International Journal of Mathematics and
Computer Science, **17**(2022), no. 4, 1553–1561

Extroversion Text Classification Model Using Logistic Regression

Sakolwan Peetaneelavat, Somjai Boonsiri

Department of Mathematics and Computer Science
Faculty of Science
Chulalongkorn University
Bangkok, Thailand

email: 6378018323@student.chula.ac.th, somjai.b@chula.ac.th

(Received May 5, 2022, Accepted June 8, 2022)

Abstract

Chatbots have become a popular alternative in customer support due to their ability to overcome human limitations. Even though it is an automatic machine approach, interaction pleasure must be considered. Many studies claim that people also get along well with those who have similar personalities, even with the machine. A chatbot can assign its talking styles through interactive language design by using keywords representing extroversion or introversion. In the interactive language design testing process, the Extroversion Text Classification Model (ETCM) is created using the Logistic Regression algorithm. This study suggests that using part of speech and unique words portion as a feature with Term Frequency-Inverse Document Frequency (TF-IDF) outperforms the model beats TF-IDF alone.

1 Introduction

Many business owners consider customer satisfaction to be the most important factor. Everyone recognizes that more satisfying products or services can earn more benefits. Interaction with the customers is one of the service processes. Many psychology researchers claim that individuals get along with

Key words and phrases: Text classification, Machine learning, Personality type, Logistic Regression.
AMS (MOS) Subject Classifications: 68T20, 68T50.
Somjai Boonsiri is the corresponding author.
ISSN 1814-0432, 2022, http://ijmcs.future-in-tech.net

persons who have similar personalities or lifestyles. Before or during the contact we may make educated guesses about the clients personalities. Another prominent service platform for overcoming human constraints is chatbots since they can provide customer information anytime, causing fewer errors. According to research about making conversations with the chatbot more personalized [2], it was found that matching a customer with a chatbot having a similar talking style can bring more satisfaction during interaction with the chatbot. The simplest way to assign a talking style for a chatbot is by designing its responding languages, but that still has some challenges since we could not be sure whether those designed responding languages represent the kinds of personality: Extrovert and Introvert. So, we need to develop a machine learning model to perform the text classification for text style prediction to be the testing phase of designing the responding languages of the chatbot. Research shows how the Extroversion Text Classification Model (ETCM) can be improved by employing Term Frequency-Inverse Document Frequency (TF-IDF) and unique word terms portion as selected features. The suggested work's chosen machine learning model is logistic regression, with accuracy, precision, recall, and F1 score as model assessment metrics.

The structure of the remaining section of this paper is as follows: In Section 2, we provide related theories and technologies. In section 3, we present presents the methodology for developing the proposed Extroversion Text Classification Model (ETCM). In Section 4, we demonstrate the proposed model's experiment in terms of performance and evaluation. Finally, we conclude this study in Section 5.

2 Literature Reviews

Many studies show that people who have similar personalities get along well. These can be retrieved from the language used in daily conversation. Beukeboom et al. [1] found the style of words typically used in text messages and speaking languages. A person with a high level of introversion is more likely to use fewer words overall and to use more negative keywords. Extroverts, on the other hand, are more prone to use adjectives and sociality terms. Furthermore, extroversion language is shown to be more abstract than that of other personality types.

Since artificial intelligence machines have become more popular these days, many brands or businesses have decided to utilize chatbots as customer care representatives and so the user experience is one of the most important elements to be considered. Shumanov and Johnson [2] demonstrated the engagement experiment between chatbot and users and found that matching users with congruent chatbots, which are manipulated to be more personalized by designing the interactive language as an extrovert or introvert person, can increase more rates of acceptance and adoption from the users. It is vital to recognize the sentence features and word choice of an extrovert and introvert person, when building the interactive language of chatbots, which can use machine learning in conjunction with Natural Language Processing (NLP). Many research papers have studied personality prediction based on text classification using machine learning methods, such as Naive Bayes, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). Pratama and Sarno [3] demonstrated that combining these traditional methods increased performance accuracy. The research on Emotion Recognition by Textual Tweets Classification [4] also proposed the voting text classifier, which predicted emotion using the text using an average of each labeled class probability of two machine learning algorithms, Logistic Regression (LR) and Stochastic Gradient Descent (SGD). It comes up with better performance using Term Frequency-Inverse Document Frequency (TF-IDF). However, research on the Thai language is still scarce. It is one of the languages with a unique structure and many limitations and so implementing the specific Natural Language Processing library for Thai is required. PyThaiNLP is chosen in this work since it is a python package for text processing and linguistic analysis focusing on the Thai language [5].

3 Proposed Methods

In this section, the methology of ETCM is demonstrated.
There are 5 essential processes of the proposed methods shown in Figure 1.

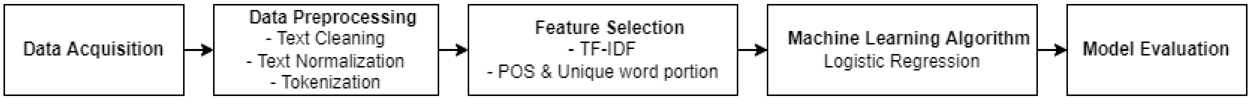


Figure 1: Essential Processes of the proposed methods

3.1 Data Acquisition

The following three columns appear in 600 records of Train data, 200 records of Validation data, and records of Testing data: content, content_len, and personal_type. The posting and comments from the public MBTI Facebook group make up the material collection. The length of the content string is represented by content_len. Finally, the column personal_type denotes the personality type of each contents posters which E being Extrovert while I denotes Introvert.

3.2 Data Preprocessing

After data acquisition, all data in column content is preprocessed by text cleaning and normalization, likely to noise removal processes. Aids in the removal of extraneous symbols such as the asterisk (*) and question mark (?), as well as hashtags, punctuation, and separator (\n or \t). Text normalization restores the usual form of words that have been misspelled or reordered. The contents are then tokenized using word segmentation and stored in the content_token. Another important stage in preprocessing which has not been done in this study is the elimination of stop words. Stop words are the group of regularly used terms in a language that is generally eliminated from the original material since they have no bearing on the content's meaning [6]. However, in this work, the emotive meaning of the text is not the major goal, but the set of stop words is one of the most important word values for text extroversion categorization.

3.3 Feature Selection

3.3.1 Term Frequency-Inverse Document Frequency (TF-IDF)

Term Frequency (TF) measures how many times terms of words appear in the document. The TF(t) calculation is the number of term t that appears in the document divided by the total number of terms in the document. Meanwhile, Inverse Document Frequency (IDF) measures the term importance. The IDF(t) calculation is log(e) of the total number of the document divided by the number of documents that term t appeared there. Then, the TF-IDF score of each word is computed from its TF values multiple by its IDF value [7].

3.3.2 Part of Speech and Unique words portion

The amount of each word recognized as each considered kind of word is divided by the total number of words in the content and their sum is used for part of speech and unique words scoring.

$$\sum \frac{Number\, of\, marked\, word\, as\, each\, word\, type\, in\, each\, content}{Number\, of\, words\, in\, content} \tag{3.1}$$

3.4 Machine Learning model: Logistic Regression

Logistic regression is one of the most popular machine learning algorithms for text classification tasks, especially for binary classes. It performs probabilities estimating of a discrete outcome given an input variable [8] [9]. In this work, we develop two versions of the Extroversion Text Classification Model (ETCM), each with a distinct feature selection technique. The first model just has TF-IDF as a characteristic, but the second adds part of speech and unique words to the equation.

3.5 Model Evaluation

In order to compare the performance of the suggested models, the confusion matrix (see Figure 2) is utilized to compute accuracy, precision, recall, and F1-score. The formulas of accuracy, precision, recall, and F1-score are applied as follows:

$$Accuracy = \frac{TPs + TNs}{P + N} \tag{3.2}$$

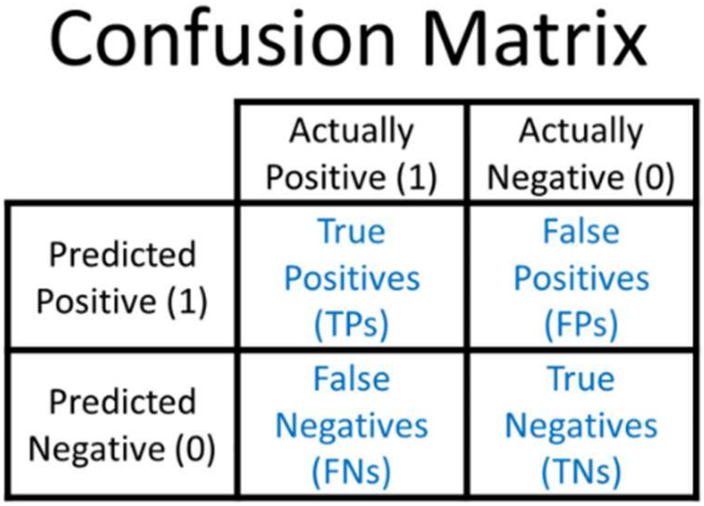
$$Precision = \frac{TPs}{TPs + FPs} \tag{3.3}$$

$$Recall = \frac{TPs}{TPs + FNs} \tag{3.4}$$

$$F1 - score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{3.5}$$

4 Results and Discussion

In this section, we give details of the performance of both versions of ETCM and compare their efficiency.



Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 2: Confusion matrix [10]

4.1 Performance of ETCM using only TF-IDF

The confusion matrix of ETCM is depicted in Figure 3 using the TF-IDF feature. The top left rectangle represents true positives (TPs) indicating that the model correctly predicts that content owners are extroverted people. Similarly, the bottom right one represents true negatives (TNs) indicating that the model correctly predicts that content owners are introverted individuals. Meanwhile, the top-right and bottom-left columns represent false positives (FPs) and false negatives (FNs). Both of these sections indicate the amount of time the model predicted incorrectly the extraversion of testing contents. When tested with 200 testing sample data, ETCM with TF-IDF as a feature has TPs, TNs, FPs, and FNs of 52, 66, 45, and 37, respectively.

4.2 Performance of ETCM using TF-IDF and Part of Speech and Unique words portion

Similarly, Figure 4 illustrates another version of the ETCM model's confusion matrix including the Part of Speech and Unique words portions as selected features. When tested with 200 testing sample data, the version of ETCM has TPs, TNs, FPs, and FNs of 56, 66, 45, and 33, respectively.

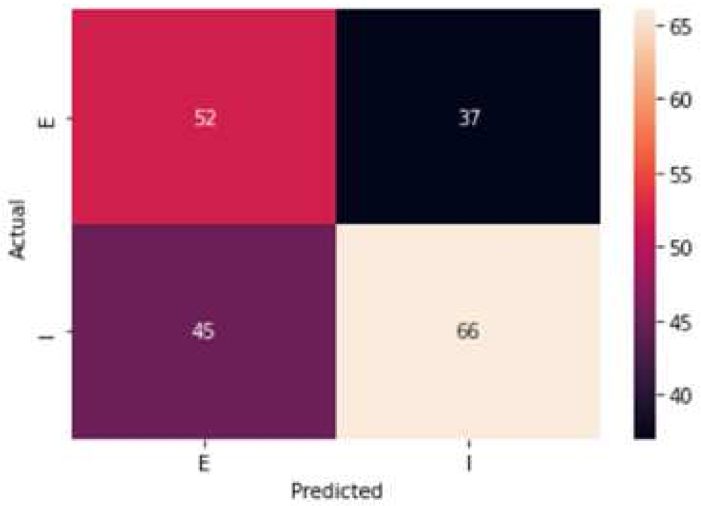


Figure 3: Confusion Matrix of ETCM using TF-IDF as a feature and be tested with 200 sample data

Table 2: Comparison of performance between two versions of ETCM

	ETCM using TF-IDF only	**ETCM using both**
Accuracy	0.59	0.61
Precision	0.584	0.629
Recall	0.536	0.554
F1-Score	0.559	0.589

Table 2 demonstrates that the second version of ETCM adding part of speech and unique words portion as another feature gets a sightly better score in terms of the confusion matrix. It was found that considering the word choice used in the text can help the model predict extroversion from the text. This study has some limitations in the amount of sample data and unique dictionary of parts of speech, and unique words that the authors needed to gather manually. However, the result shows that it is still working and it improves the performance of ETCM by having more accuracy. Hence, if we keep developing this dictionary (the model learns more vocabulary) and test with more sample data, ETCM would certainly provide better performance.

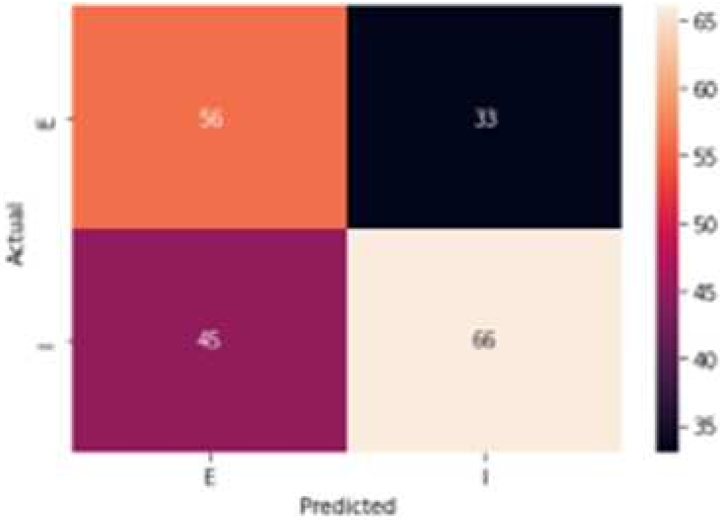


Figure 4: Confusion Matrix of ETCM using TF-IDF and Part of Speech and Unique words portion as selected features and be tested with 200 sample data

5 Conclusion

The authors have developed the Extroversion Text Classification Model (ETCM) to aid the talking styles for the interactive language of the customer support chatbot. Using the Logistic Regression model, TF-IDF, and part of speech unique words portion score, we have built the ETCM. The proposed methods have five essential steps. To begin with, data was collected from posts in public Facebook groups. Data preparation and feature selection are the next phase. The logistic regression model is then used, and the model's performance is assessed using the confusion matrix. Consequently, we discovered that our suggested technique outperforms the model employing TF-IDF as a feature alone in accuracy, precision, recall, and F1-score. As we anticipate improved techniques for our ETCM in the future, we will apply additional machine learning algorithms and overcome the existing dataset and part of speech vocabulary in the Thai language.

References

[1] Camiel Beukeboom, Martin Tanis, Ivar Vermeulen, The Language of Extroversion Extraverted People Talk More Abstractly, Introverts Are More Concrete, Journal of Language and Social Psychology, **32**, (2013), 191–201.

[2] Michael Shumanov, Lester Johnson, Making conversations with chatbot more personalized, Computers in Human Behavior, **117**, (2021), 106627.

[3] B. Y. Pratama, R. Sarno, Personality classification based on Twitter text using Naive Bayes, KNN and SVM, International Conference on Data and Software Engineering, (2015), 170–174, doi: 10.1109/ICODSE.2015.7436992

[4] A. Yousaf et al., Emotion Recognition by Textual Tweets Classification Using Voting Classifier (LR-SGD). In IEEE Access, **9**, (2021), 6286–6295, doi: 10.1109/ACCESS.2020.3047831.

[5] PyThaiNLP, (2022), [Online], Available: https://pythainlp.github.io/, [Accessed: Apr. 13, 2022]

[6] Kavita Ganesan, What are Stop Words?, (2019), [Online], Available: https://www.opinosis-analytics.com/knowledge-base/stop-words-explained/, [Accessed: Apr. 20, 2022]

[7] D. E. Cahyani, A. F. Faishal, Classification of Big Five Personality Behavior Tendencies Based On Study Field with Twitter Analysis Using Support Vector Machine, 7th International Conference on Information Technology, Computer, and Electrical Engineering, (2020), 140–145, doi: 10.1109/ICITACEE50144.2020.9239130.

[8] Ravinder Ahuja, Aakarsha Chug, Shruti Kohli, Shaurya Gupta, Pratyush Ahuja, The Impact of Features Extraction on the Sentiment Analysis, Procedia Computer Science, **152**, (2019), 341–348, https://doi.org/10.1016/j.procs.2019.05.008.

[9] Thomas W. Edgar, David O. Manz, Research Methods for Cyber Security, 2017.

[10] Pagon Gatchalee, Confusion Matrix, an important tool for evaluating predictive outcomes in machine learning, (2019), [Online], Available: https://medium.com/@pagongatchalee/, [Accessed: Apr. 23, 2022]