The Search of Non-Standard Words in the Documents Written in Indonesian Language with Nazief and Adriani Algorithm

Dewi Soyusiawaty
Program Studi Informatika
Fakultas Teknologi Industri
Universitas Ahmad Dahlan Yogyakarta

Oko Carono
Program Studi Informatika
Fakultas Teknologi Industri
Universitas Ahmad Dahlan Yogyakarta

ABSTRACT
Indonesian language has a variety of affixed words used in a document. The words or sentences in a document must be written based on the Great Dictionary of Indonesian Language (KBBI). Errors often occur when writing a word in the document such as errors in writing the standard words. To find out the standard and non-standard forms of an affixed word needs the root. One of methods to find the root of an affixed word is by using Nazief & Adriani Stemming Algorithm. Searching the root words in a document by checking them one by one will take a long time and is not efficient. Therefore, an application that can search the root words is required to make the quick and more efficient search. This research is an implementation of the search of the root and standard words in the documents written in Indonesian language to ease in determining the standard and non-standard words. The method used is by checking the words in the documents then implementing Nazief & Adriani algorithm to find out the root words then checking in KBBI to determine the non-standard words and implementing spell checker method to recommend the standard ones. The testing used in this research is the accuracy testing by using 50 documents written in Indonesian language with 28,023 numbers of words and the result of the accuracy testing is 96.74%.

Keywords
Indonesian language, Nazief & Adriani Algorithm, Non-Standard Words, Stemming

1. INTRODUCTION
A document is one of communication media used by human beings usually in the form of written language. Writing errors in the document are common problems such as errors in writing the standard words that have not matched those in the Great Dictionary of Indonesian language (KBBI). Spelling errors often occur due to the writer's lack of understanding towards spelling, finger slip (mistyping), or keyboard errors. Stemming is a method used to increase the performance of Information Retrieval by changing the words or sentences in a document into its roots. Stemming is a process of removing all affixes attached to a word or a sentence consisting of prefix, infix, suffix and confix or the combination of prefix and suffix [4]. The process of stemming for texts written in Indonesian language is more complicated and more complex because there are various affixes that need to remove to obtain the root of a word or sentence. To minimize misunderstanding due to poor communication, it needs a translator or KBBI since KBBI is a media in learning good and correct Indonesian language. Since KBBI does not provide affixed words, to learn the language with affixed words, stemming algorithm is needed to find out the roots of the affixed words. Nazief & Adriani algorithm is an algorithm that is mostly used for stemming particularly in Indonesian language. Based on the problem discussed above, this research developed an application to search the standard and non-standard words according to KBBI in the text documents written in Indonesian language with Nazief & Adriani algorithm for affixed words.

2. LITERATURE REVIEW
2.1 Natural Language Processing
Natural Language Processing (NLP) is a field of artificial intelligence, computer science, and linguistics that studies interaction between computers and human natural language with all grammatical and semantic rules; also in particular how to change the language into a formal representation that can be processed by the computer. NLP is a way of analyzing texts with computerized means. NLP involves gathering of knowledge on how human understands and use language. This is done to develop an appropriate tool and technique that can make computer systems understand and manipulate natural language to carry out a variety of tasks [6]. NLP is a critical component of text mining and subfield of artificial intelligence and computational linguistics. Text mining uses natural language processing to input the structure into a collection of texts [11]. Text mining uses pre-processing text for searching, abstracting, and word categorization in a document.

2.2 Text preprocessing
Processing aims at obtaining dataset that can be processed quickly and result in appropriate conclusion, one of data processing processes that can be done is feature selection [5]. There are several stages in the feature selections such as:

1. Tokenizing
Tokenizing is a stage of cutting a word or sentence based on each word. The word processed is called token or term and that to save in database is term, then indexing is done to carry out the search. For an example is the sentence "*Aku makan bakso urat*". The result of tokenizing of the sentence is four tokens: "Aku", "makan", "bakso", "urat" [9].

2. Filtering
Filtering is a stage of selecting the essential words from the result of tokenizing. The process of filtering can use the stoplist algorithm (removing less essential words) or the wordlist algorithm (keeping the essential words). Stoplist or stopword is non-descriptive words that can be removed in the

approach of bag-of-words. The examples of stopwords are "yang", "dan", "di", "dari" and others [7].

3. Stemming

Stemming is a way to increase the Information Retrieval Performance by transforming the words in the text documents into the roots. Stemming is a process used to find out the roots from the affixed words by removing all affixes consisting of prefix, infix, suffix, and confix or combination of prefix and suffix [4].

2.3 Nazief & Adriani Algorithm

This algorithm was developed by Bobby Nazief and Mirna Adriani from Faculty of Computer Sciences Universitas Indonesia in 1996. This algorithm uses the Indonesian morphological rules that group an affixed word such as allowed affixes and disallowed affixes. Nazief & Adriani algorithm is an algorithm to change a word with suffix, prefix and confix to be its root [12][13]. Nazief & Adriani algorithm uses the root words as the dictionary for restructuring the words that experience over stemming. These root words are highly required to check whether the stemming process has run correctly or not. Nazief & Adriani algorithm has several stages as following [1]:

1. Find the word to be stemmed in the database. If it is found, it can be assumed that the word is a correct stem or root word.
2. Remove the inflectional suffixes by removing particles ("-lah", "-kah", "-tah", or "-pun"), then remove the inflectional possessive pronoun suffixes ("-ku", "mu", or "nya"). Check the word in the database, if it is found, the algorithms stops, if it is not found, continue to the next stage.
3. Remove the derivational suffixes ("i" or "-an"). If the word is in the database, stop. If it is not, continue to the stage of 3a:
   a. If suffix "-an" has been removed and the last letter of the word is "-k", "-k"is removed. If the word is in the database, stop. If it is not, continue to 3b.
   b. Remove suffix if ("-i", "- an" or "-kan") after that, continue to stage 4.
4. Remove Derivational Prefix ("be-", "di-", "ke-", "me-", "pe-", "se-" and "te-"). If the word is found in the database, the process stops. If it is not, continue to recoding. This stage will stop if several conditions has met the requirements as following:
   a. There is a disallowed combination of prefix and suffix
   b. The prefix is similar to that previously removed
   c. Three prefixes have been removed.
5. If the root word is not found in the database after following all stages, this algorithm will restore the word to that before stemming.

2.4 Spell checker

Spell checker is a machine or tool to check typing errors in a document. Spell checker helps the users who need to check their documents and can ease them to repeat searching and identify errors in spellings, characters or standard words. Spell checker aims at recommending the correct words based on the difference between the tested word and the correct word in dataset used to make particular system (citation). The purpose of this feature is to give the ease or anticipation for users who will search for particular words or sentences so the users can choose some alternative recommendations [2].

2.5 Standard Words of Indonesian Language

Standard words are both spoken and written words which are appropriate to standardized language rules. The rules are General Guidelines for Indonesian Spelling (EYD/ PUEBI), standard word grammar, and KBBI. In the context of standard language variety, both spoken and written languages use the standard words [8]. The followings are several spelling examples of Indonesian standard words which barely have the characteristics of both regional and foreign languages [3]:

Table 1. Samples of Regional Language Vocabulary

Non-standard	Standard
rapet	rapat
cuman	cuma
dudu'	duduk
gubug	gubuk

Standard viewed from the spelling means all words written not based on the spelling rules in EYD/PUEBI are non-standard ones and vice versa. The followings are samples of non-standard and standard words in EYD:

Table 2. Samples of vocabularies in EYD

Non-standard	Standard
ekpres	ekspres
kompleksistem	komplekssistim
do'a	doa
Jum'at	jumat

3. RESEARCH METHODOLOGY

3.1 Data Collection

This research used some datasets available such as:

1) Indonesian root words from the dataset with 28,743 numbers of words in txt format and can be accessed via link https://github.com/nolimitid/nolimit-kamus.
2) *Stopwords* (unessential words) from *sastrawi* library with 126 numbers of words.
3) Standard and non-standard words from kbbi library with 127,036 numbers of words.

3.2 System Requirement Analysis

1. Data requirement analysis

Data used were 50 student's documents written in Indonesian language, root words, and standard words based on kbbi.

2. Analysis of user requirements

The level of users in this research is communities who have interest with documents particularly to identify non-standard words.

3. System requirement

To determine the roots of affixed words in Indonesian language and find out the standard form of the roots.

4. RESULT AND DISCUSSION

4.1 System Design

The system design stage is carried out to describe in detail how the system works, in order to meet the needs and produce software that meet the user's need. Data used for this system design was the stemming system design in the documents written in Indonesian language with Nazief & Adriani algorithm. Figure 1 shows the system design. The followings are the stages:

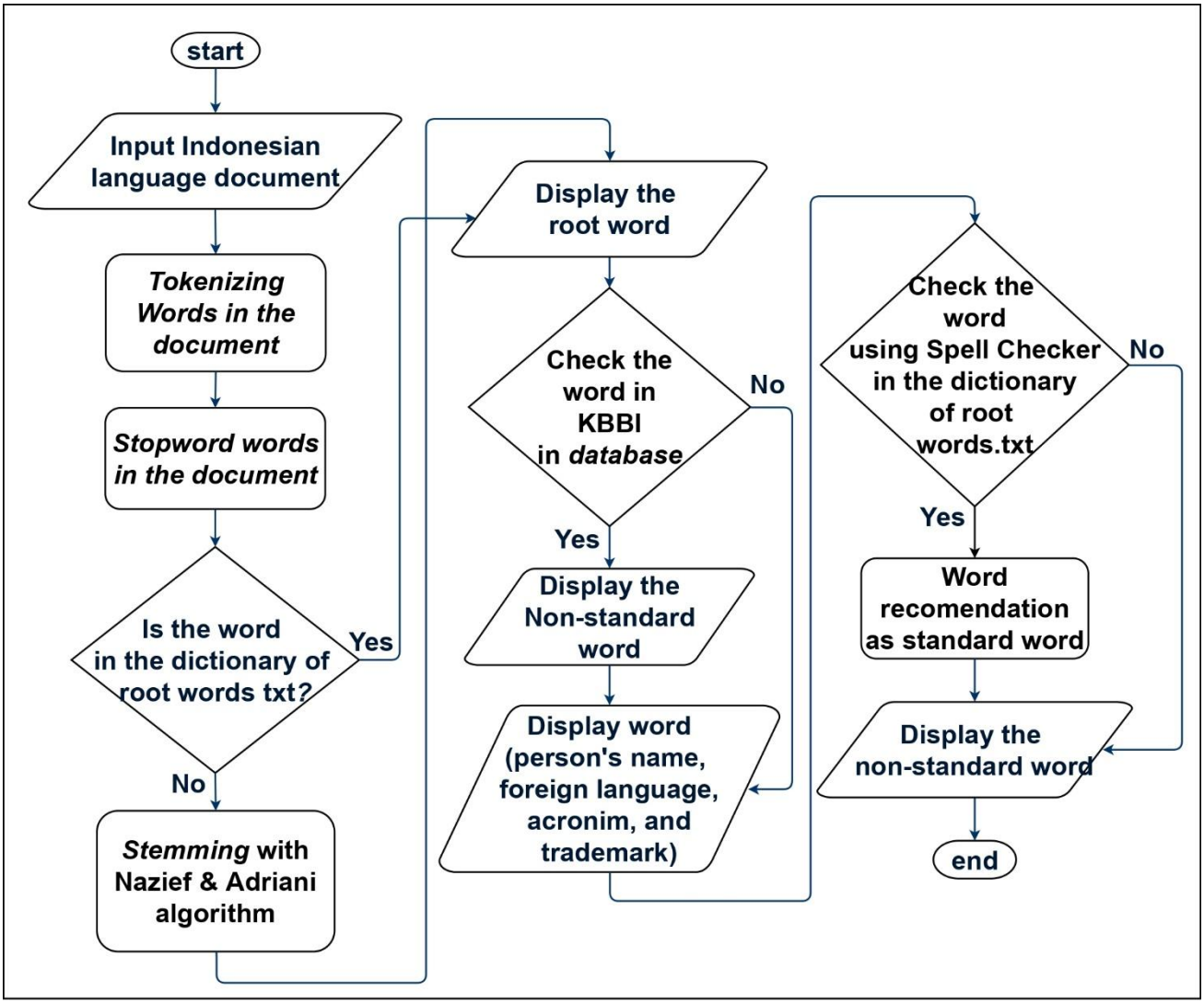

Figure 1. System Design

1. Inputting the documents written in Indonesian languages
2. *Tokenizing* the words in the inputted documents
3. The system will check whether the words in the documents are in the database or not. If the words in the documents are in the database, the system will display the root words, but if they are not, the system will carry out stemming by using Nazief & Adriani algorithm to find out the root words
4. If the root have been found, the system will check by using kbbi library to determine the non-standard words
5. If the non standard words have been found, spell checker will check them to obtain the recommendation of the standard form that the word with similar structure is taken in the database of root words.

The following figure shows the simulation of the system design:

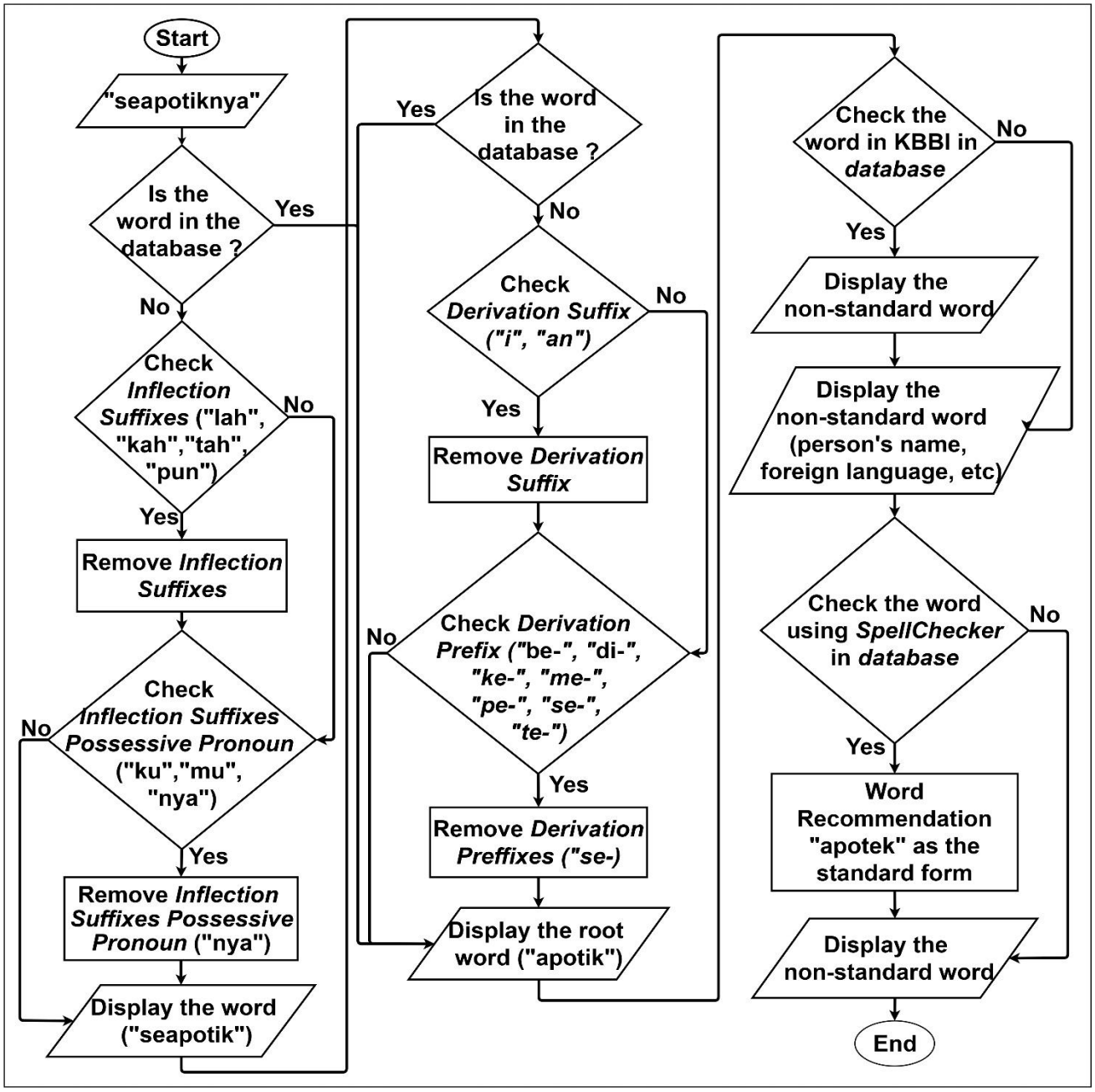

Figure 2. Simulation of System Design

The followings are the explanation of the system design simulation:

1. Check the word "seapotiknya" in the database, the word "seapotiknya" is not found in the database.
2. Check the Inflection Suffixes. In this Inflection Suffixes checking process, there is no particle so it directly continues to the Inflection Suffixes Possessive Pronoun and the Inflection Suffixes Possessive Pronoun ("nya") is removed, so the word becomes "seapotik". Check the word "seapotik" in the database, the word "seapotik" is not found in the database.
3. Check the Derivation Suffixes. Since there is no Derivation Suffixes, so it continues to check Derivation Prefixes and removes Derivation Prefixes ("se"). Check the word "apotik" in the database, if the word is found, it is displayed, and if it is not found it displays the root word.
4. The stemming process ends and the result is the word "apotik".
5. Check the word "apotik" in KBBI saved in the database. To check the standard form of the word "apotik", the word "apotek" is found in the database and is given as a recommendation of the standard form of "apotik". Therefore the root of the word "seapotiknya" is "apotik" and the standard form is "apotek".

4.2 System Testing

The system testing is carried out after the program has been developed to determine the accuracy level of the system. Testing in this research used the accuracy test in the forms of the words in the documents written in Indonesian language to test the accuracy level of the system in finding the standard and non-standard words. The accuracy testing used 50 documents of scientific papers in the fields of Informatics, Religions, Biology, Law and Health written in Indonesian

language. The testing results of the documents are shown in the following table 3:

Table 3. Accuracy Testing on the System

No	Document Tested	Total of words stemmed	Number of standard word	Number of non-standard word	Total of correct non-standard words manually	Total of incorrect non-standard words manually	Incorrect non-standard words manually
1	KP_1300018045	746	726	20	17	3	Aga Terlalu perhati
2	KP_1400018199	562	546	16	14	2	terkadang rinci
3	KP_1400018232	1033	982	51	48	3	Peluang Pegawai perhati
4	KP_1500018068	345	326	19	18	1	mitigasi
5	KP_1500018156	278	258	20	20	0	-
6	KP_1600018027	442	428	14	13	1	sekali
7	KP_1600018100	499	484	15	15	0	-
8	KP_1600018119	442	425	17	17	0	-
9	KP_1600018220	558	528	30	30	1	perhati
10	KP_1700018190	572	557	15	15	0	-
11	Islam dan demokrasi	414	403	11	9	2	Sesuatu kalang
12	Jual beli online	354	328	26	24	2	Sesuatu sekali
13	Kekerasan	409	398	11	11	0	-
14	Koeksistensi damai	621	582	39	39	0	-
15	Kontroversi hadis	608	577	31	29	2	Sekali Ketimbang
16	Makna sukses	745	703	42	41	1	Karakter
17	Pendidikan agama	660	632	28	27	1	Karakter narkoba
18	Pendidikan moral	461	446	15	14	1	narkoba
19	Konsep literasi	490	458	32	30	2	Internet literasi
20	Tiga jalan islam politik	669	630	39	37	2	Tebal karakter
21	Distribusi dan pola	671	645	26	25	1	sitasi
22	Pemanfaatan media	619	604	15	14	1	terlalu
23	Pengembangan bahan	993	962	31	31	0	-
24	Multimedia	822	791	31	31	0	-
25	Pengobatan tradisional	420	405	15	15	0	-
26	Pentingnya asesmen	296	288	8	8	0	-
27	Pertumbuhan	693	664	29	29	0	-
28	Pertumbuhan ikan mas	490	455	35	34	1	dinas
29	Profil kemampuan	661	619	42	41	1	karakter
30	Validitas perangkat	321	306	15	14	1	karakter
31	Gagasan struktur	388	361	27	26	1	sitasi
32	Harmonisasi hukum	477	459	18	17	1	internet
33	Kebijakanhukum	520	490	30	28	2	Kokoh narkoba
34	Kepastianhukum	662	633	29	28	1	holistik
35	Membangunhukum	354	341	13	12	1	karakter
36	Penerapan deversi	391	361	30	30	0	-
37	Peranan bahasa	491	475	16	16	0	-
38	Perlindungan hukum	290	281	9	9	0	-
39	Problemtika hukum	516	483	33	32	1	kalang
40	Zakat profesi	619	591	28	26	2	Familiar literatur
41	Efektifitas penyuluhan	428	405	23	22	1	internet
42	Penyuluhan peergroup	838	795	43	42	1	genetik
43	Elfika-faktor	630	603	27	26	1	genetik

44	Faktor risiko diabetes	696	657	39	38	1	genetik
45	Faktorygberhubungan	614	582	32	32	0	-
46	Ketersediaan	465	430	35	35	0	-
47	Kepemimpinan	495	470	25	25	0	-
48	Pengaruh pendidikan	1078	1013	65	65	0	-
49	Pengaruh self help	604	577	27	27	0	-
50	Terapirelaksasi	573	539	34	33	1	Sampling
Total		28023	26702	1321	1278	43	

The formula to determine the percentage of the system accuracy uses the math calculation as following [10]:
Number of tested word = 1321 (Number of non-standard word)
Number of correct word = 1278 (Total of correct non-standard word manually)
Number of incorrect word = 43 (Total of incorrect non-standard word manually)

$$\text{Percentage of Correct Word} = \frac{number\ of\ correct\ words}{number\ of\ tested\ words}\ X\ 100\%$$

$$\text{Percentage of Correct Word} = \frac{1278}{1321}\ X\ 100\% = 96.74\%$$

$$\text{Percentega of Incorrect Word} = \frac{number\ of\ correct\ words}{number\ of\ incorrect\ words}\ X\ 100\%$$

$$\text{Percentage of Incorrect Word} = \frac{43}{1321}\ X\ 100\% = 3.26\%$$

In the accuracy testing by using the calculation formula above, it can be concluded that the level of accuracy is 96.74% and the level of inaccuracy (errors) is 3.26%.

Table 4. Incorrect Non-Standard Words Grouping List

No	Typos	Stemming errors	the word not found in the database
1	Aga (1x)	Terkadang (1x)	Terlalu (1x)
2	Perhati (3x)		Rinci (1x)
3			Peluang (1x)
4			Pegawai (1x)
5			Mitigasi (1x)
6			Sekali (3x)
7			Sesuatu (2x)
8			Kalang (1x)
9			Ketimbang (1x)
10			Karakter (6x)
11			Narkoba (3x)
12			Internet (3x)
13			Literasi (1x)
14			Tebal (1x)
15			Sitasi (1x)
16			Dinas (1x)
17			Internet (2x)
18			Kokoh (1x)
19			Holistik (1x)
20			Familiar (1x)
21			Literatur (1x)
22			Genetik (3x)
23			Sampling (1x)
Total	4x	1x	38x

Based on 43 incorrect non-standard words that were found manually and after checking the website https://kbbi.kemdikbud.go.id/, those 43 words were included as standard words. This is because these words are not contained in the standard word database which is used as a system reference in displaying non-standard words. The 43 words found incorrectly were grouped into 3 categories, namely: stemming errors, typos and words not found in the standard word database. Table 2 states incorrect non-standard word grouping list.

Percentage of errors in writing words =

$$\text{Percentage of errors because of typos} = \frac{number\ of\ typos}{number\ of\ incorrect\ words}\ X\ 100\%$$

$$\text{Percentage of errors in writing words} = \frac{4}{43}\ X\ 100\% = 9.3\%$$

$$\text{Percentage of errors in stemming errors} = \frac{number\ of\ errors\ in\ stemming}{number\ of\ incorrect\ words}\ X\ 100\%$$

$$\text{Percentage of errors in stemming} = \frac{1}{43}\ X\ 100\% = 2.3\%$$

$$\text{Percentage of errors because the word not found in the database} = \frac{number\ of\ \text{the word not found in the database}}{number\ of\ incorrect\ words}\ X\ 100\%$$

$$\text{Percentage of errors because the word not found in the database} = \frac{38}{43}\ X\ 100\% = 88.3\%$$

Figure 3 shows incorrect non-standard words grouping list and the biggest errors is because the word not found in the database, which means that the number of standard vocabulary words in the database is still limited.

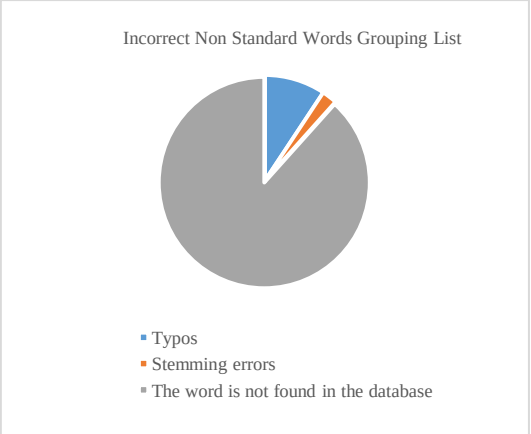

Figure 3. Group of Incorrect Non Standard Words

5. CONCLUSION

5.1 Conclusion

1. This research has developed a search application of non-standard words in the documents written in Indonesian language.
2. From the accuracy testing, the success percentage of Nazief & Adriani method for stemming and spell check for searching the standard and non-standard words is 96.74% of correct result from 50 documents written in Indonesian language. The system has not reached 100% accurate since the system only read Indonesian words.

5.2 Recommendation

Several recommendations for further development of the system are as following:

1. The amount of vocabulary in the database has not been complete, so it can be completed to make it more accurate.
2. The system can be developed by enhancing it with another spell checker to increase the accuracy.
3. Application of this search system for non-standard words is still in the localhost format and can be developed in the mobile format for easier access in the use.
4. The system can be enhanced to increase the speed of the process.

6. REFERENCES

[1] Asian, J., Williams, H. E., & Tahaghoghi, S. M. M. (2007). Stemming Indonesian A confix-stripping Approach. Conferences in Research and Practice in Information Technology Series, 38(January), 307–314. https://doi.org/10.1145/1316457.1316459

[2] Bhaire, V. V., Jadhav, A. A., Pashte, P. A., & P.G, M. M. (2015). Spell check. Nursing Times, 5(4), 38–40. https://doi.org/10.12968/sece.2007.5.260

[3] Chaer, Abdul. (2011). KesantunanBerbahasa. Jakarta: RinekaCipta.

[4] Dini Nopiyanti, K. A. S. (2014). Aplikasi pencarian kata dasar dokumen berbahasa indonesia dengan metode stemming porter menggunakan php & mysql. Kommit, 8(Kommit), 215–222.

[5] Herwijayanti, B., Ratnawati, D. E., & Muflikhah, L. (2018). Klasifikasi berita online denganmenggunakan pembobotan TF-IDF dan cosine similarity. Jurnal PengembanganTeknologi Informasi Dan Ilmu Komputer, 2(1), 306–312.

[6] Joseph, S. R., Hloman, H., Letsholo, K., & Sedimo, K. (2016). Natural Language Processing: A Review. International Journal of Research in Engineering and Applied Sciences, 6(3), 1–8.

[7] Juang, D. (2016). Analisis spam dengan menggunakan naïve bayes. Jurnal Teknovasi, 03(1998), 51–57.

[8] Kosasih, E. dan Hermawan, Wawan. (2012). Bahasa Indonesia Berbasis Kepenulisan Karya Ilmiah dan Jurnal. Bandung: Thursina.

[9] Oeyliawan, R. F., & Gunawan, D. (2017). Aplikasi rekomendasi buku pada katalogperpustakaan Universitas Multimedia Nusantara menggunakan vector spacemodel. ULTIMATICS, Vol. IX, 97–105.

[10] Riyanto, (2014). VALIDASI & VERIFIKASI METODE UJI. Yogyakarta. Deeppublish.

[11] Sulhan, M., & Kurniawan, R. (2014). Metode Stemming Sebagai Preprocessing Pada Filter Kata Porno Melalui Aspek Pendidikan. Seminar Nasional Teknologi Informasi Dan Komunikasi, 2014(Sentika), 52–60.

[12] Soyusiawaty, Dewi, Anna Hendri Soleliza Jones, and Nora Lestari Lestariw. 2020. "The Stemming Application on Affixed Javanese Words by Using Nazief and Adriani Algorithm." IOP Conference Series: Materials Science and Engineering 771(1).

[13] Wibowo, J. (2016). Aplikasi Penentuan Kata Dasar Berimbuhan Pada Kalimat Bahasa Indonesia Dengan Algoritma Stemming. Jurnal Riset Komputer (JURIKOM), 3(5), 346–350.