# A Literature Research on Machine Learning Techniques used for Training Annotated Corpus

**Fitrah Rumaisa, Halizah Basiron, Zurina Saaya, Noorli Khamis**

***Abstract—The development of research in the annotation area is growing. Researchers perform annotation task using various forms of datasets such as text, sound, images, and videos. Various algorithms are used to perform tasks. The purpose of this survey is to find out algorithms that are often used by researchers to perform annotation tasks, especially on text data. The literature surveys thirteen research papers on text annotation from the last 5 years. The results of this review indicate that SVM is the algorithm used for all three annotation methods: manual, automatic and semi-automatic annotation, with a significant accuracy above 80%. The result of this survey will be referred by the authors as the basis for subsequent research that will be conducted, especially in the semi-automatic annotation method.***

*Index Terms: Annotation, Algorithm, Text Survey, Semi-automatic*

## I. INTRODUCTION

The Internet is a wonderful resource. By using the Internet we can find various information in the form of text, sound, pictures, and even videos. Such forms of information become the layer of information that is being communicated. In addition, the language is the content that makes the Internet users' understand what is delivered on web content and also able to connect content with other media.

The more varied data available on the Internet, more and more users are using data from the internet for research, especially in the linguistic area. The demands of the Internet users in the understanding of the content are also higher so that every data that is researched always requires the annotation stage in order to obtain maximum results. The data should be prepared as needed so that computers can more easily find patterns and conclusions. This is usually done by adding additional relevant information to the data set. Each tag of the data used to flag the dataset element is called an annotation of the input. The purpose of the annotation activity is to add syntactically differentiated text descriptions of the text and can then be used to add information about the desired visual presentation, or semantic information that the machine can read. In addition, the algorithm can learn efficiently and effectively, annotations performed on the data must be accurate, and relevant to the task requested by the machine to perform [1].

There are three types of methods used in annotation, namely manual annotation, automatic annotation, and semi-automatic annotation. In this study, we will classify algorithms that are used to perform the three methods.

The purpose of this survey is to find out algorithms that are often used by researchers for annotations in the form of text and which algorithm has the results of the analysis with a high percentage average. The results of this study will be used as the basis for the use of algorithms in subsequent research that will be conducted, especially in the semi-automatic annotation method. The literature surveyed is 18 papers from the last 5 years, because starting in 2010, the International Standards Organizations (ISO) began to identify and implement text writing format used for the text annotation process [1].

The next section will describe most of the algorithms used for text annotation is under the machine learning approach. Then will be presented a review of researches that use algorithms for the process of annotation training and divided into 3 (three) subsection that is a manual, automatic and semi-automatic annotation.

## II. MACHINE LEARNING

According to Pustejovsky et.al (2013), the data discovered by machine learning algorithms are natural language, and most often text. The data is then annotated using tags that focus on specific features that match the learning task. There are three basic types of machine learning algorithms:

a. Supervised Learning is each technique whose output is a function mapping from input to fixed label set. Usually, metadata has been provided before by giving an annotation tag to the corpus for data training

b. Unsupervised Learning is a technique that must find the structure of input data that has not been labeled

c. Semi-supervised Learning is a technique whose output is a function mapping from a variety of labeled and non-labeled dataset inputs.

Table 1 shows an overview of algorithms and some annotation tasks that are often used to duplicate [1].

Revised Version Manuscript Received on August 19, 2019.

**FitrahRumaisa**, Information Technology Department, UniversitasWidyatama, JlCikutra 204 A, Bandung 40125, Indonesia.

**HalizahBasiron**, FakultiTeknologiMaklumatdanKomunikasi (FTMK), UniversitiTeknikal Malaysia Melaka (UTeM), Melaka Malaysia

**ZurinaSaaya**, FakultiTeknologiMaklumatdanKomunikasi (FTMK), UniversitiTeknikal Malaysia Melaka (UTeM), Melaka Malaysia.

**NoorliKhamis**, PusatBahasadan Pembangunan Insan (PBPI), UniversitiTeknikal Malaysia Melaka (UTeM), Melaka Malaysia

**Table 1. Annotation task and ML Algorithm (Pustejovsky et al. , 2013)**

| Algorithm | Tasks |
|---|---|
| Decision Tree | Semantic types or ontology class task, coreference resolution |
| Clustering | Genre classification, spam labeling |
| Naive Bayes | Sentiment classification, semantic type or ontological class assignment |
| Maximum Entropy | Sentiment classification, semantic type, or ontological class assignment |
| Support Vector Machine | Sentiment classification, semantic type, or ontological class assignment |
| K-Nearest Neighbour | Classification objects based on the closed training example in the feature space |
| Hidden Markov Model | POS tagging, sentiment classification, word sense disambiguation |
| Conditional Random Field | POS tagging, sentiment classification, word sense disambiguation |

In addition to the above algorithm, there are several other algorithms that can be used in corpus annotations such as K-Nearest Neighbour and Support Vector Machine (supervised learning); Hidden Markov Model, Maximum Entropy Markov Models, Conditional Random Field (sequence induction models). As for clustering and unsupervised learning does not require algorithm in the process of data annotation, because according to Perez-rosas there is no explicit role for annotated data. While for semi-supervised learning can use algorithm commonly used by supervised learning.

Pustejovsky et.al also creates a list of algorithm groupings based on the task to achieve the desired goal as outlined in Table 2.

**Table 2. Suggested ML Algorithm**

| Task | Suggested Algorithm |
|---|---|
| Determine the category of words (POS tagging) | HMMs, CRFs, or possibly SVMs |
| Determine the topic of articles, emails, or web pages | Latent Dirichlet Allocation (LDA), Probabilistic Latent Semantic Analysis (PLSA) |
| Mood, affect, or classification of sentiments of text or speech | Naïve Bayes, MaxEnt, and SVM |
| Determine the type of semantic or ontological class to a word or phrase | Naïve Bayes, MaxEnt, decision trees (e.g., C4.5), and SVMs |
| Word sense disambiguation | SVM, memory-based learning (MBL) |
| Temporal and event recognition | Naïve Bayes, decision trees, or MaxEnt |
| The semantic role for the event participants in a sentence | SVMs, MaxEnt |
| NE identification | Naïve Bayes, MaxEnt, SVMs, CRFs, MEMMs, and even MBLs |
| Coreference resolution | Decision tree induction, CRFs |

In the annotation task, there are two ways to enter data, which are manually and automatically. Automatic annotations can be operated on many documents compared to those done by humans, but sometimes the results are less precise. While manual annotations are considered more appropriate but require a solid process and this process is often used to train automatic annotation machines [2]. Over time, the researchers also combined these two techniques to annotate data, thus called semi-automatic annotations.

The next section will discuss the algorithm used by the researchers in accordance with how to enter the data, namely manual, automatic, and semi-automatic annotation.

## III. TEXT ANNOTATION ALGORITHM

This section will discuss research using manual annotation techniques (Perez-rosas et al., Inkpen et al., Tarasov, Kiritchenko et al., Hamdan, Mozetic et al., Niu et al.), automatic annotation (Xiang et.al, Samejima et al., Wang et al., Volodina et al.), dan semi-automatic annotation (Smatana et al., Elanwar et al., Fu et al., Liu et al., Névéol et al., Sadoun, Koncz et al.), along with the resulting accuracy. The automatic annotation only discussed 4 researchers, because most who use automatic techniques using video or image dataset.

*Manual Annotation*

The research that will be discussed first is done by Perez-rosas et.al (2012), they present a framework for obtaining lexicon sentiment in target languages annotated manually and automatically. Data taken from electronic resources are easy to find, such as English. In the process of data training, the researchers use the SVM algorithm, both for manual or automatic annotation. The results show that the accuracy of manual annotations is 90%, while the automatic annotation is 74% [3].

Furthermore, Inkpen et.al (2017), discusses a similar task like Named Entity Recognition (NER) but focuses on entity location. In contrast to NER, they propose a more detailed task that classifies detected locations into the city, province/states, and country names to map them physically. The data obtained are trained using Conditional Random Fields (CRF) that can detect the location. They use 4 rules to train data:

1. All steps (adjacent location, global context, and adjacent location + global context). Nothing is disabled
2. Disable adjacent locations
3. Disabling the global context
4. Disable adjacent locations + global context

When the researcher passed all the steps (no steps were deactivated), 95.5% accuracy was found. When nearby locations are disabled, the accuracy decreases to 93.7%, indicating that the rule is useful. When the global context rule is disabled, the accuracy increases to 98.2%, which indicates that the rule is useless. When both rules are disabled, and only saves the rules that ensure the correct candidate type and the rule that selects the default location with the largest population when no other rules apply, they achieve an accuracy lower than 96.4%. Thus, the best results are obtained when using all rules except global context rules (98.2%) [4].

**Table 3. Result of Location Disambiguation**

| Deactivated Steps | Accuracy |
|---|---|
| All steps (none deactivated) | 95.5% |
| Deactivating adjacent locations | 93.7% |
| Deactivating global context | 98.2% |
| Deactivating adjacent locations + global context | 96.4% |

The research conducted by Tarasov (2015) has studied the application of different Recurrent Neural Network architectures including uni and bi-directional Elman and Long Short-Term Memory (LSTM) models for content-based sentiment analysis that included extraction of aspects. In the Russian language, the dataset obtained the best results at extracting all aspects based on proportional size (60%) and the best result in extracting all aspects on car dataset according to exact size (74.8%) while maintaining the second best result in the restaurant dataset 71.4 %. In the English dataset, they obtained a fairly good result (79.80%), equivalent to the sixth best result on this dataset (79.6%). Of all RNN models, best results were obtained with a deep bidirectional LSTM with 2 hidden layers (74.8%) [5].

**Table 4. SentiRuEval test dataset result**

| **Method** | **SentiRuEval Restaurant dataset** | | | |
|---|---|---|---|---|
| | ***Proportional*** | | ***Exact*** | |
| | ***Explicit*** | ***All*** | ***Explicit*** | ***All*** |
| BRNN | 67.2 | 52.2 | 57.5 | 64.5 |
| LSTM | 71.9 | **60.0** | 62.6 | **66.8** |
| LSTM, Depth 2 | - | - | - | - |
| Other systems best result | 72.8 | 59.6 | 63.1 | 59.5 |
| **Method** | **SentiRuEval Cars dataset** | | | |
| | ***Proportional*** | | ***Exact*** | |
| | ***Explicit*** | ***All*** | ***Explicit*** | ***All*** |
| BRNN | 71.7 | 70.4 | 61.7 | 59.9 |
| LSTM | - | - | - | - |
| LSTM, Depth 2 | **74.8** | **71.4** | 65.1 | 63.0 |
| Other systems best result | 73.0 | 65.9 | **67.6** | **63.6** |

Kiritchenko et.al (2016) explores sentiment compositions in phrases that have at least one positive word and at least one negative word for example phrases like "happy accident" and "best winter break". They collect sets of data from opposite polarity phrases and manually annotate them with the actual score of sentiment associations. Using this dataset, they analyzed the linguistic pattern present in the opposite polarity phrase. Then the dataset is trained using a Support Vector Machine classifier with RBF kernel for binary classification tasks and SVM regression model with RBF kernel for regression assignment using LibSVM package. The best result is 82.6% [6].

**Table 5. Automatic systems performance**

| **Features** | **Binary (Acc.)** | |
|---|---|---|
| | ***2-gr*** | ***3-gr*** |
| **Baselines** | | |
| a. Majority label | 56.6 | 60.8 |
| b. Last unigram | 57.2 | 59.3 |
| c. Most polar unigram | 66.9 | 69.8 |
| d. POS rule | 65.6 | 63.8 |
| **Supervised classifier** | | |
| e. POS + sent. label | 65.7 | 64.2 |
| f. POS + sent. score | 74.9 | 74.8 |
| g. Row f + uni | 82.0 | 81.3 |
| h. Row f + emb(avg) + emb (max) | 78.2 | 79.5 |
| i. Row f + emb (conc) | 80.2 | 76.5 |
| j. Row f + emb(conc) + uni | **82.6** | 80.9 |
| k. POS + emb (conc) + uni | 76.3 | 80.2 |

The research conducted by Hamdan (2016), analyze the sentiments of social media text such as tweets or customer opinions. There are two main tasks to focus on detecting the sentiment polarity (positive, negative, neutral) and opinion target extraction to find out customer expression at the aspect-based level. In the supervised learning stage, the researcher uses three classification methods: SVM, Logistic Regression and some classification methods that are proposed. In this paper, only the first two classification methods are highlighted. Results of training accuracy data generated are SVM of 69.97% and LR of 71.65% [7].

**Table 6. Results of the SVM and LR algorithms, and also the model that is proposed**

| Model | Twitter | Laptop |
|---|---|---|
| SVM | 52.35 | 69.97 (664/949) |
| LR | 53.38 | **71.65** (680/949) |
| ne | 42.26 | 72.29 (686/949) |
| pmi | 53.15 | 71.13 (675/949) |
| orr | 54.11 | 72.4 (687/949) |
| cpd | 19.20 | 67.65 (642/949) |
| kl | 57.30 | 69.02 (655/949) |
| rf | 24.46 | 71.65 (680/949) |
| dbidf | 13.25 | 8.32 (79/949) |
| zd | 49.24 | 69.65 (661/949) |
| wllr | 45.11 | 50.68 (481/949) |
| ngl | 50.27 | 69.65 (661/949) |

Similar to Hamdan, Mozetic et.al (2016) also analyze a large number of tagged twitter data (1.6 million) from various languages (Albanian, Bulgarian, English, German, Hungarian, Polish, Portuguese, Russian, Ser / Cro / Serbian, Croatian, and Bosnian), Slovak, Slovenian, Spanish, and Swedish). From the experimental results seen that there is no significant difference resulting from some top classification model, such as 5 kinds of SVM model and Naive Bayes Classifier [8]. Figure 1, the comparison of the 6 classification methods.

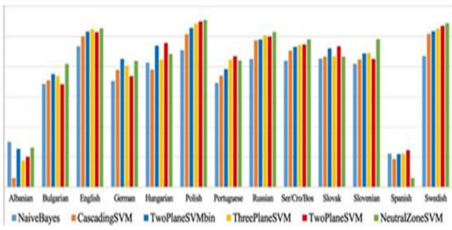


**Figure 1. Comparison of six Classification Method**

Then the following research is very interesting. Research conducted by Niuet. Al (2016) discusses the understanding of user sentiment towards tweets that combine short images and text on Twitter. Therefore, the researcher introduced a multi-view sentiment analysis (MVSA) dataset using manual annotations obtained from Twitter. The experimental results show that the quality of work can be improved by using textual and visual views simultaneously.

**Table 7. Accuracy text dataset results**

| Method | Accuracy | F-positive | F-negative | F-Average |
|---|---|---|---|---|
| SentiWordNet | 0.603 | 0.640 | 0.557 | 0.598 |
| SentiStrength | 0.632 | 0.628 | **0.636** | 0.632 |
| TF | **0.719** | **0.791** | 0.569 | **0.680** |
| TF-IDF | 0.692 | 0.767 | 0.542 | 0.655 |

Table 7 further lists the results of the lexicon-based approach. In comparison, TF and ID-IDF results are also included. It can be seen that the highest accuracy of the TF method is 71.9% [9].

*Automatic Annotation*

Reviewing the algorithm used for automatic annotation with text dataset is quite difficult because the literature using automatic annotation uses more video or image datasets. Therefore, in this section, only the two studies are reviewed which are by Xiang et.al (2012), Samejima et.al (2015), Wang et al. (2012) andVolodina et al. (2014) respectively.

The first will be discussed is the research conducted by Xiang et.al (2012). They implement an approach that is able to exploit language regularity in gross language through statistical topic modeling of corpus twitter which contains offensive tweaks using automatic features. In their research, the top four machine learning algorithms which are SVM, Logistic Regression, and Random Forest (RF) have been adopted. However, among the four algorithms, the LR algorithm outperforms the other algorithms. The result of the approach is True Positive 75.1% above 4029 test tweet using Logistic Regression, a significant increase of 5.4% from the baseline [10].

**Table 8. The results of F1 logistic regression use a threshold of 0.5 of probability**

| | | #topics learned by LDA on the training data | | | | |
|---|---|---|---|---|---|---|
| **ML algorithm** | Lexicon feature | 10 | 20 | 30 | 40 | 50 |
| **Logistic Regression** | NO | 0.65 | 0.712 | 0.745 | 0.739 | 0.746 |
| | YES | 0.825 | 0.834 | 0.835 | 0.841 | 0.849 |
| **Keyword matching baseline** | | 0.787 | | | | |

The second research of the automatic annotation conducted by Samejima et al. (2015). In their study, they emphasize on the opinion of problems and solutions for automated system facilitation. An automated facilitation system is required for effective discussion of case methods without a facilitator. The facilitation system automatically captures learners' opinions by voice recognition with a microphone and provides facilitation based on opinions. The proposed method using SVM integrates classification results based on reliability annotations. of the experimental results show that the "problem" attribute is annotated at the recall rate of 83% and the precision is 81%, but the "solution" attribute is annotated at the 81% recall rate and at the precision rate of 39% [11].

Most social media used to conduct research in the NLP area uses datasets taken from Twitter. In contrast to the studies discussed above, the research conducted by Wang et al. (2012) identifies comprehensive coverage of emotional situations. According to them a previous study that discussed emotional issues only uses relatively few data sets. Therefore, they conducted research with a large labeled dataset of around 2.5 million tweets. In classifying training data, they use two different algorithms namely LIBNEAR and Multinomial Naive Bayes (MNB) to identify emotions. In addition, the algorithm is to analyze the effectiveness of various combinations of features and effects from the measurement results of training data. The experiments they performed showed variations in unigrams, bigrams, and sentiments per word that contained emotions. The highest accuracy obtained is 65.57% from the LIBNEAR algorithm [12].

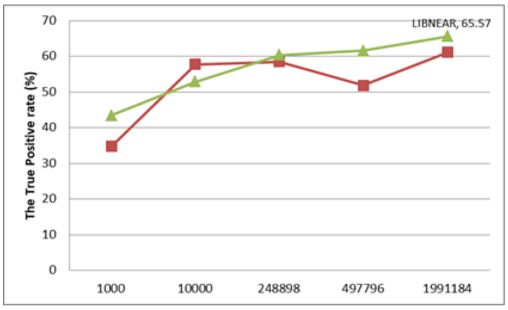


**Figure 2. Accuracy results of LIBNEAR and MNB**

The latest research for automatic annotations discussed in this study is the research conducted by Volodina et al. (2014). They took an approach that was able to identify the level of understanding of Swedish using automatic training from the corpora. In their study, they unified the methods and knowledge of using machine learning from rule-based studies of the dictionary Good Examples and from the second language learning syllabus. The proposed selection method has also been applied as a module in a free web-based language learning platform. They obtain a readability classification accuracy of 71%, which uses the performance of other models used in the same task. Furthermore, empirical evaluations with teachers and students, about seven of the ten sentences chosen are considered understandable, the rule-based approach slightly outperforms the method of combining machine learning models [13].

**Table 9. Full-featured classification results**

| Classifier | Acc | F1 | B1 Prec | B1 Recall |
|---|---|---|---|---|
| **Baseline** | 0.50 | 0.66 | 0.50 | 1.00 |
| **SVM** | 0.71 | 0.70 | 0.73 | 0.68 |

*Semi-Automatic Annotation*

This subsection discusses the algorithm used for research in semi-automatic annotation, among others by Smatana et.al (2013), Elanwar et.al (2013), Fu (2014), Liu et.al (2015), Névéol et.al (2011), Sadoun (2016).

The first study discussed is from Smatana et.al (2013), they propose a semi-automatic annotation for active learning to improve the effectiveness of document annotations using the hotel evaluation domain. The Naive Bayes classifier is used to train aspect-based sentiment classification. In the paper can be seen continuous improvement of F1-measure. The average F1 size for all aspects achieved after the annotation of all 270 sentences is over 62%. When compared the difference between annotations using active learning and those without active learning, after the annotation of 270 sentence samples, it is more than 6%. It only takes 190 sentences using active learning to get the same quality of 270 sentences without using active learning. Sentiment accuracy for both annotation tools was also measured, where the accuracy of the active learning annotation tool increased by more than 60% and the accuracy of the annotation tool without active learning was about 57% [14].

The second study to be discussed is research from Elanwar et.al (2013). They present a semi-automatic annotation tool for Arabic online handwriting datasets. This research produces a number of tools and utilities that are capable of handling handwritten data segmentation and data explanation for the training and evaluation of the word identifier. The tool performs word extraction based on the classification of the white gap between or intra-word using validation data. The SVM classifier algorithm is used for the initial word extraction proposal. After applying the test data to the system, Gap Classification Rate (GCR) was 88.4% and Word Extraction Rate (WER) 71.5% [15].

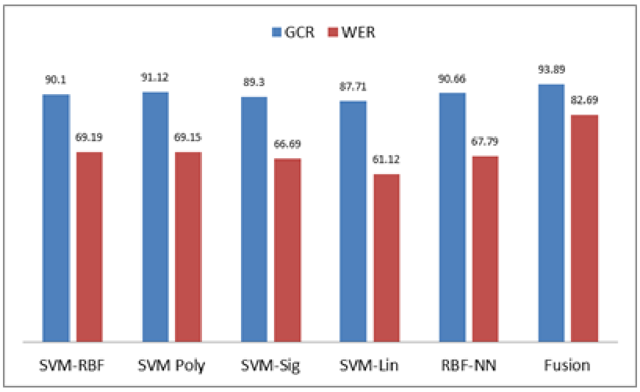


**Figure 3.Comparison of the performance of a single classification of GCR with WER**

Similar to research conducted by Smatana, Fu et.al (2013) also conducts research on semi-automatic annotation for active learning. The dataset used is in the form of Chinese language text. The focus of this research is on uncertainty-based sampling and query-based sampling algorithms to evaluate informative examples. The corpus Chinese language events (collected from various search engines such as Sina, Yahoo, Sohu and so on) are used as datasets. The corpus contains a set of L (hand labeled articles) and U sets (unlabeled articles). Set L is used to train on the initial CRF model, then the CRFs model is applied to U by using active learning algorithm (Leaf Confidence (LC) and Sequence Vote Entropy (SVE)). Both algorithms are able to improve the performance accuracy of 8.20% for LC and 8.14% for SVE, which is equal to 73% [16].

Subsequent research predicts a small set of annotation data in the forum. The results of annotations conducted by Liu et. al (2015) allows the MOOC manufacturer to summarize the state of the forum. In addition, for researchers, it is possible to better understand the role of the forum in learning. In this study, several machine learning methods were applied such as Multiclass Logistic Regression, Bayesian Model, Random Forest Model, Support Vector Machine and Kernel Method to get the maximum Cohen κ value. Of the five machine learning methods, the Random Forest Model by Randomization produces the best κ value of 0.57. Then the relevance predictions are extracted using 10 features and get the Root Means Squared Error (RMSE) as the value of accuracy. After several regression methods were tested, SVM and Linear Kernel were used and produced the lowest RMSE of 0.96. At the end, the researchers tested Comprehensibility predictions in the same way as the results of Penalized Linear Regression as the most efficient method [17]. Details of these values are detailed in table 10.

**Table 10. Cohen κ value of a number of machine learning methods for classification**

| Method | Cohen κ |
|---|---|
| Logistic Regression | 0.35 |
| Bayesian Model | 0.40 |
| Random Forest Model | **0.57** |
| SVM | 0.42 |

Research on the annotations conducted by Koncz et al. (2017) in the area of analysis sentiment said that many methods of sentiment analysis based on machine learning depend on manual annotations. But this method takes a lot of time. According to them, the active learning method can select classification tasks that are more informative and can also be used to improve the effectiveness of work annotations. Therefore, researchers conducted a survey of several active learning strategies that existed in annotating analytical sentiments.

In his research, several active learning strategies were used, namely SVM, Naïve Bayes Classifier, external models, external dictionaries, offline generated dictionaries, and online generated dictionaries. If the initial two strategies mentioned above use the technique of classifying annotated training corpora data, on external model active learning based on a set of pre-existing annotation documents to analyze sentiment. The designed model will be used to calculate the uncertainty of classification in the previous strategy. While Active learning based on external dictionaries is based on the use of dictionaries of positive and negative words. Active learning based on offline generated dictionaries is intended that the dictionary of positive and negative words is made before active learning is done. Whereas Active learning based on online generated dictionaries is formed together with the results of the annotation taking place based on the results of each iteration evaluation.

The results achieved verify the efficiency of active learning methods in the document annotation process for the needs of sentiment analysis. From active learning strategies based on uncertainty, classification of SVM-based active learning has been shown as the best performing strategy, which is 90.2% [18].

## IV. DISCUSSION& RESULTS

Based on the above review, it can be concluded that several algorithms are often used to make annotation of text data. Table 8 outline the list of the algorithm used according to the annotation methods and the accuracy results respectively.

Based on Table 8, the highest accuracy performance is the CRFs algorithm which is 98.2%. However, CRF is used on the manual annotation task only. The algorithm that is applied to all annotation methods is SVM. In fact, SVM outperforms the other algorithms in manual and semi-automatic annotation methods with a fairly high average accuracy of over 80%. Automatic annotation is not included because of the lack of literature, so it cannot be taken into account in this study. This indicates that SVM has sufficient performance to perform annotation, either manually or semi-automatic, especially for text data.

**Table 11.Review Text Annotation Algorithm**

| Annotation | Algorithm | Top Accuracy per Algorithm |
|---|---|---|
| Manual | SVM | 95% |
| | CRFs | 98.2% |
| | RNN | 69,7% |
| | TF | 71.9% |
| Automatic | Logistic Regression | 75.1% |
| | SVM | 83% |
| | LIBNEAR | 65.57% |
| Semi-Automatic | Naive Bayes Classifier | 62% |
| | SVM | 88.4% |
| | LC | 73% |
| | SVE | 73% |
| | Random Forest Model | 57% |

## V. CONCLUSION AND FUTURE WORK

Research on annotations is increasingly being done, especially in the form of text data. This study reviews algorithm always used by researchers, manual, automatic and semi-automatic. The result of the review of the menu shows that the SVM algorithm can be used for the two kinds of annotation model with high accuracy result.

Very many opportunities for researchers who want to find the accuracy value of annotation data, because there are still many algorithms as described in section introduction not yet used in research. So a better algorithm than SVM is likely to be obtained, or the possibility of a specific algorithm suitable for each annotation model.

The result of this study will be used as the basis for the use of algorithms in subsequent research that will be conducted, especially in the semi-automatic annotation method of the BahasaMelayu and Bahasa Indonesia corpus which has the same vocabulary but different meanings and polarities. The SVM algorithm will be used in the study by the authors.

## VI. ACKNOWLEDGMENT

This research work is funded by GLUAR/MPM/2017/PBPI-CTED/I00028. Thank you to the University Widyatama, Indonesia, UniversitiTeknikal Malaysia Melaka (UTeM), and Center of Advanced Computing Technology (C-ACT) for the support that is always given in conducting this research. Then I want to say many thanks to parents, husband and my two children, for your prayers and understanding were given.

## REFERENCES


1. J. Pustejovsky and A. C. Stubbs, "Natural Language Annotation for Machine Learning," pp. 1–343, 2013.
2. M. Petrillo and J. Baycroft, "Introduction to Manual Annotation," no. April, 2010.
3. V. Perez-Rosas, C. Banea, and R. Mihalcea, "Learning Sentiment Lexicons in Spanish," Proc. Eighth Int. Conf. Lang. Resour. Eval., pp. 3077–3081, 2012.
4. D. Inkpen, J. Liu, A. Farzindar, F. Kazemi, and D. Ghazi, "Location detection and disambiguation from twitter messages," Springer Sci. Media New York 2017, 2017.
5. D. S. Tarasov., "Deep recurrent neural networks for multiple language aspect-based sentiment analysis of user reviews," Proc. 21st Int. Conf. Comput. Linguist. Dialog-2015, Vol. 2, pages 77-88, 2015.
6. S. Kiritchenko and S. M. Mohammad, "Sentiment Composition of Words with Opposing Polarities," Naacl, pp. 1102–1108, 2016.
7. H. Hamdan, "Sentiment Analysis in Social Media," no. May, 2016.
8. I. Mozetič, M. Grčar, and J. Smailović, "Multilingual Twitter Sentiment Classification: The Role of Human Annotators," PLoS One, vol. 11, no. 5, p. e0155036, 2016.
9. T. Niu, "Sentiment Analysis on Multi-view Social Data," 2016.
10. G. Xiang, B. Fan, L. Wang, J. Hong, and C. Rose, "Detecting offensive tweets via topical feature discovery over a large scale twitter corpus," Proc. 21st ACM Int. Conf. Inf. Knowl. Manag. - CIKM '12, p. 1980, 2012.
11. M. Samejima, D. Hisakane, and N. Komoda, "Automatic annotation method on learners' opinions in case method discussion," Interact. Technol. Smart Educ., vol. Vol. 12, no. 2, pp. 90–99, 2015.
12. W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth, "Harnessing Twitter ' Big Data ' for Automatic Emotion Identification," 2012.
13. E. Volodina and R. Johansson, "Rule-based and machine learning approaches for second language sentence-level readability," pp. 174–184, 2014.
14. M. Smatana, P. Koncz, P. Smatana, and J. Paralič, "Active learning enhanced semi-automatic annotation tool for aspect-based sentiment analysis," SISY 2013 - IEEE 11th Int. Symp. Intell. Syst. Informatics, Proc., pp. 191–194, 2013.
15. R. Elanwar, M. A. Rashwan, and S. A. Mashali, "A Semi-Automatic Annotation Tool for Arabic Online Handwritten Text A Semi-Automatic Annotation Tool for Arabic Online," Int. J. Islam. Appl. Comput. Sci. Technol., vol. 1, no. 1 July, pp. 19–31, 2013.
16. J. Fu, Ni. Liu, and S. Wang, "Active Learning Based Semi-automatic Annotation of Event Corpus." 2014.
17. W. Liu, L. Kidzinski, and P. Dillenbourg, "Semi-automatic annotation of MOOC forum posts," Lect. Notes Educ. Technol., pp. 399–408, 2015.
18. P. Koncz et al., "Active Learning Enhanced Document Annotation for Sentiment Analysis," HAL Arch., 2017.

## AUTHORS PROFILE

**Fitrah Rumaisa,** Information Technology Department, Universitas Widyatama Jl. Cikutra 204A, Bandung, 40125, Indonesia Email: fitrah.rumaisa@widyatama.ac.id. PhD student at Universiti Teknikal Malaysia Melaka.

**HalizahBasiron,** Faculty of Information And Communications Technology UniversitiTeknikal Malaysia Melaka (UTeM), Melaka, Malaysia, halizah@utem.edu.my

**ZurinaSaaya,** Faculty of Information And Communications Technology UniversitiTeknikal Malaysia Melaka (UTeM), Melaka, Malaysia, zurina@utem.edu.my