Accredited by National Journal Accreditation (ARJUNA) Managed by
Ministry of Education, Culture, Research and Technology with Second Grade (Sinta 2)
since year 2021 to 2026 according to the decree No. 158/E/KPT/2021

Published online on the journal's webpage: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 5 No. 6 (2021) 1044 - 1051 ISSN Media Elektronik: 2580-0760

Hate Speech Detection on Twitter in Indonesia with Feature Expansion Using GloVe

Febiana Anistya[1], Erwin Budi Setiawan[2]
[1,2]Informatics, School of Computing, Telkom University
[1]febiananistya@student.telkomuniversity.ac.id, [2]erwinbudisetiawan@telkomuniversity.ac.id*

Abstract

Twitter is one of the popular social media to channel opinions in the form of criticism and suggestions. Criticism could be a form of hate speech if the criticism implies attacking something (an individual, race, or group). With the limit of 280 characters in a tweet, there is often a vocabulary mismatch due to abbreviations which can be solved with word embedding. This study utilizes feature expansion to reduce vocabulary mismatches in hate speech on Twitter containing Indonesian language by using Global Vectors (GloVe). Feature selection related to the best model is carried out using the Logistic Regression (LR), Random Forest (RF), and Artificial Neural Network (ANN) algorithms. The results show that the Random Forest model with 5.000 features and a combination of TF-IDF and Tweet corpus built with GloVe produce the best accuracy rate between the other models with an average of 88,59% accuracy score, which is 1,25% higher than the predetermined Baseline. The number of features used is proven to improve the performance of the system.

Keywords: Feature Expansion, GloVe, Hate Speech

1. Introduction

Of the top five countries, Indonesia is one of the countries that invest in social media in general, especially Twitter [1]. Many netizens use the Twitter platform as a channel of opinion in the form of criticism and suggestions. But it is often netizens who misinterpret criticism with hate speech. Criticism could be a form of hate speech if the criticism implies attacking something (an individual, race, or group) [2]. The hate speech crime has been included in the ITE Law Number 11 of 2008 Article 45 Paragraph 2 [3].

In the detection process, the use of inappropriate vocabulary makes sentences uploaded in the form of Tweets challenging to understand without context [4], which can be overcome by word embedding. Word embedding is a step used to find the vector of the word and its context in the corpus to be matched with specific criteria. Word2vec was used for feature expansion in the previous study [4]. In addition to these methods, feature expansion can be carried out using Global Vectors for word representation (GloVe). GloVe is said to be an efficient and effective method for the representation learning process vector of words. GloVe is a log-bilinear global regression model for unsupervised learning of word representation that outperforms other models in analogy, word equations, and named entity detection developed by Stanford University. In this study, the choice of GloVe as the word embedding method was because GloVe consistently outperformed word2vec; by achieving better and faster results, the best results are also obtained regardless of the speed [5].

In research [6], hate speech detection using 16K annotated tweet dataset is the first research to use a deep learning architecture to learn semantic word embeddings to handle this complexity, outperforming the N-gram word method with ~18 F1 points. Research has also previously been conducted to detect Indonesian hate speech [7]–[9]. In the previous study [9], Random Forest Decision Tree (RFDT) with Label Power-set (LP) as a transformation method provides the best accuracy with fast computational time in general. The research [8] used the Latent Dirichlet Allocation (LDA) algorithm, and the F-measure of 93,5% was achieved when using the word n-gram feature with Random Forest. Word N-gram outperformed the character n-gram in research [7].

Accepted by Editor: 22-09-2021 | Final Revision: 22-11-2021 | Online Publication: 30-12-2021

Hate detection using GloVe has been carried out with Deep Belief Network (DBN) algorithm [10], which weighs the GloVe feature to improve accuracy before classification with 86% accuracy and 85,42% Fl-Score. The superiority of the newly trained GloVe model was also demonstrated in the study [11], outperforming the pre-trained word embeddings model (5,9% higher, 69,13% compared to 63,2%).

Several studies on Feature Expansion have been carried out previously using word2vec, intended for topic classification [4] and Twitter sentiment analysis [12]. In research [4], feature expansion with Google News datasets can improve performance consistently when using LR. The performance of LR classification with feature expansion was also obtained with an accuracy rate of 98,81% compared to Naïve Bayes (82,4%) and SVM (92,1%) [12].

This research's main objective and focus are to implement feature expansion to reduce vocabulary mismatches in hate speech on Indonesian-language Twitter using GloVe. The researchers' steps included implementing feature extraction using Boolean features and TF-IDF, expanding features with GloVe, and selecting features related to the best model using Logistic Regression (LR), Random Forest (RF) algorithms, and Artificial Neural Networks (ANN). The limitation of the problem in this study is that the data used is Indonesian tweet data. Harsh words that lead to an individual or oneself are included in this study's definition of hate speech.

2. Research Method

The system plan of the hate speech detection is shown in Figure 1.

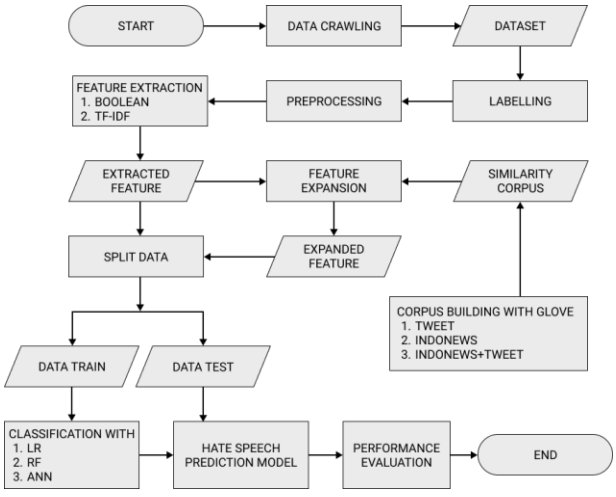

Figure 1. Hate Speech Detection System

2.1. Hate speech

Hate speech is all actions, both direct and indirect, based on hatred based on specific groups and incitement to individuals or groups through various means [13]. In Indonesia, crimes regarding hate speech have been included in UU ITE Number 11 of 2008 Article 45 Paragraph 2, imprisonment for a maximum of 6 years or a fine of one billion rupiahs [3]. Based on the Circular Letter of the Chief of Police Number: SE/6/X/2015 section 2f, hate speech is a criminal act in the form of insults, defamation, blasphemy, unpleasant actions, provoking, inciting, and spreading false news or hoaxes [13].

Based on descriptive research that has been done on Facebook, the form of hate speech in the context of speech is found to be the most common form of hate speech regarding blasphemy, and in the comment's column, it is found that the condition of insult is reproachful [14].One of the mediums to express hate speech is social media networks. With the rapid circulation of data and information on social networks, it is easier for individuals to push specific issues and spread hate speech which will cause a commotion among netizens. With that, the anonymity and mobility facilitated by the Internet have made harassment and hate speech easy to express in an abstract landscape and beyond the realm of law enforcement systems to control. By combining legal intervention with technology and regulatory mechanisms, the harm caused by online hate speech could be reduced [15].

2.2. Data Crawling

The dataset used is derived from the crawling results of Twitter in Indonesian using the Application Programming Integration (API) Key that the Twitter Developer has provided. In the crawling process, tweets with keywords are taken based on topics. The tweet topics used are determined based on trending topics during the crawling data period (October 2020 - June 2021), such as Omnibus Law, Religion, and Controversial Figures. Explicit words are determined to be topics based on harsh words as one of the characteristics of hate speech. From the crawling results, 20.601 tweets were collected.

Table 1. Crawling Keyword List

Topic	Keyword	Total
Religion	*Agama, FPI*	6.657
Explicit words	*Anjing, babi, bajingan, bangsat, gila, goblok, kontol, lonte, pantek, tai, tolol.*	9.186
Controversial Figure	*Tirta hudhi, selebgram*	3.584
Omnibus Law	*Omnibuslaw*	1.174

2.3. Data Labelling

Data labeling is carried out on the dataset before the data is preprocessed. Tweets are detected as hate speech intended to attack individuals (oneself or others) and groups containing abusive words. Each data in the dataset will be labeled with Hate Speech (HS) or Not Hate Speech (NHS), where HS means that the tweet

DOI: https://doi.org/10.29207/resti.v5i6.3521

contains hate speech, and the NHS does not contain hate speech. The data labeling process was carried out by four people in which three people were responsible for one class produced. Classes are determined by most of the votes of the three people involved. Assume that the first person detects a tweet is hate speech. If the second and third people detect that the tweet is not hate speech, the class will be determined not to be hate speech. Table 1 shows the labeling distribution of the class.

Table 2. Labeling Distribution

Label	Total
Hate Speech	10.101
Not Hate Speech	10.500

2.4. Data Preprocessing

Data preprocessing is one of the methods in data processing. The quality of the data will be improved by going through a series of "cleaning" methods to ensure that the results of the classification process can be more accurate [16]. Tweet data from crawling is a text with no structure that usually contains significant noise (data that does not collect valuable information for existing analysis). In this research, six stages are passed in data preprocessing. (1) Data Cleaning: eliminates noise, including emoticons, punctuation marks, and numbers to reduce unnecessary data information. (2) Case folding is the conversion of capital letters to lowercase letters [17]. (3) Stop Words Removal removes or filters words with no importance or irrelevant in the classification process [18]. (4) Normalization or normalization is standardizing words that have almost the same meaning. (5) Stemming: word substitution into a basic word by removing the affix. (6) Tokenizing or tokenization is the process of breaking sentences into words, phrases, and symbols called tokens, where the tokens generated will assist in parsing and processing data [19].

Figure 2. Word Cloud of Preprocessed Data

2.4. Feature Extraction

In the feature extraction stage, a tweet will be represented. The representation of tweets in this study uses a fixed-length Boolean vector feature, with each feature indicating the presence or absence of a word in the tweet [4].

2.5. N-gram

N-gram is a statistical modeling language for processing a text consisting of a sequence of items, where n indicates the length of the series (if n = 1 is called a Unigram, n = 2 is called a Bigram, n = 3 is a Trigrams) [20]. The tweet representation process begins with modeling the N-gram feature, including Unigram, Bigram, and Trigram.

2.6. Term Frequency-Inverse Document Frequency (TF-IDF)

Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical approach that functions in giving weight to a word according to its level of importance in the corpus of the document [21]. In the feature extraction stage, tweets based on keywords are given weighting with a TF-IDF score. After calculating the W weights for each document, the W weights are sorted to determine the degree of similarity between the documents and keywords (the greater the W value, the higher the similarity level). In this case, TF-IDF is used in the second vectorization method. The weight W in the TF-IDF calculation can be formulated in equation 1, with t_{fij} as the number of words searched in a document and $Id\,f_j$ as Inverse Document Frequency.

$$W_{ij} = t_{fij} \times Id\,f_j \quad (1)$$

2.7. Global Vectors (GloVe)

Global Vectors (GloVe) is a global log-bilinear regression model for unsupervised learning of word representation that outperforms other models in analogy, word equations, and named entity detection by Stanford University. GloVe is a model that takes advantage of the benefits of count data while simultaneously capturing meaningful linear substructures common in log-bilinear prediction-based methods. Compared to word2vec (when its corpus, vocabulary, window size, and training time are the same), GloVe consistently outperforms word2vec; in addition to achieving better and faster results, the best results are also obtained regardless of the speed [5]. In a study comparing GloVe with other word-embedding models such as Continuous bag-of-words, Skip-gram, and Hellinger PCA, it was found that GloVe is the best model compared to other models because GloVe can be scaled to a large corpus with good performance (even with a small corpus), thereby improving the quality of the learned representation by normalizing the sum and log smoothing [22].

2.8. Building Similarity Corpus

The corpus is created based on each word contained in the tweet data. From each of these words, a similarity corpus is built with GloVe. To find the similarity between words, tweet data, IndoNews, and a combination of the two are used. IndoNews data was previously used in a study [4] that included Indonesian

DOI: https://doi.org/10.29207/resti.v5i6.3521

media articles. In this study, thirty epochs and four threads were used in training the GloVe model. The corpus similarity was developed three times with different datasets, namely Tweet data, IndoNews data, and a combination of the two. Table 3 shows examples of vocabulary similar to "LGBT" in the similarity corpus built from the IndoNews dataset.

Table 3. Top 10 Vocabulary Similar To "LGBT"

LGBT		
Rank	Word	Value
1	*Transgender*	0,8144
2	*Lesbian*	0,7483
3	*Biseksual*	0,7316
4	*Gay*	0,6309
5	*Perilaku*	0,6234
6	*Simpang*	0,5538
7	*Kaum*	0,5306
8	*Anti*	0,5287
9	*Propaganda*	0,5121
10	*Rasisme*	0,5116

Table 3 explains that the ranking is obtained from the similarity value generated by GloVe for the highest Rank-1 to Rank-10 with the lowest value. Table 4 shows the number of vocabulary in each corpus that has been built.

Table 4. Number of Vocabulary in Corpus

Corpus	Word Count
Tweet	19.385
IndoNews	278.347
Tweet, IndoNews	286.484

2.9. Feature Expansion

The feature expansion method is used to solve the problem of data distribution in corpus-based supervised Word Sense Disambiguation. Feature expansion can effectively fix the low retrieval efficiency caused by word ambiguity in short queries [23]. The concept of feature expansion is to identify missing words in the tweet representation, substituted with semantically related words [4]. This research implementation of feature expansion is based on research [4], [12]. The following algorithms show the steps of the feature expansion based on the prior study.

Algorithm 1. Feature Expansion Boolean Version [4]

```
Input: Text Vectors, Similarity Corpus
Output: Expanded Text Vector
Initialization i, j
Get max
  max=size(Text Vectors)
  for i = 0 to max do
    for j = 0 to size(Text Vectors[i]) do
      if Text Vectors[i][j] = 0:
          check = checkSimilarity(features[j])
          if check == True:
                Text Vectors[i][j] = 1
          end if
        end if
      end for
    end for
```

Algorithm 2. Feature Expansion TF-IDF Version

```
Input: Text Vectors, Similarity Corpus
Output: Expanded Text Vector
  Initialization i, j
  Get max
    max=size(Text Vectors)
    for i = 0 to max do
      for j = 0 to size(Text Vectors[i]) do
        if Text Vectors[i][j] == 0.0:
          check = checkSimilarity(features[j])
          if check != Null:
            Text Vectors[i][j] = weight(check)
          end if
        end if
      end for
    end for
```

2.10. Logistic Regression (LR)

Logistic Regression (LR) is a type of regression that connects the independent (independent) and dependent (category) variables. LR and ANN are currently the most widely used biomedical models (based on the number of publications indexed on MEDLINE: 28,500 for LR, 8500 for ANN). Both come from different communities (statistics and computer science) but have much in common [24]. LR can predict the presence of a characteristic/outcome based on the value of a set of predictor variables, like Linear Regression, and is suitable for the dichotomous dependent variable model (nominal data scale with two categories) [24]. The following is a class membership probability formula for one of the two categories in the data set in the LR model, with P as the logistic function value and x as the input data value.

$$P\,(0 \mid x, \alpha) = 1 - P\,(1 \mid x, \alpha) \quad (2)$$

$$P\,(1 \mid x, \alpha) = \frac{1}{1 + e^{-(\alpha \cdot x)}} \quad (3)$$

As for parameters, , we conducted a trial in the LR model with C = 1.0, 100 maximum iteration, *newton-cg* as the solver, and multinomial logistic regression.

2.11. Random Forest (RF)

Random Forest is a combination of tree predictors. Each tree depends on the value of a random vector whose sample is obtained with a uniform distribution independently for all trees in the forest [25]. Random Forest was introduced by Ho (1995) by combining many trees in the training data to produce a high level of accuracy [26]. The starting point of the tree is the root node, while the end where the chain ends is called the leaf node. A node represents a particular characteristic, whereas a branch represents a range of values [27]. In the RF partition, we divide the datasets into test and training sets. Each tree will form in-bag data with a subset of the training data and out-of-bag from the remaining parts [28].

DOI: https://doi.org/10.29207/resti.v5i6.3521

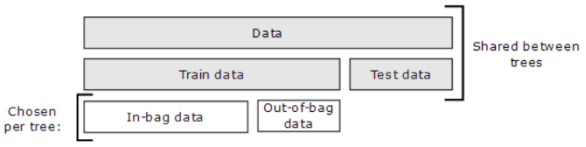

Figure 3. Data Partition of RF [28]

In this study, we use the *Scikit-learn* library to build RF. It uses an optimized version of the CART algorithm to build decision trees. Binary trees are constructed in the CART algorithm using threshold and the feature which yield the largest information gain at each node. As for parameters used in this study, we conducted a trial in which we did not give maximum depth of the tree; thus, the nodes expanded until it contained less than the minimum number of samples required to split an internal node. We used the bootstrap samples when building trees.

2.12. Artificial Neural Network (ANN)

Artificial Neural Network, commonly referred to as ANN, is a neural network model that is a branch of artificial intelligence, consisting of many interconnected simple processors (neurons) that work in parallel in the network [29]. ANN teaches systems to perform tasks instead of programming computational approaches to perform specific tasks. The teaching mode can be either supervised or unsupervised. Neural Networks learn in the presence of noise [30]. In this study, the Multi-layer Perceptron (MLP) model is used as a class of ANN [31]. MLP consists of three or more layers (input and output layers with one or more hidden layers) of nonlinearly activating nodes. Each node in one layer relates to a certain weight to each node in the next layer.

As for parameters, we set a trial with the ANN model's parameters that used hidden layer sizes = (8, 8, 8) (three hidden layers of 8, 8, 8 units respectively), with alpha = 1e-5 and stochastic gradient-based optimizer by Kingma, Diederik, and Jimmy Ba as the solver for weight optimization.

2.13. Performance Evaluation

Confusion Matrix represents how often a behavior is correctly detected and classified as a class [32]. In the confusion matrix, a result correctly classified in the positive class is called True Positive (TP) and correctly categorized into the negative class True Negative (TN). Meanwhile, the positive class is classified as False Negative (FN) and the negative class as positive False Positive (FP). From the frequency of the four components, an indicator of the classifier's performance in detecting a given class can be obtained by calculating accuracy, precision, recall, and F1-Score in the built algorithm. In this study, accuracy and F1-Score were obtained through the average of the program execution results in five iterations. Here are the equations of accuracy, precision, recall, and F1-Score.

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (4)$$

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP+TN}{TP+FN} \quad (6)$$

$$F1-Score = \frac{2*Precision*Recall}{Precision+Recall} = \frac{2*TP}{2*TP+FP+FN} \quad (7)$$

3. Result and Discussion

This research is divided into three test scenarios for each classification model with LR, RF, and ANN. Accuracy results are obtained through the average of the results of program execution five times. The first scenario implements feature extraction using Boolean features and TF-IDF. The second application of feature expansion with corpus similarity was built with GloVe. The similarity corpus consists of three types (Tweet, IndoNews, and a combination of the two) and their sub-combinations using rankings of one (Top 1), five (Top 5), and top ten (Top 10) in the ranking of similarity between words. The third scenario applies feature selection to compare data with 5.000, 10.000, 15.000, and 19.370 feature vectors.

3.1. Results

In the first scenario, feature extraction is performed using boolean features as baseline and TF IDF. Table 5 is the result of evaluating the performance ratio of the boolean features of the RF, LR, and ANN models with each test size ratio of 0,1, 0,2, and 0,5 with 19.370 features.

Table 5. Baseline Ratio Performance Value

Model	Test Size	Accuracy (%)
LR	0,5	86,63
	0,2	87,20
	0,1	**86,94**
RF	0,5	86,29
	0,2	86,95
	0,1	**87,15**
ANN	0,5	81,63
	0,2	82,20
	0,1	**82,36**

Table 5 shows the highest accuracy obtained at the test size ratio of 0,1 or 10% of the overall tweet data. The next step is to determine the optimal n-gram at the Baseline. The evaluation is limited to 5.000 features for a sample to overcome the runtime memory usage, which is quite large. The following are the results of the evaluation of the performance of the N-gram compared to the Baseline with a test size of 10% and a training size of 90%.

Table 6. N-gram Performance

N-gram	Accuracy (%)		
	LR	RF	ANN
Unigram	**86,94**	**87,34**	**83,34**
Bigram	78,43	77,90	76,99
Trigram	72,90	72,45	73,17

DOI: https://doi.org/10.29207/resti.v5i6.3521

From Table 6, it can be concluded that Unigram with a test size of 0,1 in each classifier proved to have the highest accuracy compared to Bigram and Trigram, respectively 86,94% for LR, 87,34% for RF, and 83,34% for ANN. Unigram will be applied with a test ratio of 0,1 as the Baseline for the following scenario. Furthermore, in feature extraction, weighting is carried out with TF-IDF on the baseline vector with the experimental results in Table 7.

Table 7. *Baseline* + TF-IDF Performance

Classifier	Accuracy (%)	F1 Score
LR	87,18 **(+0,24)**	0,8716
RF	88,03 **(+0,69)**	0,8801
ANN	81,92 (-1,42)	0,8192

Table 7 shows increased accuracy with the application of TF-IDF on the LR (0,24%) and RF (0,69%). There is a decrease in accuracy of 1,42% on the ANN.

The second scenario applies feature expansion with corpus similarity consisting of three types (corpus tweet, IndoNews, and a combination of the two) and their sub-combinations using Top 1, Top 5, and Top 10 similarity between words.

Table 8. GloVe Performance with Baseline on LR

Rank	Accuracy (%)		
	Baseline + Tweet	Baseline + IndoNews	Baseline + Tweet, IndoNews
Top 1	86,49 (-0,45)	87,01 **(+0,07)**	86,71 (-0,23)
Top 5	85,87 (-1,07)	86,10 (-0,83)	86,71 (-0,23)
Top 10	85,68 (-1,26)	86,10 (-0,83)	86,37 (-0,57)

Table 8 shows the performance of GloVe on the LR classifier. The decrease in accuracy occurs when using the entire Tweet corpus, Top 5 & 10 on the IndoNews corpus, and the entire combination of Tweet and IndoNews corpus. The highest increase of 0,07% was achieved by Top 1 with the IndoNews corpus.

Table 9. GloVe Performance with Baseline on RF

Rank	Accuracy (%)		
	Baseline + Tweet	Baseline + IndoNews	Baseline + Tweet, IndoNews
Top 1	87,07 (-0,08)	86,11 (-1,04)	87,47 **(+0,32)**
Top 5	87,25 **(+0,10)**	85,61 (-1,54)	87,14 (-0,01)
Top 10	87,59 **(+0,44)**	86,25 (-0,90)	87,16 **(+0,01)**

Table 9 shows the performance of GloVe on the RF classifier. The decrease in accuracy appears when using the Top 10 in the IndoNews corpus and a combination of Tweet and IndoNews corpus. The highest increase of 0,44% was achieved by Top 10 with Tweet corpus.

Table 10. GloVe Performance with Baseline on ANN

Rank	Accuracy (%)		
	Baseline + Tweet	Baseline + IndoNews	Baseline + Tweet, IndoNews
Top 1	83,43 **(+1,07)**	82,13 (-0,22)	83,70 **(+1,34)**
Top 5	84,15 **(+1,80)**	83,52 **(+1,16)**	84,51 **(+2,15)**
Top 10	84,43 **(+2,08)**	84,56 **(+2,20)**	84,73 **(+2,37)**

Table 10 shows the performance of GloVe on the ANN classifier. The decrease in accuracy only occurs when using Top 1 in the IndoNews corpus. The highest increase of 2,37% was achieved by Top 10 with the combination of Tweet and IndoNews corpus.

Table 11. GloVe Performance with Baseline, TF-IDF on LR

Rank	Accuracy (%)		
	Baseline + Tweet	Baseline + IndoNews	Baseline + Tweet, IndoNews
Top 1	87,35 **(+0,41)**	86,62 (-0,32)	87,18 **(+0,24)**
Top 5	86,89 (-0,05)	86,97 **(+0,03)**	86,95 **(+0,01)**
Top 10	87,27 **(+0,33)**	87,13 **(+0,19)**	87,01 **(+0,07)**

Table 11 shows the performance of GloVe on the TF-IDF and LR classifiers. The decrease in accuracy occurs when using the Tweet corpus in the Top 5 and the IndoNews corpus in the Top 1. The highest increase of 0,41% was achieved by Top 1 with Tweet corpus.

Table 12. GloVe Performance with Baseline, TF-IDF on RF

Rank	Accuracy (%)		
	Baseline + Tweet	Baseline + IndoNews	Baseline + Tweet, IndoNews
Top 1	87,77 **(+0,62)**	87,04 (-0,12)	87,18 **(+0,03)**
Top 5	87,39 **(+0,24)**	86,74 (-0,41)	87,82 **(+0,67)**
Top 10	87,55 **(+0,40)**	86,20 (-0,95)	88,01 **(+0,85)**

Table 12 shows the performance of GloVe on the TF-IDF and RF classifier. Decreased accuracy occurs when using the IndoNews corpus. The highest increase of 0,85% was achieved by the Top 10 combined Tweet and IndoNews corpus.

Table 13. GloVe Performance with Baseline, TF-IDF on ANN

Rank	Accuracy (%)		
	Baseline + Tweet	Baseline + IndoNews	Baseline + Tweet, IndoNews
Top 1	82,40 **(+0,04)**	81,71 (-0,65)	83,24 **(+0,88)**
Top 5	82,23 (-0,13)	81,32 (-1,04)	81,80 (-0,56)
Top 10	81,87 (-0,49)	81,84 (-0,51)	82,91 **(+0,55)**

Table 13 shows the performance of GloVe on the TF-IDF and ANN classifiers. The decrease in accuracy occurred when using the IndoNews corpus, Top 5 combination of Tweet and IndoNews corpus, Top 5 and Top 10 in the Tweet corpus. The highest increase of 0,88% was achieved by the Top 10 combined Tweet and IndoNews corpus.

The third scenario applies feature selection, where there is a comparison between data with 5.000, 10.000, 15.000, and 19.370 TF-IDF feature vectors using RF. After that, we apply feature expansion to the number of features with the highest accuracy.

Table 14. Performance Comparison on Number of Features

Classifier	Accuracy (%)	F1 Score
5.000	**89,08**	**0,891**
10.000	88,03	0,880
15.000	87,02	0,870
19.370	88,03	0,880

DOI: https://doi.org/10.29207/resti.v5i6.3521

Based on Table 14, it can be concluded that the combination of TF-IDF with 5.000 features has the highest accuracy. Table 15 describes the performance values from the Baseline with 5.000 features.

Table 15. *Baseline* with 5.000 Features Performance

Classifier	Accuracy (%)	F1 Score
LR	87,10	0,8710
RF	87,34	0,8734
ANN	83,34	0,8333

Table 16. GloVe with Baseline, TF-IDF on LR (5,000 Features)

Rank	Accuracy (%)		
	Baseline + Tweet	Baseline + IndoNews	Baseline + Tweet, IndoNews
Top 1	87,56 **(+0,46)**	86,59 (-0,51)	87,67 **(+0,56)**
Top 5	87,41 **(+0,31)**	87,53 **(+0,43)**	86,97 (-0,14)
Top 10	87,41 **(+0,31)**	87,25 **(+0,15)**	87,11 **(+0,01)**

Table 16 shows the performance of GloVe on the TF-IDF and LR classifiers. A decrease in accuracy occurs when using the Top 1 corpus of IndoNews and the Top 5 corpus with the combination of Tweet and IndoNews corpus. The highest increase of 0,56% was achieved by Top 1 with the combination of Tweet and IndoNews corpus.

Table 17. GloVe with Baseline, TF-IDF on RF (5,000 Features)

Rank	Accuracy (%)		
	Baseline + Tweet	Baseline + IndoNews	Baseline + Tweet, IndoNews
Top 1	88,59 **(+1,25)**	87,30 (-0,04)	88,05 **(+0,72)**
Top 5	88,12 **(+0,79)**	86,43 (-0,90)	87,42 **(+0,09)**
Top 10	87,54 **(+0,20)**	86,83 (-0,50)	87,42 **(+0,50)**

Table 17 shows the performance of GloVe on the TF-IDF and RF classifier. There is no increase in accuracy when using the IndoNews corpus. The highest gain in accuracy of 1,25% was achieved by Top 1 with Tweet corpus.

Table 18. GloVe with Baseline, TF-IDF on ANN (5,000 Features)

Rank	Accuracy (%)		
	Baseline + Tweet	Baseline + IndoNews	Baseline + Tweet, IndoNews
Top 1	83,00 (-0,34)	83,94 **(+0,60)**	83,18 (-0,16)
Top 5	83,46 **(+0,13)**	83,44 **(+0,10)**	83,30 (-0,04)
Top 10	83,20 (-0,14)	83,43 **(+0,09)**	83,30 **(+0,19)**

Table 18 shows the performance of GloVe on the TF-IDF and ANN classifier. There is no increase in accuracy when using the Top 1 and Top 10 of the IndoNews corpus and the Top 1 and Top 5 of the Tweet corpus, IndoNews. The highest increase of 0,6% was achieved by Top 1 with the IndoNews corpus.

3.2. Discussion

Based on the results of the tests, the RF and ANN classifiers most often experience an increase in accuracy after feature expansion, which is 16 increases compared to LR with 15 increases. RF achieves higher accuracy than other classifier models. The highest increase in accuracy in the feature expansion of the ANN model occurred in the combination of 19.370 features on Baseline + IndoNews, with an accuracy value of 84,73% and an increase of 2,37%. Meanwhile, the highest accuracy was achieved by RF in Top 1 with GloVe Tweet corpus, TF-IDF, and 5.000 features at 88,59%. Therefore, the classification algorithm that worked better and had the most influence on the feature expansion are RF and ANN. This result proves that feature selection and the weighing method with TF-IDF is responsible for the RF model achieving the best accuracy compared to the Baseline with the same similarity corpus and rank (87,07%). On the contrary, ANN performs better when we don't implement it.

The combination of the Tweet and IndoNews is the similarity corpus with the most increase in accuracy compared to the Baseline (19 increased accuracy). Meanwhile, the Tweet corpus has the most accuracy increase against the Baseline with 5.000 features (7 improvements in accuracy), followed by the combined corpus (6 improvements) and the IndoNews corpus (5 improvements). Thus, the dataset used as similarity corpus that has the most influence on the overall Baseline is the combination of the Tweet and IndoNews. After we implemented the feature selection, the most influential corpus in increasing accuracy is the Tweet similarity corpus. The implementation of feature selection has proven to improve the system's performance scaled by each accuracy compared to the Baseline.

4. Conclusion

In this study, research on the detection of hate speech on Indonesian Twitter has been carried out. The researcher applies feature expansion using a word embedding Global Vectors (GloVe) to overcome vocabulary mismatches. Researchers apply this approach using a collection of 20.601 Indonesian tweet data. A corpus of similarity was developed, which was needed in the feature expansion process with the tweet and IndoNews data with GloVe. The implementation of feature extraction uses Boolean features and TF-IDF. After that, we perform feature expansion and selects features related to the best model using Logistic Regression (LR), Random Forest (RF), and Artificial Neural Network (ANN) algorithms.

The results show that the Random Forest model with 5.000 features and a combination of TF-IDF and Tweet corpus built with GloVe produce the best accuracy rate between the other models with an average of 88,59% accuracy score, which is 1,25% higher than the predetermined Baseline. The highest increase of average accuracy was obtained by ANN with 84,73% accuracy,

DOI: https://doi.org/10.29207/resti.v5i6.3521

gaining 2,37% accuracy with the Top 10 from the combination of Tweet and IndoNews similarity corpus built with GloVe compared to the Baseline. The corpus that has the most influence on the overall Baseline is the combination of the Tweet and IndoNews corpus. Meanwhile, after the feature selection, the most influential corpus in increasing accuracy is the Tweet similarity corpus. Based on the research results, in the feature expansion of the RF and ANN classifiers, the accuracy increases the most after the feature expansion, with RF achieving higher accuracy than the others. The number of features used proven to improve the performance of the system.

References

[1] T. Shi and Z. Liu, "Linking GloVe with word2vec," vol. arXiv prep, p. 1, 2014, [Online]. Available: http://arxiv.org/abs/1411.5595.

[2] B. Heller and L. Magid, "Parent's and Educator's Guide to Combating Online Hate Speech | ConnectSafely." https://www.connectsafely.org/hatespeech/ (accessed Nov. 22, 2020).

[3] Republik Indonesia, "Undang-Undang ITE," 2008. https://peraturan.go.id/common/dokumen/ln/2008/UU 11 Tahun 2008.pdf (accessed Nov. 22, 2020).

[4] E. B. Setiawan, D. H. Widyantoro, and K. Surendro, "Feature expansion using word embedding for tweet topic classification," in *Proceeding of 2016 10th International Conference on Telecommunication Systems Services and Applications, TSSA 2016: Special Issue in Radar Technology*, 2017, p. 1, doi: 10.1109/TSSA.2016.7871085.

[5] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," 2014, doi: 10.3115/v1/d14-1162.

[6] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," 2019, doi: 10.1145/3041021.3054223.

[7] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, "Hate speech detection in the Indonesian language: A dataset and preliminary study," 2018, doi: 10.1109/ICACSIS.2017.8355039.

[8] T. Febriana and A. Budiarto, "Twitter Dataset for Hate Speech and Cyberbullying Detection in Indonesian Language," 2019, doi: 10.1109/ICIMTech.2019.8843722.

[9] M. O. Ibrohim and I. Budi, "Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter," 2019, doi: 10.18653/v1/w19-3506.

[10] I. Z. Muhammad, M. Nasrun, and C. Setianingsih, "Hate Speech Detection using Global Vector and Deep Belief Network Algorithm," 2020, doi: 10.1109/ibdap50342.2020.9245467.

[11] B. Vidgen and T. Yasseri, "Detecting weak and strong Islamophobic hate speech on social media," *arXiv*. 2018.

[12] E. B. Setiawan, D. H. Widyantoro, and K. Surendro, "Feature expansion for sentiment analysis in twitter," 2018, doi: 10.1109/EECSI.2018.8752851.

[13] Komnas HAM, "Buku Saku Penanganan Ujaran Kebencian (Hate Speech)," in *Komisi Nasional Hak Asasi Manusia*, 2015, pp. 9, 13.

[14] D. J. Ningrum, S. Suryadi, and D. E. Chandra Wardhana, "KAJIAN UJARAN KEBENCIAN DI MEDIA SOSIAL," *J. Ilm. KORPUS*, p. 1, 2019, doi: 10.33369/jik.v2i3.6779.

[15] J. Banks, "Regulating hate speech online," *Int. Rev. Law, Comput. Technol.*, p. 238, 2010, doi: 10.1080/13600869.2010.522323.

[16] A. Krouska, C. Troussas, and M. Virvou, "The effect of preprocessing techniques on Twitter sentiment analysis," 2016, doi: 10.1109/IISA.2016.7785373.

[17] M. A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, "Improving Text Preprocessing for Student Complaint Document Classification Using Sastrawi," 2020, doi: 10.1088/1757-899X/874/1/012017.

[18] G. Angiani *et al.*, "A comparison between preprocessing techniques for sentiment analysis in Twitter," 2016.

[19] S. Saha, J. Yadav, and P. Ranjan, "Proposed Approach for Sarcasm Detection in Twitter," *Indian J. Sci. Technol.*, 2017, doi: 10.17485/ijst/2017/v10i25/114443.

[20] E. Fehn Unsvåg and B. Gambäck, "The Effects of User Features on Twitter Hate Speech Detection," 2019, doi: 10.18653/v1/w18-5110.

[21] I. G. M. Putra and D. Nurjanah, "Hate Speech Detection In Indonesian Language Instagram," in *2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Oct. 2020, pp. 413–420, doi: 10.1109/ICACSIS51025.2020.9263084.

[22] S. Bhoir, T. Ghorpade, and V. Mane, "Comparative analysis of different word embedding models," in *International Conference on Advances in Computing, Communication and Control 2017, ICAC3 2017*, 2018, p. 3, doi: 10.1109/ICAC3.2017.8318770.

[23] N. L. Tsao, D. Wible, and C. H. Kuo, "Feature expansion for word sense disambiguation," in *NLP-KE 2003 - 2003 International Conference on Natural Language Processing and Knowledge Engineering, Proceedings*, 2003, p. 1, doi: 10.1109/NLPKE.2003.1275882.

[24] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: A methodology review," *J. Biomed. Inform.*, 2002, doi: 10.1016/S1532-0464(03)00034-0.

[25] L. Breiman, "Random forests," *Mach. Learn.*, p. 1, 2001, doi: 10.1023/A:1010933404324.

[26] T. K. Ho, "Random decision forests," 1995, doi: 10.1109/ICDAR.1995.598994.

[27] C. R. Sekhar, Minal, and E. Madhu, "Mode Choice Analysis Using Random Forrest Decision Trees," in *Transportation Research Procedia*, 2016, p. 6, doi: 10.1016/j.trpro.2016.11.119.

[28] N. Kuznetsova, "Random forest visualization Eindhoven University of Technology Master Thesis Random Forest Visualization," *Wald Lect. II, Dep. ofStatistics, Calif. Univ.*, 2014.

[29] G. F. Hepner, T. Logan, N. Ritter, and N. Bryant, "Artificial neural network classification using a minimal training set: comparison to conventional supervised classification," *Photogrammetric Engineering & Remote Sensing*. 1990.

[30] S. K and S. S, "Review on Classification Based on Artificial Neural Networks," *Int. J. Ambient Syst. Appl.*, 2014, doi: 10.5121/ijasa.2014.2402.

[31] P. Lewicki and T. Hill, "Statistics : Methods and Applications - A comprehensive reference for science, industry and data mining," in *StatSoft Inc.*, vol. 1, 2006.

[32] S. Ruuska, W. Hämäläinen, S. Kajava, M. Mughal, P. Matilainen, and J. Mononen, "Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle," *Behav. Processes*, 2018, doi: 10.1016/j.beproc.2018.01.004.

DOI: https://doi.org/10.29207/resti.v5i6.3521