Set of Frequent Word Item sets as Feature Representation for Text with Indonesian Slang

Dian Sa'adillah Maylawati[1,3], **G. A. Putri Saptawati**[2,4]
[1]Departement of Informatics, State Islamic University of Bandung, Indonesia
[2]Informatics, Bandung Institute of Technology, Indonesia
[3]Email: diansm@uinsgd.ac.id
[4]Email: putri@informatika.org

Abstract. Indonesian slang are commonly used in social media. Due to their unstructured syntax, it is difficult to extract their features based on Indonesian grammar for text mining. To do so, we propose Set of Frequent Word Item sets (SFWI) as text representation which is considered match for Indonesian slang. Besides, SFWI is able to keep the meaning of Indonesian slang with regard to the order of appearance sentence. We use FP-Growth algorithm with adding separation sentence function into the algorithm to extract the feature of SFWI. The experiments is done with text data from social media such as Facebook, Twitter, and personal website. The result of experiments shows that Indonesian slang were more correctly interpreted based on SFWI.

1. Introduction

Social media acts as a medium for the emergence of a variety of slang, including Indonesian slang. The use of Indonesian slang results in difficulties for social media mining since it has unstructruced syntax, and since there is no complete and official source of Indonesian slang [9]. Moreover, Indonesian slang is increasingly deviate from the rules, it called 'Alay' language. There have been available text mining research, none has pre-processing phase for handling Indonesian slang.

Compared with bag of words, Multiple of words is considered as good text representation [5] [8] since it can semantically maintain the meaning of sentence by maintaining the relationship between words, phrases, clauses and sentences [1] [5] [10]. Frequent Word Sequence (FWS) [2] [4] [10] and Set of Frequent Word Sequence (HFWS) [10] which are text features based on sequential pattern, have showed their ability in semantically maintaining document meaning. However, they work well for formal text (documents). Meanwhile, Frequent Word Item sets (FWI) which is feature based on frequent pattern [2] [3], is not work well for formal text [10]. FWI does not consider the sequence words (phrases, clauses, moreover sentence) to maintain the text meaning [2] [10]. Therefore, frequent pattern can be used as a structured representation of the text that is not formal.

In this paper, we propose Set of Frequent Word Item sets (SFWI) which based on frequent pattern as text features for Indonesian slang. In description above, FWI can not maintain the text document meaning well because FWI does not consider sequence of words. But intead, in SFWI, the sequence of sentence is payed, because the sentence contain the meaning of text. In our work, feature of SFWI was produced by Frequent Pattern Growth (FP-Growth) algorithm [7] with adding separation sentence function into algorithm.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

2. Features Extraction for text with Indonesian slang

We begin this section by explaining the general concept of SFWI and how it is used as features for text with Indonesian slang. Then, we explain how to prepare text with Indonesian slang before conducting feature extraction based on SFWI.

2.1. SFWI representation for text with Indonesian slang

FWI is a structured representation for the text which are text based on frequent pattern [2] [10]. FWI has the following syntax structure: <(w1,w2),(w3,w4),…>, where (w1,w2) is FWI i, (w3,w4) is FWI i+1, etc. All items in FWI always appear together, but do not mean consecutive. For example, (*cerita, seru*) will be treated the same as (*seru, cerita*).

To overcome the problem of unstructured syntax in Indonesian slang, we propose to use Set of FWI (SFWI) instead of FWI. SFWI representation has the following syntax structure: << (w1, w2), (w3, w4)>, <(w5, w6) >>, where (w1, w2) is FWI i, (w3, w4) is FWI i + 1, (w5, w6) is FWI i + 2. Each item in FWI does not mean consecutive, instead it means that all items always appear together. <(w1, w2), (w3, w4)> is called SFWI i with two FWI in sentence i, means that FWI i, FWI i + 1 and FWI i + 2 are consecutive. Moreover, <(w5, w6)> is SFWI i + 1 with one FWI in sentence i + 1. Sentence i and Sentence i+1 are consecutive too. For example, <<(*drama, korea, hipnotis, gue*), (*asli, seru, abis*))>>, <(*baper, parah*)>> is not equal to <<(*baper, parah*)>, <(*drama, korea, hipnotis, gue*), (*asli, seru, abis*)>>. But equal to <<(*gue, hipnotis, drama, korea*), (*seru, abis, asli*)>, <(*parah, baper*)>>. It means that the appearance of FWI in SFWI is consecutive depend on the sequence of sentence in text.

As mentioned before, Indonesian slang does not have structured syntax by which their meaning is not represented by the sequence of words, but by the sequence of sentence. Accordingly, SFWI is more applicable compared with FWS and more applicable than FWI because using sequence of sentence for maintain the meaning of text.

2.2. Handling Indonesian slang before the establishment SFWI

Prior to the extraction of SFWI we need to address some Indonesian slang as follow:

- Slang words which are exist in the slang dictionary will be recognized. If slang word is not exist in the slang dictionary, SFWI will still recognize the slang word if it appears frequently. For example, from 10 documents with minimum frequency is 5, we get SFWI <<(*doi, gantung, gue*)>, <(*galau, tingkat, dewa*)>> with frequency is 6. So, SFWI is still able to keep the meaning from those 10 documents.
- Slang that change the prefixes and suffixes that are not match with Indonesian (eg. the prefix ‘me-’ into ‘nge-’, the suffix ‘-kan’ to be ‘-in’, for example “mengerjakan” to be “ngerjain”. For handling this problem, the word will be changed to its basic word with stemming process. To do stemming process, it is prepared stem word dictionary. We used Porter algorithm for stemming with several modifications. Such as checking the word which is the stem word. Stemming process is only done for the words with affixes. Next, we add prefixes and suffixes which are appeared in Indonesian slang.
- There is a case in Indonesian slang that the slang word is the same as Indonesian word but has different meaning. For example the word “secara”, in Indonesian slang it means “confirmation” but in Indonesian it means “in a way”. Another example is the word “cumi”, in Indonesian it means “squid as animal” but in Indonesian slang it means “mockery”. Therefore, SFWI cannot represent the proper meaning of sentences if they use some Indonesian words for expressing different meaning in Indonesian slang. The word “cumi” in Indonesian slang and the word “cumi” in Indonesian will be considered as the same word by SFWI.
- In Indonesian slang, there is sound of word which is usually used conversation is used to be word. For example, the word “lihat”, in Indonesian slang is written “liat” (use the sound of the word). The problem is in Indonesian, “liat” it means “something soft, flexible” but in Indonesian slang it

means “lihat (see)”. If the sound of word which is written as word is the same as Indonesian word, then SFWI cannot differentiate them.

- Since Indonesian slang is common to shorten words such as “makan” to be “mkn”, “mkan”, “makn”. So, it is necessary to translate them into the complete words before extracting SFWI. This is a part of Natural Language Processing (NLP) that recognize an unfamiliar words or containing words or structures that have not been seen before. As consequences, it is required to recognize a complete collection of abbreviations for all possibilities of shortened words. SFWI will be able to keep the meaning as long as word acronym has been changed into the complete word.
- Excessive use of punctuation marks can be removed. Whilst the punctuation which is not appropriate will affect the process of separating sentences. The affect is the different SFWI with the same item. But as long as the word or sentence is frequent, the meaning is kept in SFWI.
- Slang for excessive or "Alay" and slang that mix with foreign or local language are only a few in this works, so they are not considered to be frequent. For handling foreign of local language, of course we need dictionary of all foreign and local language to translate them to Indonesian. Then, for “Alay” language which is mix character numbers and letter still difficult to distinguish. If all number characters are changed to be letter, such as “4” to “a”, “5” to “s”, “2” to “z”, “3” to “e”, “13” to “b”, “12” to “r”, and so on, for example “6u3 131n6un9 54m4 tul154n 6u3 53nd1121” to be “gue ieingung sama tulisan gue sendiizi” or “gue bingung sama tulisan gue sendiri”. However, because of "Alay" language is often used by particular group such as teenager, it is not considered to be frequent.

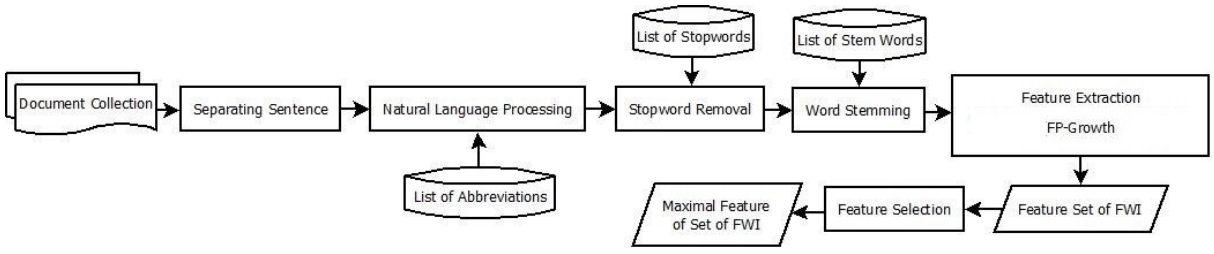


Figure 1. Process of creating the SFWI representation

2.3. The process of producing SFWI feature

This process starts with separating sentence. Followed by NLP tackling the characteristics of slang which use abbreviations. Next processes are eliminating stop words and word stemming. Finally, the process of generating and selecting SFWI features are conducted. Figure 1 describes the overall process in more detail. Detail of each phase for the process of extracting SFWI feature will describe as follow (the example available in Indonesian language in Figure 2):

- First, collection of text document will be separated by sentence in accordance with the SFWI’s structure. SFWI looks carefully at sequence of sentence, because a sentence has a complete meaning. Sentence separating process is done with simple algorithm, over the contents of document is read, if it finds the end of the sentence punctuation such as point (.), exclamation (!), or question (?), then it will be separated as different sentence.
- After separating sentence process, the content of document will be checked weather it contains a natural language such as abbreviation word. If it finds, it will be changed to be an original form. We used abbreviation dictionary which save abbreviation and its original form. Content of abbreviation dictionary such as *ak* to *aku*, *rame* to *ramai*, *lby* to *lebay*, *cys* to *ciyus*, *jys* to *jayus*.
- Next, the stop words will be removed. If the content of document contains stop word which exist in stop words dictionary, it will be removed. Stop words in Indonesian, including slang, such as *apakah, apalagi, adalah, atau, ataukah, ataupun, bagai, bagaikan, sebagai, sebagainya, bagaimana, dan, yang, ini, itu, loe, gw, deh*, etc

- Then, after removing the stop words, the words that contain an affixes (prefix or suffix) will be stemmed. We used Porter algorithm which is modified. We added stem word checking process at the first step. It the word is the stem word that exist in the stem dictionary, then it will not be stemmed. We also changed the prefix and suffix in accordance with the Indonesian grammar and Indonesian slang. Stem words in Indonesian, including slang, such as *belajar, sapu, mampu, ada, keles, bingit, rempong, lol, unyu, kepo, php, geje, gokil*, etc.
- Next step is feature extraction. This step is produced SFWI features from document text. We used FP-Growth for create SFWI. FP-Growth was selected because it produce frequent item set which is accordance with FWI representation. Frequent item sets which is produced by FP-Growth is the same as FWI's structure that it is not represent sequence of words. Just need to add the separating sentence process before use FP-Growth so that SFWI can be created.
- The last step is feature selection for selecting redundant features. For example if an item set has frequented, automatically sub item set has frequented too. We used maximum approach for this work. A feature f is called maximum if no more feature f' which sub feature of feature f.

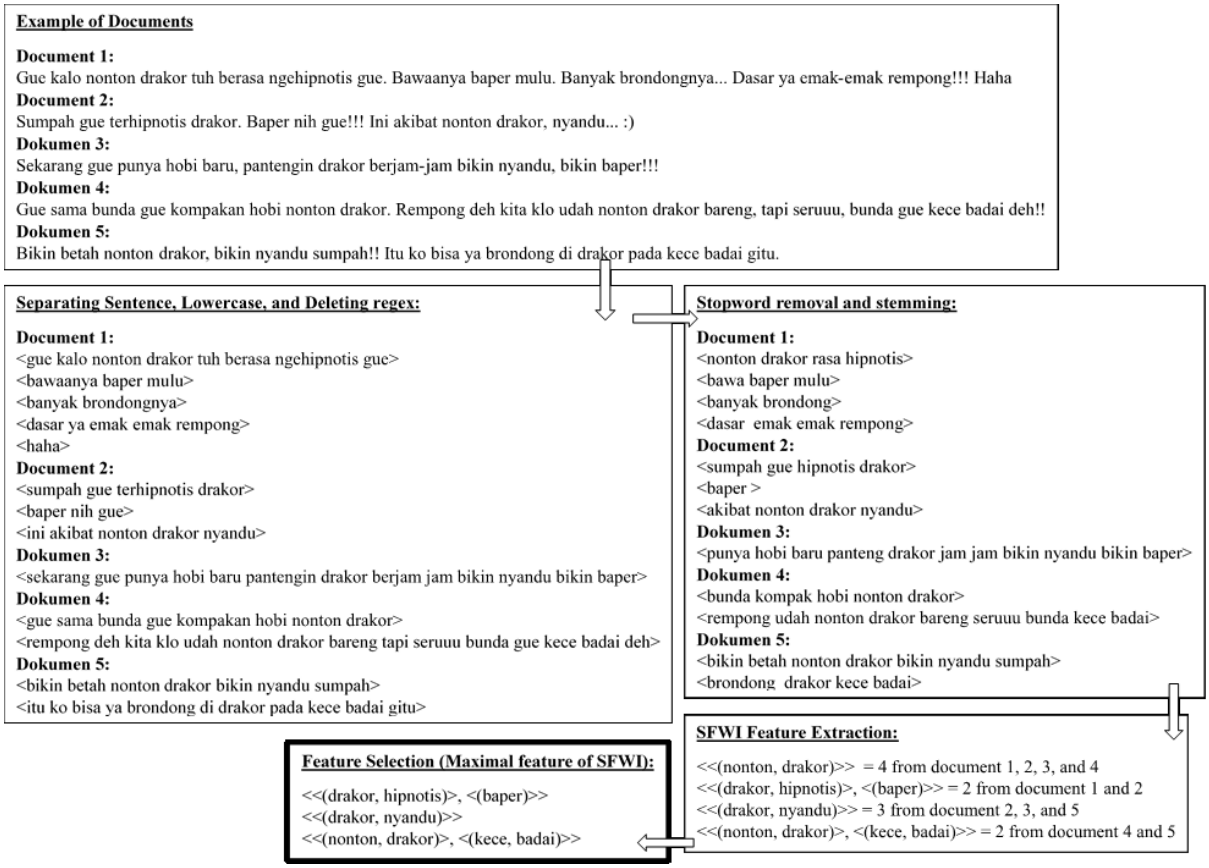


Figure 2. Example of producing SFWI with 40% minimum support, 1 maximum gap, and 1 maximum length

3. The experiment and result

In accordance with the needs of research, text documents taken from tweet of Twitter, some of the blogs that use Indonesian slang and Facebook. Text documents are grouped into directories according to the topic. SFWI representations are stored in a file. We build the text preprocessing library for the needs of this experiment. We also use Sequential Pettern Mining Framework (SPMF) Library for FP-Growth Algorithm [6]. The experiments were conducted based on several test scenarios as follows:

- Examine the meaning of Indonesian slang by SFWI representation. Text data is taken from Facebook (168 of data, with five category, among others: "galau", "hiburan", "gaul", "statustereliye", and "cinta"), Twitter (681 of data, with five category, among

others:“kampanye”,“tahunbaruislam”, “kabinetkerja”, “gayahidup”, and “artiskorea”), and form Blog (191 of data, with five category, among others: “anakgaul”,“cirihamilmuda”, “masalahrumahtangga”, ””“nasehatcinta”, and “pengobatantradisional”). This test aims to see accuracy of keeping slang meaning by SFWI representation. The combination of parameters entered for this test is the support of at least 10-100%, maximal gap 5 and 10, and a minimum length of 1, 2, and 3. The test that produces SFWI accordance with the input data topics rated 1 and the results are not appropriate topics or do not produce SFWI rated 0.

- Seeing the comparison of accuracy between SFWI representation and FWI representation in maintaining the meaning of Indonesian slang. The same test scenario for examining the meaning of Indonesian slang by SFWI representation will be used for FWI representation.

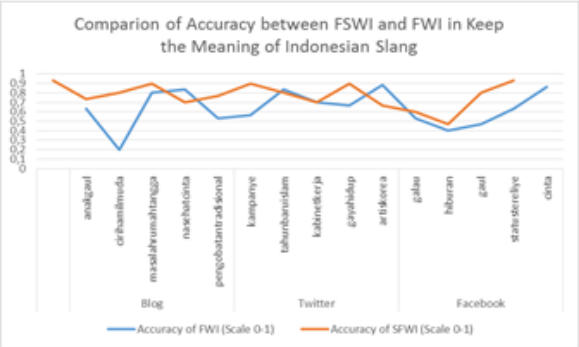


Figure 3. Comparison of Accuracy between SFWI and FWI

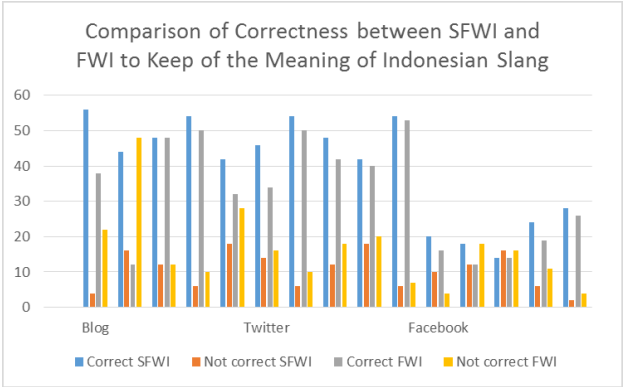


Figure 4. Comparison of Correctness between SFWI and FWI

Table 1. Result of keep the meaning of slang between Set of FWI and FWI representation

No	Dataset	Tot. File	Tot. Test	Correct SFWI		Not correct SFWI		Accuracy (Scale 0-1)		Time Average (ms)	
				SF WI	FWI	SF WI	FWI	SF WI	FWI	SFWI	FWI
1	anakgaul	72	60	56	38	4	22	0,93	0,63	15536	9125
2	cirihamilmuda	28	60	44	12	16	48	0,73	0,2	3092	2311
3	masalahrumahtangga	24	60	48	48	12	12	0,8	0,8	3773	2429
4	nasehatcinta	44	60	54	50	6	10	0,9	0,83	55161	14696
5	pengobatantradisional	23	60	42	32	18	28	0,7	0,53	1864	3206
6	kampanye	450	60	46	34	14	16	0,77	0,57	8363	2363
7	tahunbaruislam	100	60	54	50	6	10	0,9	0,83	1497	2369
8	kabinetkerja	41	60	48	42	12	18	0,8	0,7	1471	1652
9	gayahidup	23	60	42	40	18	20	0,7	0,67	1157	657
10	artiskorea	67	60	54	53	6	7	0,9	0,88	1531	778
11	galau	25	30	20	16	10	4	0,67	0,53	2649	3001
12	hiburan	22	30	18	12	12	18	0,6	0,4	2429	987
13	gaul	41	30	14	14	16	16	0,47	0,47	2468	2409
14	statustereliye	50	30	24	19	6	11	0,8	0,63	2575	2095
15	cinta	30	30	28	26	2	4	0,93	0,87	4807	4632
Total Time Average of SFWI : 7225,411922; Total Time Average of FWI : 5014,5890574											
Total Accuracy Average of SFWI : 0,773; Total Accuracy Average of FWI : 0,6367											

The evaluation of the results can be summarized as follows:
SFWI representation can keep the meaning of slang with an average percentage of accuracy is 77.33% (contained in Table 1), while FWI representation has 63.67% average percentage of accuracy (contained in Table 2). Even though in the case of the same slang word but the meaning is different,

SFWI, moreover FWI, cannot tell the difference, this experiment proves that SFWI pretty well in keeping the meaning of slang compared with FWI (Figure 3 and Figure 4). Only one case from all test case that FWI has the same total correctness and same accuracy. Several of the experiment show that SFWI has an accuracy that does not differ much with FWI, but most of them has a higher accuracy. And also the experiment proves that pay attension to the sequence of sentence can keep the meaning of unstructure language such as Indonesian slang well. The time process to extract FWI representation is faster than SFWI because there is no sentence separation process.

4. Conclusion

- In accordance with the results of the analysis and experiment, SFWI representation can be used for keeping the meaning of unstructured language such as Indonesian slang. SFWI has a structure that similar to structure of slang that it is not pay attention to sequence of words. However, SFWI is still pay attention to sequence of sentence which contain of meaning. Even though, SFWI not be able to keep the meaning of slang that contains a lot of meaning. Such as the same word with different meaning. The processes before producing SFWI, from separating sentence until stemming can affect the result of SFWI. Including completeness of stop words, abbreviation, and stem words which are owned.
- Testing and evaluation for keeping slang meaning with FWI representation use a collection of documents from Facebook, Twitter, and Blog. The percentage of accuracy of SFWI is 77.33%. This is higher than FWI that has the precentage of accuracy 63.67%. This accuracy is obtained based on the comparison of the correct and uncorrect value.
- For the future work, the accuracy of SFWI can be improved by increasing vocabulary of slang and abbreviation, because slang language always increases and a lot of new vocabulary. The FP-Growth algorithm which is used to extract SFWI can be replaced by another frequent pattern algorithm. And also the problem of SFWI which can not keep the meaning of word that contains a lot of meaning must be resolved to keeping the meaning better.

5. References

[1] Adityarani, Kania. 2009. *Aplicating CFWS* Algorithm for Document Grouping with Text representation Set of Frequent Word Sequence and Text Representation Qord Frequent itemset*. Informatic, Bandung Institute of Technology.

[2] Agrawal, Rakesh, et al. 1999. *Mapping Words, Phrases using Sequential Pattern to Find User Specific Trends in a Text Database – Patent Document*. International Business Machines Corporation, Armonk, NY.

[3] Agrawal, Rakesh. Ramakrishnan Srikant. 1995. *Mining Sequential Patterns*. IBM Research Center,

[4] Ahonen, Helena. 1999. *Knowledge Discovery in Document by Extracting Frequent Word Sequence*. Library Trends. Vol. 48, No. 1, Summer 1999, pp. 160-181.

[5] Doucet, Antoine. 2005. *Advance Document Description, a Sequential Approach*. Accademic Dissertation, Departement of Computer Science, Series of Publication A. Report A-2005-2, University of Helsinki, Finland.

[6] Fournier-Voger, Philippe. 2008-2016. *Sequential Pattern Mining Framework* (SPMF) *Data Mining Software*. *Tools.* Available in: http://www.philippe-fournier-viger.com/spmf

[7] Han, Jiawei. Kamber,Micheline. 2006. *Data Mining: Concepts and Techniques Second Edition*. Simon Fraser University. Morgan Kaufmann Publisher.

[8] Hoonlor, Apirak. July, 2011. *Sequential Patterns and Temporal Patterns for Text Mining.* A Thesis Submitted to Graduate Faculty of Rensselaer Polytechnic Institute in Partial Fulfillment of the Requirement for Degree of Doctof og Philisophy. Major Subject: Computer Science. NewYork.

[9] Muliady, Wahyu. 2012. *Generating Indonesian Slang Lexicons from Twitter*. International Conference on Uncertainty Reasoning and Knowledge Engineering.

[10] Nalistia, Rindhu Astri. 2008. *Feature-Based Clustering for documents grouping with the Structured Representation of Text*. Informatic, Bandung Institute of Technology.