Sym	NN	NNC	…	RP	Pi
W1	0	0	2	1	…	0	0.333
W2	0	0	2	1	…	0	0.333
W3	0	0	3	0	…	0	1
W4	0	0	3	0	…	0	1
W5	0	0	2	1	…	0	0.333
W6	0	0	3	0	…	0	1
W7	0	0	2	0	…	0	0.333
…	…	…	…	…	…	…	…
W618	0	0	0	0	…	0	1
Total	0	28	181	0		0	
pj	**0**	**0.015**	**0.098**	**0**		**0**	

Based on Table 9 the value is known n=3, N= 618, k=28, sum of Pi = 484.33 and sum of cells= 618. Then the calculation of $\bar{P}$, $\bar{P}_e$ and k (kappa) as follows:

$$\bar{P} = \frac{1}{618}(484.33) = 0.783$$

$$k = \frac{0.783 - 0.021}{1 - 0.021} = 0.778$$

$$P_e = 0^2 + 0.015^2 + 0.098^2 + \cdots + 0.015^2 + 0.098^2 + 0.004^2 + 0.015^2 + 0^2 = 0.021$$

Then the result of kappa value for Bahasa Melayu data from tagset attribute on <KATA> element is 0.778 or 77.8%.

Thus, the results of annotation evaluations using the BMBI annotation scheme model can be summarized as in Table 10.

Table 10: BMBI Fleiss Kappa Agreement Results for Different Tasks

No	Task	Agreement Score	Level of Agreement
1	Language Identification	73.9%	Moderate agreement
2	BI polarity (on the <KATA> element)	93.8%	Almost Perfect
3	BM polarity (on the <KATA> element)	73.3%	Moderate agreement
4	BI polarity (on the <KALIMAT> element)	81.4%	Strong agreement
5	BM polarity (on the <KALIMAT> element)	71.7%	Moderate agreement
6	BI tagset (on the <KATA> element)	75.3%	Moderate agreement
7	BM tagset (on the <KATA> element)	77.8%	Moderate agreement

As shown in Table 10, the results of the evaluation of the annotation process using the BMBI annotation scheme model are as expected, which is at least 70% or in the sense of a Moderate agreement.

6. CONCLUSION

This study develops a corpus of Bahasa Indonesia and Bahasa Melayu where the vocabulary is the same but has different meanings. Besides having different meanings, these words have different Part of Speech-tags and sentiments, thus allowing the results of sentiment analysis research to be contradict.

Because of this, the first contribution to this study was the formation of an annotation scheme BMBI model consisting of the elements <KATA>, <KALIMAT>, <HOLDER>, <TARGET>, <MODIFIER> and <APPRAISAL GROUP>. The BMBI model can identify differences in language, sentiment and tagset from Bahasa Indonesia and Bahasa Melayu from both the word level and sentence level. This model successfully carried out the annotation process for Bahasa Indonesia and Bahasa Melayu which was one of the members of the language group from similar languages.

From the BMBI annotation scheme model produced a second contribution, namely the formation of the BMBI corpus consisting of 300 words which are words that have the same vocabulary but have different meanings. This corpus BMBI also contains the tagset and sentiment values that have been annotated using the BMBI annotation scheme model.

The BMBI corpus before, before being declared feasible to be used for further research, must go through an evaluation process using a human annotator which is the third contribution of this study. The results of the agreement between annotators consisting of 6 annotators who annotated 618 sentences combined between Indonesian and Malay. Calculation of the agreement between annotators focused on 7 tasks namely language identifier, BI polarity on <KATA> element, BM polarity on <KATA> element, BI polarity on <KALIMAT> element, BM polarity on <KALIMAT> element, BI tagset and BM tagset; on average produces a Moderate agreement value which shows that the agreement results are feasible to be used as a basis for further research.

Overall, this research can be considered successful in forming the BMBI annotation scheme model, BMBI corpus and evaluating using human annotators.

The results of this research will form a new corpus named BMBI corpus. This corpus contains the results of annotations that have been carried out using the BMBI model as explained in this paper. Furthermore, the corpus will be conducted training, testing and evaluation using the SVM algorithm in accordance with the results of a survey paper that has been done previously [17]. The results of this training and testing determine the validity of the corpus for machine learning.

ACKNOWLEDGEMENT

The authors would like to thank the Universitas Widyatama, Indonesia, Fakulti Teknologi Maklumat dan Komunikasi (FTMK), Universiti Teknikal Malaysia Melaka (UTeM), Center of Advanced Computing Technology (C-ACT) and Computational Intelligence and Technologies laboratory (CIT Lab) research group for their incredible supports in this project.

REFERENCES

[1] M. Syamala and N. . Nalini, "A Deep Analysis on Aspect based Sentiment Text Classification Approaches," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 5, pp. 1795–1801, 2019. https://doi.org/10.30534/ijatcse/2019/01852019

[2] B. Liu, "Sentiment Analysis and Opinion Mining," no. May, pp. 1–108, 2012.

[3] A. M. Hasan, N. M. Noor, T. H. Rassem, and A. M. Hasan, "Knowledge-based semantic relatedness measure using semantic features," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 2, pp. 914–924, 2020. https://doi.org/10.30534/ijatcse/2020/02922020

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

[5] M. M. S. Missen *et al.*, "OpinionML-Opinion markup language for sentiment representation," *Symmetry (Basel).*, vol. 11, no. 4, pp. 1–37, 2019. https://doi.org/10.3390/sym11040545

[6] R. Ingria *et al.*, "TimeML: Robust Specification of Event and Temporal Expressions in Text," *New Dir. Quest. answering*, 2003.

[7] J. Pustejovsky, K. Lee, H. Bunt, and L. Romary, "ISO-TimeML: An international standard for semantic annotation," in *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, 2010.

[8] L. Robaldo and L. Di Caro, "OpinionMining-ML," *Comput. Stand. Interfaces*, vol. 35, no. 5, pp. 454–469, 2013.

[9] M. Di Bari, "Improving multilingual sentiment analysis using linguistic knowledge," no. September, 2015.

[10] S. M. Missen, M. Attik, A. Doucet, and C. Faucher, "SentiML ++ : An Extension of the SentiML Sentiment Annotation Scheme," *Lect. Notes Comput. Sci. 9341*, pp. 1–4, 2015.

[11] J. Pustejovsky and A. C. Stubbs, "Natural Language Annotation for Machine Learning," pp. 1–343, 2013.

[12] J. L. Fleiss, "Measuring Nominal Scale Agreement Among Many Raters," *Psychol. Bull.*, vol. 76, no. 5, pp. 378–382, 1971.

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

[14] 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.

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

[16] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochem. Medica*, vol. 22, no. 3, pp. 276–282, 2012. https://doi.org/10.11613/BM.2012.031

[17] F. Rumaisa, H. Basiron, Z. Saaya, and N. Khamis, "A literature research on machine learning techniques used for training annotated corpus," *Int. J. Recent Technol. Eng.*, 2019.