A Conceptual Paper of Building A Modern Standard Corpus for A Minority Language (Indonesia) from The Web

Teamsar M Panggabean, Albert K Hutapea, Fujiatma P Napitupulu, Dwide E Sembiring
School of Informatics, Del Institute of Technology, Sitoluama-Laguboti, Indonesia

Abstract
This is a conceptual paper for designing Indonesian political corpus. Some new approaches have been implemented in particular circumstances, evaluated and resulted in better sentiment analysis output. This study will review and study the process of making an Indonesian corpus that focuses on politics. As playing an essential role in natural language processing research, building corpora is mandatory. There is an increase in language-engineering related research using Indonesia corpus. In fact, no major development for this minority language has been shown. Recent developments done by some researchers are focusing on various topics (but not politics) and the data are limited. The necessity of a relevant corpus is the key for any objective research.

In this paper, we present the results of experiments in building a modern standard corpus for a minority language (Indonesia) by employing data available on the world wide web (WWW). To narrow our research area down, we build a corpus for sentiment analysis for the next Indonesia's presidential election that will have happened in 2019. Therefore, we limit the data crawled only on political news. Considering the amount of data availability, several trusted Indonesia's online mass media will have been selected as the samples.

There are three contributions to this paper, first we handle Indonesian misspellings in which we employ EDB normalizer, stands for Edit Distance and Bigram; second, we build an Indonesian annotated slang-words dictionary, and last, we address exaggerated words shortening problem that frequently occurred in comments by a simple algorithm.

Keywords: Corpus, Minority Language, News Extraction, News Polarity Classification

1 Introduction

Politics has an important role in a country and government. Politics has a close relationship with all aspects of community life and is applied to everything, in accordance with its omnipresent nature, politics is everywhere around social life. Consciously or not, like it or not, politics has also influenced our lives, both as individuals and as members of society and therefore nothing special, as long as ethics, economics, and society remain united and not manifested in structural differentiation. Politics are dynamic and forcing people to contribute directly or indirect, for instance posting an argument in every policy proposed by the government. Driven by increasingly rapid technological advances, the spread of information amongst people's lives is also getting easier and faster. The Indonesian President, Joko Widodo, has eagerly encouraged people to use technology to accelerate productivity. As policy makers, the president regularly publishes information via Twitter, Facebook and online mass media.

Technological developments characterized by the easiness and fast dissemination of information provide both positive and negative impacts to the community. Information that have been posted often contain useful knowledge. There are a plenty of technologies to support public in expressing their feelings, and one of the most preferable media is Twitter. Several studies have been conducted using Twitter to gather the informations, including applications for detecting, tracking, and visualizing real-time events developed by (McMinn et al., 2014) where millions of tweets were collected. An earthquake reporting system developed by (Sakaki et al., 2013) also uses Twitter as a sensor for its application. He claimed that by using Twitter, it was likely that the earthquake had been detected. Twitter was also successfully used for political forecasting, as demonstrated by (Tumasjan et al., 2010). Previous work on tweets polarity classification has been done by (Nakov et al., 2013; Hu et al., 2013; Kouloumpis et al., 2011; Agarwal et al., 2011; Pak and Paroubek, 2010)

The use of technology to broadcast information not only from social media, but also online mass media. It can be an option to spread the information, whether it is an expression or knowledge. In Indonesia, there are various online mass media that broadcast information quickly, and some of them are trusted social media. They not only provide reliable information, but also have access to post their ideas using a commentary dialogue. This allows them to convey additional information or express their opinions.

Comment also has positive and negative polarity, in which it can be a media to change the point of view of a reader regarding a public figure, for instance. Comments may contain some debatable ideas. Often, a comment does not clearly state its intentions, giving rise to multiple meanings (ambiguity), other than that opinions or responses conveyed by a reader can be different from other readers so that it will be difficult to get a conclusion or assess the information that has been submitted to the online mass media.

It is essential to identify and analyse expressions from the public. Based on these needs, in recent years many researchers have studied the problem, which is then called sentiment analysis. Sentiment analysis is a method used to analyze opinions, sentiments, evaluations, attitudes, and emotions towards an entity such as products, services, organizations, individuals, problems, events, topics, and attributes.

Nowadays, there are many aspects that utilize sentiment analysis, such as education, hoax analysis and people’s satisfaction of the government services. Sentiment analysis in education can be used to find out feedback from students from lessons that have been given or the curriculum applied in the educational environment, by using sentiment analysis the conclusions can be drawn by which

will be then used to make a decision that is appropriate to the student's needs. This was previously done by Nabela Altrabheh, Mohamed Medhat Gaber, and Mihaela Cocea under the title Sentiment Analysis for Education.

In the field of government, sentiment analysis is used to discover people's views on the current government system. In governments around the world, social media is often used to get closer to the community. These things can provide insight to the people's desires about the government. Therefore, the government is currently trying to implement the citizen-centric model. That is a system of government in which all the priorities and services will be driven according to the needs of the community rather than the ability of the government (Arunachalam and Sarkar, 2013). Community needs can obviously be obtained through aspirations conveyed to social media and the community comments given to a particular problem. Through this, the government will more easily find out the needs of the community through the sentiments given to these media.

Sentiment analysis can also be used to judge whether a news is a hoax or a fact. In Indonesia, a hoax news was circulated and distributed by an Indonesian artist who had been involved in the theater scene and was also an activist in social organizations. In the midst of the warm political problems in Indonesia ahead of the biggest political feast, the general election of the President and Vice President in 2019. The dissemination of hoax news is often associated with the political problems that are happening. The election that held will be followed by two pairs of candidates namely Joko Widodo and Ma'aruf Amin and Prabowo Subianto and Sandiaga Uno. At present, the political parties become the biggest and most talked in both social media and mass media. In various online mass media that discuss these issues, the community also participates in expressing their opinions. They voice their response to all political problems which then become closely related to the 2019 election.

This problem becomes our background to conduct this research, that is to build an Indonesian corpus focused on politics. According to (Baker 2010: 93) corpus is a collection of texts both oral and oral written on a computer. Baker defines the corpus in electronic media only. In addition, there are three aspects that are considered in understanding the concept (Baker, 1995). First, the corpus is basically a collection of text that is produced electronically and can be analyzed automatically or semi-automatically. Second, the corpus does not only contain a collection of written texts, but also includes utterances. Third, the corpus also includes texts in numbers that come from and come from a variety of sources, for example from various writers and speakers and on various topics. From Baker's opinion it can also be stated that there are four criteria in understanding the corpus in broad concepts, namely form, size, representative, and open-close. The significance of corpus has been increasing in the natural language processing community, specifically for opinions analysis.

We have looked for some information about political sentiment analysis in Indonesia that use online mass media to build the corpus, but no previous work(s) found. Moreover, Indonesian online dictionary that we hopefully help us much on this research is in fact very limited while scrapping the data (approx. 500 words per day). We also looked around (Le, T.A et al, 2016) research but we found some mistakenly annotated words polarity, such as: "*korupsi*" (corruption) annotated as *"positive"* when the positive score is around 0.75 while negative is just 0.125, other examples are "*sogokan*" (bribing), "*kekejian*" (atrocity), and "*membunuh*" (killing) which have positive scores 0.625, 0.625 and 0.75 respectively. We do use incomplete KBBI circulated online to help us out from data deficiency.

To achieve our goal in building political corpus, we perform three steps, first off, we select all political news, including the comments posted on each topic, scrapped from online mass media (detik.com, kompas.com, and cnnindonesia.com). We do manual annotations to the news not to the public expressions, whether politics or general information, by performing Cohen Kappa analysis which reaches the value 0.81. The rationale behind this is that detik.com and kompas.com do not specifically categorize the news. Moreover, comments rarely contain any specific information regarding politics.

The second stage is to determine the polarity of the filtered comments. Using machine learning methodology is believed not a good approach, since the data are not labeled, and asking people to do manual labelling is time-consuming and labor-intensive due to the massive scale and rapid growth of political comments. To address this problem, we construct a small set of sentiment lexicon with positive and negative subjectivity. We use *OpinionFinder* and *SentiWordNet* with *strong positive* and *strong negative* polarity. We also construct Indonesian slang words (around 4700 words) dictionary, and do manual polarity to each word, since the people frequently use slang words in their utterance and in expressing their feelings through online media, and in fact there is no Indonesia slang words publicly opened and ready-used annotated slang words. We ultimately will make this project as freely and as openly as possible for the next improvement. The latter task is to undertake sentiment analysis using comments.

Even though this research is partly conceptual, we contribute some works to this paper by applying some new algorithms to deal with any natural-processing-language related problems, such as building annotated Indonesian slang words, misspelling correction, and words without spaces.

2 Related Works

There have been extensive works on sentiment analysis. To extract sentiment automatically, there are two approaches can be performed. First, using lexicon-based approach (Turney 2002) that involves calculating orientation for a document from the semantic orientation of words or phrases in the document. Lexicon-based classification refers to classification of rules in which documents are assigned labels based on the count of words from lexicons associated with each label (Taboada et al. 2011). The lexicon-based classification has been widely used in the industrial and academic fields, indicated by studies that related to sentiment and opinion mining classifications (Pang and Lee 2008; Liu 2015) and the other researches that related to psychological analysis and ideological texts (Laver and Garry 2000; Tausczik and Pennebaker 2010). In Lexicon based technique a dictionary with annotated word will be used with the semantic word orientation, or polarity. Semantic orientation refers to the polarity and strength of words, phrases, or texts. The dictionary that will be used in the lexicon technique can be created manually (Stone et al. 1966; Tong 2001) or automatically by using seed words to expand the list of other words that will be used (Hatzivassiloglou and McKeown 1997; Turney 2002; Turney and Littman 2003). There is also another proposed methods for lexicon expansion, such as work on handling multi-word phenomena such as negation (Wilson, Wiebe, and Hoffmann 2005; Polanyi and Zaenen 2006), and discourse (Somasundaran, Wiebe, and Ruppenhofer 2008; Bhatia, Ji, and Eisenstein 2015). The last approach is a statistical or machine-learning methodology.

Tweet polarity classification of Indonesia telephone provider companies had been made by (Calvin Setiawan 2014). They focused to presents their study in automatically building a training corpus for the sentiment analysis on Indonesian tweets. In their research, they built the corpus automatically to perform two different tasks, which are extracting opinions derived from tweets and also classifying the polarity of tweets using various machine learning approaches. The experiment relies on a small set of domain-dependent opinionated words.

Research on opinion mining and analytical sentiments has also been widely undertaken, as explained in (Pang and Lee, 2008). Their study covers a number of techniques and approaches that allow searching system based on information can be done directly. Associated with sentiment analysis, in his research (Liu 2007) said that the most important indicators in sentiments are sentiment words or also called opinion words. These words are used to express positive sentiments or negative sentiments towards a particular object. The example is *good, great*, and *extraordinary* are examples of words with positive sentiment, while *bad*, *terrible* and *horrible* are words with negative sentiment. In addition, many people use idiom words when expressing their opinions such as *the cost of someone's arm and leg*. Therefore, even though sentiment words or sentences are the most important thing when doing analytical sentiments, it is not enough to only use them. There are several problems that arise if only use lexicon sentiments in sentiment analysis, including words with positive or negative sentiments that can have an orientation that is congested if used in a different domain. A sentence that contains sentiment words in it can also not express any sentiment, this often happens in several types of sentences such as question sentences. In addition, sentences that mean sarcasm are also difficult to assess their sentiments and other sentences that actually do not contain sentiment words but also imply opinions. This sentence is often used to convey some information that contains facts in it.

Even though we performed similar works, we have two different points. First, we focus on building a corpus which domain is political areas especially political environment in Indonesia. The data also were scrapped from online media mass and not from Twitter. Through online mass media, the people can freely express their opinions or comments on a particular topic. Unlike Twitter which limits its users to submit their comments. The sentiments posted online are more diverse and the number is more numerous. Second, we do political news extraction and sentiment analysis using comments. By doing this, we know the sentiment given by readers based on the polarity of each word.

3 Literature Review

There are several steps taken in building a political Indonesian corpus. In this section we will explain some of the theories that used in building Indonesian political corpus. Data collection is done by scraping 19329 news from detik.com, kompas.com, and CNN websites. In addition to the news, comments from each topic news are also collected and inserted into a database. The number of comments that were obtained was 559234 comments. This data will then be used in the development of the corpus.

3.1 Stemming

The collection of comments will go through a process of stemming or lemmatization. In Indonesian, stemming is done by removing the additions from a word, such as: prefixes, suffixes, inflections and words absorption.

In this process an existing stemmer is used, namely Sastrawi. Stemmer Sastrawi applies Nazief and Adriani algorithms (see on github.com/sastrawi), then it is enhanced by the CS (Confix Stripping) algorithm which will be then improved by the ECS (Enhanced Confix Stripping) algorithm then upgraded again with Modified ECS. It is inevitably finding irregular or slang words during stemming, in fact, many of them contain polarity. We cannot ignore those words as they will affect to the accuracy while performing sentiment analysis using comments. We address this issue by employing Indonesian annotated slang words dictionary. Other issues as follows:

- Complex rewards, for example, prefixes (suffix in front of words), suffix (additions at the end of words), confix (add in front and end of words), inflections (add in the middle of words), additions from foreign languages, for example: final- isasi and sosial-isasi, and changes in prefixes such as (*me-*) become (*meng-, mem-, men-, meny-).*
- Overstemming
- Understemming
- Words that are not standardized (slang words), for example, such as '*gak*' which means **no**, '*sy*' which means **I**, 'wkwkwkwk' which states that the author **laughs** and more.

3.2 POS-Tagging

After stemming process performed, we do Part of Speech Tagging (POS-Tagging). POS-Tagging is a process of automatically assigning a word class label to a word in sentence (Jurafsky and Martin, 2000). In this paper the authors undertake POS-Tagging by scrapping data on the KBBI website along with the tags of each word. We then map these data to the news and comments from online media. There are some building blocks when performing POS-Tagging, first not all the words in the news and comments exist in KBBI (due to limitation while scrapping words), such as: "*serbaguna*" (traduce), "*terintimidasi*" (intimidated), "*dihormati*" (respected), "*mempesona*" (mesmerized), "*mondar-mandir*" (moons), and more. Hence, the writer needs to do manual tagging to these words or insert a new word if unavailable. The second is that Indonesian slang words which have been addressed by mapping to Indonesian annotated slang words dictionary. The last one is that misspelling word that is handled by implementing EDB normaliser.

3.3 Words Polarity

In this study, we use sentiments lexicon that are widely used by other researchers, *OpinionFinder* and *SentiWordNet*. We select adjective words with highest objectivity score (above 0.7). The objectivity score can be calculated: $ObjScore = 1 - (PosScore + NegScore)$.

Table 1 : Polarity Rules

Rules	Polarity
Positive > Negative	Positive
Negative > Positive	Negative
Positive > Negative	Neutral

There are 107047 words that taken from *SentiWordNet*. 14730 have positive polarity, 13871 negative, and 78441 neutral. This is the examples of the words that have been taken before.

Table 2 : Polarity of Words (SentiWordNet)

Words	Positive	Negative	Polarity
Mampu (able)	0.125	0	Positive
Kekejian (abomination)	0	0.125	Negative

Pembenci (hater)	0	0.25	Negative
Dengan akurat (accurately)	0.25	0	Positive

The main purpose of giving this polarity is to calculate the overall polarity of a comment that has been collected.

4 Data Collection and Annotation

We collected political opinions from several trusted online mass media, such as: detik.com, kompas.com, and cnnindonesia.com, from July 2017 to August 2018. We do not consider a news that is not politic and has no comment. We successfully gathered approximately 20 thousand political news and with around 500 thousand comments (a news may have many comments). We classified news (whether politics or non-politics) manually. The agreement of two annotators by using Cohen Kappa method. Cohen kappa is used to assess inter-reability when observing qualitative variables / categories. Kappa was first created by Jacob Cohen in 1960 to create a more accurate measurement between two raters in determining decisions about how an object or unit is assigned to a particular category. And we reached the level of Kappa value 0.81, which is considered as a satisfactory agreement. Below is our general process to construct political corpus:

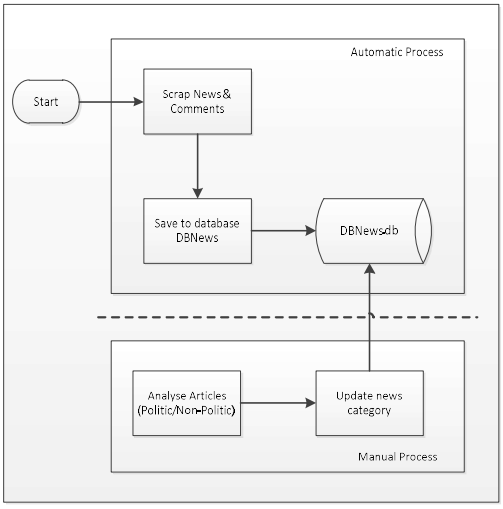

Figure 1. Political Corpus Construction Flow

4.1 Language issues in opinion text

Sentiment analysis refers to extracting subjectivity and polarity from text (potentially also speech) (Taboada, M. et. al, 2011). As our goal is to construct a corpus for political domain, building a complete collection of opinions from online mass media are challenging, since people frequently use informal words (grammatical mistakes, misspelling, and noisy such as words without spaces, slangs, abbreviations, and acronyms) and emoticons. Most of informal words and emoticons contain positive or negatives sentiments, and thus we count them as opinion words and need to be normalized.

Be able to identify the informal nature of the language in opinions and normalize OOV (out of scope vocabulary) terms would improve the accuracy of natural language processing techniques (Han et al., 2013). There are 5 categories of lexical variations (Imran et al., 2016 and Yadav et al., 2013):

1. Typos/misspellings: e.g. “maruk” should be “kemaruk” (in English “greedy”), “masalh” should be “masalah” (in English “problem”).
2. Single-word abbreviation/slangs: e.g. “sdkt” should be “sedikit” (in English “few”), “wkwkw” refers to “tersenyum” (in English “smiling”).
3. Multi-word abbreviation/slangs: e.g. “sksd” refers to “sok kenal sok dekat” (in English “act like you know me”), “EGP” refers to “emang gue pikirin” (in English “ignorant”).
4. Phonetics subtitutions: e.g. “se7” should be “setuju” (in English “agree”), “gtw” should be “ga tau” (in English “Do not know”).
5. Words without spaces: e.g. “kurangperhatian” should be “kurang perhatian” (in English “lack of attention”).
6. Exaggerated word shortening: e.g. “tidaakkk” should be “tidak” (in English “No”), “hancuuurrr” should be “hancur” (in English “mess up”).

4.2 Seed Words Annotation

As we intend to build Indonesian sentiment lexicon, we seed words as system input that have been labeled with positive and negative polarity to help us expand the large-scale annotated political corpus. We translate words from both *OpinionFinder* and SentiWordNet and remove any terms that do not have correspondences to Indonesian terminologies. We conduct two steps of manual evaluation to gain high precision, such as translation evaluation which performs words elimination that have no correlation to Indonesia terms, duplicate translations and mistranslated words. The last one is subjectivity evaluation in which we perform manual check whether translated words contains similar polarity with the English words. Below is our grand design of seed words annotation:

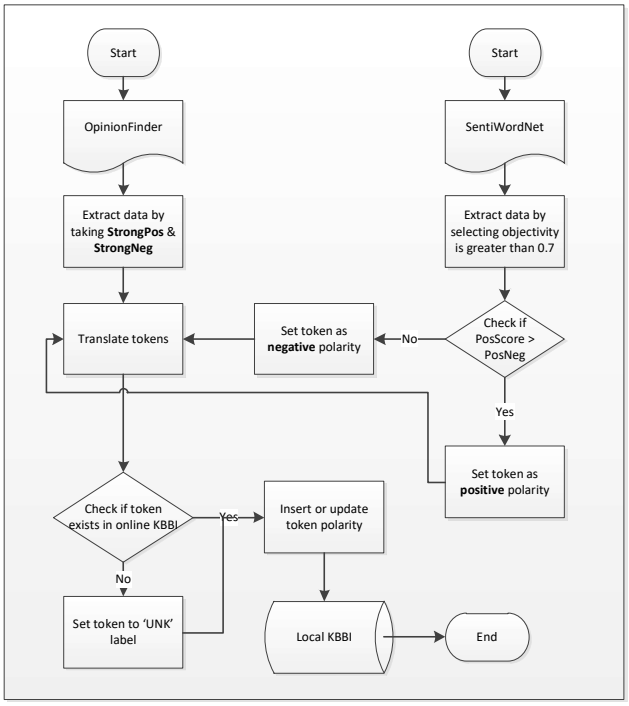

Figure 2. Seed Words Annotation Process

4.3 Sentiment Analysis Process

The obvious challenges that we found while undertaking sentiment analysis are vague and mixed. It is likely to have one or more expressions posted, and even negation of one opinion. For example, "Not many contributions that Jokowi implements during his presidency". It is even worse if some of words of an opinion are in irregular forms or using slang words. To address this issue, we combine ED and Bigrams to get the best probability of a fixed word. Below is our grand design of sentiment analysis process:

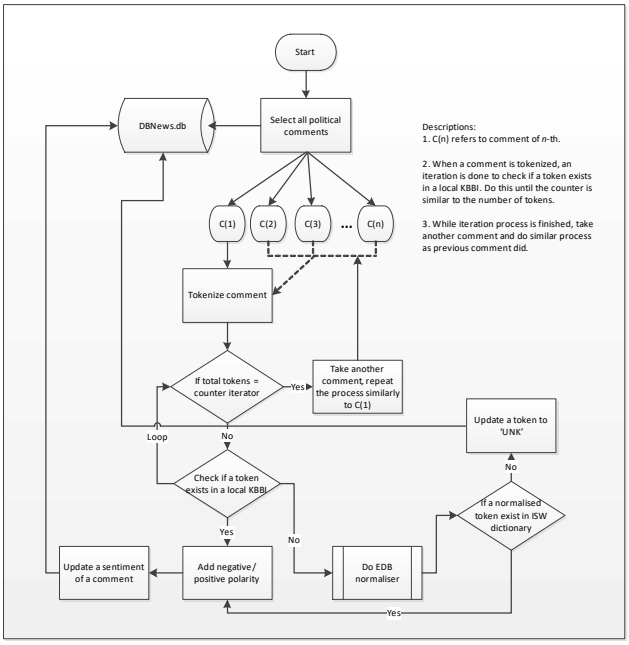

Figure 3. Sentiment Analysis Process

4.4 EDB Normalizer on Sentiment Analysis Process

EDB normalizer is two traditional approaches that are combined together and is believed to be able to address a misspelt word. Based on our experiment, it is inadequate to conduct sentiment analysis using ED, as doing sentiment analysis we extract subjectivity and polarity from a text. A text may consist of two or more words, and thus performing ED to a text will return the distances of all words that form the text. It is also possible to have k similar distances in just one text. This puzzles us by which we are not able to decide which one is the best word for such irregular token. On the other hand, it is very difficult to know the true probability of an arbitrarily sequence of words. But this does not mean impossible to do; we create an N-Grams language model (on this research we choose bigrams). We choose bigram model as we do not focus on capturing the contextual meaning of an information, instead we prefer to have high probability of the occurrence of successor word, given a preceding word. Given a token (see Figure 3) that does not exist in the local KBBI, we do EDB normalizer to that token. Our concept for EDB normalizer is as follows:

(1) Given irregular word, we naively filter words from a local KBBI by selecting all words that have pre-post characters criterion. We are doing this is to save the computation. See the illustration below:

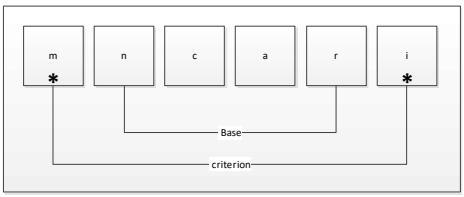

Figure 4. An illustration of misspelt word need to be normalized

(2) The filtered words are then looped until they are empty in the list. For any word given through the looping process will be a "hoping" normal token w.r.t a misspelt word. The calculation output is recorded into a dictionary. We then employ that dictionary as a system input to the next process. See the ED algorithm below:

```
EDITDISTANCE(s1, s2)
 1  int m[i, j] = 0
 2  for i ← 1 to |s1|
 3  do m[i, 0] = i
 4  for j ← 1 to |s2|
 5  do m[0, j] = j
 6  for i ← 1 to |s1|
 7  do for j ← 1 to |s2|
 8     do m[i, j] = min{m[i − 1, j − 1] + if (s1[i] = s2[j]) then 0 else 1fi,
 9                      m[i − 1, j] + 1,
10                      m[i, j − 1] + 1}
11  return m[|s1|, |s2|]
```

Figure 5. Edit Distance Algorithm (Stanford)

(3) After completing step (2), we will have a dictionary $\boldsymbol{D}$ with $\boldsymbol{W}$ possible words, and $\boldsymbol{K}$ for all tokens in the text comment. However, there is only one word $\boldsymbol{w}$ that is appropriate to the text $\boldsymbol{T}$. Picking naively to the $\boldsymbol{i}$-th lowest distance $\boldsymbol{s_i}$ in $\boldsymbol{D}$, where $i = \{1, \dots, n\}$, is not a good approach, instead we compute the probability of each possible word $\boldsymbol{P(w_i)}$ in $\boldsymbol{D}$ (as successor) given preceding word of an irregular word from the text comment $\boldsymbol{k_{j-1}}$, where $\boldsymbol{k_j}$ is a misspelt word and $j = \{1, \dots, m\}$ and $\boldsymbol{i \neq j}$. This will be done $\boldsymbol{o}$ times, where $\boldsymbol{o \leq len(D)}$. It is arbitrarily to set $\boldsymbol{o}$ value. The higher its value the longer its computation for the next process, which is finding the highest probability by employing bigrams. Below is the standard equation for calculating bigram and the algorithm to obtain the appropriate normalized token:

$$P_{MLE}(k_i|k_{i-1}) = \frac{C(k_{i-1}, k_i)}{C(k_{i-1})} \quad (1)$$

Since the preceding word will be from the text comment, we need to have slightly changed from equation (1), thus:

$$P_{MLE}(w_i|k_{j-1}) = \frac{C(k_{j-1}, w_i)}{C(k_{j-1})} \quad (2)$$

From equation (2) we compute the probability of occurring word $\boldsymbol{w_i}$ given a previous word $\boldsymbol{k_{j-1}}$ where $i \neq j$ and $\boldsymbol{w_i} \cong \boldsymbol{k_j}$. The algorithm to obtain appropriate normalized token as follows:

$j \leftarrow arbitrary\ number$
for $i \leftarrow 0\ to\ \boldsymbol{o}$
    $output[w_i] = P_{MLE}(w_i|k_{j-1})$
sort and return the $output$

From that computation, the highest probability score will be then picked as the best assumption.

Table 3 : EDB Normalizer Benchmarking

Misspelt Words	After EDB Norm	Expected Words (true/false/prob)
mncari	mencari	True/11.54×10^{-4}
baerani	berarti	False/13.6×10^{-4}
mgkn	mungkin	True/6×10^{-3}

kenesjganan	Kesenjangan	True/7.32×10^{-5}
kmeuankifan	kemakmuran	False/7×10^{-3}
mmperetahakkan	mempertahankan	True/18×10^{-4}
berfikir	berpikir	True/5×10^{-4}

From the table above, we show the significance of implementing EDB normalizer while performing sentiment analysis. We cannot ignore these irregular words as it may contain positive/negative polarity. Doing spelling correction using ED will only return n possible words, where n equals to the number of tokens available on the corpus. We have to admit that the probability of an occurring word is heavily relying on corpus.

5 Conclusions and Future Work

After conducting some experiments to build a corpus, in particular for under-resourced language, we can conclude that constructing a new corpus for sentiment analysis is not an objective task. For example, the existence of slang words is tremendously important, since people tend to use slang words frequently in their utterance and even in expressing their feelings through online media. In fact, some of slang words have strong positive or negative polarity which affect to the accuracy while doing sentiment analysis. Aside from that, spelling corrections also contribute significantly to gain better quality of sentiment analysis.

Many considerations need to be applied to this research. Even though this research is partly conceptual, but some of fundamental works have been applied, such as: constructing a political corpus and Indonesian slang words dictionary which consists of around 4000 informal words. We also annotate the polarity of each slang word manually; we do this since there is no such dictionary available and the urgency of making this accessible is significant. We have also fleshed out an incomplete KBBI database semi-automatically as our basis to do this research. An incomplete KBBI data has reached approx. 107.000 words, and most of them have been annotated their polarity. We also propose a combined technique (ED and Bigrams), EDB normalizer, and this has been implemented in a specific scenario, just to know whether EDB normalizer help much on spelling correction, and the result is better (see table 3). Even though exaggerated shortening problem does not take much time to solve, we also address this issue using a simple algorithm, which is removing any occurring duplicate chars on a suspected word.

Further research is that how to deal with a single-word of Indonesian abbreviation/slang problem. To the best of our knowledge, this issue can be handled by phonetic analysis.

6 Acknowledgements

We thank to Waseda University for joining us to ISIPS 2018, and for financial support on this occasion. We also thank to Dr. Inggriani Lieam and Dr. Arlinta Christy Barus, ST., M.InfoTech for the suggestions.

References

(1) Baker, Mona. "Corpora inTranslation Studies. An Overview and Suggestions for Future Research". Target, 7(2).1995. pp. 223-243.

(2) Baker, P. (2010). Corpus Methods in Linguistics. In Litosseliti, Lia. 2010. Research Methods in Linguistics. New York: Continnum International Publishing Group.

(3) Bhatia, P.; Ji, Y.; and Eisenstein, J. 2015. Better document level sentiment analysis from rst discourse parsing. In Proceedings of Empirical Methods for Natural Language Processing (EMNLP).

(4) Bing Liu. 2007. Web Data Mining: Exploring Hyper-links, Contents, and Usage Data. Data-Centric Sys-tems and Applications. Springer

(5) Hatzivassiloglou, Vasileios and Kathleen McKeown. 1997. Predicting the semantic orientation of adjectives. In Proceedings of 35th Meeting of the Association for Computational Linguistics, pages 174–181, Madrid.

(6) Imran, M., Mitra, P. and Castillo, C., 2016. Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages. arXiv preprint arXiv:1605.05894

(7) Jurafsky, Daniel, and James H. Martin. 2009. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd edition. Prentice-Hall.

(8) Laver, M., and Garry, J. 2000. Estimating policy positions from political texts. American Journal of Political Science 619–634.

(9) Liu, B. 2015. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge University Press.

(10) Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. Foundations and trends in information retrieval 2(1-2):1–135.

(11) Somasundaran, S.; Wiebe, J.; and Ruppenhofer, J. 2008. Discourse level opinion interpretation. In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, 801–808. Association for Computational Linguistics.

(12) Stone, Philip J., Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. The General Inquirer: A Computer Approach to Content Analysis. MIT Press,Cambridge, MA.

(13) Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K.; and Stede, M. 2011. Lexicon-based methods for sentiment analysis. Computational linguistics 37(2):267–307.

(14) Tausczik,Y.R.,andPennebaker,J.W. 2010. Thepsychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology 29(1):24–54.

(15) Tognini-Bonelli, E. 2001. Corpus Linguistics at Work, Amsterdam: Benjamins.

(16) Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M., 2011. Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), pp.267-307.

(17) Tong, Richard M. 2001. An operational system for detecting and tracking opinions in on-line discussions. In Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification, pages 1–6, New York, NY.

(18) Turney,P. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the Association for Computational Linguistics (ACL), 417–424.

(19) Wilson, T.; Wiebe, J.; and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of Empirical Methods for Natural Language Processing (EMNLP), 347–354.

(20) Polanyi, L., and Zaenen, A. 2006. Contextual valence shifters. In Computing attitude and affect in text: Theory and applications. Springer.

(21) Yadav, V. and Elchuri, H., 2013. Serendio: Simple and Practical lexicon-based approach to Sentiment Analysis. In Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013) (Vol. 2, pp. 543-548).

(22) Le, T.A., Moeljadi, D., Miura, Y. and Ohkuma, T., 2016. Sentiment Analysis for Low Resource Languages: A Study on Informal Indonesian Tweets. In Proceedings of the 12th Workshop on Asian Language Resources (ALR12) (pp. 123-131).