3rdMalikussaleh International Conference
on Multidiciplinary Studies 2022 (3rdMICoMS 2022)

3rd MICoMS
Malikussaleh International Conference on Multidisciplinary Studies

INTERNATIONAL CONFERENCE PROCEEDINGS ISSN: xx/xxx | RESEARCHARTICLE

Obtaining Elderly Patients' Lifestyle Information from Unstructured Text Sources

Defry Hamdhana[1,2*]AsmaulHusna[2]

[1] *Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, Japan*
[2]*Informatics, Universitas Malikussaleh, Indonesia*
[*]*Corresponding author. Email: hamdhana.defry205@mail.kyutech.jp*

ABSTRACT

In this work, we made many simulations to take information from free-text notes belonging to the patient that indicates his or her habits. Detailed information about a patient's habits will allow the nurse to create a personalized daily schedule. Due to the fact that each patient has a different routine. The information is separated into five categories: dietary habits, drinking/smoking habits, excretion habits/toilet style, fashionable/colour preference/footwear, and favourite music/radio/TV shows. To realize this, we use six machine learning models, there are Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Multinomial Logistic Regression (MLR), and Gradient Boosting (GB). As a result, all models have more than 90% accuracy on both train sets and test sets.In this study, we focus on SVM with Train Set Accuracy score 93.9% and Test Set Accuracy score 94.1%, which is more resistant to overfitting issues.

Keywords:*information extraction, unstructured text, patient lifestyle, SVM*

1. INTRODUCTION

In the modern world, the number of elderly people is increasing significantly [1], which means that they need special care [2], such as facilities that are specifically designed for parents or caregivers who are assigned to assist the elderly in carrying out their daily lives and assisting them with their daily tasks. In some cases, there are conditions in which the elderly are unable to recognize what their need, have difficulty achieving daily necessities, or have a hard time communicating what they really want or need[8]. It is common for them to have this condition, which can cause them to be irritable and aggressive easily, be depressed easily, and to even be angry at people[3]. In contrast, care facilities, hospitals, or even caregivers already have their own schedules that they implement in the elderly's daily routines[4], contrary to what is stated above. This will cause confusion for the elderly and cause them to become angry and do other things that are destructive actions because the routine schedule given is not in accordance with their habits, or reminds them of trauma.

In light of this background, we conducted research to find out what habits the elderly have through various digitalsources like clinical notes, patient diaries, health records, social media, and so on in order to discover what they do. The challenge lies in the fact that these sources are still unstructured, so a model is needed to categorize this vast amount of data. In order to classify them into five categories, we decided to put them into the following: dietary habits, drinking/smoking habits, excretion habits/toilet style, clothing colour preference, style of footwear and lastly, favourite television shows/music/radio stations.

2. RELATED WORKS

With the help of Natural Language Processing (NLP), Xin Zhou et al. have been able to automatically extract Alzheimer's patient information from electronic health records (EHRs)[5]. Information from the clinical note is extracted and converted into indexed UMLS concepts. In the next step, they identify information regarding the lifestyle that Alzheimer's patients maintain. A lifestyle assessment is the first step to identifying the right intervention strategy for the patient.

In another study, Maxim Topaz et al developed an NLP algorithm that can identify common neuropsychiatric symptoms of Alzheimer's disease and related dementias (ADRD) that can be found in the free text clinical records of home healthcare clinicians (HHC)[6]. As a result, symptom clusters and the frequency of inpatient visits by symptomatic or asymptomatic patients are then described in the report.

In this study, we aimed to obtain information about patient habits from clinical notes, patient diaries, health records, social media, etc.Our aim is to classify five categories for each sentence in the source by performing several scenarios and simulations based on 6 machine learning methods in order to do this.

3. DATASET

We used a dataset in this study that was generated ourselves based on the 5 categories that we identified in the beginning of this study. In the context of health, we take a few categories from a wide range of resources that are closely related to the field of health. However, we manually generate several other categories in the online application due to difficulties in obtaining them. For more details, we can see the following table:

Table 1. The dataset of our study

No	Category	Data sources	Amount of data
1.	Dietary habits	https://sentence.yourdictionary.com/	250
2.	Drinking/smoking habits	n2c2-NLP-Research-Data-Sets[7]	129
3.	Excretion habits/toilet style	https://www.scie.org.uk/dementia/living-with-dementia/difficult-situations/using-the-toilet.asp	129
4.	Fashionable/color preference/footwear	https://sentence.yourdictionary.com/	199
5.	Favorite tv/music/radio	https://sentence.yourdictionary.com/	204

4. METHODS

We first do some feature engineering before we manage the data that will be processed with the use of some machine learning methods that we will be able to train the algorithm using the raw text documents Here are the steps:

1. Text cleaning: cleaning of special characters, down casing, punctuation signs. possessive pronouns and stop words removal and lemmatization.
2. Label coding: creation of a dictionary to map each category to a code.
3. Train-test split: 70% for training and 30% for testing.
4. Text representation (evaluation): use of TF-IDF scores to represent text.

4.1. Random Forest

Random Forest is a decision tree ensemble derived from a large number of individual decision trees that work together as a unit. Each tree in Random Forest issued the class and class predictions with the most votes to be the prediction of the proposed model. The low correlation between models is the key. Uncorrelated models can produce ensemble predictions that are more accurate than individual predictions with low correlation and combine to create a portfolio that is greater than the sum of its parts. Basically, the trees protect each other from their individual errors (as long as they don't constantly all err in the same direction). The trees are able to move in the right direction as a group even if some trees are wrong.

4.2. *Support Vector Machine*

The basic function of SVM is to find a separating line (or hyperplane) between two classes of data. Data are input into SVM, which outputs a line that separates classes if possible. The SVM algorithm finds the points closest to the line from both classes. These points are referred to as support vectors. In the next step, we calculate the distance between the line and the support vectors. This distance is called the margin. The goal is to maximize margins. Hyperplanes with maximum margins are optimal hyperplanes.

4.3. *K-Nearest Neighbors*

K-NN algorithm is based on the assumption that similar things exist in close proximity to each other. Similar things tend to be close together. It is essential that this assumption be true for the KNN algorithm to be useful. A KNN captures the concept of similarity (sometimes referred to as distance, proximity, or closeness), which we might have first learned as children when we were calculating the distance between the two points of a graph.

4.4. *Multinomial Logistic Regression*

Multinomial Logistic Regression is an extension of logistic regression that solves multiclass problems given multiple independent variables. This model predicts the probabilities of the categorically dependent variable, which may have more than one outcome class. Logistic regression is used when the dependent categorical variable has two outcome classes, such as students passing an exam or a bank manager rejecting a loan application.

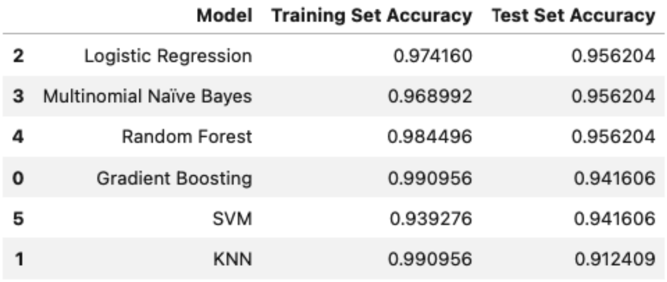
4.5. *Gradien Boosting*

One of the most powerful techniques for building predictive models is gradient boosting. Three elements are involved in gradient boosting is a loss function to be optimized, a weak learner to make predictions, an additive model to add weak learners to minimize the loss function.

5. RESULTS

In the Training Set Accuracy, Gradient Boosting and KNN had the highest score with 99.09%, followed by Multinomial Naïve Bayes with a score of 96.89%. And SVM has the lowest score with 93.92%. However, in Test Set Accuracy, the highest scores were Logistic Regression, Multinomial Naïve Bayes, and Random Forest with a score of 95.62%. While the lowest score for Test Set Accuracy is KNN with 91.24%. Here we can focus on SVM because only this model has an increased score between Training Set Accuracy to Test Set Accuracy. It is a good indicator because it is unlikely to be overfitted. For more details, see Table 2 below.

Table 2. The results of each model



	Model	Training Set Accuracy	Test Set Accuracy
2	Logistic Regression	0.974160	0.956204
3	Multinomial Naïve Bayes	0.968992	0.956204
4	Random Forest	0.984496	0.956204
0	Gradient Boosting	0.990956	0.941606
5	SVM	0.939276	0.941606
1	KNN	0.990956	0.912409

In order to gain a better understanding of Test Set Accuracy owned by SVM, here is a confusion matrix that we can see as a way to explore this subject further.

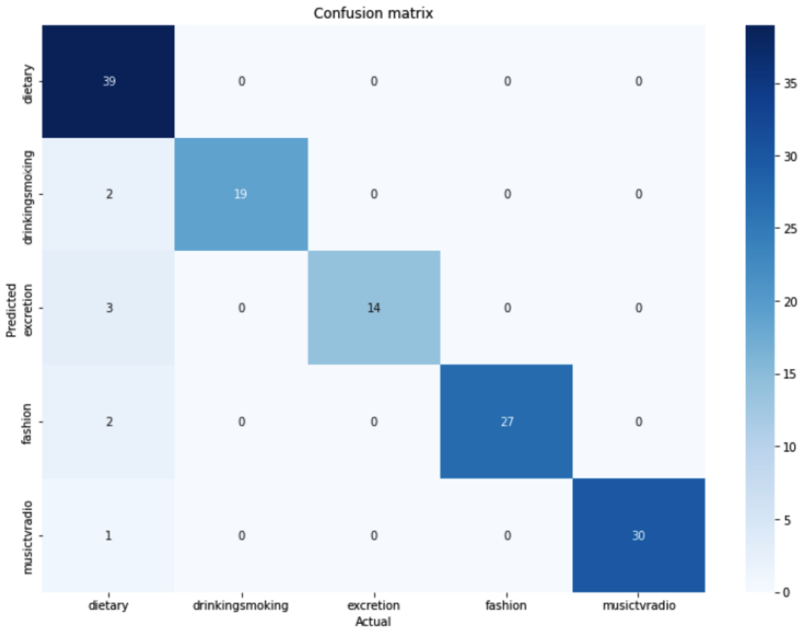
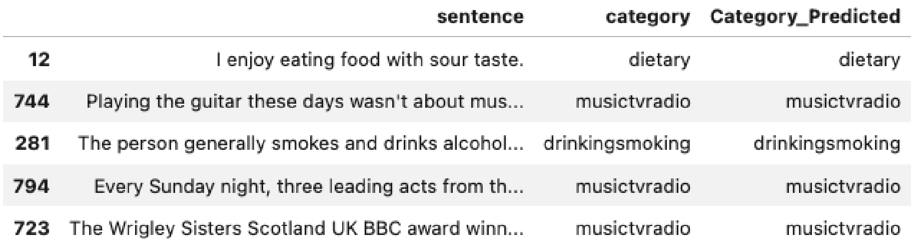


Figure 1 Confusion matrix of SVM model

The confusion matrix above indicates that all the data tested were successfully classified into five categories as a result of the classification. Although some test data are classified into the wrong category, in this case, some prediction are classified as dietary habits. However, it is still within the accepted amount.Based on the results of SVM model, the following Table 3 illustrates how the 5 categories of patient habits were correctly predicted.

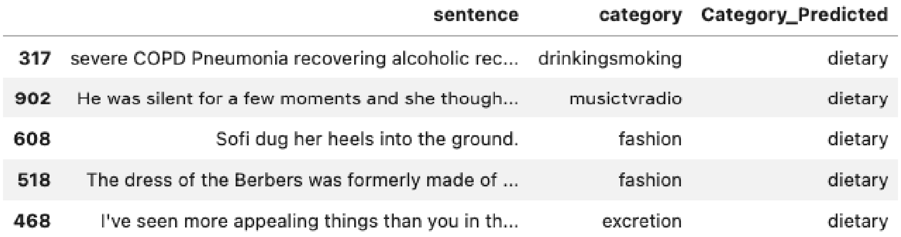
Table 3. Precise predictions using the SVM model



	sentence	category	Category_Predicted
12	I enjoy eating food with sour taste.	dietary	dietary
744	Playing the guitar these days wasn't about mus...	musictvradio	musictvradio
281	The person generally smokes and drinks alcohol...	drinkingsmoking	drinkingsmoking
794	Every Sunday night, three leading acts from th...	musictvradio	musictvradio
723	The Wrigley Sisters Scotland UK BBC award winn...	musictvradio	musictvradio

Additionally, as we can see in the Confusion Matrix image above, the SVM model sometimes makes inaccurate predictions. For the example of an incorrect prediction made by SVM is shown in Table 4.

Table 4. Incorrect predictions using the SVM model



	sentence	category	Category_Predicted
317	severe COPD Pneumonia recovering alcoholic rec...	drinkingsmoking	dietary
902	He was silent for a few moments and she though...	musictvradio	dietary
608	Sofi dug her heels into the ground.	fashion	dietary
518	The dress of the Berbers was formerly made of ...	fashion	dietary
468	I've seen more appealing things than you in th...	excretion	dietary

6. CONCLUSION

The initial step of extracting information by classifying sentences in the raw source into 5 categories is very likely to be carried out by the 6 models we propose above. This is proven by the train and test accuracy scores which all reach above 90%. Due to the limited amount of data and limited time, we chose SVM as the best model. Consequently, only SVM has a Test Set Accuracy score higher than Train Set Accuracy. In our opinion, this is a good indication that the training set is not overfitted. In order to make this issue clear, it would certainly be helpful to test with larger data sets.

AUTHORS’ CONTRIBUTIONS

In this study, our contribution is to carry out a simulation to classify random sentences into 5 patient habit categories. In this simulation, we also generated our own dataset because data for several categories such as dietary habits, fashionable/color preferences/footwear, and favorite tv/music/radio could not be found at that time. We also chose SVM to be the best model for classifying sentences in this case with the considerations from Table 2 which indicated an overfit training set.

ACKNOWLEDGMENTS

I would like to thank Sozolab for providing me with many insights into the breadth of artificial intelligence, as well as input and solutions that have always kept me motivated to study. I also want to express my gratitude to Universitas Malikussaleh for giving me the opportunity to study in my dream country, Japan. It is my pleasure to thank the co-author of this paper, Asmaul Husna, who continues to strive to understand the world of information technology with passion.

REFERENCES

[1] World Health Organization, World report on ageing and health.,World Health Organization, 2015.

[2] Nitschke, I., and I. Kaschke, Special care dentistry for dependent elderly and people with disabilities, Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz 54.9, 2011, pp. 1073-1082.

[3] Margari F, Sicolo M, Spinelli L, Mastroianni F, Pastore A, Craig F, Petruzzelli MG. Aggressive behavior, cognitive impairment, and depressive symptoms in elderly subjects, Neuropsychiatric disease and treatment, (8) 2012, pp. 347.

[4] Li C, Cheung WK, Liu J, Ng JK., Automatic extraction of behavioral patterns for elderly mobility and daily routine analysis. ACM Transactions on Intelligent Systems and Technology (TIST). 9(5), 2018, pp.1-26.

[5] Zhou X, Wang Y, Sohn S, Therneau TM, Liu H, Knopman DS., Automatic extraction and assessment of lifestyle exposures for Alzheimer’s disease using natural language processing, International journal of medical informatics, (130) 2019, pp. 103943.

[6] Topaz M, Adams V, Wilson P, Woo K, Ryvicker M., Free-text documentation of dementia symptoms in home healthcare: a natural language processing study, Gerontology and Geriatric Medicine, 2020.

[7] Healthcare P. n2c2 nlp research data sets, 2020. URL: https://portal. dbmi. hms. harvard. edu/projects/n2c2-nlp.

[8] Hamdhana, Defry, Mobile application for caregiver in collecting statistical data of BPSD attack focused on macro activities: PhD forum abstract, Proceedings of the 18th Conference on Embedded Networked Sensor Systems, 2020.