Study of Hoax News Detection Using Naïve Bayes Classifier in Indonesian Language

Inggrid Yanuar Risca Pratiwi
Malang State Polytechnic
Electrical Engineering Department
Malang, Indonesia
inggrid_yanuar@polinema.ac.id

Rosa Andrie Asmara
Malang State Polytechnic
Information Technology Department
Malang, Indonesia
rosa.andrie@polinema.ac.id

Faisal Rahutomo
Malang State Polytechnic
Information Technology Department
Malang, Indonesia
faisal.polinema@gmail.com

***Abstract*—Nowadays internet has been well known as an information source with many form including online news articles. People mostly search news in the internet. Online news articles are spreading on websites. Those articles' validity may both authentic and fake. Fake news article usually called as hoax news. Hoax news may lead the readers to feel burdened, provoked or even be in a loss. This research proposes to build an automatic hoax news detection. The research describe about hoax news article detection in Indonesian language. This research using own dataset on 250 pages of hoax and valid news articles. Three reviewers conduct manual classification for this purpose. Final tagging are obtained by voting of those three reviewers. Based on three times randomly on training and testing datasets using php-ml component library's obtained average highest on 70% training set and 30% testing set with accuracy is 78,6%, hoax precision is 67,1% valid precision is 91.6%, hoax recall is 89,4% and valid recall is 71,4. This dataset is openly so future research can replicate of dataset and comparison of the result and baseline testing.**

Keywords—Hoax news detection, dataset, naïve bayes classifier

I. INTRODUCTION

Internet has been growing as a popular information media in many aspects, comprised of news, product reviews, public services, movies and many others. Those all are presented in varied sources too, like social media, news articles, and blogs.

Websites and blogs are regarded as popular media to publish news. Due to the facts that the news' sense can be either positive, neutral, or even negative, the news articles spreading on those websites may be directed of the news writer to those senses either the articles or hoax. Fake and misled information is very possible to bring negative effects on human's mind [2].

Through internet collection, the news can be spread. The recipients information need to classify the validity the news of those articles in Indonesian language. With this problem, this study aims to make a dataset of valid and hoaxes online news articles that can be used for classification using machine learning algorithm is naive bayes and useful source for the future research.

In previous study, has been experiments of hoax news classification in Indonesian language by comparing three method i.e. C4.5 algorithm, naïve bayes and SVM [6]. Developing a hoax detection system by incorporating text matching method using Levenshtein Distance measure for emails [2][11], hoaxanalyser.com using bing as web service. The difference with the previous research is the dataset is freely for other researchers, which can be helpful for standard comparison data, regarding the fact that so far such dataset is not freely available and using google as web service.

Scientific contribution of this research are, make own datasets, the dataset can be accessed freely so future researcher can replicate of datasets, compare of the result and baseline testing, the datasets are consists of 10 topics became research in each topics, final manual tagging are obtained by voting of three reviewer, and this datasets in Indonesian Language.

This paper is structured as follows: Section 2 is hoax news detection. Section 3 is dataset development. Section 4 is evaluation and Section 5 is conclusion of the paper.

II. HOAX NEWS DETECTION

A. System Description

Following system description in this research

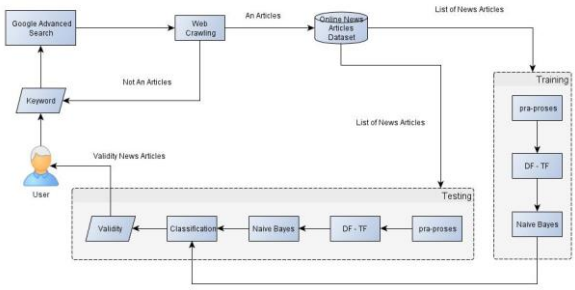

Fig. 1 System Description

User entering keyword of news articles who wants and using google advanced search to find easier a news articles. The system will web crawling based on keyword. If find a news articles, the content will be stored in database else system will return to user and user must entering a new keyword and use valid google advanced search. After database hold all the news articles content's, this dataset divide into

978-1-5386-2827-0/17/$31.00 ©2017 IEEE

training dataset and testing dataset. In training phase, dataset will be pre-process (i.e. case folding, tokenizing, and stop word removal), calculate a document frequency and term frequency then calculate with naïve bayes algorithm. In testing phase, dataset will be pre-process, calculate a document frequency and term frequency, then calculate with naïve bayes algorithm so will get a classification of testing dataset also validation from result of classification.

B. *Machine-Based Approach*

Machine learning algorithm basis needs a couple set of documents - training and testing data. The training phase covers self-training by using different texts and it is then tested by using the testing data. There are some algorithms of Machine learning, such as Maximum Entropy (ME), Naïve Bayes (NB) and Support Vector Machines (SVM), C4.5, usually used for classifying text [4]. Data testing requires a dataset in order to be able to produce a classification, which is also because Mechanics-based algorithm needs a proper study before the testing.

C. *Naïve Bayes*

In this study will using Naïve Bayes algorithm. Naive Bayes (NB) is a popular machine learning tool for classification, due to its simplicity, high computational efficiency, and good classification accuracy, especially for high dimensional data such as text [8]. The formula of Naïve Bayes :

$$X_{NB} = argmax\ P\ (c) \prod P\ (d\ |c) \qquad (1)$$

$$P\ (c|d) = \frac{P(\ d\ |\ c)\ P(c)}{P\ (d)} \qquad (2)$$

Where *d* is word, *c* is category, *P(d|c)* is word probability in category *c*, *P(c)* is probability of category *c* and *P(d)* is probability of d word.

D. *Precision*

Precision measures the percentage of the items that the system detected precision (i.e., the system labeled as positive) that are in fact positive (i.e., are positive according to the human gold labels) formula of precision [11] :

$$\text{Precision} = \frac{true\ positives}{true\ positives + false\ positives} \qquad (3)$$

E. *Recall*

Recall measures the percentage of items actually present in the input that were correctly identified by the system. Recall is defined as [11] :

$$\text{Recall} = \frac{true\ positives}{true\ positives + false\ negative} \qquad (4)$$

F. *Accuracy*

It is measured by the fraction of number of correct predictions over total number of predictions. The formula is [11] :

$$\text{Accuracy} = \frac{tp+tn}{tp+fp+fn+tn} \qquad (5)$$

tn is true negative, tf is true positive, fp is false positive, fn is false negative.

III. Dataset Development

Features in the context of opinion mining are the words, terms or phrases that strongly express the opinion as positive or negative [5]. This pre-processing data phase is to create a dataset that can produce wording arrays.

Fig. 1 Data Pre-processing Phase

A. *Web Crawler*

Web crawler is a process of searching or crawling a page or pages of information from a page. Not only crawling, but web crawlers also take the information from the page. The main function of web crawlers is to search or crawling information from a page [7].

B. *Identifying Articles*

Articles identification is to decide if a website consists of articles, for all the stages and testing phases in this study use articles. The news articles have to be in a form of chronological stories or they at least cover a two-paragraph or eight-line report. This condition is arranged for making an exception for a page that may consist of summaries (which usually has links directed to the next reading sites).

To make sure if the search results like news and blog reviews cover articles or non-articles like pdf files, documents, images, etc., is by making use of *google advanced search*. For example, we can start by entering the keyword of "Ikan lele sumber kanker". On the Google URL, it will appear detailed URLs, which then can be copied on a coding page included in a simple DOM html library for making the web crawling. This is the example of a detailed URL: Https://www.google.com/search?hl=en&safe=active&tbo=d&site=&source=hp&q=e-ktp+malang&num=20

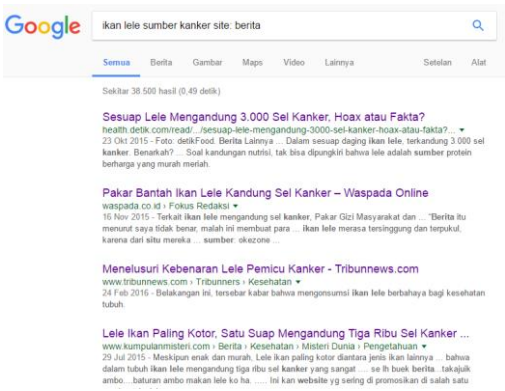

Fig. 2 Keyword Search for A valid and hoax news Article

C. HTML Tag Removal

HTML Tags and any punctuations on a set of sentences are removed by using the HTML parsers with DOM (*document Object Model*). Every HTML file can be mapped within DOM. DOM mapping is essential for analyzing an efficient web content. The DOM components include HTML Tag elements. Every file element like a paragraph tag ( ) is defined in a content block that impacts the position of relative content. The model element only affects the visual attribute contents such as font size and font colors. The HTML element put within blocks in this paper is . Other than the tags and which are removed, the system will acquire the content on tags tag until .

To decide the main text of a web article in a tag … , the character's maximum number needs to be limited. In this study, however, an article to study should at least have at least 500 characters or more, so it can be helpful to decide on which part that article is begun and ended.

Fig. 3 URL Advanced Search for
Web Crawling and HTML Tag Removal

D. Case Folding

Case folding is important to remove any characters other than the letters during the information acquisition process. It is also to omit any disturbance during such information acquisition process [6]. This process is to change the uppercase letters into the lowercase ones. Besides the letters, punctuation is regarded as a delimiter which then is omitted as well by using the syntax presented by the PHP language that is "*strtolower"*. For example, a sentence "*Dalam sesuap daging ikan lele, terkandung 3000 sel kanker*", which is processed by using the case folding, results on a sentence "*dalam sesuap daging ikan lele terkandung 3000 sel kanker*".

E. Tokenizing

Tokenizing is a fragmentation process that splits sentences into words [6]. Tokenizing process in this study splits the sentences into words based on the spaces, and those split words are compiled into arrays. For example, a case folded sentence, "*dalam sesuap daging ikan lele terkandung 3000 sel kanker*" will be split into words as follows :

TABLE I. EXAMPLE FOR TOKENIZING

Index	Words
0	dalam
1	sesuap
2	daging
3	ikan
4	lele
5	terkandung
6	3000
7	sel
8	kanker

F. Stop Word Removal

A word which occurs in 80% of the documents in the collection is useless for purposes of retrieval. Such a word is frequently referred to as stop-word and it is normally filtered out as a potential index term. Since stop-word removal also provides compression of the indexing structure, the list of stop-words might be extended in order to include words other than articles, prepositions and conjunction. For instance, some verbs, adverb, and adjective could be used as stop-words [8].

This stage really helps the classification process since there are a number of words removed, which then results on faster computing process. *Stop-wordlist* used in Indonesian language is the *Stop-wordlist* by Tala F. Z.

This stop-word removal process is done after the word indexing process in tokenizing phase. The removed word on this process is "dalam" because the word "dalam" is found in the stop-word list. Take a look at the example below, which takes the same sentences to study.

TABLE II. BEFORE STOP WORD REMOVAL

Index	Words
0	dalam
1	sesuap
2	daging
3	ikan
4	lele
5	terkandung
6	3000
7	sel
8	kanker

TABLE III. AFTER STOP WORD REMOVAL

Index	Words
0	sesuap
1	daging
2	ikan
3	lele
4	terkandung
5	3000
6	sel
7	kanker

G. Document Frequency

Document frequency is the number of documents in which a term occurs. Computed the document frequency for each unique term in the training corpus and removed from the feature space those terms whose document frequency was less than some predetermined threshold. The basic assumption is that rare terms are either non-informative for category prediction or not influential in global performance. Improvement in categorization accuracy is also possible if rare terms happen to be noise terms [8]. Document Frequency which counts how many documents a feature appears in, has been recognized as a simple yet quite effective metric in solving different text classification problems [8]. Document frequency is also calculate on training and testing data.

For example we have 10 documents in data training with two categories. There are valid news articles and hoax news

articles and four documents are valid news article and six documents are hoax news articles. So, value for document frequency of valid news articles is $\frac{4}{10}$ and hoax news articles is $\frac{6}{10}$.

H. Term Frequency

Term frequency may be considered as relatively more important, since document frequency is based on binary value of a term presence or absence in a document and it ignores the actual contribution of a word within a document. For instance, two words having term frequencies of 10 and 100, respectively, in a document will have same document frequency of 1. This means that we are unable to judge their relative importance for a document. Term frequency on the other hand considers such information which may be useful in selection of important features [8].

I. Mutual Information

Suggestion for next feature is mutual information. MI measures how much information the presence/absence of a term contributes to making the correct classification decision on class [9].

IV. Evaluation

This study covers 250 data of the total dataset number in Indonesian language. Valid and hoaxes dataset are comprised of 10 topics in which has 25 news with some topics as follows – eating catfish can allow the cancer cells to grow, acupuncture using needle can lead the patients to suffer from strokes, iPhone 6 is easily curved, *reog Ponorogo* was burned in Philippines, the sympathizers of 212 event are forbidden to go into *Istiqlal* Mosque, the tooth brushes from piggy hair, the dangerous pacifier sweets, *Pokemon* Go : I am a Jewish, the clouds during *Uje's* funeral, and *Munarman* the Freeport's Lawyer.

The following stage is to give the tagging manually by the reviewers. Each dataset is scored for having valid or hoaxes. This scoring process of each dataset requires at least 3 people for it is considered helpful to count the average score. If the scores are taken from only two people or from an even number of people, it may result on the same scoring grade which may even complicate the average-counting process.

A. Dataset

These are some examples of raw dataset before it is attached to the manual tagging.

TABLE IV. LINK EXAMPLES

Link of Dataset
http://health.detik.com/read/2015/10/23/154341/3051784/763/sesuap-lele-mengandung-3000-sel-kanker-hoax-atau-fakta?l992205755
http://www.tribunnews.com/tribunners/2016/02/24/menelusuri-kebenaran-lele-pemicu-kanker
http://regional.kompas.com/read/2015/10/28/09123241/Lele.Disebut.Mengandung.Ribuan.Sel.Kanker.Pembudidaya.Tersinggung

An example which has been manually tagged by the reviewers is on the topic of "Makan Ikan Lele Mengandung Sel Kanker".

TABLE V. EXAMPLE FOR MANUAL TAGGING FROM 3 REVIEWER

Link Berita	Review-er 1	Review-er 2	Review-er 3	Result
http://health.detik.com/read/2015/10/23/154341/3051784/763/sesuap-lele-mengandung-3000-sel-kanker-hoax-atau-fakta?l992205755	Valid	Valid	Hoax	Valid
http://regional.kompas.com/read/2015/10/28/09123241/Lele.Disebut.Mengandung.Ribuan.Sel.Kanker.Pembudidaya.Tersinggung	Hoax	Valid	Hoax	Hoax
http://www.tribunnews.com/tribunners/2016/02/24/menelusuri-kebenaran-lele-pemicu-kanker	Valid	Valid	Hoax	Valid

This is result of initial manual tagging from 250 pages dataset as follows :

TABLE VI. INITIAL MANUAL TAGGING OF DATASET

Tagging	Total
Valid News Article	155
Hoax News Article	95

Example for term frequency datasets' :

TABLE VII. EXAMPLE OF TERM FREQUENCY

Index	Words	Term Frequency
0	sesuap	7
1	daging	21
2	ikan	302
3	lele	333
4	terkandung	8
5	3000	8
6	sel	97
7	kanker	94

B. Evaluation

Using the datasets with three times randomly on training and testing datasets using php-ml component library. To evaluate performance model of classifier, this research using accuracy and precision. This is result of steps as follows :

TABLE VIII. ACCURACY, PRECISION AND RECALL

Time	Train : Test (%)	Accuracy (%)	Precision (%)		Recall (%)	
			Hoax	Valid	Hoax	Valid
1	70 : 30	82,6	72,1	96,8	96,8	72,1
2		72	56,4	88,8	84,6	65,3
3		81,3	72,9	89,4	87	77

Time	Train : Test (%)	Accuracy (%)	Precision (%)		Recall (%)	
			Hoax	Valid	Hoax	Valid
1	80 : 20	76	50	93,3	83,3	73,6
2		72	56	88	82,3	66,6
3		80	70,3	91,3	90,4	72,4
1	60 : 40	70	44,8	94,1	88	64
2		64	46,8	79,2	66,6	62,6
3		76	61,1	93,4	91,6	67,1

TABLE IX. AVERAGE ACCURACY, PRECISION AND RECALL

Train : Test (%)	Accuracy (%)	Precision (%)		Recall (%)	
		Hoax	Valid	Hoax	Valid
70 : 30	78,6	67,1	91,6	89,4	71,4
80 : 20	76	58,7	90,8	85,3	70,8
60 : 40	70	50,9	88,9	82,1	64,5

TABLE X. TESTING DOCUMENT ID

70 : 30

Time 1				Time 2				Time 3			
No	Id	No	Id	No	Id	No	Id	No	Id	No	Id
1	38	39	8	1	31	39	154	1	161	39	56
2	134	40	167	2	215	40	154	2	42	40	250
3	116	41	168	3	58	41	219	3	208	41	62
4	51	42	73	4	194	42	140	4	233	42	227
5	240	43	153	5	155	43	22	5	74	43	59
6	248	44	160	6	2	44	63	6	65	44	138
7	35	45	163	7	120	45	240	7	160	45	217
8	15	46	72	8	210	46	229	8	63	46	208
9	31	47	196	9	87	47	148	9	20	47	190
10	174	48	115	10	232	48	148	10	39	48	221
11	92	49	17	11	214	49	49	11	147	49	162
12	160	50	175	12	121	50	152	12	144	50	174
13	212	51	84	13	18	51	40	13	228	51	79
14	2	52	196	14	209	52	41	14	32	52	116
15	75	53	7	15	122	53	203	15	248	53	118
16	70	54	200	16	84	54	175	16	193	54	33
17	102	55	119	17	11	55	106	17	35	55	156
18	47	56	223	18	15	56	149	18	174	56	123
19	140	57	94	19	228	57	58	19	172	57	228
20	188	58	224	20	50	58	179	20	139	58	81
21	134	59	161	21	59	59	240	21	75	59	243
22	103	60	236	22	121	60	86	22	2	60	107
23	125	61	155	23	171	61	49	23	100	61	166
24	37	62	86	24	80	62	141	24	81	62	248
25	48	63	234	25	143	63	117	25	75	63	225
26	17	64	34	26	135	64	249	26	163	64	238
27	18	65	200	27	108	65	43	27	5	65	230
28	112	66	198	28	190	66	183	28	104	66	137
29	17	67	120	29	227	67	175	29	230	67	189
30	23	68	97	30	215	68	102	30	130	68	38
31	42	69	44	31	16	69	201	31	72	69	178
32	34	70	213	32	102	70	198	32	209	70	98
33	26	71	163	33	196	71	105	33	197	71	54
34	72	72	195	34	204	72	62	34	167	72	62
35	63	73	186	35	238	73	87	35	126	73	211
36	32	74	32	36	85	74	245	36	199	74	103
37	232	75	199	37	126	75	183	37	6	75	55
38	15			38	14			38	95		

80 : 20

Time 1				Time 2				Time 3			
No	Id	No	Id	No	Id	No	Id	No	Id	No	Id
1	80	26	192	1	165	26	111	1	100	26	226
2	30	27	173	2	25	27	218	2	208	27	20
3	110	28	88	3	128	28	156	3	179	28	159
4	203	29	138	4	156	29	77	4	34	29	191

80 : 20

Time 1				Time 2				Time 3			
No	Id	No	Id	No	Id	No	Id	No	Id	No	Id
5	169	30	28	5	76	30	206	5	70	30	156
6	175	31	166	6	62	31	207	6	113	31	129
7	61	32	68	7	9	32	144	7	52	32	199
8	25	33	234	8	198	33	246	8	112	33	140
9	137	34	105	9	97	34	197	9	63	34	211
10	177	35	210	10	98	35	181	10	146	35	65
11	244	36	220	11	145	36	83	11	153	36	32
12	169	37	115	12	35	37	239	12	60	37	75
13	54	38	155	13	246	38	170	13	49	38	237
14	65	39	45	14	185	39	147	14	29	39	55
15	59	40	91	15	170	40	40	15	230	40	98
16	112	41	207	16	74	41	165	16	250	41	68
17	11	42	180	17	140	42	211	17	236	42	86
18	74	43	239	18	37	43	148	18	226	43	226
19	63	44	241	19	206	44	3	19	31	44	173
20	229	45	104	20	4	45	85	20	227	45	158
21	85	46	107	21	48	46	51	21	159	46	208
22	155	47	51	22	33	47	153	22	157	47	88
23	58	48	92	23	24	48	3	23	160	48	29
24	108	49	112	24	32	49	84	24	112	49	150
25	103	50	201	25	93	50	200	25	232	50	250

60 : 40

Time 1				Time 2				Time 3			
No	Id	No	Id	No	Id	No	Id	No	Id	No	Id
1	15	51	105	1	9	51	87	1	125	51	11
2	125	52	242	2	233	52	40	2	55	52	69
3	207	53	169	3	245	53	28	3	249	53	149
4	157	54	193	4	208	54	69	4	175	54	40
5	96	55	98	5	168	55	214	5	71	55	218
6	165	56	69	6	89	56	174	6	24	56	221
7	220	57	117	7	51	57	46	7	87	57	105
8	75	58	108	8	122	58	201	8	213	58	131
9	58	59	44	9	21	59	217	9	133	59	129
10	36	60	207	10	9	60	17	10	229	60	178
11	11	61	188	11	11	61	207	11	200	61	226
12	181	62	89	12	204	62	204	12	176	62	250
13	161	63	203	13	245	63	92	13	227	63	163
14	30	64	157	14	229	64	106	14	21	64	157
15	50	65	15	15	9	65	194	15	62	65	50
16	163	66	151	16	114	66	223	16	22	66	181
17	109	67	49	17	86	67	134	17	242	67	249
18	29	68	230	18	60	68	25	18	109	68	5
19	229	69	173	19	137	69	36	19	185	69	126
20	108	70	87	20	178	70	177	20	94	70	229
21	56	71	178	21	3	71	199	21	44	71	234
22	154	72	186	22	215	72	250	22	102	72	53
23	131	73	168	23	3	73	231	23	110	73	190
24	190	74	223	24	82	74	218	24	27	74	102
25	74	75	15	25	93	75	59	25	42	75	168
26	177	76	154	26	115	76	140	26	86	76	142
27	77	77	210	27	168	77	32	27	140	77	46
28	80	78	10	28	66	78	33	28	3	78	6
29	126	79	171	29	107	79	55	29	182	79	21
30	192	80	226	30	7	80	165	30	49	80	177
31	246	81	69	31	246	81	142	31	241	81	85
32	75	82	28	32	194	82	93	32	233	82	35
33	69	83	233	33	174	83	155	33	132	83	247
34	242	84	161	34	28	84	214	34	27	84	69
35	216	85	149	35	19	85	2014	35	223	85	199
36	83	86	83	36	131	86	19	36	20	86	104
37	121	87	33	37	32	87	242	37	191	87	245
38	33	88	169	38	98	88	102	38	146	88	203
39	172	89	104	39	48	89	58	39	4	89	246

60 : 40											
Time 1				Time 2				Time 3			
No	Id	No	Id	No	Id	No	Id	No	Id	No	Id
40	203	90	177	40	168	90	111	40	85	90	12
41	80	91	106	41	155	91	193	41	120	91	111
42	166	92	151	42	108	92	222	42	200	92	238
43	219	93	58	43	146	93	84	43	49	93	109
44	205	94	110	44	126	94	36	44	30	94	133
45	48	95	130	45	83	95	78	45	237	95	17
46	57	96	85	46	82	96	110	46	144	96	171
47	16	97	186	47	168	97	74	47	119	97	229
48	65	98	29	48	55	98	23	48	114	98	112
49	130	99	61	49	22	99	125	49	149	99	139
50	81	100	61	50	218	100	192	50	28	100	14

Based on three times randomly on training and testing datasets using php-ml component library obtained average highest on 70 : 30 train : testing's with accuracy is 78,6%, hoax precision is 67,1% valid precision is 91.6%, hoax recall is 89,4% and valid recall is 71,4%.

This dataset can be used for classification with machine learning algorithm and can be accessed freely on http://indonesian-ir.org.

V. Conclusion

Nowadays internet has been well known as an information source with many form including online news articles. Those articles' validity may both authentic and fake. Fake news article usually called as hoax news. This research proposes to build an automatic hoax news detection.

Text processing needs a dataset which is then developed through this study includes several stages produce 250 pages of hoax and valid news articles. Three reviewers conduct manual classification for this purpose. Final tagging are obtained by voting of those three reviewers. This dataset is openly and can be used for future research.

The dataset in this study is expected to be used for classification which are used learning machine algorithm such as naïve Bayes, support vector machine (SVM), C4.5 or the Maximum Entropy (ME). In this study proposes using naïve bayes algorithm for classification method.

This dataset was good enough to be used for classification because not only the dataset affects but amount of percentage of training data and testing data also affects the classification results.

Based on three times randomly on training and testing datasets using php-ml component library obtained average highest on 70 : 30 train : testing's with accuracy is 78,6%, hoax precision is 67,1% valid precision is 91.6%, hoax recall is 89,4% and valid recall is 71,4.

References

[1] TANG, Huifeng; TAN, Songbo; CHENG, Xueqi. A survey on sentiment detection of reviews. Expert Systems with Applications, 2009, 36.7: 10760-10773.

[2] ISHAK, Adzlan; CHEN, Y. Y.; YONG, Suet-Peng. Distance-based hoax detection system. In: Computer & Information Science (ICCIS), 2012 International Conference on. IEEE, 2012. p. 215-220.

[3] RANA, Shweta; SINGH, Archana. Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques. In: Next Generation Computing Technologies (NGCT), 2016 2nd International Conference on. IEEE, 2016. p. 106-111.

[4] Jain, Anuja P; Asst. Prof Dandannavar Padma. Application of Machine Learning Techniques to Sentiment Analysis. In: Applied and Theoretical Computing and Communication Technology (iCATccT), 2016 2nd International Conference on. IEEE, 2017.

[5] HADDI, Emma; LIU, Xiaohui; SHI, Yong. The role of text pre-processing in sentiment analysis. Procedia Computer Science, 2013, 17: 26-32.

[6] RASYWIR, Errissya; PURWARIANTI, Ayu. Eksperimen pada Sistem Klasifikasi Berita Hoax Berbahasa Indonesia Berbasis Pembelajaran Mesin. Jurnal Cybermatika, 2016, 3.2.

[7] MUZAD, Aad Miqdad Muadz, RAHUTOMO, Faisal. *Korpus Berita Daring Bahasa Indonesia Dengan Depth First Focused Crawling.* Seminar Nasional Terapan Riset Inovatif, 2016, 01.

[8] BAEZA-YATES, Ricardo, et al. Modern information retrieval. New York: ACM press, 1999.

[9] MANNING, Christopher D., et al. *Introduction to information retrieval*. Cambridg-e: Cambridge university press, 2008.

[10] BENGIO, Yoshua; GRANDVALET, Yves. No unbiased estimator of the variance of k-fold cross-validation. *Journal of machine learning research*, 2004, 5.Sep: 1089-1105.

[11] JURAFSKY, Daniel, MARTIN, James H. Speech and Language Processing Naïve Bayes and Sentiment Classification. 2016.

[12] RAHUTOMO, Faisal; KITASUKA, Teruaki; ARITSUGI, Masayoshi. Test collection recycling for semantic text similarity. In: *Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services*. ACM, 2012. p. 286-289.

[13] CHEN, Yoke Yie; YONG, Suet-Peng; ISHAK, Adzlan. Email Hoax Detection System Using Levenshtein Distance Method. *JCP*, 2014, 9.2: 441-446.