International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2019): 7.583

Incremental Machine Learning to Detect Fake New Using Support Vector Machine

Edwin F Wicaksono[1], Ruddy J Suhatril[2], Matrissya Hermita[3]

[1, 2]Gunadarma University, Faculty of Computer Science, Jl. Margonda Raya 100, Depok, West Java 16424, Indonesia

[3] Gunadarma University, Faculty of Psychology, Jl. Margonda Raya 100, Depok, West Java 16424, Indonesia

Abstract: ***As one of the biggest online market worldwide, 97.4% of internet users in Indonesia are active on social media [1]. Facebook, one of the most common social media in Indonesia, provides a platform to spread information throughout its user. However, fact-based information is not the only one circulate on the Internet. The number of fake news shared throughout the Internet, especially social media, is concerning. Investigating fake news requires considerably longer time in collecting the data to compare. In addition, humans naturally are not very good at differentiating between real and fake news [3]. It makes machine learning becomes advantageous in dealing to this problem. However, the rapid changes of news throughout time requires machine learning to be able to train its model dynamically. Incremental machine learning is proposed to solve this problem. As much as 6757 labeled data containing both fake and factual news provided by George McIntire, Politifact, and Buzzfeed are set to be primary data in this study. In addition, over 30.000 crawled news from various reliable sources are prepared to observe the most efficient data ratio to train the model. Based on model-selection approach, Support Vector Machine outperformed the other models with the initial accuracy of 0.889. Along with feature extraction, parameter tuning, and feature selection, performance of the incremental machine learning can reach over 96% accuracy.***

Keywords: Machine Learning, Fake News, Support Vector Machine

1. Introduction

Indonesia is listed in the biggest population both in the world and in the Internet. The number of internet users in Indonesia had reached 132.7 million users in 2016. As much as 24.4% and 29.2% were in the range of 25-34 and 35-44 years old respectively with 62% of the total are in the rage of productive age group. According to the association of Internet providers in Indonesia, to be updated with the latest information (25.3%) is the main reason of using internet. In addition, the percentage of active media social users in Indonesia is as high as 97.4% which makes Indonesia to become the largest social media active users in South East Asia and the fifth in the world sitting right in between Brazil and Japan [1].

From the given facts, it can be concluded that most of Indonesian, in productive age group, search for the latest information on the Internet and also active on social media. However, the ease that technology has been brought forth has its drawbacks. Fact-based news is not the only one that circulate on the Internet. The number of fake news shared throughout the Internet, especially social media, is also concerning. The information that appears on the social media can be relayed among users with no filtering, fact-checking, or editorial judgment. Anyone can alter any information and share it as easy as left-clicking on the mouse or tap on the screen.

Statista, the online statistics portal, showed that more than 30% of social media users in U.S. stated that they see more than one fake news in the Internet in one day. Out of 100% of the respondents, only 8% that perceived less often than once a week of fake news exposure on the Internet.
The phenomenon of United States presidential election in 2016 also became the momentum for the supporters of the parties to attack their opposition through fake news. It is said that if one fake news were about as persuasive as one TV campaign ad, it would have changed vote shares by an amount on the order of hundredths of a percentage point in 2016 election [2].

Based on a survey done by BuzzFeed, it is reported that as many as 75% of Americans adults who were familiar with a fake news headline considered the stories as accurate (Silverman et al., 2016). It had been discovered also that the consumers are likely to believe the fake news even though it does not fit their ideological bias. Therefore, fake news is a serious concern for the information technology era globally.

However, investigating fake news requires considerably longer time in collecting the data to compare. In addition, humans naturally are not very good at differentiating between real and fake news [3]. One can investigate the fake news by doing research on various resources. The more the information comes up in the investigation, the less tendency of fake news the information will be. To achieve that, machine learning comes in handy by providing automatic filtering on numerous sources.

The automation that machine learning offers might solve the problem of fake news spreading on the internet. To address these challenges writer wants to create the hoax detector based on machine learning.

2. System Design

Figure 1 shows the machine learning flow design in this study. The flow includes the dataset preparation, model selection, performance forging, testing, and resulting in final model for the machine learning. The explanation of each nodes in the design will be explained later in this chapter.

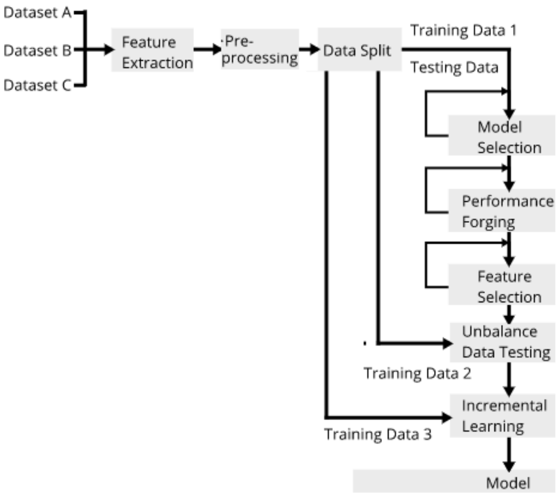

Figure 1: Machine Learning Flow Design

3. Dataset

Contrary to the simplicity of its term, fake news refer to a range of phenomena. It varies from deliberately misleading attempts to undermine election to national security at one end of the continuum to any view that challenges consensus "group think" on the other. Tambini (2017) mentioned that there are six types of fake news that require difference responses, including: Alleged foreign interference in domestic elections through fake news, ad-driven invention, parody and satire, bad journalism, news that is ideologically opposed, and news that challenges orthodox authority [4]. However, all categories of the news share the same technique in spreading it which is telling the distorted facts that could mislead reader [2]. Consequently, all types of the fake news will be treated equally in this study. The different responses that is mentioned by Tambini (2017) is not the concern of the study because the response is actively done by human that gain the information from this machine learning [4].

The datasets used in this research sourced from BuzzFeed and PolitiFact datasets which consist of both fake news and real news. As much as 91 fake news and 91 real news are provided by BuzzFeed dataset in the form of JSON and another 120 fake news data and 120 real news data are also provided by PolitiFact in the same form as the dataset provided by BuzzFeed [5][6]. Consequently, a total of 422 data is gathered through the internet as the dataset for the machine learning. Apart from Politifact and Buzzfeed dataset, as much as 6335 data consist of fake news and actual news is provided by George McIntire through github (source: https://github.com/GeorgeMcIntire/fake_real_news_dataset. The dataset is shaped in four columns that consist of id, title, text, and label. The label of this provided dataset consist of two varieties: FAKE and REAL. Combining all the dataset results in 6757 data which consist of 3377 real news data and 3380 fake news data.

In order to observe the best ratio of data label to classify fake news, another dataset consist of more than 30,000 news from different reliable sources are crawled. The crawled dataset consists of URL, published date, title, and content. All the news crawled from giant media such as BBC, CNN, ABC News, etc. Consequently, every single news will be labeled as real news in the dataset. The purpose of this dataset is to be added to the primary dataset (Buzzfeed, Politifact, and George McIntire datasets) incrementally to observe the trend of the evaluation metrics as the real news outnumbered fake news.

4. Feature Extraction

Apart from the words on the content and title, there are numerous of variable that contribute to fake news classification. In order to improve the performance, additional features are added. Table 1 shows the additional features that was sourced from various study [7][8].

Table 1: Additional Features

Category	Quantity	Description
Character-related	10	Count of average letter, percentage of character, percentage of uppercase letter
Word-related	6	Count of word, count of long and short word, average word count in sentence
Punctuation and Special Character	80	Count of punctuation, count of special character, presence of punctuation, presence of special character
Part-of-speech	22	Count of part-of-speech

Crowdsource will only affect the machine learning model if the number of inputs is big enough. In doing so English is picked due to its number of speakers.

Having single language as the input has made lemmatization and stemming becomes possible [9]. Lemmatization works by comparing each word into a predefined dictionary and replace the corresponding words based on it. On the other hand, stemming does not need a predefined dictionary, instead it removes prefix and suffix on each word. Both methods will be examined in this study to determine the most effective and efficient one to be applied in the machine learning.

5. Text Representation Model

After normalizing the content, the next step will be transforming it into something that can be measured, in this case numeric representation. There are several model representations in doing so, bag of words and TfIdf are the most applicable methods available. Both representation models will be tested to discover the best model for this particular problem of fake news detection.

6. Model Selection

As in the future research mentioned by Shu et al. (2017) about the research directions which consist of: Data-oriented, Feature-oriented, Model-oriented, and Application-oriented, this study will focus on the model-oriented. The model-oriented study enables a more effective and practical model for the fake news classification [5][6]. In the model-oriented study, the processed dataset will be applied to many models to obtain the most suitable model for the corresponding dataset that will be utilized for the analysis.

The first step of model selection is defining data and the purposes. From the obtained dataset and the problem hypothesis, it can be concluded that the dataset is discreet which is not affected by the variable of time. Accordingly, the analysis can be done by binomial classification. Basic possible models that suits for binomial classification will be: Decision tree, Neural network, Support vector machine (SVM), Naïve Bayes, and K Nearest Neighbor (KNN).

7. Evaluation Metrics

There are various evaluation metrics that can be applied to classification machine learning including: Accuracy, F1, Precision, and Recall. In the evaluation process, the obtained dataset will be divided in the ratio of 80:20 where 80% are the training data and the rest are the test data. From this division, the accuracy, precision, recall, and F1 of the machine learning will be measured as a threshold of the model selection and the post-processing manipulation.

8. Feature Selection

Ikonomakis et al. (2005) stated that one of the problems in text classification is the number of the feature [10]. It can reach tens of thousands by only hundreds article into account. Consequently, dimension reduction method is necessary to be applied to the feature data. Feature selection is one of the methods of dimension reduction. Feature selection works by getting rid of the least affecting features on the model. The parameter it needs is the threshold of the evaluation metrics represent by floating point. In this study, a set of thresholds with the increment of 0.05 will be applied to the model to seek for the most optimum result with least features being analyzed. The result will be faster processing time with the chance of better performance.

9. Incremental Learning

As news is associated to time, using static dataset will be inappropriate to determine whether a news is fake. Historical data is not able to entirely represent current fake news as the method of delivering the message might divergent in time. As a consequence, enabling online or incremental learning is necessary. By enabling online/incremental learning, the model can be improved following the evolution of both fake and factual news thought out time.

Crowdsourcing allows normal people to annotate news content. It has been applied to several technology such as Fiskkit and Line. Both technologies rely upon user for growth of its dataset so that it can follow the style of fake news at one point.

In doing so, after having the most suitable model and the post processing applied to the machine learning, as much as 1000 data containing both fake and factual news will be added iteratively as the simulation of the accumulation of data insertion that will be done by the user through the feedback. From the test, it can be concluded how the incremental learning affect the base model of machine learning.

10. Result

10.1 Data Preparation and Visualization

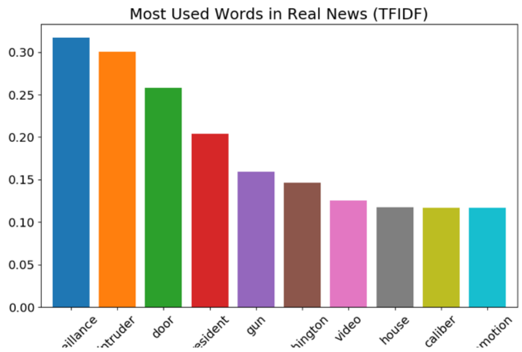

Figure 2: Top 10 Most Used Words in Factual News using TFIDF

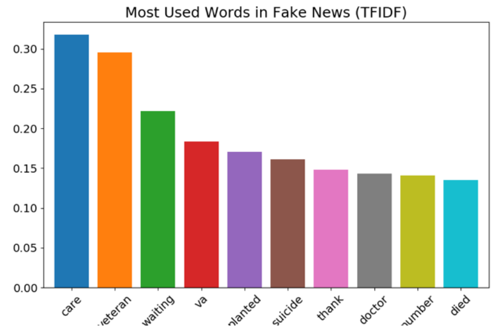

Figure 3: Top 10 Most Used Words in Fake News using TFIDF

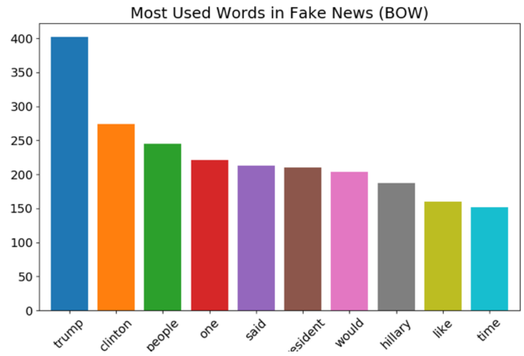

Figure 4: Top 10 Most Used Words in Fake News using BOW

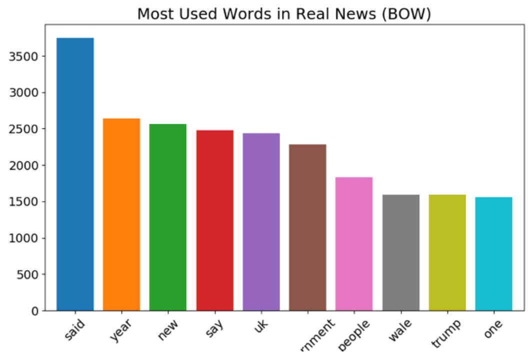

Figure 5: Top 10 Most Used Words in Factual News using BOW

Figure 2,3,4, and 5 shows the visualization of the datasets regarding to the fake news based on two different text representation models. From the data visualization it can be concluded that TfIdf (Figure 2 and Figure 3) is more representable to be used as a model because the occurred word in TfIdf is more distinct. In contrast, the Bag of Word (Figure 4 and Figure 5) model has more probability of showing the same word being in both classes, in this case the word 'said' occurred both in factual and fake news, since it only counts the occurrence of words in text. Hence, the bag of words text tokenization is eliminated in this study. After applying TfIdf to both content and title as well as a set of text pre-processing mentioned in chapter 3, as much as 71,798 features are reserved.

10.2 Model Selection

After applying the dataset into the selected models, the results show as in table 2. In the table 2, K-Nearest Neighbors is excluded from the list due to the processing time that is over the limit surpassing the others by over 35 times of the processing time with Support Vector Machine. Apart from the cost of the processing time, table 2 also shows that K-Nearest Neighbor has the least accuracy, precision, and F1 measurement among the others. Even though the recalls show the perfect value of 1 or 100 per cent, it has to be noted that it is resulted in the imbalance of confusion metrics. The value of confusion metrics shows 659, 1, 572, and 80 for TP, FP, FN, and TN respectively that means the model has 572 failed prediction over 652 actual news (87.73%).

Based on the evaluation metrics to all potential models, Support Vector Machine the best results as in shown in Table 2. In addition, the processing time of Support Vector Machine model is also the fastest among all which is 32 seconds which is over 28% better than Decision Tree and over 43% better than Naive Bayes with given training data.
It is resulted from the fact that Support Vector Machine is one of the models that works well with text classification [11][12][8]. In addition, Support Vector Machine also is able to cope with several types of model violation and outlier as well as having efficient computational processing time compared to the other [13]. Accordingly, Support Vector Machine model is chosen as the base model in this machine learning.

Table 2: Evaluation Metrics on Different Models

Model	Acc	Prec	Rec	F1	t(s)
Naive Bayes	0.753	0.825	0.648	0.726	62
SVM	0.889	0.849	0.948	0.896	32
Decision Trees	0.753	0.751	0.742	0.755	45
K-Nearest Neighbor	0.519	0.511	1	0.676	1138

10.3 Performance Forging

Epoch determine the number of data passes each iteration. The more data passes to the epoch the more memory will be used and the more consistent the result will be. Figure 6 shows different value of epoch being passed to the model. Based on Table 3, The epoch value of 300 gives the best cost amongst all. Even though the tighter variance occurred while the value of epoch is set to be 500, the processing time increases as much as 66%. Besides that, the variance is also still around the error tolerance which is 0.005 except the recall. Accordingly, the value of epoch is set to be 300 to get consistent result every time the model is trained with the least amount of cost.

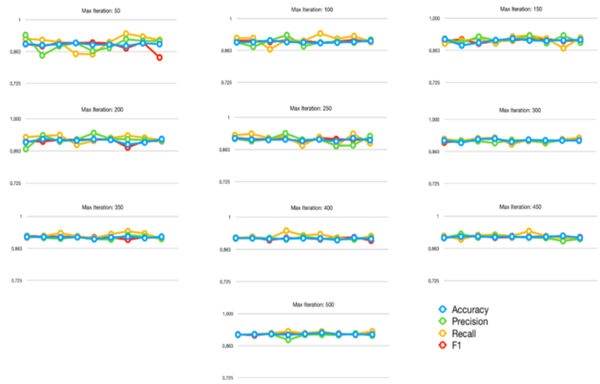

Figure 6: Evaluation Metrics on Various Epoch Value Settings

Table 3: Variance Comparison to Iteration

i	Acc	Prec	Rec	F1
1	0.0610	0.0765	0.0139	0.0841
50	0.0041	0.0273	0.0307	0.0211
100	0.0034	0.0149	0.0189	0.0035
150	0.0088	0.0127	0.0172	0.0057
200	0.0079	0.0187	0.0131	0.0107
250	0.0043	0.0167	0.0208	0.0052
300	0.0054	0.0043	0.0101	0.0058
350	0.0032	0.0066	0.0108	0.0047
400	0.0029	0.0038*	0.0110	0.0051
450	0.0027	0.0083	0.0102	0.0030
500	0.0025*	0.0078	0.0058*	0.0028*

10.4 Regularization

Regularization prevents the coefficients to fit so perfectly to overfit. Table 4 shows the evaluation metrics on each regularization term tested on the model. Based on the processing time model without regularization has the fastest processing time which is 600 seconds while Elasticnet is the most expensive one on the processing time cost which reaches 2035 seconds. Based on the evaluation metrics L2 regularization has the most impact from the other in terms of accuracy, precision, and F1 measurement. In addition, L2 normalization processing time cost 22.16% more than the model without regularization.

Table 4: Regularization Comparison

Penalty	Acc	Prec	Rec	F1	t(s)
None	0.887	0.893	0.877	0.884	600
L1	0.893	0.894	0.900	0.898	1872
L2	0.916	0.907	0.877	0.911	733
Elasticnet	0.901	0.923	0.898	0.907	2035

10.5 Alpha

Figure 7 shows the comparison of alpha channel value given as a model parameter. From given chart, it can be concluded

International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2019): 7.583

that the value of 0.0001 has the most value on accuracy, precision, and f1 with the processing time of 751 seconds.

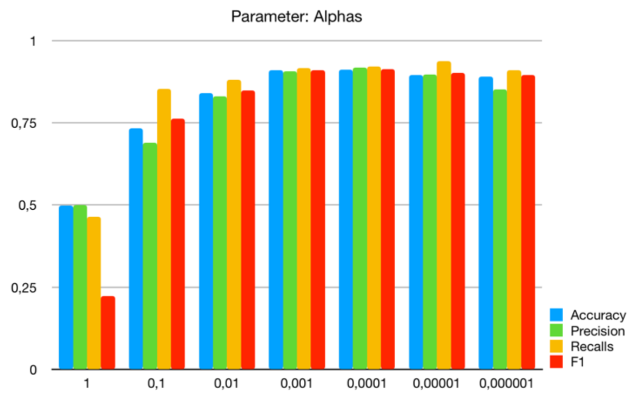

Figure 7: Alpha Channel Comparison

10.6 Additional Features Impact

Table 5 shows the evaluation metrics for both model with and without the additional extracted features. It shows that as many as 118 extracted features increase as much as 0.027, 0.067, -0.009, and 0.024 points on accuracy, precision, recalls, and F1 measurement respectively.

Table 5: Comparison of Evaluation Metrics on Model either with or without Extracted Features

Extracted	Acc	Prec	Rec	F1	t(s)
False	0.871	0.832	0.928	0.878	682
True	0.898	0.899	0.919	0.902	1085

10.7 Data Ratio Testing

As the model with its additional features is ready to be implemented with all its parameter and tuning, the next step is to find out the best label ratio to increase the accuracy of the model. More than 25.000 additional factual news are gradually inserted to the dataset to see model behavior towards unbalanced data. However, due to the lack of computational machine, it stops at the 6th increment which is 24518 factual news and 3380 fake news. Table 6 shows the evaluation metrics of the increment model. As much as 3500 data are added to the dataset on each increment. Different from previous evaluation metrics, the evaluation metrics on data ratio testing grows accordingly to the number of data with the same ratio as 80:20.

Table 6 shows that as the number increases, the accuracy increases as well. However, the cost of improves from the accuracy are the down fall of the precision, recalls and F1, and also increases in processing time. This price contributes in determining the ratio of data to be implemented in the model.

Apart from the evaluation metrics, the confusion metrics also plays big role in determining performance of machine learning. Table 7 expose the TP, FP, FN and TN of the confusion metrics. On that table it can be seen that three degree of confusion metrics has up and down on three of them but the false negative which relatively stable. The TN increase as much as 6.5 times with cost of decreasing on TP as much as 7.9% on the sixth increment. The most affected degree in confusion metrics in FP which increases as much as 139% on the 6th increment. However, the increase of FP has occurred since the first increment with 83% increase that makes the increment from 1st to 5th increment relatively stable.

Seeing the result as a whole, the accuracy starts to improve slowly on the fourth increment onwards which is less than 0.5%. In addition, the gap of processing time on 4th and 6th increment towards original state is also wide. The processing time on 4th increment increases as much as 26% while the 6th increment offers 270% of the original state. Consequently, the most efficient data ratio for this study is set to be 1:5.2 which is the 4th increment.

Table 6: Evaluation Metrics on Incremental Data

i	N	Acc	Prec	Rec	F1	t(s)
0	3518	0.898	0.899	0.919	0.902	1085
1	7018	0.916	0.886	0.825	0.862	1554
2	10518	0.945	0.905	0.889	0.882	1702
3	14018	0.946	0.884	0.818	0.857	2301
4	17518	0.961	0.916	0.830	0.876	2453
5	21018	0.964	0.897	0.824	0.864	3016
6	24518	0.965	0.872	0.828	0.855	5182

Table 7: Confusion Metrics on Incremental Data

i	N	TP	FP	FN	TN
0	3518	647	36	47	650
1	7018	583	66	49	1382
2	10518	624	62	37	2057
3	14018	585	71	39	2785
4	17518	597	64	47	3472
5	21018	607	75	51	4147
6	24518	596	86	47	4851

10.8 Feature Selection

The implementation of the feature selection is applied to the machine learning with support vector machine as the base model. Table 8 shows the comparison of the base model applied with feature selection with different threshold. Feature selection will get rid of features that contribute less than the given threshold. The more threshold value given, the less feature will pass.

On the processing time, the more threshold put on the feature selection, the lesser processing time it takes because of the number of features that gets smaller. However, when the evaluation metrics take accounts, the most and the least threshold listed does not show the best evaluation metrics. Looking at the accuracy, it was started by 0.958 on 0 threshold and ended by 0.9.44 with the maximum of 0.961 on thresholds 0.05. It can be concluded that the best threshold lays on the middle of the list.

The purposes of features selection can vary. Based on table 8, the gap of the accuracy is not wide. The maximum value and median have only 0.003 points difference. In result, the accuracy will not become the determinant on feature selection threshold. It leaves precision, recalls, and F1, and processing time on account. Precision, recalls, and F1

measurement shows random distributed value on each threshold with variance of 0.013, 0.036, 0.019 respectively.
To simplify the selection process, max and median of the accuracy, precision, recalls, and f1 are obtained. The max' are 0.691, 0.899, 0.864, 0.874, and the median are 0.957, 0.878, 0.840, 0.860 for accuracy, precision, recalls, and F1 respectively. Based on the obtained max and median, feature selection with the threshold of 0.1 and 0.3 has evaluation metrics score above median value. Both feature selection with threshold of 0.1 and 0.3 are tie in the performance. However, the processing time of both feature selection shows different result with feature selection with threshold of 0.3 resulting in better processing time with 110 seconds faster than the 0.1 threshold. Consequently, the feature selection threshold in this study is set to be 0.3 which means feature selection will eliminate features with score less than 0.3 resulting in 1472 distinct features to be fit into model.

Table 8: Evaluation Metrics on Feature Selection

Thd	Fn	Acc	Prec	Rec	F1	t(s)
0	71916	0.958	0.899	0.818	0.873	2426
0.05	16225	0.961	0.891	0.855	0.860	1402
0.1	7824	0.950	0.872	0.788	0.844	1250
0.15	4868	0.955	0.879	0.827	0.868	1197
0.2	3200	0.960	0.868	0.850	0.874	1168
0.25	2139	0.958	0.884	0.864	0.873	1150
0.3	1472	0.954	0.885	0.830	0.859	1140
0.35	1088	0.960	0.867	0.852	0.853	1135
0.4	792	0.953	0.877	0.858	0.856	1134
0.45	570	0.944	0.852	0.752	0.812	1130

10.9 Incremental Data Testing

In this section, the impact obtained from the user input is recorded. It is crucial to be evaluated since the model supposed to be able to cope with user input. Table 9 shows the evaluation metrics after inserting number of data incrementally. Each increment contains as much as 50 actual and 50 fake additional news data. For the record, the test set used in incremental data testing is the same test set in each increment to see the dynamic of learning curve with the same sample applied onto it.

Table 9: Evaluation Metrics on Incremental Learning

TP	FP	FN	TN	Acc	Prec	Rec	F1
517	48	52	3363	0.960	0.898	0.884	0.861
520	45	59	3356	0.962	0.877	0.877	0.848
523	42	60	3355	0.961	0.874	0.821	0.864
529	36	63	3352	0.962	0.881	0.856	0.863
530	35	66	3349	0.961	0.866	0.851	0.862
533	32	76	3339	0.961	0.858	0.846	0.866
530	35	68	3347	0.960	0.890	0.883	0.863
531	34	72	3343	0.962	0.877	0.840	0.870
535	30	83	3332	0.961	0.853	0.861	0.869
539	26	96	3319	0.962	0.883	0.860	0.868
538	27	91	3324	0.962	0.863	0.856	0.855

At the table 9 first row indicates the base evaluation metrics before incremental data are tested. It can be seen at table 9 that the evaluation metrics from incremental learning are relatively stable. The value of the variance is less than 0.02 on all metrices. In addition, the variance for accuracy is not more than 0.001 which means that it is almost consistent to its value.

Besides the accuracy, precision, recalls, and F1 measurement, table 9 also indicate the value of confusion metrics. On each iteration, it can be seen that the value of true positive and false negative is raising gradually while the value of false positive and true negative are declining. In other words, the performance to predict fake news gets better while the performance to predict actual news gets worse. However, with the total of 1000 additional data containing equal fake and actual news, the divergence it builds is less than 5% on fake news and even smaller on actual news which is 1.1%. It can be the result of the size of the data which makes factual news has more resistance to the incoming alteration than fake news.

11. Conclusion

Support vector machine turns out to be the most efficient algorithm considering the processing time. Moreover, among the other model such as Naive Bayes, decision tree, and k-nearest neighbor, support vector machine has the most score for its evaluation metrics.

Along the way of determining the model, there are several findings such as the best parameter for support vector machine and the feature selection ratio. As for the parameter, the best combination is the epoch of 300, to get the most stable result yet the fastest computation, regularization of L2 which gives the best evaluation metrics, and the alpha of 0.0001. As for the feature selection, the threshold of 0.6 gives the best result and the processing time as well.

Apart from the suitable model for classifying fake news, finding ratio of the dataset is also becomes the focus of this study. The ratio testing shows that the number of one side label does not affect much on confusion matrix even though the evaluation metrics shows downfall eventually. The true positive shows the relatively stable result while true negative shows 250% increase with the cost of 40% and 28% on false positive and the false negative respectively. Apart from the evaluation metrics, there is an unintended finding on this study.

It turns out that at some ratio point, the processing time can exponentially increase. With the test of gradually increasing the number of factual news as many as 3500, the processing time increase exponentially on the seventh increment. It means that the model needs more time in calculation unbalance data ratio when it reaches about 1:8.

The last objective of this study is to develop incremental machine learning which is also successfully done in this study. As much as 1000 data consisting of both fake news and actual news are tested to be the additional incremental dataset testing and the score reach over 0.96 accuracy. The variance of incremental testing was 0.001, 0.014, 0.019, and 0.006 for accuracy, precision, recall, and F1 respectively which means the model has solid fundamental that is not easily altered by the input. It was resulted from the number of the training data that is proven by the factual news that has more resistant to input than the fake news which has lesser data.

12. Future Study

Since the dataset is limited to time, it is beneficial to have a set of labeled data in multiple years. The main focus of the study is to create a dynamic model that can follow the trending fake news. The number of data as the starting quality of the model is crucial. Having a set range of dataset can boost the model and enclose the timing gap between training data and test data.

In addition, having a model that can represent not only one language but multiple language can also be beneficial. It is obvious that fake news is not only published in English but also other language. It is time consuming to find a platform to classify fake news based on the language the news originated from. Instead of having a word representation, syllable corpus representation can be applied to the pre-processing. The challenge is getting the right proportion of languages as a training dataset to build a reliable model.

References

[1] Statista. Countries with the highest number of internet users as of june 2017, 2017.
[2] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. Journal of Economic Perspectives, 31(2):211-236, 2017.
[3] Andrew Ward, Lee Ross, Edward Reed, Elliot Turiel, and Terrance Brown. Naive realism in everyday life: Implications for social conflict and misunderstanding. *Values and knowledge*, pages 103–135, 1997.
[4] Damian Tambini. Fake news: Public policy responses. In *Media Pol- icy Project*. London School of Economics and Political Science, London, 2017.
[5] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake News Detection on Social Media: A Data Mining Perspective, volume 19. ACM New York, NY, USA, 2017.
[6] Kai Shu, Suhang Wang, and Huan Liu. Exploiting tri-relationship for fake news detection. arXiv preprint arXiv:1712.07709, 2017.
[7] Shivam B. Parikhand Pradeep K. Atrey. Media-rich fake news detection: A survey. IEEE Conference on Multimedia Information Processing and Retrieval, 2018.
[8] Sadia Afroz, Michael Brennan, and Rachel Greenstadt. Detecting hoaxes, frauds, and deception in writing style online. IEEE Symposium on Security and Privacy, pages 461-475, 2012.
[9] Marin Vukovi´c, Krešimir Pripuži´c, and Hrvoje Belani. An intelligent au- tomatic hoax detection system. *International Conference on Knowledge- Based and Intelligent Information and Engineering Systems*, 5711:318– 325, 2009.
[10] M Ikonomakis, Sotiris Kotsiantis, and V Tampakas. Text classification using machine learning techniques. WSEAS transactions on computers, 4(8):966-974, 2005.
[11] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Sup- port vector machines. IEEE Intelligent Systems and their Applications, 13 (4):18-28, July 1998. ISSN 1094-7167. doi: 10.1109/5254.708428.
[12] Victoria L. Rubin, Niall J. Conroy, Yimin Chen, and Sarah Cornwell. Fake news or truth? using satirical cues to detect potentially misleading news. In Computational Approaches to Deception Detection at the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, California, June 2016.
[13] Ingo Steinwart and Andreas Christmann. *Support Vector Machine*. Springer, 2008.

Author Profile

Edwin Fachri Wicaksono received the Bachelor of Information Engineering and Master of Information System Management degrees from Gunadarma University in 2016 and 2019, respectively. During 2016-2018 he stayed in Center of Human and Technology Laboratory to study various research associated in the application of Information Technology in Psychology. He is now with PT Asuransi Atra Buana.