PERFORMANCE ANALYSIS OF CLASSIFIER ON FACEBOOK DATA USING UNIGRAM & BIGRAM COMBINATIONS

Yudha Sulistiyo Wibowo [1], Mohammad Reza Faisal[2], Ahmad Rusadi[3], Dodon T nugrahadi[4], Muhammad Itqan Mazdadi[5]

[12345]Ilmu Komputer FMIPA ULM

Jl. A. Yani Km 36 Banjarbaru, Kalimantan Selatan

Email : J1f115227@mhs.ulm.ac.od

Abstract

This research on sentiment analysis uses the random forest method as classification. Tf-idf is a weighted feature and feature combination of n-grams is unigram and bigram as feature words. In this research tf-idf used for the extraction feature, this test uses facebook comment data about the sports news. In this study, datasets were used as much as 1000 data divided into 2, namely data testing and training data. Achieved high accuracy performance results in unigram features with an accuracy of 83.67% of 2757 features, bigram produces 58% with features as much as 8457.

Keyword: sentiment analysis, random forest, n-gram, unigram, bigram

1. Introduction

The growth of online media such as social media facebook began to grow tremendously. According to the survey conducted by APJII in 2016 recorded 54% of internet users in Indonesia [3]. While according to a survey conducted by statista.com facebook users in Indonesia in the year 2019 ranked 4th in the world with 120 million active users spread across Indonesia. Facebook users can create a status and can also provide comments made. Comments made are not always positive, but also negatively impacting public opinion in user-generated comments, from those comments can be identified in both positive and negative classes.

Research conducted on negative comments in social media, using the random forest decision tree classification, naive bayes, support vector machine, and bayesian logistic regression and the extraction feature is the n-gram word. One of them is hate speech detection in the Indonesian language: a dataset and, preliminary study by Alfina (2017) [1].

Tf-idf is used to make a weighted one of the research that uses tf-idf is by the title of application of cosine similarity and the installation of tf-id in the document classification system obtained the accuracy level of systematic classification of 98% by Wahyuni (2017) [5]. The research under the title Rachmat & Lukito (2016) in the study under the title classification the sentiment of political commentary from a facebook page using naive nayes, from the results of implementation and testing with the method naive bayes as well as using the weighing of tf-idf has an accuracy classification level sentiment reached more than 83%.

In research by Basari (2012) under the title opinion mining of movie review using the hybrid method of support vector machine and particle swarm

optimization, in the study used the svm method using n-grams compared to feature-weighted. 3 types of n-gram features are used namely, unigram, bigram, and trigram. With the weighing of features is tf and tf-idf, the results of the accuracy obtained are 73.13% (tf) and 72.20% (tf-idf) [2].
In the research that will be done, that is to see the performance analysis of the feature word unigram and bigram with the weighted use tf-idf which will be classified with random forest.

2. RESEARCH METHODS

This research procedure is as follows:
Below is a groove in the form of diagrams occurring in Figure 1.

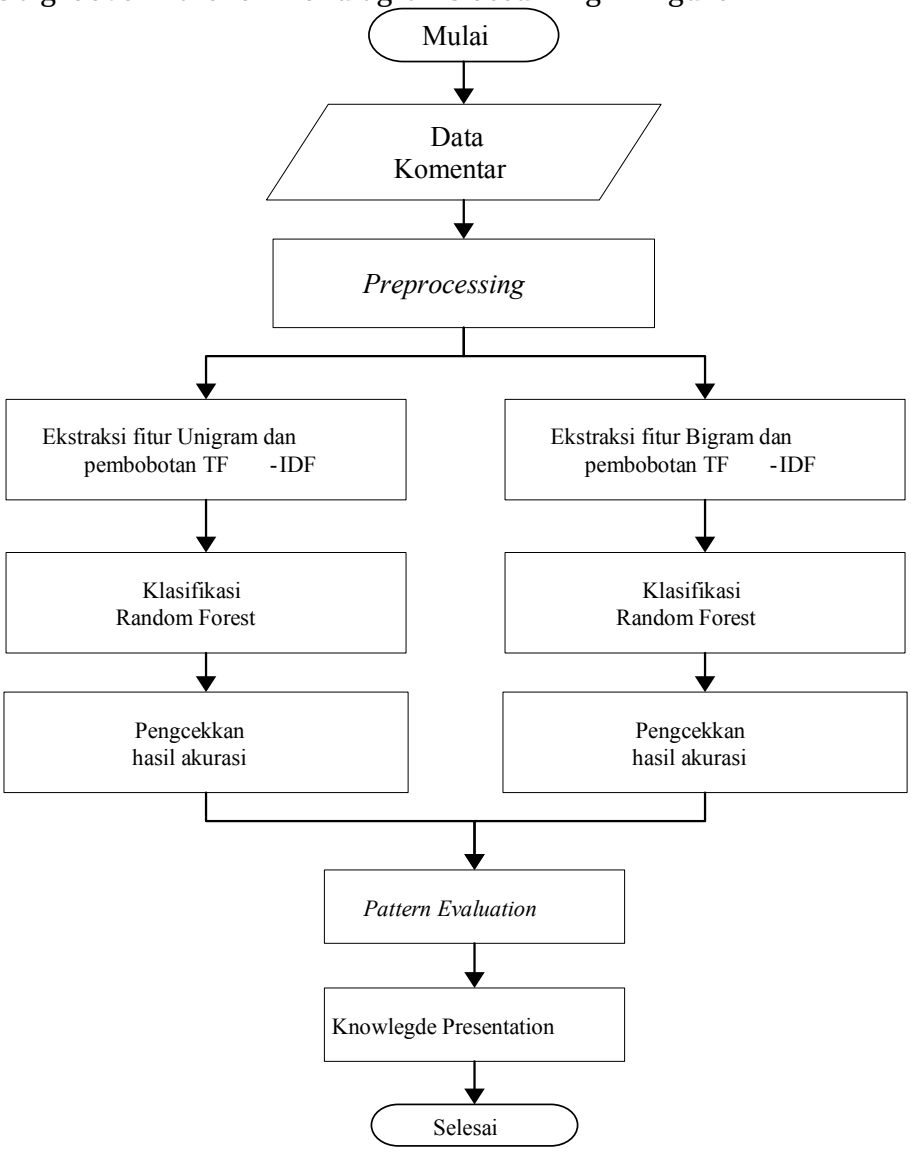


Figure 1. Research plot

The research procedures conducted in this research are as follows:

2.1 Comment Data Collection

On the data research used is the facebook comment data about the sports news from Hidayat research (2017). Where this data uses as many as 1000 comments on sports news.

2.2 Preprocessing

Preprocessing is a process for preparing data to be structured by converting data into an easy-to-process form. In this study, the preprocessing stage used is stemming & remove punctuation.

2.3 The process of weighted

The weighted for this study using tf-idf with its word feature is unigram and bigram

2.4 The random forest classification process

In this stage classification process is done to get the best accuracy results. The classification process is done after the preprocessing stage and the weighted feature has been done first.

2.5 Evaluation

Assessment of the results of data mining techniques that have been done before.

2.6 Knowledge presentation

Stages of the results of the analysis are formulated. This step is performed visualization and presentation of knowledge obtained.

3. RESULTS AND DISCUSSION

3.1 Results

3.1.1 Data collection

The data used contains comments on the facebook page compass on sports news which is data from Hidayat's research (2017). Data obtained as much as 1000 comments, and is divided into training data as many as 700 comments and test data (tests) as many as 300 comments. The comments have positive labels and negative and balanced labels between positive sentiments and negative sentiments. The comment Data is in table 1 below.

Tabel 1. Data collection

Id	From/ id	From/name	Created time	Message	Label
3078	10202 33577 43308 51	Fauzan Amri Oesman	2016-12-18	udah nampak di pertandingn melawan vietnam kebobolan dulu karena kebobolan dahulu masih juga bertahan maen lawan thailand, cepat berpuas diri dengan hasil 2-1... akhirnya kebobolan 2 gol..	Negatif
3093	49761 34104 03210	Elly Egi Tanjung	2016-12-18	Reformasi TIMNAS..!! Pecat Riedl..!! Cari coach baru yg Bagus.!	Negatif
3201	61243 33222 36734	Rivaldo Iglesias	2016-12-18	TERIMA KASIH OPA RIDEL ! YOU STILL THE BEST COACH FOR INDONESIA !	Positif
3028	17079 1 69261 53886	M Semper	2016-12-18	Udah terlambat pak tua Pensiun aja. Istirahat!!!	Negati ve
3222	49446 53751 878	Haryono Wawan	2016-12-18	Betul....jgn gampang bongkar pasang tim....kalo diksh wkt lbh panjang tim ini bisa lbh solid. Dg kondisi skrg aja bisa no 2... Bisa jadi tahun depan lbh bagus...bravo tim nas!	Positif
3248	10201 30359 69110 22	Sammar Samna	2016-12-18	Semangat yg luar biasa,ayo anak2 bangsa jgn kalah semangat dgn riedl ..	Positif
3253	45841 22709 84292	Rikki Indriantoo	2016-12-18	tetap semangat opa, bangunlah timnas sepakbola indonesia sekompak dan sesukses mungkin	Positif
3257	10857 39784 81055 9	Oge Sarioa Piters	2016-12-18	Dukung opa alfred untuk latih timnas terus,biar bisa bentuk timnas yg lebih bagus n kuat	Positif

3.1.2 Preprocessing

3.1.2.1 stemming & Remove Punctuation.

Stemming is the process of getting a basic word by eliminating the suffix that exists in each text. Remove punctuation aims to remove all non-alphabet characters such as dots (.), commas (,), spaces, and others.

Below is the result of the preprocessing stage found in table 1.

Tabel 1. Preprocessing Comment Data

Data Komentar		**Label**
Input	**Output**	
sudah tampak di tanding lawan vietnam bobol dulu karena bobol dahulu masih juga tahan main lawan thailand cepat puas diri dengan hasil dua 1 akhir bobol dua gol	sudah tampak di tanding lawan vietnam bobol dulu karena bobol dahulu masih juga tahan main lawan thailand cepat puas diri dengan hasil dua satu akhir bobol dua gol	Negatif
Reformasi TIMNAS..!! Pecat Riedl..!! Cari coach baru yg Bagus.!	Reformasi timnas pecat riedl cari latih baru yang bagus	Negatif
TERIMA KASIH OPA RIDEL ! YOU STILL THE BEST COACH FOR INDONESIA !	terima kasih opa riedl you still the best latih for Indonesia	Positif
Udah terlambat pak tua Pensiun aja. Istirahat!!!	Sudah lambat bapak tua pensiun saja istirahat	Negatif
Se7	Tuju	Positif

aq bangga melihat timnas sampai final . padahal persiapan timnas mepet tapi bisa masuk final . terima kasih timnas indonesia sukses selalu	aku bangga lihat timnas sampai final padahal siap timnas mepet tapi bisa masuk final terima kasih timnas indonesia sukses selalu	Positif
Sy dukung opa Riddle melatih kembali.. beri beliau kesempatan jangka panjang menangani timnas ingat prestasi	saya dukung opa riedl latih kembali beri beliau sempat jangka panjang tangan timnas ingat prestasi	Positif
butuh waktu bukan instant.	butuh waktu bukan instant	
Saya sependapat dengan teman2 semua.lebih baik coach riedl di pertahn dlm dua tahun kedepan..meskipun terkendala dengam 2 pemain yg di setujui club nyatanya riedl mmpu mencapai puncak mskipun juara.	saya dapat dengan teman semua lebih baik latih riedl di tahan dalam dua tahun depan meski kendala dengan dua main yang di tuju klub nyata riedl mampu capai puncak meski juara	Positif
Dengan pemain seadanya Bisa membawa timnas ke final Sudah pencapaian yg bagus...	dengan main ada bisa bawa timnas ke final sudah capai yang bagus	Positif
saya setuju riedl tetap menukangi timnas,saya yakin timnas bisa juara.	saya tuju riedl tetap tukang timnas saya yakin timnas bisa juara	Positif

Harus di beri kepercayaan lg tuk melatih timnas dan beri keluasan penuh tuk menentukan pemain....bravo Riedle.....PSSI	harus di beri percaya lagi untuk latih timnas dan beri luas penuh untuk tentu main bravo riedl pssi	
Orang indonesia itu sukanya instan. Tapi sebenernya semua butuh persiapan. Om reild tolong bantu timnas untuk menyambut gelar juara	orang indonesia itu suka instan tapi benar semua butuh siap om riedl tolong bantu timnas untuk sambut gelar juara	
Saya sependapat dengan teman2 semua.lebih baik coach riedl di pertahn dlm dua tahun kedepan..meskipun terkendala dengam 2 pemain yg di setujui club nyatanya riedl mmpu mencapai puncak mskipun juara.	saya dapat dengan teman semua lebih baik latih riedl di tahan dalam dua tahun depan meski kendala dengan dua main yang di tuju klub nyata riedl mampu capai puncak meski juara	Positif
Tangan dingin Alfred Riedl meskipun hanya beberapa bulan menangani timnas sungguh mengagumkan ,bisa mencapai final walau hrs kalah dengan agregat 3-1.Terima kasih oom.	tangan dingin alfred riedl meski hanya beberapa bulan tangan timnas sungguh kagum bisa capai final walau harus kalah dengan agregat 3 1 terima kasih om,positif	Positif

3.1.2.2 The process of weighted

Tram frequency inverse document frequency (tf-idf) as a weighted term drawn from how many occurrences of the term in a document and n-gram used is unigram and bigram for word feature on comment data. See table 2 and 3 below.

Tabel 2. Tf-idf- weighted process in unigram

Bilang	bina	Binasah	bintang	Bis	Bisa	Label
0	0	0	0	0	0	Negatif
0	0	0	0	0	0	Negatif
0	0	0	0	0	0	Negatif
0	0	0	0	0	0	Negatif
0	0	0	0	0	0.246.	Negatif
0	0	0	0	0	0	Positif
0	0	0	0	0	0	Positif
0	0	0	0	0	0	Negatif
0	0	0	0	0	0	Negatif
0	0	0	0	0	0	Negatif
0.688	0	0	0	0	0	Negatif
0	0	0	0	0	0	Negatif

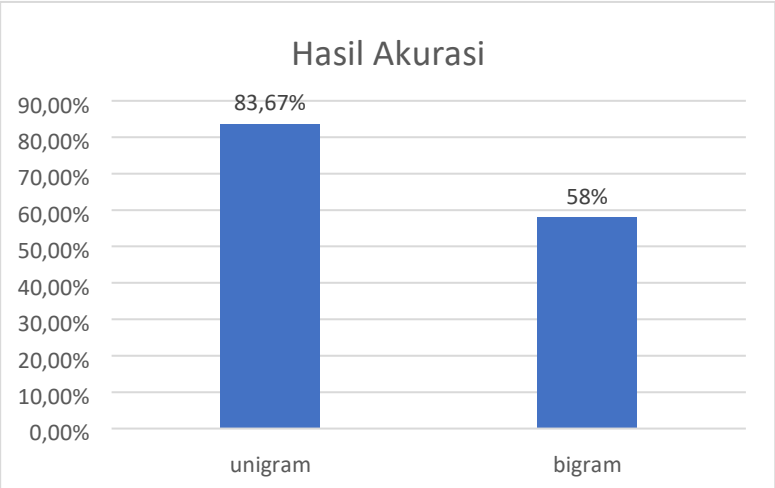
In the unigram test, the number of features tested to get the classification results using the unigram feature earned 83.67% of the results and features as much as 2757.

Tabel 3. Tf-idf- weighted process on bigrams

ada sama	ada sarana	mental tarkam	yg tetap	Zulham kalau	Label
0	0	0	0	0	Negatif
0	0	0	0	0	Negatif
0	0	0	0	0	Negatif
0	0	0	0	0	Negatif
0	0	0	0	0	Positif
0	0	0	0	0	Positif

0	0	0	0	0	Positif
0	0	0	0	0	Negatif
0	0	0	0	0	Negatif
0	0	0	0	0	Negatif
0	0	0	0	0	Negatif
0	0	0	0	0	Negatif

In the second test was conducted classification using bigram feature with the number of features on bigram as much as 8457. In the second test, the classification result is 58%. Where the work of the bigram feature is that each word on the document will advance one word forward and generate bigram features.



3.2 Discussion

In this research, the data used is data that already has positive and negative labels that have been done in previous research. The amount of comment data on this research is that 1000 comment data already has a label. Where positive labels totaled 500 and 500 for negative labels. From that data, the prefix is preprocessing the data. At this stage preprocessing there are 2 stemming & remove punctuation. After preprocessing is done then the next stage is to perform the extraction of features using the term frequency inverse document frequency (tf-idf) so that the data can be done next. By extracting a feature using tf-idf is to calculate the number of occurrences of a term on the data.

This research uses a classification of random forest methods as a method for classifying documents. The next stage is 2 times testing, n-gram is used as a feature word in this study. The n-gram is divided into 2 unigrams and bigrams are used to create a word feature which will make 2 comparative word features as a result of accuracy.

This test was done using data that had previously passed the preprocessing stage. Using as much as 1000 data. To perform this test it is divided into training data as much as 700 and 300 test data. In 700 the training data has 350 data with positive labels and 350 data with a negative label. While the 300 test data consists of 150 positive labeled data and 150 negative labeled data. The first test is that without using the n-gram merge feature. Where preprocessing data is further done classification using random forest. The classification stage for the word feature the unigram accuracy results obtained is 83.67% with a number of features as much as 2757. As for the accuracy results in the second test, using a bigram of accuracy results obtained by 58% with a number of features as much as 8457.

The highest accuracy result obtained unigram with a yield of 83%. Because of the data on the unigram, noise is not too much then the results at unigram are higher than testing with bigram features. The main problem that occurs in the data tested is noise or the features in the data are not appropriate, which should be wasted but are still legible and processes in the data. This study also did not have a feature selection first so as to make the results of accuracy decreased.

4. The conclusion

The research was concluded by using 2 n-gram test features of unigram and bigram. That the best accuracy result is using the unigram word feature with the highest yield of 83.67%. While the reduced accuracy results occur in the second test with the bigram word feature that only results in an accuracy of 58%. It can be seen in the words feature bigram decreased accuracy results due to the data noise that occurred and did not do the first feature selection that makes the result down.

BIBLIOGRAPHY

[1] Alfina,ika. 2017. "**Hate Speech Detection in the Indonesian Language: A Dataset and Preliminary Study**". FMIPA, Universitas Indonesia.

[2] Basari, Abd Samad Hasan dkk, 2012. "**Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization**". Universitas Dian Nuswantoro, Semarang.

[3] Hidayat, Sarif.dkk. 2017 **"Pengaruh Media Sosial Facebook Terhadap Perkembangan E-Commerce di Indonesia**". Vol 8 No 2.

[4] Rachmat, Antonius & Yuan Lukito,2016. "**Klasifikasi Sentimen Komentar Politik dari Facebook Page Menggunakan Naive Bayes**". Seminar Nasional ke-9: Rekayasa Teknologi Industri dan Informasi.

[5] Wahyuni, Rizki Tri. 2017. "**Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi".** Universitas Negeri Semarang. Vol.9, No.1 page 1411-0059.