Prototype Analysis of Basic Words in Social Media in Indonesia

Wahyu Wijaya Widiyanto, Sofyan Pariyasto and Dwi Iskandar

1Polytechnic Indonusa Surakarta, Street K.H Samanhudi No.31, Mangkuyudan, Central Java, Indonesia.
2PT. Durio Indigo, Yogyakarta, Indonesia.

ABSTRACT

Aims: The use of social media in generation Y and generation Z as a forum for communication is commonly used in everyday life. The use of language as a means of communication in social media also attracts attention, especially the use of Indonesian, as a national language and as a unifying language of the nation, Indonesian must of course be maintained. Especially in writing and using it. The frequency of writing words on social media often becomes a problem in Indonesian as a national language. Both consciously and unconsciously, Indonesian writing is often found that is not in accordance with the structure of grammar and vocabulary in social media, causing concerns about the use of Indonesian in the future that has strayed far from the general guidelines of Indonesian spelling. Researchers conduct research related to the use of basic words in social media.

Study Design: The analysis process is carried out using an approach that starts from data collection, stemming and grouping data from random social media (WhatsApp, Instagram and Facebook).

*Corresponding author: E-mail: wahyuwijaya@poltekindonusa.ac.id;
1. INTRODUCTION

The use of information technology is increasing rapidly. Likewise, the use of social media as a means of communication is used in everyday life. The language used as an expression of ideas or the delivery of information on social media is very free to use without regard and follow the rules that should be. The Ministry of Communication and Information [1] revealed that internet users in Indonesia currently reach 63 million people. Of this number, 95 percent use the internet to access social networks [1]. According to research by We Are Social, a British media company that works with Hootsuite, the average Indonesian spends three hours 23 minutes a day to access social media [2]. From a report entitled "Essential Insights into Internet, Social Media, Mobile and E-Commerce Use around the World" published in 2019 shown in Fig. 1.

Of Indonesia's total population of 265.4 million, active users of social media reach 130 million with 49 percent penetration. As many as 120 million Indonesians use mobile devices, such as smartphones or tablets to access social media, with 45 percent penetration. Within a week, online activity on social media via smartphones reached 37 percent. Based on the most downloaded applications, social media companies under Mark Zuckerberg dominate in the top three. In order from the first position are WhatsApp, Facebook, Instagram and only followed by social media made in South Korea, Line. Based on the average site traffic per month, Facebook is the most visited social media with more than 1 billion million monthly visitors. The average Facebook visitor spends 12 minutes 27 seconds to access the social network. 92 percent access Facebook via mobile with a percentage ratio based on gender as much as 44 percent for women and 56 percent are male users. Facebook users are dominated by the age group of 19-24 years with a percentage of 20.4 percent are women and 24.2 percent are men. While total monthly active Instagram users in Indonesia reach 53 million with a percentage of 49 percent women and 51 percent are men.

Total users reach 150 million users, this means the majority of the use of the internet to socialize through social media. The number of social media users reaches 56% of the total population of Indonesia, with mobile-based users reaching 130 million. No wonder all social media platforms finally focus on optimizing their applications on mobile [3]. The top 5 (five) favorite social media sites are YouTube, WhatsApp, Facebook, Instagram and Line [4].

Based on the data above there are problems in posting status or commenting in Indonesian, various errors in the use of basic words that are not in accordance with the proper grammatical and vocabulary structure. For example the symptoms of hypercorrect in the use of basic words that come from words uptake. Absorption words that are found in social media can be in the form of absorption words that come from regional languages or absorption words that come from foreign languages. As a result of this hypercorrect symptoms make the basic words become the wrong writing so that the resulting basic words are not standard [5].
The use of basic words in social media can be classified into two parts [4]. The first part is in accordance with the General Guidelines for Indonesian Spelling, while the second classification is that which is not in accordance with the General Guidelines for Indonesian Spelling and which is not in accordance with Indonesian writing conventions. Far from the historical side of Indonesian language has been through various long struggles before finally being designated as the National Language. Indonesian is the National Language that is used as a unifying tool for the nation. Indonesian was established as the National Language at the 1975 national language political seminar and later confirmed at the national language political seminar precisely in 1999 [6].

Humans as social creatures of course in everyday life always use language as a means of interaction between people. Language is an inseparable part of a culture. Through language can be assessed culture, character and mindset of a society. Through language can be assessed mindset as well as a reflection of one's way of thinking [7].

In social media, of course, it can indirectly draw conclusions about a person's character based on the language he uses. The use of good and right language and politeness reflects the character of someone who has intellectual intelligence. Conversely the use of Indonesian language that is not according to the rules reflects how high a person's intellectual intelligence even though it is not [8].

2. MATERIALS AND METHODS

Error in the use of Indonesian is something that cannot be avoided either intentionally or unintentionally. Writing errors can be found in writing and in pronunciation. Some errors in the use of Indonesian include [9]:

2.1 Word Formation Error

Error formation of words in a sentence is as follows:

\[ \text{Toni sedang menyuci baju.} \]

Word writing menyuci must be replaced so that the correct sentence is as follows:

\[ \text{Toni sedang mencuci baju.} \]

2.2 Missing Error Because Hypercorrect

Pronounced errors in writing words can be caused by the influence of local languages and foreign languages [10]. Examples of word writing errors faham, pasal, fikir, fihak transformed into paham, pasal, pikir, pihak.
2.3 Error Analysis

Mistakes in the use of basic words can be grouped into three parts:

2.3.1 Error in writing prepositions

Mistakes in writing prepositions can be found in writing words that state the place. The example of writing the preposition that is not appropriate is as follows:

*Saya belajar dirumah bersama Tono dan Badu.*

Preposition writing "di" in the example sentence above it is not in accordance with the general guidelines of Indonesian spelling because of the word "di" considered as one word. So that the correct sentence writing is: *Saya belajar di rumah bersama Tono dan Badu.*

2.3.2 Hypercorrect

The word hypercorrect, which can be interpreted too correctly, is the writing of a word that goes beyond the correct and appropriate boundary so that it becomes wrong. The example sentences containing the word hypercorrect are as follows:

*Jaka sudah mendapatkan ijin cuti dari perusahaan.*

Error writing hypercorrect words found in words "ijin". Word ijin comes from Arabic, where in Arabic letters "j" in the word permit is written in letters "dzal" (ژ). So writing the right sentence is as follows [6].

*Jaka sudah mendapatkan izin cuti dari perusahaan.*

2.3.3 Error writing words (Typo)

Mistakes in writing words (typo) that is intended is an error in the process of writing words that are usually not realized. This form of error can be in the form of excess letters in writing, lack of letters, and wrong letters in writing words [11]. Examples of writing the wrong word are as follows:

Word writing "rumaha" which should "rumah".

Word writing "Septu" which should "sepatu".

Word writing "pakian" which should "pakaian".

3. RESULTS AND DISCUSSION

To be able to find out the use of words that are not in accordance with the correct Indonesian language rules and in accordance with the General Guidelines for Indonesian Spelling in social media, this research process is carried out. The use of Indonesian in social media used in communicating on social media can be grouped into three parts namely [10]:

i) The words used are usually irregular and cannot be formulated with certainty.

ii) The word used tends to be abbreviated.

iii) Using particles in writing.

With research related to word writing errors in social media it can be educated about writing good and right words.

3.1 Research Flow

The research process was carried out with several stages as shown in Fig. 2.

The research flow is broken down into seven parts while each explanation from each section is as follows:

3.1.1 Determination of keywords

Keywords are a group of words that are combined into one that represents the main meaning that the author will convey in an article. Keywords are the main words that will be used to gather important parts of an article that have a relationship with keywords [12]. One key to success in finding data is of course having to use the appropriate keywords. The process of determining keywords will certainly be a problem if you cannot map precisely what kind of data will be searched. There are ten methods that can be used in the process of determining keywords (keywords) in finding appropriate data, among others, as follows:

i) Using synonyms, words that have similar or similar meanings to keywords.

ii) Using acronyms, which are abbreviated words that can be pronounced. Example: *Cipularang* (*Cikampek, Purwokerto, Padalarang*).

iii) Use abbreviations, which are commonly abbreviated results. Example: *MPR (Majelis Permusyawaratan Rakyat).*

iv) Using other languages, what is meant by other languages is what has the same
meaning as keywords. Example: student words with the word student.

v) Using broad terms, it means broad terms related to keywords. Example: Bahasa Indonesia merupakan istilah luas dari pragmatik, sintaksis, dan semantik.

vi) Using scientific language (Latin). The use of scientific language in the process of searching data can be done to get more accurate results.

vii) Using plural or singular forms. Searching in English certainly has to pay attention to the words used both plural and singular. Example: Car dan cars, people dan peoples.

viii) Use spelling variations. In determining the keywords will also produce more accurate results.

The initial stage of the research process is related to the analysis of errors in the use of basic words, namely the determination of keywords to be used. Keywords will be a reference in the process of gathering articles in social media. In this study keywords were used "belajar", so that every article in social media that contains the word learning will automatically be stored in a dataset.

3.1.2 Data collection

Data collection is one of the stages that cannot be separated from the research process [13]. Data collection was carried out to obtain data in accordance with the needs of the study [14,15].

The process of collecting data is done by using bots that have been set up for search keywords. The data collected is limited to as many as 3000 articles containing words "belajar" taken from social media. This is done so that the inspection process can be more leverage.

3.1.3 Stemming

Stemming is one of the stages in text data processing. In the process of stemming data text is broken down and returned to form the basic words. This process is carried out by removing suffixes in the form of beginning, insert, and suffix [14,16].

Stemming is the process of breaking an article into basic words. From each article that has been collected in the previous stage will then be solved per word. After each word is broken down, each base word will be taken from each word. The process of taking a base word is done by removing the prefix and suffix that is in each word, then turning it into an original root word.

Stemming (or perhaps more precisely lemmatization?) Is the process of turning an affixed word into a basic word. Language rules are applied to remove the affixes [17].

Example:

membetulkan -> betul
berpegangan -> pegang

There are many problems encountered in the Indonesian stemming process, including:

The addiction to Indonesian is quite complex, consisting of:

i) Prefix, affix before the word: ber-tiga

ii) Suffix, an affix at the end of the word: makan-an

iii) Conflict, affixes before and at the end of the word: per-ubah-an

iv) Inf, affix in the middle of a word: kemilau

v) Affixes from foreign languages: final-isasi, sosial-isasi

vi) Rules for changing prefixes, such as (me-) to (meng-, mem-, men-, meny-)

Word-Sense Ambiguity, i.e. one word can have two meanings (such as homonyms), and come from different basic words. Example:

i) Berikan -> Ber-ikan

ii) Berikan -> Beri-kan

Overstemming: The word give is based on the rules of beheading, it can be beheaded into Ber-i-kan. Become a basic word i. To prevent overstemming, the algorithm requires a basic word list. If the decapitated word is in the root word, stop the decapitation process.

Understanding: Mengecek -> become meng-cek, should meng-cek. This can be due to the basic word dictionary, ecek is also a basic word.

Dependence on dictionaries/basic word lists

To prevent overstemming, the algorithm becomes dependent on the root word. The existence of deficiencies or strengths in the basic words can cause overstemming or understemming.
Indonesian users are not consistent in determining stem (manually).

We, humans, also sometimes differ in opinion in determining the stem of a word. For example, is "adalah" an influence word from the word "ada"? Is "bagian" a word that affects the root word "bagi"?

i) Plurals = buku-buku (basic words of buku)
ii) Absorption words from foreign languages = mengakomodir -> meng-akomodir
iii) Writing error = penambahan, it must be an penambahahan, so it cannot be systemized.
iv) Acronym = pemilu can not stem becomes to pe-milu
v) Proper Noun (Item Names), for example people's names, city names: Abdullah can not stem becomes to abdul, it should not be stem.

Sastrawi Stemmer: As written in Readme, Sastrawi stemmer is a simple stemmer library that is designed to be used easily. Even so, Sastrawi stemmer tries to solve the above problems.

There are several Indonesian Stemming algorithms, all of which face similar challenges. Literature stemmer applies an algorithm based on Nazief and Adriani, then enhanced by the CS Algorithm (Confix Stripping), then enhanced again by the ECS algorithm (Enhanced Confix Stripping), then enhanced again by Modified ECS.

By using these algorithms, many stemming problems have been successfully overcome:

i) Prevent overstemming with the basic word dictionary
ii) Prevent understemming with additional rules
iii) Succeeding plural form words: Buku-buku -> buku

3.1.4 Basic word analysis most often appears

The process of stemming porters for Indonesian is as follows [17]:

i) Removing particles
ii) Removing pronouns (Possessive Pronoun)
iii) Delete the first prefix
iv) Removes the second prefix
v) Remove the first suffix
vi) Remove the second suffix

3.2 Evaluation and Error Analysis

The research process was carried out using 5243 data taken from social media using the keyword "belajar". The data that has been collected is stored in a CSV file which is then stored in a text file for processing, for the following process code:
$token = $this->multiexplode(array(" ",","n"),$casefolding);

/* Punctuation Removal and Stopword Removal */
$ps = $this->punctuationRemovalandstopwardRemoval($token);

/* Stemming */
$hasil = $this->stem($ps);
return implode(" ",$hasil);
}

function multiexplode ($delimiters,$string) {
    $ready = str_replace($delimiters, $delimiters[0], $string);
    $launch = explode($delimiters[0], $ready);
    return $launch;
}

function punctuationRemovalandstopwardRemoval($string){
    include "stopword.php";
    foreach ($string as $key => $value) {
        $delimiter = array(' ','.',',','"','\n','\r','\t',':','\','"','@','#','$','%','&','\(','\)'\?'\'>','<','[','\'],'\','\','\','=','\n','\r','\t','\n');
        $value = str_replace($delimiter,"\n","\n"$value);
        if((ctype_alpha($value) || ctype_alnum($value)) && !empty($value)){
            if(!in_array($value, $stopword)){
                $data[] = $value;
            }
        }
    }
    return $data;
}

function stem($string)
{
    $stemmerFactory = new \Sastrawi\Stemmer\StemmerFactory();
    $stemmer = $stemmerFactory->createStemmer();
    foreach ($string as $key => $value) {
        $a = $stemmer->stem($value);
        if(empty($a)){
            $data[] = $value;
        }else{
            $data[] = $a;
        }
    }
    return $data;
}
?>

The display file that has been tidied text looks like the Fig. 3 random data collection seen when data retrieval on social media, Fig. 4 stemming results after sprucing up:
The tidied text data is then converted into its basic word form (stemming) by removing the affixes attached to the word. Text data that has become a basic word looks like the following Fig. 4.

Stemming text data is data that will be processed for word analysis in accordance with the general Indonesian spelling guidelines and incorrect words. At this stage also obtained information that there are 32092 words.
3.2.1 Basic word analysis most used

From the process of stemming obtained information that there are 32092 words. However, from this data there are still words that are repeated, so it is necessary to do the process of deleting duplicate data and counting the words that appear most often until they appear the least. Look at Table 1 to see the most used words:

From Table 1, it can be seen that said "ajar" is the most commonly used word, followed by the word "yg" and word "orang". From Table 1, it can also be described the appearance of the word graph as shown in Fig. 5, for vertical lines is the amount of data and horizontal lines are words that often appear.

3.2.2 Analysis of the wrong basic words

Basic words that are not in accordance with the general guidelines of Indonesian spelling and are not in accordance with the grammatical structure and vocabulary structure are the basic words that are considered wrong. From the results of the research, the most frequently used basic words look like the following Table 2.

From Table 2 it can be concluded that the word "yg" is the word that appears most often and the possibility of correction for the word "yg" is a word "ya" or "yang" with 559 appearances.

### Table 2. Wrong basic words that often appear

<table>
<thead>
<tr>
<th>Wrong word</th>
<th>The number</th>
<th>Should</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yg</td>
<td>559</td>
<td>y, yu, ye, g, ya</td>
</tr>
<tr>
<td>Ga</td>
<td>230</td>
<td>a, ya, ka, gai, ha, oga, gah, iga, ua, za, gas, gar, ca, da, g, ia, gap, sa, na, fa, gaz, ba, gam, wa, gua, gu</td>
</tr>
<tr>
<td>Smoga</td>
<td>147</td>
<td>moga, sog</td>
</tr>
<tr>
<td>Arteria</td>
<td>90</td>
<td>arteri</td>
</tr>
<tr>
<td>la</td>
<td>74</td>
<td>a, lam, lab, fa, lat, na, vla, ya, wa, l, ia, sa, lak, da, ala, lah, lap, ca, ba, ka, lo, las, ua, ela, lai, za, ha nanar, nanas, nyana, mana, naka, lana, kana, pana, fana, nada, bana, nara, rana, nanap, nama, nang, nanah, wana, ana, sana nawa, naga, nala, dana, hana, nan, gana</td>
</tr>
<tr>
<td>nana</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>gak</td>
<td>67</td>
<td>gar, gauk, gas, gam, pak, gaz, uak, mak, wak, gap, vak, rak, agak, bak, cak, gaek, hak, gaok, gah, gai, sak, tak, nak, lak, goak, kak</td>
</tr>
<tr>
<td>pake</td>
<td>64</td>
<td>pakem, pak, pace, pike, pale, pak, pakde, paket, pare, sake</td>
</tr>
<tr>
<td>nonton</td>
<td>56</td>
<td>ponton, tonton</td>
</tr>
<tr>
<td>faham</td>
<td>55</td>
<td>paham, daham, waham, saham, baham, gaham</td>
</tr>
</tbody>
</table>

![Fig. 6. Nonstandard words that often appear](image)
From Table 2 can also graph the appearance of the most frequent words as shown in Fig. 6 below, for vertical lines is the amount of data and horizontal lines are words that often appear.

4. CONCLUSION

From the results of the research conducted it can be concluded that:

i) The use of words that are not standard and not in accordance with the general guidelines of Indonesian spelling are words that are dominated by abbreviated words.

ii) Abbreviations like "yg" should "yang" rank first, while the word "ga" who should "tidak" rank second.

iii) The writing of words caused by hypercorrection is also one of the causes of errors in writing words.

iv) Hypercorrection words that appear include "zenius", "cuba", and "smoga".

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

https://kominfo.go.id/index.php/content/detail/3415/Kominfo+%3A+User+Intern+et+in+Indonesia+%3A+63+Million+People+0/Berita_satker
digital-2019-media-social/
4. YES YES. Social Media and & Social Network (Differences and classification); 2018.
16. Magriyanti AA. Analysis of the development of porter stemming algorithms in Indonesian, Attrib. 4.0 Int.; 2018.

© 2020 Widiyanto et al.; This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sdiarticle4.com/review-history/54649