Automatic Semantic Orientation of Adjectives for Indonesian Language Using PMI-IR and Clustering

Dewi Riyanti, M. Arif Bijaksana, and Adiwijaya
School of Computing, Telkom University, Jl. Telekomunikasi No.1, Bandung, 40257, Indonesia
E-mail: dewiriyanti@telkom.co.id, arifbijaksana@telkomuniversity.ac.id, adiwijaya@telkomuniversity.ac.id

Abstract. We present our work in the area of sentiment analysis for Indonesian language. We focus on building automatic semantic orientation using available resources in Indonesian. In this research we used Indonesian corpus that contains 9 million words from kompas.txt and tempo.txt that manually tagged and annotated with of part-of-speech tagset. And then we construct a dataset by taking all the adjectives from the corpus, removing the adjective with no orientation. The set contained 923 adjective words. This systems will include several steps such as text pre-processing and clustering. The text pre-processing aims to increase the accuracy. And finally clustering method will classify each word to related sentiment which is positive or negative. With improvements to the text preprocessing, can be achieved 72% of accuracy.

1. Introduction
The need for information encourages the development of research and technology to can answer those needs. The required information is experienced the development of general information to more information specific and specific. Obtain the right and fast information will help the organization to be able to make an answer of the problems. Research in sentiment analysis driven by a thought that information is a sentiment of data which is important and necessary. Sentiment is related to assessment of a context or discourse. Positive sentiment states giving good value to context in text and negative sentiments declared the opposite.

The sentiment analysis which is part of opinion mining [17], is the process of understanding, extracting and processing textual data automatically for information [18]. Conducted to view opinions on a problems, or can also be used to identify trends in the market [1]. Analysis of sentiments in the study this is the process of classifying textual documents into two classes, namely positive and negative sentiment class. The magnitude the effects and benefits of sentimental analysis, leading to research or application of sentiment analysis growing rapidly, even in America approximately 20-30 companies that focus on analytics services sentiment [17]. Basically the sentiment analysis is classified, but the reality is not as easy as the classification process ordinary due to language usage. Where there is ambiguity in the use of words, the absence of inner intonation a text, and the development of the language itself.

As one of the ways to convey, language can be viewed as the tool to express individual subjective valuation that may include the way toward considering, feeling, or a blend of both. The expression can appear as examination, feeling, or assessment, that is communicated towards elements. Sentiment analysis intended as a field of study to analyze the sentiment on the text,
or in other words the grouping the polarity of text either in the document level, sentences or features/aspects of level n-gram (noun, verb, adverb, clause, phrase)-whether opinions expressed are positive or negative (there are also neutral). Sentiment analysis aims to determine the attitude of an author of some topics.

The term opinion mining and sentiment analysis are considered interrelated and interchangeable [11] because in the objective opinion mining is gathering public opinions for analysis. Whether opinions expressed in a document, the phrase or entity features/aspects of positive or negative.

One of the techniques of sentiment analysis is semantic orientation or polarity of words. Polarity is binary value of positive or negative sentiment in a word. For identify word polarity can be categorized into positive and negative categories. Two types of techniques have been used in the literature for semantic orientation-based approach for sentiment analysis: corpus based and dictionary or lexicon based. In this study, we explore the corpus-based semantic orientation approach for adjectives. Corpus-based semantic orientation approach requires large dataset to detect the polarity of the words. The main problem with this approach is that it relies on the polarity of the words that have emerged in the training corpus since polarity is computed for the words that are in the corpus.

Research on sentiment analysis continues to grow, each study using data, language, features, different weighting algorithms and calculations. Common corpus data used in sentiment analysis among others, online news web, the wall street journal [3], social media [1], web pages Altavista [5], and movie reviews [14] [15].

Many words communicate the speaker’s evaluation of the item that is under discussion as desirable or undesirable. This evaluative character a word is called its semantic orientation. A word with a positive semantic orientation conveys the evaluation that the item is desirable (e.g., nice, good) and a negative orientation conveys the evaluation that the item is undesirable (e.g., bad, ugly).

Sentiment analysis, also known as opinion mining, refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. Sentiment analysis aims to determine the attitude of a writer with respect to some topic. The attitude may be his/her judgement or evaluation based on his/her experience. A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. In most statistical classification methods, the neutral class is ignored under the assumption that neutral texts lie near the boundary of the binary classifier, several researchers suggest that, as in every polarity problem, three categories must be identified.

Research on sentiment focused on adjectives or adjective phrases as the primary source of subjective content in a document has been gaining a lot of attention in the recent years. The goal is to identified automatically semantic orientation of words and allow a system to further refined the retrieved semantic similarity relationship. Automatic creation of polarity lexicons is a crucial issue to be solved in order to reduce time and efforts in the first steps of sentiment analysis.

For English, the current state-of-the-art automatic creation of polarity lexicons system achieve around 82% accuracy score for adjectives in the semantic orientation identification task [3].

It is said lexicon based because it uses lexicon and phrases (even idioms) that have been stored in the dictionary/lexicon. The use of a dictionary / lexicon is based on the assumption that sentiment orientation in a sentence context can be determined from the sentiment orientation of the vocabulary or the word equivalent in it.

The use of lexicon when conducting sentimental analysis research can be constructed on its own or utilizing dictionaries already built by others [12]. Whether building your own or utilizing
an existing dictionary equally has its advantages and disadvantages. By building your own
dictionary, it means it will take up a lot of time and energy having to collect from scratch.
But the advantages, the contents of the dictionary in accordance with the needs of research. If
using an existing dictionary, it is obviously more time-saving and labor-intensive to only modify
the language and the dictionary is considered valid because the dictionary is used for academic
affairs, as well as the fact that the dictionary is built and used by academics. The drawback,
the contents of the dictionary is not necessarily according to the needs of research.

The conceptual framework of sentiment lexicon is described as follow:

![Sentiment Lexicon Process](image)

**Figure 1. Sentiment Lexicon Process**

Dataset is collected, then data preprocessing stage is prepared, this process is include
cleansing, case folding, and parsing. Feature extraction and feature weighting selects the related
attribute. Next process is learning and testing process, using soft clustering methods with to
train a set of base learner to result predictions with each one of them, and then do voting to
these predictions. This research is expected to predict the polarity of the lexicon.

2. **Theory**

This section discusses the literature studies and theories of this study. It is divided into two
sections: (1) the related work; and (2) the Existing Method.

2.1. Related work

For the purpose of building such lexicon automatically, a lot of studies have researched
unsupervised method of learning polarity of words. In 1997, V. Hatzivassiloglou and K.R.
McKeown used 21 million words of Wall Street Journal corpus to predict the semantic orientation
of adjective. In this paper, the authors hypothesized that two adjectives joined by and have
the same semantic orientation while two adjectives joined by but have the opposite one. They
used this idea along with a loglinear regression model and a set of supplementary morphological
rules in text preprocessing and clustering. The method detected the word polarity based on the
strength of co-occurrence with the seeds to predict whether a pair of adjectives joined by any
of these conjunctions has the same or different semantic orientation. Once pairs of adjectives
are extracted, they utilized a clustering method to separate all the adjectives conjoined into two
clusters, placing as many words of the same orientation as possible into the same subset. The
cluster with more elements was labeled as positive adjectives and the other as negative. The
strengths of this method are this method fully unsupervised because the algorithm started from
nothing and the result was convincing. This work achieved accuracy score between 78% and
92% in the classification of positive and negative adjectives.

P.D. Turney presented a general strategy for learning semantic orientation from semantic
association. The unsupervised classification method based on distributional semantics. The
method involves issuing queries to a customer reviews and using pointwise mutual information (PMI) to measure an adjective's tendency to appear in positive or negative vicinity. The algorithm is empirically evaluated using a training corpus of approximately 5000 reviews. Tested with 410 reviews, the algorithm attains an average accuracy of 74%.

P.D. Turney and M.L. Littman in 2002 showed the accuracy of SO-PMI-IR with a training corpus of 350 million Web pages (at least one hundred billion words). These are the English Web pages that are indexed by AltaVista. Semantic orientation was used as a measure of confidence that the word will be correctly classified. An accuracy of 80% is attained, using an unlabeled corpus of about one hundred billion words.

Y. Matsuo, K. Uchiyama, T. Sakaki and M. Ishizuka proposed an unsupervised method for word clustering based on a word similarity measure by web check. Each pair of words is queried to a search engine, which produced a co-occurrence matrix. By computing the similarity of words, a word cooccurrence graph is obtained. Newman clustering, a new kind of graph clustering algorithm was applied for distinguishing word clusters. Evaluations are made on two arrangements of word groups determined from a open directory and WordNet. In this research they compared two clustering methods: Newman clustering and word clustering. The word co-occurrence graph is created using PMI, Jaccard and Chi-square measures.

S. Vazquez, M. Padro, N. Bel and J. Gonzalo proposed bootstrapping algorithm to automatically extracts all of the polar adjectives joined by y (and) or pero (but) in a given corpus. And they utilized the adjectives that were in the seed polarity lexicon as input for their algorithm to find new adjectives joined and identifying the priority. They propose that polar adjectives and their corresponding polarity values can be automatically identified if they are found in a coordinate construction with the appropriate conjunctions and with other adjectives that were not in the seed lexicon. The process continued until any adjective of lexicon is not found join with any new adjective or until is no more conjunctive relation of this type.

2.2. Sentiment lexicon

Sentiment lexicon uses the phrases and words (even idioms) that have been stored in the dictionary / lexicon. The use of a dictionary / lexicon is based on the assumption that the orientation of sentiment in a word, phrase or sentence context can be determined from the sentiment orientation of the vocabulary or the word equivalent in it.

The main component of the sentiment lexicon is the list of words and orientation that relate to the word. In this study there are two categories of orientation: positive and negative. The next stage after determining the type of emotion that is determining the words seed. Words seed are a number of words that are directly related to one particular kind of orientation. We constructed this seed by taking all adjectives appeared in Kompas and Tempo corpus 20 times or more, then removed adjectives that have no orientation.

The word list in words seed consists of only a few words, while the sentiment lexicon requires large numbers of words. The word in the lexicon is expanded based on the words seed. After the expansion of words seed, the next step is to determine the weight of each word. This study uses SO-PMI-IR. Then, classify the adjectives in to positive or negative orientation.

2.3. Part-of speech tagging

Part-of-Speech tagging is an NLP task. Most of the activities done in the NLP field such as Information Extraction, Question-Answering, Speech Recognition, Intelligent Tutoring System, Parser, and others require this POS Tagging system for initial processing. Unfortunately, for the Indonesian language the system is few, the publication is lacking, and no one provides for download.

Part-of-speech tagging is a system that automatically labels words to a sentence. In this study used pos tagger method has proposed in [7]. They designed a linguistically motivated
POS tagset for the Indonesian language. The design process is divided into two phases: (1) define initial POS tagset and (2) test and revise POS tagset which involve manually tagging the Indonesian sentences in the IDENTIC corpus. The corpus that manually tagged consists of 10,000 sentences, containing 262,330 lexical tokens.

Suppose, there are sentences "Saya pergi ke sekolah." and there are labels PRP = personal pronouns, VBI = verbs, IN = preposition, NN = noun. The system will accept the input of the sentence, the output is:

Saya/PRP pergi/VBI ke/IN sekolah/NN

The tags used are different depending on the language used in tagging. For Indonesian, below is a table that contains tags with descriptions and examples that has proposed in [7].

2.4. Adjectives
Adjective is a word class that changes a noun or pronoun, by explain and make the word more specific that can explain the quantity, adequacy, order, quality, or emphasis of a word. In Indonesian language, adjectives can be formed due to several processes that occur, as follows:

1. Formed from basic word, for example: tua, muda, tinggi, and rendah.
2. Formed from additional affix (suffix, prefix), for example: tertampan, tercantik.
3. Formed from word repetition, for example: compang-camping, gelap-gulita.
4. Formed from another language, for example: kreatif, and legal.
5. Formed from group of words, for example: baik hati, lemah jiwa and keras kepala.

2.4.1. Adjectives pair coordinated by conjunction To achieve coherence or cohesion in a paragraph, it is necessary to connect words or also called conjunctions. It is conjunction that makes a sentence in the paragraph logical, grammatical and concerted. Conjunction itself has a sense of the word task or phrase that connects two equivalent language units, for example

![Diagram](image-url)
Figure 3. Indonesian Tagset

between words with words, clauses with clauses, phrases with phrases, and sentences with sentences.

Coordinated conjunctions are connecting words that connect two sentences or more in a paragraph having an equal position. Below are the various coordinated conjunctions and their examples

1. Dan: this conjunction is a marker of the relationship between words.
2. Atau: this conjunction is a marker of election relationships in a paragraph or sentence and usually has the opposite meaning.
3. Tetapi or tapi: this conjunction is a marker of inter-sentence resistance in the paragraph.

Adjectives pair coordinated by conjunction is merging or pairing adjectives with conjunction present in the middle. It is part of part-of-speech tagging (POS). POS tags will identify simple adjectives. Below is an example sentence with a POS tag.

Kita/PRP menginginkan/VBI pembangunan/NN yang/RP damai/JJ dan/CC demokratis/JJ untuk/IN Indonesia/NNP.

The tagged words are then extracted based on the adjective pair pattern, connecting words and adjectives (JJ-CC-JJ). The first word identified is the adjective (JJ), the second word identified is the connector (CC), and the third word identified is the adjective (JJ).

2.5. Pointwise mutual information

Pointwise mutual information (PMI) or point mutual information, is a measure of association used in information theory and statistics [9]. Pointwise Mutual Information (PMI) between two adjectives as follows:

$$PMI(x, y) = \log_2\left(\frac{P(x, y)}{P(x)P(y)}\right).$$  \hspace{1cm} (1)

Where $p(x, y)$ is the probability that word$_1$ and word$_2$ co-occur, $p(x)$ is the probability that word$_1$ co-occur, and $p(y)$ is the probability that word$_2$ co-occur. If there is a true relationship
between word1 and word2, then the probabilities that occur in \( p(\text{word}_1 & \text{word}_2) \) will be much larger than \( p(\text{word}_1)p(\text{word}_2) \). And the consequently is \( \log(\text{word}_1/\text{word}_2) >> 0 \). The log of this ratio is the amount of information obtained about the presence of one word when observing the other.

An adjective can be known the orientation by SO-PMI-IR

\[ SO - PMI - IR(\text{word}) = PMI(\text{word, positiveparadigms}) - PMI(\text{word, negativeparadigms}) \] (2)

When the result of PMI is positive then the adjective word has a positive semantic orientation and a negative orientation when the PMI value is negative.

2.6. Clustering

Clustering is a method of data analysis, which is often incorporated as one of the methods of Data Mining, whose purpose is to group data with the same characteristics into an identical ‘region’ and data with different characteristics to the other ‘territories’.

There are several approaches used in developing clustering methods. The two main approaches are clustering with partition and clustering approaches with a hierarchical approach. Clustering with partitioning approach or often called partition-based clustering grouping data by sorting through the data that is analyzed into existing clusters. Clustering with hierarchical approach or often called hierarchical clustering classifies data by creating a hierarchy of dendograms where similar data will be placed in adjacent hierarchies and not in a remote hierarchy. In addition to both approaches, there is also clustering with the approach of automatic mapping (Self-Organizing Map/SOM).

Self-Organizing Map (SOM) is an Artificial Neural Networks type that is training in an unsupervised manner that has been proposed in [13].. SOM generates folders consisting of outputs in low dimensions (2 or 3 dimensions). This folder attempts to locate the properties of the data input. The input and output composition in SOM is similar to the composition of the feature scaling process (multidimensional scaling).

Although the learning process is similar to Artificial Neural Networks, but the process of assigning data input to the map is more similar to K-Means and kNN Algorithm. The procedure taken in doing clustering with SOM is as follows:

1. Determine the weight of the input data randomly. Select one of the data inputs.
2. Calculate the level of similarity (with Euclidian) between the data input and the weight of the input data and select the input data that has similarities with the existing weight (this data is called Best Matching Unit (BMU)).
3. Renew weight from input data by bringing the weight to BMU with formula:

\[ W_v(t + 1) = W_v(t) + \Theta(v, t) \times \text{Alpha}(t) \times \text{D}(t) \times W_v(t) \] (3)

Where \( W_v(t) \) is weight at time t. \( \Theta(v, t) \) is the neighborhood function depends on the Lattice distance between BMU and neuron v. Generally a value of 1 for neurons close enough to BMU, and 0 for the opposite. The use of Gaussian functions is also possible. \( \text{Alpha}(t) \) is Learning Coefficient is reduced monotonically and \( \text{D}(t) \) is input data.

2.7. K-Cross fold validation

Cross Validation is one technique to assess or validate the accuracy of a model built on a particular dataset. The model usually aims to make predictions and classifications of new data that may never have appeared in the dataset. The data used in the model development process is called training data, while the data to be used to validate the model is called the test data.
One of the popular cross-validation methods is K-Fold Cross Validation. In this technique the dataset is divided into a number of K partitions at random. Then a number of experimental K-times were conducted, each experiment using K-partition data as data testing and utilizing the rest of the partition as training data. As an illustration, if we do 5-Fold Cross-Validation then design the experimental data as follows:

![Figure 4. Dataset](image)

| Experiment | Training Data | Testing Data |
|------------|---------------|--------------|
| 1          | K1            | K2, K3, K4   |
| 2          | K2            | K3, K4, K5   |
| 3          | K3            | K4, K5, K1   |
| 4          | K4            | K5, K1, K2   |
| 5          | K5            | K1, K2, K3   |

![Figure 5. Data experiments](image)

To obtain the value of accuracy of the experimental results that we do, can be taken the average value of the entire experiment.

2.8. Method selection
The are some techniques to estimate word co-occurrence frequencies from an available corpus. Based on [18], gave a comparative examination of co-occurrence frequency calculation. The research reported that PMI is useful for computing semantic orientation and performed best on average. Another advantages of PMI are fully unsupervised and the result in nominal context makes adjective semantics more interpretable [5]. Then clustering algorithm creates the graph structure to produce two groups of adjectives.

3. Research Methodology
This chapter discusses methodology of the study, it is divided into five sections, namely 1) Research Design; 2) System Implementation; 3) Experiment Scenario; 4) Instrumentation and Data Collection; and 5) Tools for Data Analysis.

3.1. Research design
The design process of this research is describes as follows,

3.1.1. Data preprocessing The dataset we used in this research is 9 million words corpus, collected from Kompas and Tempo articles. The steps performed on the preprocessing data are
1. Cleansing, the process of cleaning documents from words that are not needed to reduce noise. That word omitted are HTML characters, stopwords, and url (http://situs.com)
2. Case folding, change letters as well as deletion of numbers and punctuation. In this case that is used only Latin letters between a to z.
3. Parsing, ie the process of breaking a document into a word. This is in accordance with the feature used ie unigram.
3.1.2. Feature extraction  Here is a selection process and feature extraction that will be used as the basis of the clustering process.

1. Part of Speech (POS) Tagger, i.e. the process of giving classes to words. The selected word class is an adjective (adjectives), adverbs, nouns and verbs, in accordance with Dinakaramani [7].

2. Stemming, aimed at reducing variations of words that have the same basic word. In this research, stemming are used in scenario 3.

3.1.3. Adjective filtering  Extract from POS TAG corpus (9 million words) adjective words. Adjective words labelled with JJ annotation. From this process, we obtained 923 unique adjective words. The adjective words labelling by empowering a number of human coders to provide positive and negative orientation. The study used the average result of labeling done by 4 people who had acted as human coders independently. The final set contained 923 adjectives (407 positive and 516 negative terms). From 923 adjective word, we construct a seed words. We constructed this set by taking all adjectives appearing in our corpus 20 times or more. The seed word contain 59 adjective words (39 positive and 20 negative). This seed word will be used in SO-PMI-IR calculation.

3.1.4. Expand adjective  Extract from POS tagged corpus (9 million words) and expand adjective seed words coordinated by 'dan', 'tetapi', 'tapi' and 'atau' with another adjective
words. We filter from the corpus words ‘adjectives seed words (JJ)’ + conjunction (CC) + ‘adjective words(JJ)’

3.1.5. Clustering This results in graph with same or different links between adjectives. Cluster into two orientations, positive and negative placing as many words of the same orientation as possible into the same subset. Cluster with higher overall frequency is labelled positive.

3.1.6. Performance and measurement analysis Accuracy showed information about how many successful final decisions that the system can analyze against human labeling results.

\[
\text{Accuracy} = \frac{TP}{N} \times 100\% \tag{4}
\]

Where \(TP\) is true predicted words and \(N\) is number of test words.

3.2. System implementation Here are the hardware and software specifications that were used to implement the prediction system.

3.2.1. Hardware specifications The minimum hardware specification to run this system is the same as with the minimum specifications required by Matlab2012b 64-bit:

a. Processor : Intel Pentium 4 and above, Intel Celeron**, Intel Xeon, Intel Core, AMD64 ** Processor must support SSE2 instruction set.

b. Memory : 2 GB

c. Hard disk : 7 GB* (MATLAB only)

Disk space requirement varies depending on size of partition. The MATLAB installer will inform you of the hard disk space requirement for your particular partition. Installation size will be determined by the installer and can vary for NTFS and FAT formats.

3.2.2. Software specifications The software specifications that were used are:

a. Operating system: Microsoft Windows 8

b. Matlab version R2016b

c. Microsoft Excel 2007

3.3. Experiment scenario Experiment scenario runs in this study is to split into several sub-systems as shown in the table. Each sub-system will be tested several variations of variables to compare the results. Basically this study will examine the effect of applying different preprocessing on sentiment classification accuracy results. Three methods of preprocessing will be used and compared, they are:

3.3.1. Scenario 1 The first scenario has propose to testing and investigate the performance of the basic preprocessing. We consider the accuracy of classification which is use clustering from the graph of adjective pairs.

3.3.2. Scenario 2 In this scenario, we add SO-PMI-IR to the clustering process and analyze the effect of SO-PMI-IR to the accuracy of classification which is use clustering from the graph of adjective pairs.

3.3.3. Scenario 3 This scenario has improvement by adding feature extraction stemming . This scenario also use SO-PMI-IR in process.
3.4. Instrument and data collection
The dataset used in this research is 9 million words corpus which was collected from Kompas and Tempo. From Tempo contained 6,022,980 words and Kompas contained 3,697,852 words.

4. Experiment Result
This section show the result of three scenario. The first scenario has propose to testing and investigate the performance of the basic conjunction extraction without preprocessing. In the second scenario, we add PMI to the clustering process and analyze the effect of PMI to the accuracy of classification which is use clustering from the graph of adjective pairs. The last scenario has improvement by adding preprocessing before conjunction extraction. This scenario also use PMI in clustering process.

4.1. Scenario 1
In this scenario, the total links of graph is 1180 links. From this graph we clustering the node to two class, positive adjective and negative adjective, which is shown in in Fig. 10. The blue line is for positive adjective class and the red line is for negative adjective class. The percentage of positive adjective is \(\frac{1112}{68} \times 100\% = 94.23\%\), it is shown that the positive adjective class is 94.23% from the total adjective word. The comparison link between positive adjective and negative adjective is shown in Table 1. with total link show in Table 1. The first scenario aims to show the performance of clustering. Accuracy was measured to see the performance of prediction with Accuracy. The result was shown in the table below. There were use basic preprocessing at this first scenario because it just wanted to show the real result of baseline method to build a prediction model and measure the accuracy value. The experiments were run 5 and 10 times. The lower data set provides the lower accuracy in cross fold validation with \(k = 5\).
Figure 8. Scenario 2

| No | Word                | Total Links |
|----|---------------------|-------------|
| 1  | Positive Adjective  | 1112        |
| 2  | Negative Adjective  | 68          |

Table 1. Total link 1st scenario

| Data Set | K Cross Fold (%) | Accuracy (%) |
|----------|------------------|--------------|
|          | 1    | 2       | 3       | 4       | 5       | 100% | 75% | 50% | 25% |
| 100%     | 65.97 | 63.61  | 63.57  | 65.92  | 65.29  | 64.87 |
| 75%      | 64.15 | 65.25  | 62.88  | 62.03  | 65.25  | 63.92 |
| 50%      | 65.64 | 64.74  | 68.59  | 64.36  | 67.95  | 66.26 |
| 25%      | 67.95 | 62.51  | 66.67  | 59.56  | 58.97  | 62.33 |

The average accuracy of testing was 60% to 68%. The result of this scenario still had a small value of accuracy. The small value of accuracy was caused by some adjective word that can not be detected because they had affixed. So in the scenario 3, we added some preprocessing technique.

4.2. Scenario 2

In this scenario, the total links of graph is 1644 links. From this graph we clustering the node to two class, positive adjective and negative adjective, which is shown in in Fig. 11. The blue line is for positive adjective class and the red line is for negative adjective class. Compared with the graph of first scenario, in this scenario the link of graph is more than link in first scenario.
because the effect of PMI. The percentage of positive adjective is \( \frac{1518}{1702} \times 100\% = 92.33\% \), it is shown that the positive adjective class is 92.33\% from the total adjective word. The comparison link between positive adjective and negative adjective is shown in Table 3. The second scenario aims to show the performance SO-PMI-IR with basic preprocessing and clustering. The result was shown in the table below.

| No | word           | Total links |
|----|----------------|-------------|
| 1  | Positive Adjective | 1518        |
| 2  | Negative Adjective | 126        |

The second scenario showed the performance of the clustering using data with preprocessing and combined with SO-PMI-IR. The experiments were run 5 and 10 times. Compared to without SO-PMI-IR implementation in scenario 1, overall accuracy slightly increased 2\%, compared to without SO-PMI-IR. The accuracy without SO-PMI-IR is 68\% and with SO-PMI-IR is 70\%.

![Diagram](dummy.png)

**Figure 9.** Scenario 3

Table 3. Total link 2nd scenario

| Data Set | K Cross Fold (%) | Accuracy (%) |
|----------|------------------|--------------|
| 100\%    | 1 | 2 | 3 | 4 | 5    | 69.24   |
| 75\%     | 72.29 | 72.93 | 70.38 | 65.92 | 64.65 | 69.24 |
| 50\%     | 72.03 | 64.83 | 75.00 | 69.07 | 67.37 | 69.66 |
| 25\%     | 69.23 | 69.23 | 64.10 | 66.67 | 66.03 | 68.59 |

**Table 4.** Accuracy of Scenario 2 k Fold=5

The second scenario showed the performance of the clustering using data with preprocessing and combined with SO-PMI-IR. The experiments were run 5 and 10 times. Compared to without SO-PMI-IR implementation in scenario 1, overall accuracy slightly increased 2\%, compared to without SO-PMI-IR. The accuracy without SO-PMI-IR is 68\% and with SO-PMI-IR is 70\%.
4.3. Scenario 3

The graph of 3rd scenario is shown in Fig. 12. In this scenario, the total links of graph is 3132 links. From this graph we clustering the node to two class, positive adjective and negative adjective. The blue line is for positive adjective class and the red line is for negative adjective class. Compared with the graph of first scenario and second scenario, in this scenario the link of graph is the highest because the effect of PMI and preprocessing. The comparison link between positive adjective and negative adjective is shown in Table 5. with total link show in Table 5 The third scenario aims to show the performance of proposed preprocessing combine with SO-PMI-IR and clustering. The result was shown in the table below.

Compared to scenario 1, in overall proposed preprocessing method combined with SO-PMI-IR improves accuracy by 8%. Compared to scenario 2, proposed preprocessing combined with SO-PMI-IR improves accuracy by 6%. From the table above shown that the lower data set provide the lower accuracy in cross fold validation with $k = 5$. It is indicates that the proposed method has better performances with higher data set.

| No | word                | Total links |
|----|---------------------|-------------|
| 1  | Positive Adjective  | 2315        |
| 2  | Negative Adjective  | 817         |

Figure 10. graph and cluster of 1st scenario.
**Figure 11.** graph of 2nd scenario.

**Table 6.** Accuracy of Scenario 3 k Fold=5

| Data Set | K Cross Fold (%) | Accuracy (%) |
|----------|-----------------|--------------|
|          | 1 | 2 | 3 | 4 | 5 |                |
| 100%     | 75.90 | 72.30 | 70.14 | 64.57 | 69.96 | 70.58 |
| 75%      | 79.33 | 69.95 | 71.88 | 68.03 | 64.18 | 70.67 |
| 50%      | 77.70 | 67.99 | 67.99 | 71.94 | 69.06 | 70.94 |
| 25%      | 74.64 | 70.29 | 72.46 | 60.87 | 68.84 | 69.42 |

### 4.4 Summary of findings

From section 4, we resume the summary of accuracy in all scenarios, these is shown in Figs. 13 and 14.

From Figs. 13 and 14, its shown that lower dataset has lower accuracy. Compared by each scenario, the best scenario is third scenario.

### 5. Conclusions

This research proposed the combination of graph clustering and PMI the classificate the adjective word from adjective pairs. The performance of classification models was compared between developed by not using preprocessing, without preprocessing and using combined graph clustering and PMI, and using preprocessing and using combined graph clustering and PMI. From the experiment result, some conclusions of this research are as follows:
1. There are differences in system results and human labeling. This can happen because every human being has a different behavior and understanding in giving the sentiments of a lexicon.

2. The lower data set provides the lower accuracy in cross fold validation with \( k = 5 \) and \( k = 10 \). It indicates that the proposed method has better performance with higher data set.

3. The implementation of combined graph clustering and PMI is not improved the performance of accuracy. Compared to the third scenario, which is the best performance of all scenario, the effect of preprocessing makes better performance than without using preprocessing.
4. The overall accuracy result is 72%. These results indicate that not too many successful results have been generated by the system with Indonesian Corpus. This result occurs because the terms in the default dictionary are not complete or not fully in accordance with the needs of Indonesian vocabulary. This can also happen because there has been no special treatment for some features.

References
[1] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: Sentiment Classification Using Machine Learning Techniques. In EMNLP02: Proc. Of The ACL-02 conf on Empirical methods in natural language processing, Pages 79-86, ACL, 2002.
[2] T. OKeefe and O. Katrinska. Feature Selection and Weighting Methods in Sentiment Analysis. In Proceedings of the Australasian document computing symposium, 2009.
[3] V. Hatzivassiloglou and K.R. Predicting the semantic orientation of adjective. In Proceeding of the 35th annual meeting on Association for Computational Linguistic, pages 174-181, Doi:10.3115/976909.979640, 1997
[4] S. Vazquez, M. Padro, N. Bel and J. Gonzalo. Automatic Extraction of Polar Adjectives For The Creation of Polarity Lexicons. In Proceeding of COLING 2012: Posters:24th International Conference on Computational Linguistics COLING 2012, Pages 1271-1280, ACL, 2012
[5] P. Turney and M.L. Littman. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. nparc.cisti-icist.nrc-cnrc.gc.ca. 2002
[6] F. Rashel, A. Luthfi, A. Dinakaramani and R. Manurung. Building an Indonesian Rule-Based Part-of-Speech Tagger. In International Conference on Asian Language Processing (IALP), 2014
[7] A. Dinakaramani, F. Rashel, A. Luthfi and R. Manurung. Designing an Indonesian Part of Speech Tagset and Manually Tagged Indonesian Corpus. In International Conference on Asian Language Processing (IALP), 2014
[8] F. Tala, J. Kamps, K. Muller, and M. de Rijke. The Impact of stemming on Information Retrieval in Bahasa Indonesia(Abstract). In 14th Meeting of Computational Linguistics in the Netherlands (CLIN-2003), 2003
[9] K.W. Church and P. Hanks. Word Association Norms, Mutual Information and Lexicography. In Proceedings of the 27th Annual Conference of the Association of Computational Linguistics (pp. 76-83), 1989
[10] P.D. Turney. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In Proceedings of the Twelfth European Conference on Machine Learning Berlin: Springer-Verlag,2001
[11] B. Liu. Sentiment Analysis and Opinion Mining. In Chicago: Department of Computer Science, 2012.
[12] M. Taboada, et al. Lexicon-Based Methods for Sentiment Analysis. In Association for Computational Linguistics Volume 37, Number 2, 2010.
[13] L.V. Fausett. Fundamental of Neural Network: Architectures, Algorithm, And Application. In Prentice Hall, 1st edition. ISBN-13: 978-0133341867.
[14] M.S.H. Wicaksono. Unsupervised Approach for Sentiment Analysis on Indonesian Movie Review In 6th Conference of Indonesian Students Association in Korea (CISAK), 2013
[15] R.F. Manurung Machine Learning-based Sentiment Analysis of Automatic Indonesian Translations of English
Movie Reviews In Proceedings of the International Conference on Advanced Computational Intelligence and Its Applications (ICACIA), 2008.

[16] Y. Muhamad. Analisis Sentimen pada Dokumen Berbahasa Indonesia dengan Pendekatan Support Vector Machine. In PhD thesis, Bina Nusantara University.

[17] I. Sunni and D. Widyantoro. Analisis Sentimen dan Ekstraksi Topik Penentu Sentimen pada Opini Terhadap Tokoh Publik. In Jurnal Sarjana Institut Teknologi Bandung Bidang Teknologi Elektro dan Informatika, vol. 1, no. 1, pp. 200-206, 2002.

[18] E. Terra, and C.L.E. Clarke. Frequency Estimates for Statistical Word Similarity Measures. In Proceedings of the 2003 ACL Main Paper pp 165-172, 2003.

[19] Y. Matsuo, K. Uchiyama, T. Sakaki, and M. Ishizuka. Graph Based Word Clustering Using a Web-search Engine. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2006.