2021 International Conference Advancement in Data Science, E-learning and Information Systems (ICADEIS)

978-1-6654-3709-7

2021 International Conference Advancement in Data Science, E-learning and Information Systems (ICADEIS) | 978-1-6654-3709-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICADEIS52521.2021.9702073

Implementation of Support Vector Machine on Text-based GERD Detection by using Drug Review Content

Asty Nabilah 'Izzaturrahmah
School of Computing
Telkom University
Bandung, Indonesia
astynabilah@student.telkomuniversity.ac.id

Fhira Nhita
School of Computing
Telkom University
Bandung, Indonesia
fhiranhita@telkomuniversity.ac.id

Isman Kurniawan
School of Computing
Telkom University
Bandung, Indonesia
ismankrn@telkomuniversity.ac.id

***Abstract*— GERD or Gastroesophageal Reflux Disease is a situation when the reflux of stomach contents leads to unpleasant symptoms and/or complications. The prevalence range of GERD is approximately 18.1% to 27.8% in North America, 8.8% to 25.9% in Europe, 2.5% to 7.8% in East Asia, 8.7% to 33.1% in the Middle East, 11.6% in Australia, and 23.0% in South America. The numbers may seem small, but GERD will lead to several complications including esophagitis, peptic stricture, and Barrett's esophagus if left untreated. The most common diagnostic test for the assessment of GERD along with its possible complications is the upper gastrointestinal endoscopy, or esophagogastroduodenoscopy (EGD). However, endoscopy has several risks. Disease detection using machine learning can be done and is needed due to the increment in medical data, new detection, and diagnostic modalities being developed. One of the machine learning algorithms often used in text classification is Support Vector Machine (SVM). This research applies SVM to do text-based classification, classifying data into two classes, namely "GERD patient" and "not GERD patient," using drug review data. The best model has 91.32% accuracy, 91% f1-score, and 91.32% AUC score with unigram as the n-gram range, and RBF with C is 1000, and gamma auto as the SVM kernel.**

Keywords*— GERD, SVM, GERD detection, disease detection

I. INTRODUCTION

GERD or Gastroesophageal Reflux Disease is a situation when the reflux of stomach contents leads to unpleasant symptoms and/or complications [1]. The characteristic symptoms of GERD include retrosternal burning (which is typically labeled as heartburn) and regurgitation [1]. The prevalence range of GERD is approximately 18.1% to 27.8% in North America, 8.8% to 25.9% in Europe, 2.5% to 7.8% in East Asia, 8.7% to 33.1% in the Middle East, 11.6% in Australia, and 23.0% in South America [2]. Although the number seems small, GERD will lead to several complications, including esophagitis, peptic stricture, and Barrett's esophagus [1]. Also, a study in Singapore [3] shows a significant increment in the prevalence of reflux-type symptoms in a multiracial Asian population. Early detection is essential to prevent malignant transformations [4]. The most common diagnostic test for assessing GERD, along with its possible complications, is the upper gastrointestinal endoscopy, or esophagogastroduodenoscopy (EGD) [4]. However, endoscopy has several risks [5]. An alternative for GERD detection is by using text classification, as it can also be used to detect other diseases [6] [7] [8] [9]. We can use drug review data for text classification since it is in the form of text. Also, some reviews contain symptoms to help us with disease detection.

Text classification is the task of categorizing a document into predefined categories [10]. Some text classification techniques are related to the TF-IDF approach to represent text in vector space [11]. Each feature in the text corresponds to a single word [11]. The experiments by Kadhim [12] show that TF-IDF gives a better result to feature extraction performance, with the highest value of F1-measure is 89.77. This value is more significant than its comparison, BM25, with the score of F1-measure, is 89.16 [12]. Text classification can be done by using the machine learning approach [13]. Several machine learning algorithms are often used in text classification, which is Naïve Bayes, SVM, KNN, and decision trees [10].

Text classification is used by Wu et al. to identify patients with carotid stenosis [6]. Ultrasound reports are classified using logistic regression as a baseline model [6]. Then, convolution and recurrent neural networks (CNN and RNN) are used to increase accuracy [6]. To process the texts, they create a parser that divides the raw text into fields [6]. Term frequency-inverse document frequency and bag-of-n-grams and are used as the features [6]. Their CNN model achieved 91.8% accuracy in predicting history and 93.8% accuracy in predicting the existence of carotid stenosis [6]. Meanwhile, their RNN-attention model gives 94.4% and 95.4% accuracy in predicting the history and existence of carotid stenosis [6].

A study about detecting the risk of depression was conducted by Havigerová et al. in 2020 [7]. The experiment shows that text classification performed by informal text has the highest recall and precision [7]. Another study related to disease detection and text classification was conducted by Samina Amin et al. in 2020 [8].]. Using deep learning and machine learning approaches, they use tweets data to detect and classify dengue disease [8]. The result is that LSTM has the most significant score in train accuracy, accuracy, precision, recall,and F1-Score [8]. López-Úbeda et al. also research disease detection using text data in 2020 [9]. The corresponding disease is COVID-19, and the data used is CT Scan reports [9]. By using Support Vector Machine (SVM), they got 89.15% accuracy [9].

Besides text data, GERD detection can be done by using other data types with still using machine learning algorithms [14] [15]. Disease detection using machine learning can be done because it has the potential to impact and is required due to the increase in medical data, new detection, and diagnostic modalities being developed [16]. The complexity of the data types and the importance of multimodal analysis are also increased [16]. Machine learning can equip clinicians with new tools forunderstanding high-dimensional and complex

datasets [16]. One example of disease detection using machine learning is done by Huang et al. in 2015 [14]. The research about GERD detection is conducted by using the endoscopy images dataset [14]. By using Hierarchical Heterogeneous Descriptor Fusion Support Vector Machine, they got 93.2% accuracy [14]. Another research about GERD detection was done by Horowitz et al. in 2007 [15]. By using a diagnostic questionnaire, the highest AUC score reached 0.957 [15]. The method used is a neural network with in-sample data analysis [15].

In this research, we aim to build a model of text-based GERD detection using drug review data. The classification method that is used is SVM, and the vectorization is performed by using TFIDF. We use SVM since it is one of the most efficient machine learning algorithms [17].

II. Literature Review

A. Related Works

Huang et al. researched GERD detection using the endoscopy images dataset in 2015 [14]. By using Hierarchical Heterogeneous Descriptor Fusion Support Vector Machine, they got 93.2% accuracy [14]. Another research about GERD detection was done by Horowitz et al. in 2007 [15]. By using a diagnostic questionnaire, the highest AUC score reached 0.957 [15]. The method used is a neural network with in-sample data analysis [15].

Disease detection can also be done by using text data as what is done by Wu et al. to detect patients with carotid stenosis in 2020 [6]. Ultrasound reports were classified using logistic regression as a baseline model. After that, convolution and recurrent neural networks (CNN and RNN) are used to increase accuracy [6]. To process the texts, they create a parser that divides the raw text into fields [6]. Term frequency-inverse document frequency and bag-of-n-grams and are used as the features [6]. Their CNN model gives 91.8% accuracy in predicting history and 93.8% accuracy in predicting the existence of carotid stenosis [6]. Meanwhile, their RNN-attention model gains 94.4% and 95.4% accuracy in the prediction of history and existence of carotid stenosis [6].

Havigerová et al. researched the detection of the risk of depression in 2020 by using text data [7]. In data collection, participants are asked to write four documents with four scenarios [7]. The steps that they do include outlier filtering low variability variables exclusion, assessment of normality, lowering granularity of depressive scale, exclusion of variables with no significant intergroup differences, creation of models, and the last one is evaluation of models [7]. The experiment gave insight that classification performed by informal text has the highest recall and precision [7].

Another study related to disease detection and classification is the one which was conducted by Samina Amin et al. in 2020 [8]. Using deep learning and machine learning approaches, they use tweets data to detect and classify dengue disease [8]. First, the data are extracted using Twitter API [8]. The next step is pre-processing, which includes removing URL, retweet, mentions, punctions, and stop words [8]. Tokenization and stemming are also done in this step [8]. Then, features are extracted from the username, tweet text, tweetlocation, and tweet timestamp using TF-IDF and n-gram models [8]. After that, data are labeled manually [8]. Then, the data are classified using logistic regression, support vector machine, naïve Bayes, ANN, and a combination of RNN and LSTM [8]. The last step is evaluating models, which results in LSTM has the most significant score in train accuracy, accuracy, precision, recall, and F1-Score [8].

López-Úbeda et al. also research disease detection using text data in 2020 [9]. The corresponding disease is COVID-19, and the data used is CT radiological reports [9]. Feature extraction is first done by using TF-IDF (Term Frequency-Inverse Document Frequency), TF-IDF by disabling the reweighting of the IDF, TF-IDF with a word-based n-grams model, TF-IDF with word-based n-grams model and disabled IDF, and FastText SUC [9]. They also use two types of SNOMED-based feature representations: TF and binary TF [9]. After that, mutual information is obtained, and feature reduction is made [9]. By using Support Vector Machine (SVM), they got 89.15% accuracy [9].

B. Text Processing

Text processing, according to the Oxford Dictionary, is "the manipulation of text, especially the transformation of text from one format to another." [18]. The majority of work on text processing and pre-processing is done in the context of information retrieval [19]. There are several critical phases of text processing, namely tokenization, term extraction, and normalization [20]. Tokenization is the process of converting a string of characters into a string of words (or tokens) [20].

Text classification is the task of categorizing a document into predefined categories [10]. Predefined categories have their own labels [20]. For example, when doing email classification, the emails may have the label “spam” or “not spam” [20]. Text data needs several amounts of pre-processing [20]. The common ways to pre-process raw text include text extraction, stop-word removal, stemming, case-folding, punctuation, and frequency-based normalization [20]. Text extraction is done by removing anchors, tags, unrelated content, and others [20]. Stop-words like pronouns, articles, and prepositions need to be removed [20]. Words with similar roots become a single representative [20]. For example, the word in verb-ing and the word in verb 2 are the same, and the root of both is the word in verb 1 [20]. One example of punctuation marks is hyphens, and it needs to be parsed carefully [20]. Word occurrence can be represented as boolean or the count of how many times the word appears in the document [10]. If the document length varies, normalization may be needed [10]. Frequency-based normalization weights words by the logarithm of the inverse relative-frequency of their existence in the collection [20]. It is known as inverse-document frequency (IDF) [20] and can be written as

$$idf = log10\left(\frac{N}{df_t}\right) \quad (1)$$

where df is the document frequency [21].

Feature selection is one of the text classification processes [10] which selects the subset of characteristics from a set of features [22]. It can be used to reduce the dimensionality of the dataset [23]. It removes features that are considered unimportant for classification [23]. Creating vector representation of text and applying learning algorithms are also processes in text classification [10]. Several machine learning algorithms, such as Naive Bayes, SVM, KNN, and

decision trees, are frequently used in text classification [10]. The main issue with using SVM for document classification is that text data must be transformed into numerical data [24]. This can be solved by using vectorization [24]. TFIDF is the most popular weighting method [25]. It displays documents in the Vector Space Model, especially in information retrieval problems [25]. TFIDF is a part of Natural Language Processing and can be interpreted as the normalized term frequency of each word appearing in the data [6]. TFIDF can be written as

$$w_{t,d} = tf_{t,d} \times idf_t \quad (2)$$

where tf represents the term frequency and idf represents the inverse document frequency [21]. The more frequently a word appears in a document, the more likely it is to be very relevant in this document [25]. However, in a case where a word appears too frequently in the text collection, it is not treated as a relevant word for the document [25]. If a word is frequent in all documents, it could be a stop word [25].

C. Support Vector Machine

Support Vector Machine (SVM) is one of the most efficient machine learning algorithms [17]. It is a computer algorithm that assigns labels by learning examples [26]. SVM is initially used to classify data into two classes after several developments, multi classes classification is also possible [27]. SVM is the best performing method and is more robust in doing various learning tasks [28]. SVM is also fully automatic, so there is no need to do manual parameter tuning [28]. SVM aims to find the hyperplane with the most significant margin, known as the Maximum Marginal Hyperplane (MMH) [28]. MMH can also maximize the ability of SVM to classify the previously unseen examples [29].

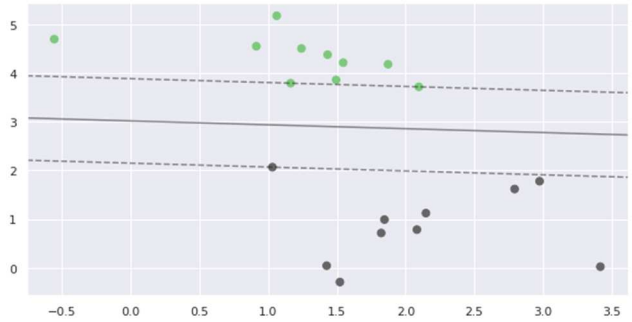

Fig. 1. Support Vector Machine

SVM classification is based on four fundamental concepts: the separating hyperplane, the maximum-margin hyperplane, the soft margin, and the kernel function [29]. Fig. 1. shows the visual representation of SVM. The dots represent the data, and the colors represent the class. The lines in the center represent hyperplane. The one line in the center of the lines is the optimum hyperplane. The hyperplane separates the two groups of points in the training set by the greatest distance [30]. A separating hyperplane is denoted as

$$W.X + b = 0 \quad (3)$$

where W represents the weight vector and b represents a scalar which is the bias [31]. Meanwhile, the two sides of the margin, or the dashed line in Fig. 1. can be written as [31]

$$H1 = w_0 + w_1x_1 + w_2x_2 \geq 1 \quad for\ y_i = +1 \quad (4)$$

$$H2 = w_0 + w_1x_1 + w_2x_2 \leq -1 \quad for\ y_i = -1 \quad (5)$$

By combining both Equation 2.4.2, the formula becomes [31]

$$y_i(w_0 + w_1x_1 + w_2x_2) \geq 1, \forall_i \quad (6)$$

The SVM algorithm can be altered by including a soft margin that allows some data points to pass through the separating hyperplane's margin without affecting the result [29]. The soft margin can be used to avoid misclassifications since it introduces a user-specified parameter that allows users to control the number of examples that are permitted to violate hyperplane and how far they can go [29]. A problem that can be faced is when there is no dividing line between the two classes, and a soft margin cannot be used to help [29]. The kernel function can solve this problem by adding dimension to the data [29]. There are four kernels, namely linear, polynomial, Gaussian RBF, and sigmoid [27].

III. Methodology

This research is done by applying several steps. The steps include data collection, data pre-processing, performing SVM model, parameter tuning, and model validation. Tokenization will also be performed to help stop word removal and lemmatization.

A. Data Collection

The first step in this research is downloading data from machine learning repository [32]. The data initially consists of data train and data test. There are 112329 rows in the data train and 48280 rows in the data test. We only use two among seven columns available, namely condition and review. The condition of patients is predicted by using *review*. *Condition* consists of GERD patients and non-GERD patients. Text in the *review* column is used to get information about the corresponding label. The example of data is shown in Table 1. Then, the data is balanced by selecting random data from non-GERD label so the percentage between the amount of data of each class are the same. There are 496 rows of data for each label in the data train. Meanwhile, in the data test, there are 219 rows of data for each label. Then, we split train data into data train and data validation using train test split function from sklearn with 7:3 ratio.

TABLE I. Example of Data in Each Label

Review	Label
"Doctor gave me this drug saying that it will help for ulcer, it did not help, pain is still unbearable, but I do not have reflux, which makes it a little bit better."	GERD
"Worsen my son. It gave him insomnia and increased the tics and he gained a lot of weight, so we stopped it."	Non-GERD
"Tried everything, been on Protonix for 5 years. I can now eat anything as spicy and as hot as I like, and do not have acid reflux."	GERD
"I have been on this medication for about 4 months. I have 3 broken areas in my lumbar spine, as well as 5 herniated discs. It works okay for me. "	Non-GERD

B. Data Pre-processing

Data pre-processing is done by removing punctuation, removing numbers, removing stop words, and lemmatization. The first thing to do is removing punctuation. Table 2. shows the example of data pre-processing. The first thing to do is removing punctuation, followed by removing numbers. We can see that full stop, comma, and quotes are removed. The number 5 is also removed. After that, stop words like 'been', 'on', 'i', 'can', 'now', 'as', 'do', 'not', and 'have' are removed. Then, lemmatization is performed. Some words are changed in this step. For example, 'tried' becomes 'try' and 'years' becomes 'year'.

TABLE II. EXAMPLE OF DATA PRE-PROCESSING

Pre-processing Step	Pre-processed
Original text	"Tried everything, been on Protonix for 5 years. I can now eat anything as spicy and as hot as I like, and do not have acid reflux."
Removing punctuation	tried everything been on protonix for 5 years i can now eat anything as spicy and as hot as i like and do not have acid reflux
Removing numbers	tried everything been on protonix for years i can now eat anything as spicy and as hot as i like and do not have acid reflux
Removing stopwords	tried everything protonix years eat anything spicy hot like acid reflux
Lemmatization	try everything protonix year eat anything spicy hot like acid reflux

Before implementing the SVM model, we transform the data into numerical form. This transformation is done by creating a vector representation of text. Vectorization is be done by using TFIDF. In this step, we train the data train and predict the data validation. We use SVM with baseline parameters from sklearn as the classifier. Then, the accuracy of the different ranges of n-gram is compared. The one with the best accuracy is chosen and applied to the model.

C. Parameter Tuning

In this step, we train the data train and predict the data test. The best parameter is searched by using GridSearchCV. There are 3 kernels in which accuracy is compared, namely linear, polynomial, and Radial Basis Function.

TABLE III. PARAMETERS AND ITS VALUE USED FOR TUNING

Parameters	Values
C	0.001, 0.01, 0.1, 1, 10, 100, 1000
Gamma	Scale, Auto
Degree	2, 3, 4, 5

As shown in Table 3., for the linear kernel, we only set the values of C. Meanwhile, for the polynomial kernel, we set the values of C, the type of gamma, and degree. On the other side, for the Radial Basis Function (RBF), we only set the values of C and the types of gamma.

D. Model Validation

The performance of the SVM model is evaluated by using, F1-score, and AUC. Accuracy can be written as[31]

$$accuracy = \frac{True\ Positives + True\ Negatives}{Positives + Negatives} \quad (7)$$

Recall, Precision, and F-factor are the common ways to evaluate the result of machine learning experiments [33]. In a medical context, recall is also considered primary, with the goal of identifying all Real Positive cases [33]. Recall or sensitivity is true positive rate [31]. Recall is defined by [31]

$$recall = \frac{True\ Positives}{Positive\ Samples} \quad (8)$$

Precision represents what percentage of tuples labeled as positive are actually positive [31]. Precision is described in [31]

$$precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (9)$$

F1 score or F-score is a single measure that uses precision and recall to count it and is defined as [31]

$$F = \frac{2\ x\ precision\ x\ recall}{precision + recall} \quad (10)$$

ROC graph can also be used to evaluate machine learning algorithms as done Spackman [34], one of the first researchers who used it [35]. Our model is also evaluated using the area under the ROC curve, abbreviated as AUC. The y-axis of this curve is the true positive rate, and the x-axis is the false positive rate [35]. AUC shows the likelihood that a randomly selected positive subject is suspected of being positive rather than a randomly selected negative subject [36].

IV. RESULTS AND DISCUSSIONS

After cleaning data, the n-gram comparison is performed. By using different n-gram ranges, validation data is predicted. The n-gram comparison gives some insights that n-gram affects the accuracy of the model. As shown in the Fig. 2., we can see that the n-gram range which contains unigram gives the highest accuracy. This result is the same as an experiment conducted by Aman and Szpakowicz [37] and it also validates their premise that unigrams can help learn vocabulary distribution to do accurate predictions. We can also see in Table 4. that unigram has the smallest number of features. This can affect the classification performance since it is somewhat difficult to select good features when there are a lot of features [38]. For the further steps, we take unigram as the n-gram range with 100% percentage of used features since as we can see in Table 4., it gives the highest accuracy.

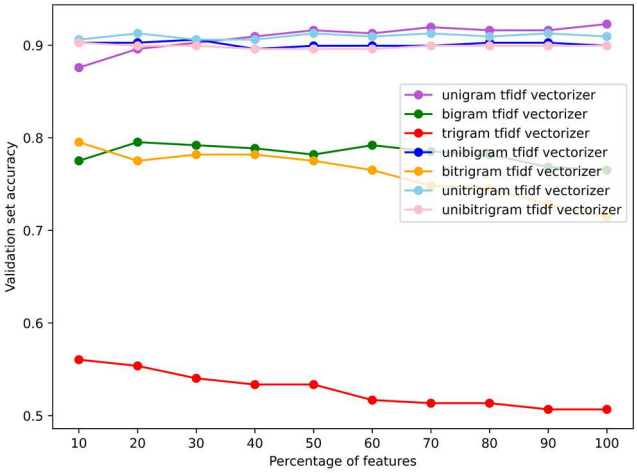

Fig. 2. Result of comparation of accuracy between n-grams and percentage of features.

TABLE IV. THE BEST ACCURACY AND THEIR BEST PERCENTAGE OF FEATURES FROM EACH N-GRAM RANGE

N-gram Range	Total Features	Best Percentage of Features	Accuracy
unigram	**3392**	**100**	**92.28%**
bigram	19276	20	79.53%
trigram	23475	10	56.38%
unibigram	22668	30	90.60%
unitrigram	2686700	10	79.53%
bitrigram	42751	20	91.28%
unibitrigram	46143	10	90.27%

Parameter tuning then performed by using GridSearchCV to see which kernel and parameters gives the highest accuracy. The best parameter for each kernel is shown in Table 5. As we can see, the best C value for linear and poly is 1, the best gamma for poly kernel is scale, and the best degree for poly kernel is 2. Meanwhile, the best C and gamma for RBF kernel is 1000 and auto respectively.

TABLE V. COMPARISON BETWEEN ACCURACY OBTAINED FROM EACH PARAMETER

Kernel	Best C	Best Gamma	Best Degree
Linear	1	-	-
RBF	1000	Auto	-
Poly	1	Scale	2

TABLE VI. FINAL RESULT OF PREDICTION

Data used	Kernel	TP	FP	FN	TN	Accuracy	Recall	Precision	F1-score	AUC
Train	Linear	348	5	2	339	98.99%	98.55%	99.41%	98.98%	98.99%
	RBF	346	4	16	328	97.12%	95.35%	98.80%	97.04%	97.10%
	Poly	**350**	**0**	**0**	**344**	**100%**	**100%**	**100%**	**100%**	**100%**
Validation	Linear	139	21	7	131	**90.60%**	86.18%	94.93%	**90.34%**	90.69%
	RBF	141	23	5	129	**90.60%**	84.87%	96.27%	90.21%	**90.72%**
	Poly	141	24	5	128	90.27%	84.21%	96.24%	89.82%	90.04%
Test	Linear	200	22	19	197	90.64%	89.95%	91.20%	90.57%	90.64%
	RBF	**208**	**27**	**11**	**192**	**91.32%**	**87.67%**	**94.58%**	**91%**	**91.32%**
	Poly	205	25	14	194	91.10%	88.58%	93.27%	90.87%	91.10%

The accuracy shown in Fig. 3. is obtained by predicting test data. As shown in Fig. 3., the accuracy of the model after tuning is done is higher than before when using the RBF kernel. The RBF kernel has the highest accuracy. We can say that RBF is the most suitable kernel for this dataset. RBF is good as the first choice in selecting SVM kernels [39]. RBF kernel is more flexible since it acts as a linear SVM [40] while it can also handle nonlinear relationships between class labels and attributes [39]. Also, RBF has fewer numerical difficulties compared to polynomial [39]. We can also see that parameter tuning is needed to increase the accuracy of the model.

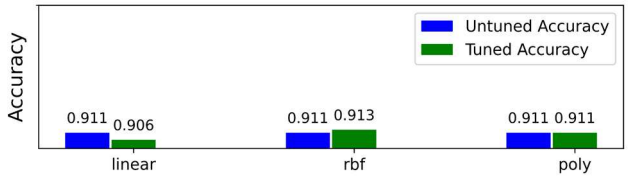

Fig. 3. Comparison of accuracy of each kernel before and after tuning

After applying the best parameters from the result of experiments, the final model validation score can be obtained and shown in Table 6. The higher accuracy, F1-score, and AUC value of train data are obtained by using poly kernel with 100% score. The poly kernel also resulting in the best True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) value for train data. The best accuracy, which is 90.60% is obtained by both linear and RBF kernel for validation data. Meanwhile, the best F1-score is obtained by linear kernel and the best AUC is obtained using RBF kernel. For test data, the most suitable kernel is RBF with 91.32% accuracy, 91% f1-score, and 91.32% AUC score.

V. CONCLUSIONS

Based on the result of the experiments, SVM successfully classifies GERD patients and non-GERD patients. The SVM model gives the best result when using unigram as the n-gram range RBF kernel. Unigram is the best n-gram range for the dataset since it gives the highest accuracy. RBF is considered as the best kernel because it has the highest value on every evaluation measurement with the accuracy, F1-score, and AUC score are 91.32%, 91%, and 91.32% respectively.

REFERENCES

[1] N. Vakil, S. V. van Zanten, P. Kahrilas, J. Dent, R. Jones, and the Global Consensus Group, "The Montreal Definition and Classification of Gastroesophageal Reflux Disease: A Global Evidence-Based Consensus," *The American Journal of Gastroenterology*, vol. 101, no. 8, pp. 1900–1920, Aug. 2006, doi: 10.1111/j.1572-0241.2006.00630.x.

[2] H. B. El-Serag, S. Sweet, C. C. Winchester, and J. Dent, "Update on the epidemiology of gastro-oesophageal reflux disease: a systematic review," *Gut*, vol. 63, no. 6, p. 871, Jun. 2014, doi: 10.1136/gutjnl-2012-304269.

[3] S. L. Lim, W. T. Goh, J.-M. J. Lee, T. P. Ng, K.-Y. Ho, and CONTRIBUTING MEMBERS OF THE COMMUNITY MEDICINE GI STUDY GROUP, "Changing prevalence of gastroesophageal reflux with changing time: Longitudinal study in an Asian population: Changing prevalence of GER in Asia," *Journal of Gastroenterology and Hepatology*, vol. 20, no. 7, pp. 995–1001, May 2005, doi: 10.1111/j.1440-1746.2005.03887.x.

[4] D. M. Clarrett and C. Hachem, "Gastroesophageal Reflux Disease (GERD)," *Missouri Medicine*, vol. 115, no. 3, p. 214, Jun. 2018.

[5] "Endoscopy: Types, preparation, procedure, and risks," Dec. 18, 2017. https://www.medicalnewstoday.com/articles/153737 (accessed Oct. 16, 2020).

[6] X. Wu, Y. Zhao, D. Radev, and A. Malhotra, "Identification of patients with carotid stenosis using natural language processing," *Eur Radiol*, vol. 30, no. 7, pp. 4125–4133, Jul. 2020, doi: 10.1007/s00330-020-06721-z.

[7] J. M. Havigerová, J. Haviger, D. Kučera, and P. Hoffmannová, "Text-Based Detection of the Risk of Depression," *Front. Psychol.*, vol. 10, p. 513, Mar. 2019, doi: 10.3389/fpsyg.2019.00513.

[8] S. Amin *et al.*, "Recurrent Neural Networks With TF-IDF Embedding Technique for Detection and Classification in Tweets of Dengue Disease," *IEEE Access*, vol. 8, pp. 131522–131533, 2020, doi: 10.1109/ACCESS.2020.3009058.

[9] P. López-Úbeda, M. C. Díaz-Galiano, T. Martín-Noguerol, A. Luna, L. A. Ureña-López, and M. T. Martín-Valdivia, "COVID-19 detection in radiological text reports integrating entity recognition," *Computers in Biology and Medicine*, vol. 127, p. 104066, Dec. 2020, doi: 10.1016/j.compbiomed.2020.104066.

[10] M. Ikonomakis, S. Kotsiantis, and V. Tampakas, "Text Classification: A Recent Overview," p. 7.

[11] Y. Zhang, L. Gong, and Y. Wang, "An improved TF-IDF approach for text classification," *J. Zhejiang Univ. Sci.*, vol. 6, no. 1, pp. 49–55, Jan. 2005, doi: 10.1631/jzus.2005.A0049.

[12] A. I. Kadhim, "Term Weighting for Feature Extraction on Twitter: A Comparison Between BM25 and TF-IDF," in *2019 International Conference on Advanced Science and Engineering (ICOASE)*, Zakho - Duhok, Iraq, Apr. 2019, pp. 124–128. doi: 10.1109/ICOASE.2019.8723825.

[13] A. Khan, B. Baharudin, L. H. Lee, and K. Khan, "A Review of Machine Learning Algorithms for Text-Documents Classification," *JAIT*, vol. 1, no. 1, pp. 4–20, Feb. 2010, doi: 10.4304/jait.1.1.4-20.

[14] C.-R. Huang, Y.-T. Chen, W.-Y. Chen, H.-C. Cheng, and B.-S. Sheu, "Gastroesophageal Reflux Disease Diagnosis Using Hierarchical Heterogeneous Descriptor Fusion Support Vector Machine," *IEEE transactions on bio-medical engineering*, vol. 63, Aug. 2015, doi: 10.1109/TBME.2015.2466460.

[15] N. Horowitz, M. Moshkowitz, Z. Halpern, and M. Leshno, "Applying Data Mining Techniques in the Development of a Diagnostics Questionnaire for GERD," *Dig Dis Sci*, vol. 52, no. 8, pp. 1871–1878, Dec. 2005, doi: 10.1007/s10620-006-9202-5.

[16] P. Sajda, "MACHINE LEARNING FOR DETECTION AND DIAGNOSIS OF DISEASE," *Annu. Rev. Biomed. Eng.*, vol. 8, no. 1, pp. 537–565, Aug. 2006, doi: 10.1146/annurev.bioeng.8.061505.095802.

[17] S. Karamizadeh, S. M. Abdullah, M. Halimi, J. Shayan, and M. javad Rajabi, "Advantage and drawback of support vector machine functionality," in *2014 International Conference on Computer, Communications, and Control Technology (I4CT)*, Langkawi, Malaysia, Sep. 2014, pp. 63–65. doi: 10.1109/I4CT.2014.6914146.

[18] "Text Processing | Definition of Text Processing by Oxford Dictionary on Lexico.com also meaning of Text Processing," *Lexico Dictionaries | English*.

https://www.lexico.com/definition/text_processing (accessed Nov. 10, 2020).
[19] V. Cho, B. Wüthrich, and J. Zhang, "Text Processing for Classification," p. 31.
[20] C. C. Aggarwal, *Machine Learning for Text*. Cham: Springer International Publishing, 2018. doi: 10.1007/978-3-319-73531-3.
[21] D. Jurafsky and J. H. Martin, *Speech and Language Processing (3rd draft ed.)*. Stanford Univ, 2019.
[22] K. J. Cios, W. Pedrycz, R. W. Swiniarski, and L. A. Kurgan, Eds., *Data mining: a knowledge discovery approach*. New York, NY: Springer, 2007.
[23] G. Forman, "An Extensive Empirical Study of Feature Selection Metrics for Text Classification," p. 27.
[24] D. Isa, L. H. Lee, V. P. Kallimani, and R. RajKumar, "Text Document Preprocessing with the Bayes Formula for Classification Using the Support Vector Machine," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 9, pp. 1264–1272, Sep. 2008, doi: 10.1109/TKDE.2008.76.
[25] P. Soucy, "Beyond TFIDF Weighting for Text Categorization in the Vector Space Model," p. 6.
[26] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, Jul. 1992.
[27] Suyanto, *Machine Learning Tingkat Dasar dan Lanjut*. Informatika Bandung.
[28] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features," in *Machine Learning: ECML-98*, vol. 1398, C. Nédellec and C. Rouveirol, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 137–142. doi: 10.1007/BFb0026683.
[29] W. S. Noble, "What is a support vector machine?," *Nat Biotechnol*, vol. 24, no. 12, pp. 1565–1567, Dec. 2006, doi: 10.1038/nbt1206-1565.
[30] Y. Lin, "Support Vector Machines for Classification in Nonstandard Situations," *SUPPORT VECTOR MACHINES*, p. 12.
[31] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques 3rd Edition*, 3rd ed. Elsevier, 2011.
[32] S. Kallumadi and F. Gräßer, *Drug Review Dataset (Drugs.com) Data Set*. 2018. Accessed: Nov. 10, 2020. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29
[33] D. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation," p. 25.
[34] K. A. Spackman, "SIGNAL DETECTION THEORY: VALUABLE TOOLS FOR EVALUATING INDUCTIVE LEARNING," in *Proceedings of the Sixth International Workshop on Machine Learning*, Elsevier, 1989, pp. 160–163. doi: 10.1016/B978-1-55860-036-2.50047-3.
[35] Jin Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 3, pp. 299–310, Mar. 2005, doi: 10.1109/TKDE.2005.50.
[36] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve.," *Radiology*, vol. 143, no. 1, pp. 29–36, Apr. 1982, doi: 10.1148/radiology.143.1.7063747.
[37] S. Aman and S. Szpakowicz, "Using Roget's Thesaurus for Fine-grained Emotion Recognition," p. 7.
[38] V. Sugumaran and K. I. Ramachandran, "Effect of number of features on classification of roller bearing faults using SVM and PSVM," *Expert Systems with Applications*, vol. 38, no. 4, pp. 4088–4096, Apr. 2011, doi: 10.1016/j.eswa.2010.09.072.
[39] V. Apostolidis-Afentoulis, "SVM Classification with Linear and RBF kernels," 2015, doi: 10.13140/RG.2.1.3351.4083.
[40] S. S. Keerthi and C.-J. Lin, "Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel," *Neural Computation*, vol. 15, no. 7, pp. 1667–1689, Jul. 2003, doi: 10.1162/089976603321891855.