Chat Topic Classification for Student Counseling using Text Mining

Akkapon Wongkoblap, Bongse Varavuddhi Muenyuddhi, Phichayasini Kitwatthanathawon, Satidchoke Phosaard, Thara Angskun, Warissadee Duangrasee and Jitimon Angskun*

Institute of Digital Arts and Science, Suranaree University of Technology, Thailand

ABSTRACT

Background and Objectives: Alongside their standard duties, providing guidance and support to students is a crucial responsibility for teachers. The effective counseling of students is enhanced by technology, which improves communication between teachers and students. This study employs text-mining techniques to analyze and categorize chat messages exchanged on social media between teachers and students, demonstrating how technology can facilitate this communication effectively.

Methodology: This research implements a text-mining technique involving natural language processing (NLP) to create a classification model for chat topics. Chat messages used in text mining are collected from 20 students and 10 teachers using the Line and Facebook Messenger apps. Messages from both platforms had similar characteristics and were combined for analysis. The 4,500 chat messages were gathered, and after the cleaning process, 2,610 messages remained. The classification model is accomplished with Term Frequency–Inverse Document Frequency (*TF-IDF*) and three machine learning methods: random forest (*RF*), support vector machines (*SVM*), and logistic regression (*LR*) to build text classifiers. The model will be used to predict the counseling objectives of students.

Main Results: The evaluation of model performance utilized a 10-fold cross-validation technique due to the small size of the dataset, which helps prevent overfitting. The experimental results showed that the model using the *RF* technique achieved the highest accuracy among all techniques, with an overall F1 score of 89.55 percent. This was followed by the *SVM* at 88.68 percent and *LR* at 88.06 percent. When analyzing the models based on chat topics, the highest F1 score was recorded for the topic titled "Leave," followed by "Urgency," "Score," and "Homework."

Discussions: The RF technique consistently yielded the highest values in all chat topics. These results indicate that the RF technique is the most effective at accurately classifying chats compared to other techniques. Moreover, the evaluation of the technique's performance in this study found that the model's errors were caused by the model identifying many duplicate

ARTICLE INFO

Article history:
Received 27 August 2024
Revised 16 December 2024
Accepted 3 January 2025

Keywords:
Chat Topic,
Natural Language Processing,
Student Counseling,
Text Mining,
Topic Classification

words across all chat topics. These words are not typically used in data analysis to identify relationships. Thus, future analyses may involve using language experts to eliminate these words.

Conclusions: The research findings can be used to categorize chats and predict their topic for student counseling. These findings can also be used to develop automated communication tools, such as integrated chatbots with e-learning. In addition, the model helps to resolve issues and streamline communication, reducing student wait times. However, the designed system has some limitations. It requires an extensive vocabulary corpus for each type of chat topic to improve the model's accuracy using text-mining techniques. Creating a vocabulary corpus for each type of chat topic necessitates linguistic experts and consumes significant time. Furthermore, the data being analyzed is collected from social media, which includes emerging vocabulary, such as chat language, that presents challenges for the model. Several improvements can be made shortly. For instance, the developed model can be improved using deep learning techniques and engaging linguistic experts to understand word characteristics and chat language better.

*Corresponding author

E-mail address: jitimon@g.sut.ac.th

Introduction

Natural Language Processing (NLP) utilizes advanced computer technology to understand and interpret human languages, which have evolved like the Thai language (Chowdhary, 2020). Since most of the data we want to analyze comprises natural language, the technology must understand and use this data effectively. The ability to process natural language is essential for bridging the communication gap between humans and technology.

Various sectors, such as machine translation, information retrieval, text summarization, and customer demand analysis, use NLP widely. It is commonly utilized for communication purposes and reducing costly burdens, such as the need for customer service staff to respond to chats. NLP involves analyzing natural language from a broader perspective and also delves into the processing of "chat language," which refers to language that deviates from standard rules, such as altered spellings and writing styles seen in informal or human language (Prachaming et al., 2017). Chat language also encompasses frequent misspellings and new slang people use to express emotions, particularly on social media platforms, which are widely used for communication.

The current use of NLP technology in education requires improvement, especially in facilitating communication between teachers and students. Communication challenges have been identified as significant problems affecting both parties. A survey and interview with ten university teachers revealed that a typical instructor teaches more than 30 students, highlighting their burdens and challenges. Some universities, like Suranaree University of Technology, have

large class sizes, with over a thousand students per class section. Consequently, communication issues significantly impact both students and teachers, ultimately impacting their work productivity. For instance, teachers' responsibilities increase when they must address multiple student inquiries or when they need to communicate information about scores and assignments. Preliminary analysis of data from social media chats, including Line and Facebook, indicated that a single conversation between a teacher and students lasts, on average, for over 8 hours because teachers or students might use the mute function to be not disturbed (Bouhnik & Deshen, 2014). This waiting time contributes to increased workloads and is detrimental to handling other tasks' efficiency.

The researchers were interested in studying how NLP technology can improve communication between teachers and students. This research gathered messages from commonly used social media platforms such as Line and Messenger. Ten thousand messages were collected and analyzed using machine learning and deep learning models. Text mining using RapidMiner ensured the accuracy and efficiency of the NLP technology. The goal is to reduce teachers' workloads, allowing them to better adhere to their schedules and help students complete tasks more effectively. This approach aims to bridge the communication gap between teachers and students, benefiting both groups.

Review of Literature

This research focuses on text mining, chat topic classification, and chatbots in educational settings. The following review covers the most relevant studies published to date.

Text Mining

The critical task for NLP is to understand the text. Text-mining techniques have been developed to extract content and knowledge from unstructured text. In the early stages of text mining, researchers primarily used term-based methods, which involved counting the number of words in a text. However, this approach encounters challenges related to polysemy (where a word like 'bank' can mean a financial institution or the side of a river) and synonymy (where different words like 'big' and 'large' share similar meanings). Phrase-based methods were developed to address these issues, which combine words into phrases. However, this method has its limitations in that some phrases have a low frequency of occurrence.

Concept-based methods have been introduced, capturing concepts' frequency rather than just words (Gaikwad et al., 2014). A well-known concept-based method is Term Frequency–Inverse Document Frequency (*TF-IDF*), which calculates the significance of the term "t" in each document "d" within a corpus. The equation for *TF-IDF* is represented in Equation 1.

$$TF - IDF(t, d) = TF(t, d) \times IDF(t) \quad (1)$$

In the given context, $TF(t, d)$ stands for Term Frequency, which indicates the number of times a specific term appears in a set of documents. In contrast, $IDF(t)$ represents Inverse Document Frequency, calculated as in Equation 2.

$$IDF(t) = log \frac{n}{1 + DF(t)'} \quad (2)$$

Where n is the total number of documents in the corpus, $DF(t)'$ denotes the number of documents containing the term t (Kim & Gil, 2019; Manning et al., 2008). Tabassum and Patil (2019) also reported that *TF-IDF* performed better than other models, such as bag-of-words and count vectorizer, because it highlighted important features and eliminated irrelevant features.

Chat Topic Classification

Topic classification is an essential part of text mining. It involves algorithms automatically assigning specific categories or types to textual content. Hingmire (2013) introduced a text classifier that uses an unsupervised generative probabilistic model called Latent Dirichlet Allocation (LDA) to assign topics to documents automatically. Although LDA is commonly used for unsupervised tasks such as topic modeling or text classification, it has drawbacks. These include generating redundant or non-specific topics, which can impact its effectiveness.

Supervised machine learning algorithms have successfully classified textual content into relevant categories. These techniques can be combined with text mining approaches to categorize text based on predefined labels (Kowsari et al., 2019). Cahyani and Patasik (2021) effectively implemented text classifiers using several supervised machine learning models with *TF-IDF* to classify text into five types of emotions. Similarly, Hendry et al. (2021) proposed a topic modeling approach to classify messages within customer service chats. Kowsari et al. (2019) reviewed text classification literature. They noted that commonly used models include random forest (RF), support vector machine (SVM), and logistic regression (LR) due to their computational efficiency, ease of implementation, and firm performance.

Chatbot in Education Settings

Rapid and effective responses can quickly improve communication between two parties. The development of chatbots in recent years has facilitated immediate online communication, allowing for quick and focused interactions. Students, teachers, and administrators often communicate via chatrooms in educational settings. However, due to other pressing tasks, teachers and administrators may only sometimes be available to respond to students' questions. To address this issue, Ali et al. (2022) implemented NEdBOT, using NLP and DialogFlow, to respond to frequently asked questions from students in real time. Calabrese et al. (2022) proposed an intelligent chatbot to suggest didactic materials to students in a Massive Open Online Course (MOOC) system. Sreelakshmi et al. (2019) proposed a specialized chatbot designed for answering questions and generating quizzes based on user-uploaded documents. Deng and Yu (2023) also found that chatbot-assisted learning has a medium-to-high positive effect on learning outcomes, particularly in reasoning, achievement, retention, and interest.

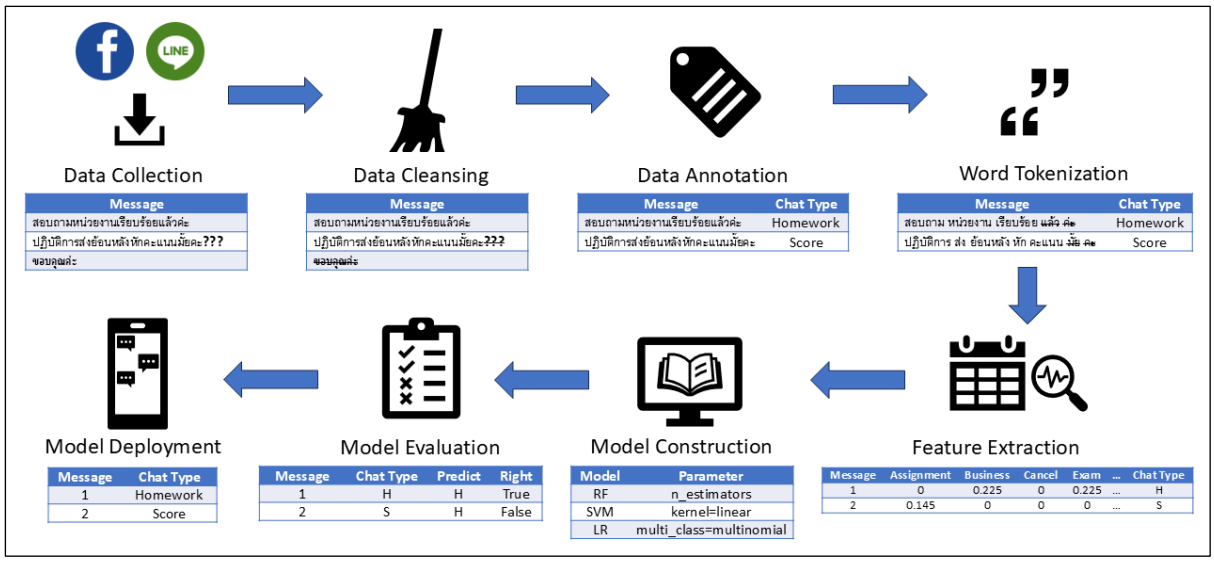
Based on a detailed survey, existing research still needs to develop a topic classifier for Thai chat messages between students and teachers. This study aims to create a text classification model to identify topics within these messages. The proposed model could enhance educational communication by applying supervised machine learning with *TF-IDF* feature extraction, supporting more effective teaching and learning.

Research Methods

The process of using text mining for chat topic classification in student counseling is visually illustrated in Figure 1. There are eight main steps involved in this process: 1) data collection, 2) data cleansing, 3) data annotation, 4) word tokenization, 5) feature extraction, 6) model construction, 7) model evaluation, and 8) model deployment. The details of each process are as follows.

Figure 1

A Model Framework for Classifying Student Counseling Conversations



Data Collection

This step involves gathering data from conversations between teachers and students. Conversation data was collected for this research from chat platforms such as Line and Facebook Messenger, which students commonly use to communicate with teachers. Chat messages were collected from 10 teachers and 20 students who communicate frequently. The chats collected included messages from several courses, as students are enrolled in multiple courses with multiple teachers. After receiving the conversation data, all messages were recorded in Microsoft Excel by saving the data as a CSV file for further model development.

Data Cleansing

This step manipulates conversation messages by removing personal details such as first and last names or student identification numbers. Conversations that learning techniques cannot use, such as punctuation marks and duplicate characters, were removed to ensure the text is in a regular word format. For example, the word "a lotttttt" would be changed to "a lot" by deleting the extra "t's," and the word "score is badddd 555555+" would be changed to "score is bad" by deleting the extra "d's" and "555555+".

Data Annotation

This step identifies the conversations between the teacher and the student. Based on the research of Bloch (2002) and this research endeavor, four main conversation types were defined: homework, score, leave, and urgency (as shown in Table 1). Categorizing the chat topics when

students send messages to the chatbot can help annotate the data. The developed model can classify the messages and automatically respond as if interacting with the teacher.

Table 1

Chat Topic by Conversation Messages

Chat Topic	Conversation Message
Homework	Consult about homework, assignment, or project
Score	Consult about homework or test scores
Leave	Consult about sick leave or absence from class
Urgency	Consult about an emergency

Word Tokenization

This step involved breaking conversational sentences into individual words and removing phrases, prepositions, pronouns, and nouns. Punctuation such as question marks (?), exclamation marks (!), and commas (,) were also removed from the text by using regular expressions. However, these punctuations are used sparingly in Thai chat conversations. Furthermore, this step includes replacing words with similar meanings with the same word. This research used the PyThaiNLP library to tokenize words in Thai sentences, convert text into space-separated words for analysis, and utilize cloud software, such as Google Colaboratory, for development purposes.

Feature Extraction

This step involved analyzing two issues. Firstly, the part of speech of given words was determined to eliminate insignificant words from sentences. Secondly, the frequency of words in each chat topic was found to understand the sender's purpose by using numerical statistics such as word frequency and inverse document frequency and applying them in text mining. This step transformed textual data by vectorization or numerical high-dimensional vectors.

Model Construction

This step involved applying machine learning techniques to create a chat topic classification model. This research chose to use three techniques and determined the parameters for each technique as follows:

1) *RF* technique: Selects the parameter value n_estimators (default value = 1000).

2) *SVM* technique: Uses the parameter Kernel = linear.

3) *LR* technique: The parameter settings will be configured for Multi-Class Classification.

Model Evaluation

This step verified the model's accuracy in classifying chats using four standard values: *Precision*, *Recall*, and *F-measure* (*F*1) as in Equation 3-5.

$$Precision = \frac{\sum_{c=0}^{N} \frac{TP_c}{(TP_c + FP_c)}}{N} \times 100 \quad (3)$$

$$Recall = \frac{\sum_{c=0}^{N} \frac{TP_c}{(TP_c + FN_c)}}{N} \times 100 \quad (4)$$

$$F1 = \frac{\sum_{c=0}^{N} \frac{2 \times (Recall_c \times Precision_c)}{(Recall_c + Precision_c)}}{N} \times 100 \quad (5)$$

Where C_i represents N chat topics when $0 \leq i \leq N$ and N equals 4.

TP (True Positive) is the number of messages on chat topic C_i that the model correctly predicts as chat topic C_i.

TN (True Negative) is the number of messages not on chat topic C_i that the model correctly predicts as not being on chat topic C_i.

FP (False Positive) is the number of messages not of chat topic C_i that the model incorrectly predicts as chat topic C_i.

FN (False Negative) is the number of messages of chat topic C_i that the model incorrectly predicts as not being of chat topic C_i.

Model Deployment

The developed model can be used in various applications, including web, mobile, and chatbots, to improve the efficiency of chat message responses. The model works by categorizing chats into topics and providing automated responses for each topic, which reduces waiting time for conversations. The results of the model deployment will be discussed later.

Experimental Results and Discussion

The research results are divided into three parts. The first part covers the design and development of the chat topic classification model. The second section presents the model evaluation results, and the final section describes the model deployment results.

Results of Model Design and Development

The experimental results are divided into five main categories based on the research process outlined in the previous section. The aspects are as follows:

1) Results of Data Collection and Cleansing: This study collected chat messages from 20 students and 10 teachers using the Line and Facebook Messenger apps. Participation was voluntary, and the data collection process followed privacy regulations. Messages from both platforms had similar characteristics and were combined for analysis. Each message represents a single communication from the sender to the recipient. Four thousand five hundred chat messages were gathered and saved in a CSV file. Subsequently, the data underwent cleaning procedures, resulting in 2,610 messages remaining after the cleaning process.

2) Results of Data Annotation: This process involves categorizing the conversation messages (or chats) between teachers and students. After cleaning, the messages were categorized into four main topics and 12 subtopics. The classification was carried out by analyzing conversations involving the input of two individuals who agreed to use them as training and test data. Messages where there was disagreement were excluded. Two thousand five hundred conversation messages or chats were categorized, as detailed in Table 2.

Table 2

Determining Chat Topics in Student Counseling: 4 Main Topics and 12 Subtopics

Chat Topic	Number of chats	Subtopic	Definition
Homework	604	(A) Inquiring about homework content	The conversation inquired about homework, including explaining the assignment and asking for help on how to do it.
		(B) Inquiring about the details of submitting homework or the submission date	The conversation inquired about the homework schedule, including the deadline for submission, the date of presentation, and the steps for submitting homework.
		(C) Inquiring about assignments	The conversation inquired about assignments or projects, like applying for university email, watching videos for e-learning lessons, and taking quizzes.
Score	405	(A) Quiz or exam scores	The conversation inquired about scores, such as pre and post-test scores and midterm and final exam scores.
		(B) Other scores	The conversation inquired about exercise or assignment scores and included messages related to inquiring about check-in and activity participation scores.
Leave	601	(A) Leave of absence	The conversation inquired about taking leave for essential business.
		(B) Sick leave	The conversation inquired about requesting leave due to a physical or mental condition that hinders studying.
		(C) Leave of absence document	Documents required for leave include: - Request for leave during studies, using both business and sick leave. - Request for leave during exams. - Other documents, such as a medical certificate.
Urgency	1000	(A) Requesting personal consultation or conversation	The conversation inquired about scheduling consultations as soon as possible, including private discussions.
		(B) Inquiring about exam information	The conversation inquired about the details of upcoming exams, including the rules and content.

Table 2
(Cont.)

Chat Topic	Number of chats	Subtopic	Definition
		(C) Notification of class date or sudden cancellation of class	The conversation inquired about upcoming teaching appointments, notifications of upcoming classes, or sudden class cancellations.
		(D) Other urgency	The conversation inquired about any other urgent matters the sender needed to notify the receiver about and requested a prompt response.

3) Results of Word Tokenization: In this step, all the collected conversation messages were processed by breaking sentences into words and removing phrases, prepositions, pronouns, and nouns. Additionally, words with similar meanings are standardized to facilitate feature extraction and data relationship analysis in the next step. An example of word tokenization is illustrated in Table 3.

4) Results of Feature Extraction: After tokenizing the text, the data was processed to extract features by analyzing two main aspects: 1) a vocabulary analysis to determine the part of speech of words in order to eliminate insignificant words from sentences, and 2) finding the frequency of words in each chat topic to indicate the sender's intention, using numerical statistics of *TF-IDF*. The results of the data analysis are presented below.

Table 3
Examples of Word Tokenization

Chat Topic	Conversation Message (in Thai)	Result (in Thai)
Homework (A)	สวัสดีค่ะอาจารย์ หนูอยากปรึกษาอาจารย์	~~สวัสดี ค่ะ~~ ~~อาจารย์~~ หนู อยาก ปรึกษา อาจารย์
Homework (B)	อาจารย์คะ ถ้าหนูขอเลื่อนส่งรายงาน	~~อาจารย์ คะ~~ ถ้า หนู ขอ เลื่อน ส่ง รายงาน
Homework (C)	สอบถามโครงงานหน่อยค่ะ	สอบถาม โครงงาน ~~หน่อย ค่ะ~~
Score (A)	ปฏิบัติการส่งย้อนหลังหักคะแนนม๊ายคะ	ปฏิบัติการ ส่ง ย้อนหลัง หัก คะแนน ~~ม๊าย คะ~~
Score (B)	หนูอยากทราบคะแนน รายวิชา นี้ ค่ะ	หนู อยาก ทราบ คะแนน รายวิชา ~~นี้ ค่ะ~~
Leave (A)	สวัสดีค่ะ หนูขออนุญาตลาไปทำธุระ	~~สวัสดี ค่ะ~~ หนู ขอ อนุญาต ลา ไป ทำ ธุระ

4.1) Results of Vocabulary Analysis that Determines the Part of Speech of a Word. This research analyzes the placement and frequency of words within sentences to determine their function. The results are presented in Table 4, which demonstrates how natural language processing is used at the syntactic level to understand sentences better. This analysis filters out insignificant word features, such as words used in introductory phrases, to get to the main content of the sender's message.

Table 4

Examples of Determining the Function of Words

Word Function	Word	Word Frequency
Opening a conversation	Lecturer	1,204
	Hello	1,045
Entering the sender's purpose	Query	109
	Excuse me	383
	Leave	295
	Exam	248

4.2) Results of Finding the Frequency of Words or Frequent Words in Each Chat Topic. This research examined the sender's intention using numerical statistics *TF-IDF*. Table 5 displays the results, indicating that while some words may appear in many chat topics, they may not accurately represent those chat topics. Therefore, *TF-IDF* statistics were utilized to identify the features of each chat topic. The *IDF* value helps eliminate words frequently found in many categories.

After that, the *TF-IDF* statistics were utilized to transform all words or text into a vector space model, as illustrated in Table 6.

Table 5

Example of Frequent Words in Each Chat Topic

Chat Topic	Frequent Words (Number of Words)
Homework	
(A) Inquiring about homework content	Requirement (44), Homework (42), Inquiry (22)
(B) Inquiring about the details of submitting homework or the submission date	Work (74), Inquiry (32), Date (23)
(C) Inquiring about assignments	Assignment (20), Project (18), Week (17)
Score	
(A) Quiz or exam scores	Score (187), Midterm (61), Final (52)
(B) Other scores	Score (196), Announcement (76), Inquiry (34)
Leave	
(A) Leave of absence	Business (115), Leave (97), Permission (48)
(B) Sick leave	Please (180), Sick (98), Leave (74)
(C) Leave of absence document	Document (51), Absence (43)
Urgency	
(A) Requesting personal consultation or conversation	Consult (122), Personal (67), Tomorrow (34)
(B) Inquiring about exam information	Exam (169), Inquiry (112), Rule (52)
(C) Notification of class date or sudden cancellation of class	Cancel (78), Date (65), Tomorrow (30)
(D) Other urgency	Tomorrow (91), Today (71), Soon (66)

5) Results of Model Construction: The data from Table 6 is used to train and evaluate the performance of models using 10-fold cross-validation. This technique was chosen due to the small size of the dataset and the elimination of overfitting. The model development involved three machine learning techniques: 1) *RF*, 2) *SVM*, and 3) *LR*.

Table 6

Example of Replacing Words with TF-IDF

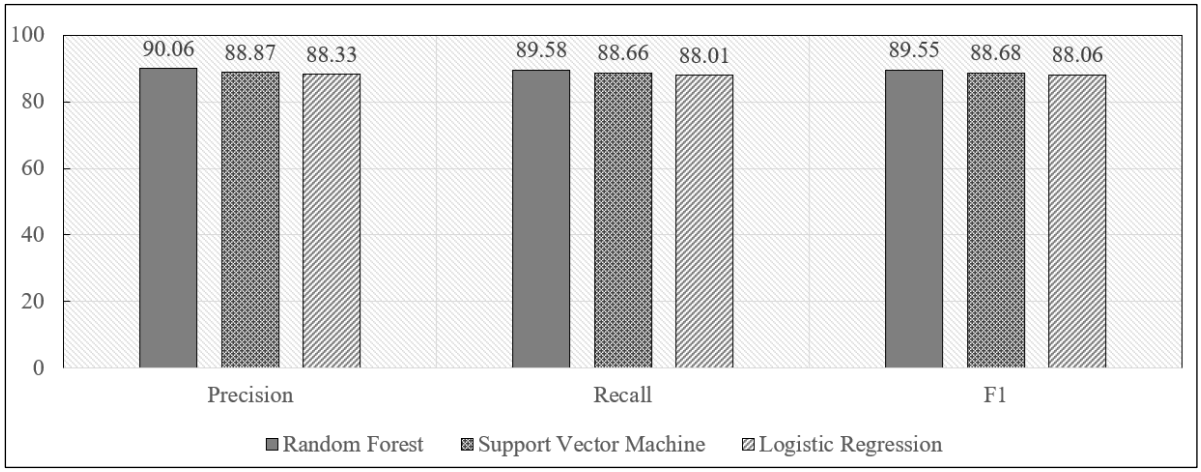
Message	Words in chats					Chat Topic
	Assignment	Business	Cancel	Exam	...	
1	0	0	0	0.225	...	Score
2	0.145	0	0	0	...	Homework
3	0.238	0.129	0	0	...	Leave
4	0	0	0.245	0		Urgency

Results of Model Evaluation

The models were evaluated using three machine-learning techniques. The results of comparing the performance of the models developed using different techniques are presented in Figure 2. The analysis revealed that the RF technique demonstrated the highest overall performance value (*F*1) at 89.55 percent.

Figure 2

Comparison Results of Model Performance Using Three Machine Learning Techniques



The results of comparing the overall performance of the models classified by chat topics are presented in Table 7. The results show that the *F*1 score for the topic titled Leave was the highest across all techniques, followed by Urgency, Score, and Homework. The RF technique consistently produced the highest values for each chat topic. These results suggest that the RF technique is the most effective in accurately classifying chats compared to other techniques. The results align with the research conducted by Meeprasert and Rattagan (2021), who analyzed comments from Twitter Shopee customers and found that the RF technique yielded the highest *F*1 score.

This study used a high-dimensional textual dataset, where RF outperformed LR and SVM. While LR and SVM are effective for low-dimensional data (Couronné et al., 2018), they struggle with high-dimensionality, which RF handles efficiently through random feature selection and bootstrapping. RF's strength is further highlighted by its ability to perform well with smaller datasets, aligning with findings by More and Rana (2017) and Luan et al. (2020). These attributes make RF particularly suited for scenarios with complex, high-dimensional data and limited sample sizes, as seen in this study.

Table 7

Comparison Results of F1 Model Performance by Chat Topics from 10-fold CV

Chat Topic	Random Forest (RF)	Support Vector Machine (SVM)	Logistic Regression (LR)
Homework	86	85	85
Score	88	87	86
Leave	93	92	92
Urgency	90	89	88
Average	89.55	88.68	88.06

Results of Model Deployment

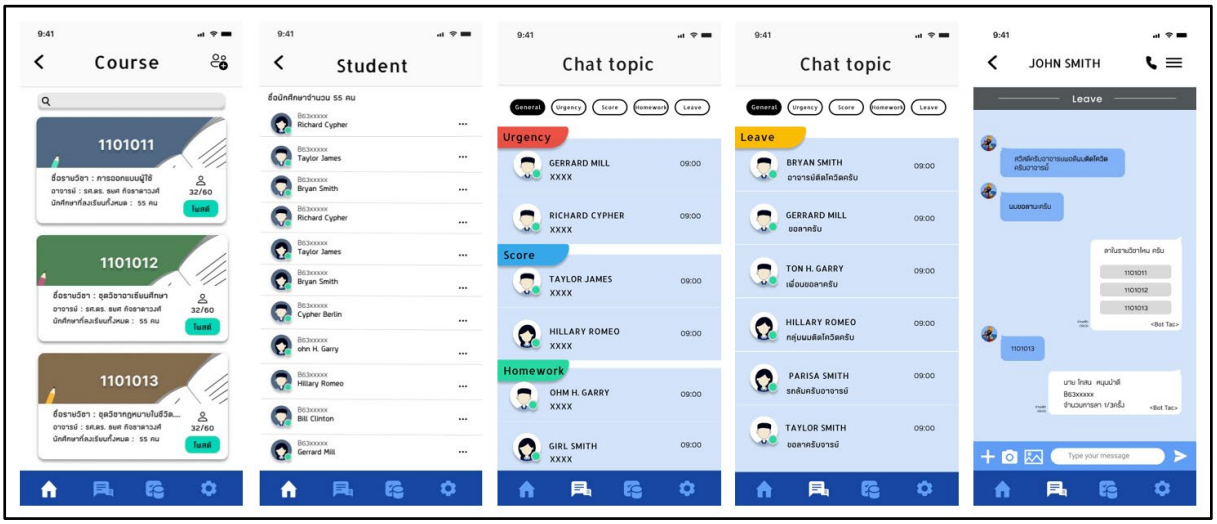
The developed model was utilized in a chatbot application. It functions by categorizing chats into topics, making it easier to respond to them. When students use the chatbot application, the categorized chat topics are saved in the application, and teachers can access them through the dashboard, as illustrated in Figure 3.

The application's database stores an appropriate reply message for each chat topic. The chatbot automatically responds if the student's discussion matches a conversation in the database. For example, if a student requests sick leave, the system recognizes it as a leave-related issue. Subsequently, the chatbot will inquire about the reason for the student's leave request. After receiving the student's response, the chatbot promptly displays how often the student has taken leave instead of the teacher responding. Implementing this topic classification model in the chatbot will expedite the response to the student's message.

In addition, when the chatbot identifies an urgent topic in the conversation, it will notify the teacher using a communication application like Line or Email. The notification will ensure that the teacher sees the message immediately. Additionally, the chatbot will send a reply message to the student.

Figure

Model Deployment in a Chatbot Application



Conclusions and Future Work

This research used text-mining techniques to analyze and categorize conversation messages between students and teachers. The study applied natural language processing and machine learning techniques to build a chat topic classification model. This model will be utilized to predict counseling objectives.

The findings revealed that the RF technique produced the highest *F*1 score compared to other methods, with 89.55%, followed by SVM at 88.68%, and LR at 88.06%. In addition, the evaluation of the technique's performance in this study found that the model's errors were caused by the model identifying many duplicate words across all types of conversations. This result aligns with the findings of Phaewattanakul and Luenam (2013), who studied suggestion mining and concluded that words with widespread distribution could affect the model's accuracy. These words are not typically used in data analysis to identify relationships. Therefore, language experts may be involved in providing insights for future analysis in eliminating these words. RF was chosen over LR and SVM due to its ability to handle the unique challenges presented by chat data, particularly its high dimensionality and informal language characteristics, such as those found in Thai chat communications. Unlike LR, which struggles with complex feature interactions in high-dimensional spaces, RF leverages an ensemble of decision trees, effectively capturing non-linear relationships and diverse patterns in the data. Similarly, while SVM performs well in low-dimensional contexts, it becomes computationally intensive and less effective as the number of features grows, a common trait in text classification tasks involving sparse word matrices.

RF includes mechanisms, such as random feature selection and bootstrapping, suitable for handling noisy and sparse datasets typical of informal chat language. These mechanisms reduce overfitting and enhance the model’s performance, even when dealing with unstructured data, including slang, abbreviations, and inconsistent syntax.

Additionally, the model developed from this research has great potential for categorizing conversations or predicting their intent in counseling students. Furthermore, the model's

capacity to streamline and decrease student wait times for communication and issue resolution is a significant advancement.

Despite its advantages, the designed system has some limitations. An extensive vocabulary corpus is required for each type of chat topic to improve the model's accuracy using text-mining techniques. However, creating a vocabulary corpus for each type of chat topic is a task that requires the expertise of linguistic professionals. Furthermore, the data being analyzed is collected from social media, which includes emerging vocabulary, such as chat language, posing challenges for the model. This experiment used a small dataset to assess the performance of the models. Therefore, collecting more data could lead to an increase in more generalized models, ensuring that they can work well on other datasets and be applied to real-world applications.

Several improvements can be made shortly, such as improving the developed model using deep learning techniques and engaging linguistic experts to understand word characteristics and chat language better.

Author Contributions

AW: Conceptualization, methodology, formal analysis, review and editing. BVM: Methodology, resources, investigation. PK: Original draft preparation. SP: Methodology, original draft preparation. TA: Validation, supervision, review and editing. WD: Conceptualization, investigation, data curation. JA: Conceptualization, investigation, data curation, visualization, original draft preparation.

Declaration of the use of AI

Artificial intelligence techniques were employed solely for research purposes in this study. Specifically, machine learning algorithms were used to analyze and model the research data. The use of these methods was part of the study's methodological framework, and all analyses, interpretations, and conclusions were conducted and verified by the authors. No generative AI tools were used to produce the scientific results or conclusions of this study.

Declaration of Generative AI

During the preparation of this manuscript, the authors used Grammarly to assist with language editing and polishing of the draft. The tool was used solely to improve clarity, grammar, and overall readability. The authors reviewed and revised the output as necessary and take full responsibility for the content of the manuscript.

Ethics

This study did not involve experiments on humans or animals. The research was conducted using secondary data from chat platforms, which were fully anonymized prior to analysis. Therefore, no human participants or animal subjects were directly involved in this study.

References

Ali, M. S., Azam, F., Safdar, A., & Anwar, M. W. (2022, November). Intelligent agents in educational institutions: NEdBOT-NLP-based chatbot for administrative support using DialogFlow. In *2022 IEEE International Conference on Agents* (*ICA*) (pp. 30-35). IEEE. https://doi.org/10.1109/ICA55837.2022.00012

Bloch, J. (2002). Student/teacher interaction via email: The social context of Internet discourse. *Journal of Second Language Writing*, *11*(2), 117-134. https://doi.org/10.1016/S1060-3743(02)00064-4

Bouhnik, D., & Deshen, M. (2014). WhatsApp goes to school: Mobile instant messaging between teachers and students. *Journal of Information Technology Education. Research*, *13*, 217. https://doi.org/10.28945/2051

Cahyani, D. E., & Patasik, I. (2021). Performance comparison of TF-IDF and Word2Vec models for emotion text classification. *Bulletin of Electrical Engineering and Informatics*, *10*(5), 2780-2788. https://doi.org/10.11591/eei.v10i5.3157

Calabrese, A., Rivoli, A., Sciarrone, F., & Temperini, M. (2022, November). An intelligent chatbot supporting students in massive open online courses. In *International Conference on Web-Based Learning* (pp. 190-201). Springer International Publishing. https://doi.org/10.1007/978-3-031-33023-0_17

Chowdhary, K. R. (2020). Natural language processing. In *Fundamentals of Artificial Intelligence*. Springer. https://doi.org/10.1007/978-81-322-3972-7_19

Couronné, R., Probst, P. & Boulesteix, AL. (2018). Random Forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics 19*, 270. https://doi.org/10.1186/s12859-018-2264-5

Deng, X., & Yu, Z. (2023). A meta-analysis and systematic review of the effect of chatbot technology use in sustainable education. *Sustainability*, *15*(4), 2940. https://doi.org/10.3390/su15042940

Gaikwad, S. V., Chaugule, A., & Patil, P. (2014). Text mining methods and techniques. *International Journal of Computer Applications, 85*(17). https://doi.org/10.5120/14937-3507

Hendry, D., Darari, F., Nurfadillah, R., Khanna, G., Sun, M., Condylis, P. C., & Taufik, N. (2021). Topic modeling for customer service chats. In *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICACSIS53237.2021.9631322

Hingmire, S., Chougule, S., Palshikar, G. K., & Chakraborti, S. (2013, July). Document classification by topic labeling. In *Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval* (pp. 877-880). https://doi.org/10.1145/2484028.2484140

Kim, S. W., & Gil, J. M. (2019). Research paper classification systems based on TF-IDF and LDA schemes. *Human-centric Computing and Information Sciences*, *9*, 1-21. https://doi.org/10.1186/s13673-019-0192-7

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, *10*(4), 150. https://doi.org/10.3390/info10040150

Luan, J., Zhang, C., Xu, B., Xue, Y., & Ren, Y. (2020). The predictive performances of random forest models with limited sample size and different species traits. *Fisheries Research*, *227*, 105534. https://doi.org/10.1016/j.fishres.2020.105534

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

Meeprasert, W., & Rattagan, E. (2021). Voice of customer analysis on Twitter for Shopee Thailand. *Journal of Information Systems in Business JISB, 7*(3), 6. https://doi.org/10.14456/jisb.2021.11

More, A. S., & Rana, D. P. (2017). Review of random forest classification techniques to resolve data imbalance. In *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)* (pp. 72-78). IEEE. https://doi.org/10.1109/ICISIM.2017.8122151

Phaewattanakul, K., & Luenam, P. (2013). Opinion mining from online social networks. *Modern Management Journal, 11*(20), 11-20.

Prachaming, S., Pimkalee, N., & Udomson, N. (2017). The language used to communicate via chat application line. *Journal of Roi Et Rajabhat University, 11*(2), 80–89.

Sreelakshmi, A. S., Abhinaya, S. B., Nair, A., & Nirmala, S. J. (2019, November). A question answering and quiz generation chatbot for education. In *2019 Grace Hopper Celebration India (GHCI)* (pp. 1-6). IEEE. https://doi.org/10.1109/GHCI47972.2019.9071832

Tabassum, A., & Patil, R. R. (2020). A survey on text pre-processing & feature extraction techniques in natural language processing. International Research Journal of Engineering and Technology (IRJET), 7(06), 4864-4867.