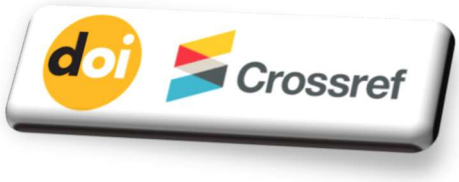
Journal, Advances in Mathematical & Computational Sciences
Vol. 10 No. 3, September, 2022
www.isteams.net/mathematics-computationaljournal



Journal of Advances in Mathematical & Computational Sciences
An International Pan-African Multidisciplinary Journal of the SMART Research Group
International Centre for IT & Development (ICITD) USA

Available online at https://www.isteams.net/ mathematics-computationaljournal.info
CrossREF Member Listing - https://www.crossref.org/06members/50go-live.html

mlChatApp: Topic Modeling in Online Chat Groups

Obiorah Philip, Onuodu Friday & Eke Batholowmeo
Department of Computer Science
University of Port Harcourt
Choba-Port Harcourt, Nigeria
E-mails: philip.obiorah@outlook.com; friday.onuodu@uniport.edu.ng;
bartholomew.eke@uniport.edu.ng

ABSTRACT

Most of the time, online messaging group users must scroll and read a large number of irrelevant posts in order to gain a clear understanding of what is being discussed in the group to which they belong. Messaging groups can get congested with unnecessary messages, causing members to miss out on important issues and information. There is a need to assist users of multi-user chat systems in understanding what the group discussion is all about at any particular time without having to read all of the posted messages. This paper describes an approach to discovering topics in online chat groups. In order to extract and categorize subjects from unseen texts in online group discussions. We developed a new multi-user chat system (ML-CHAT-APP) that automatically identifies and categorizes topics within posts/messages as they appear. We implemented a combination of a Latent Dirichlet Allocation (LDA)-based model with Multinomial Logistic Regression. The resulting model was integrated into the ML-CHAT-APP built with Python and Tkinker framework for Graphical User Interface. The results show that the application was helpful in identifying topics in text conversations and adding identified topics as labels to message posts in real-time.

Keywords: NLP, Topic Modeling, Latent Dirichlet Allocation; Logistic Regression

Obiorah P., Onuodu, F. & Eke, B. (2022): mlChatApp: Topic Modeling in Online Chat GroupsJournal of Advances in Mathematical & Computational Science. Vol. 10, No. 3. Pp 101-110 Available online at www.isteams.net/mathematics-computationaljournal.
DOI: dx.doi.org/10.22624/AIMS/MATHS/V10N2P8

1. INTRODUCTION

The widespread use of social networks has considerably enhanced people's everyday lives over the previous decade, bringing new kinds of satisfaction and forming new sorts of interactions. (Zhang et al. 2017) . The exponential growth in the usage of Multi-user chat applications in our everyday lives cannot be understated.

Due to the obvious vast range of topics and meaningless chatting, online chat groups are packed with irrelevant messages that have no connection to the group's sole purpose. Most of the time, users must scroll and read a large number of irrelevant posts in order to gain a clear understanding of what is being discussed in the group to which they belong. Online chat groups are sometimes clogged with irrelevant messages, causing users to miss key subjects and information. Topic discovery in online group chat applications becomes a vital task to aid users of online chat groups in comprehending the topic of the group discussion at any one time without having to read every message posted.

Chats are self-contained narratives that emerge from the textual interaction of a varied group of people who obey few rules and frequently appear chaotic as a result(Kolenda, Hansen, and Larsen 2002). Client-based chat software connects users directly to a central directory server, which announces each client's availability. Since chat conversation is not disseminated through a central server, there is no central authority to monitor or impose appropriate behaviour. Online chat allows you to communicate with people by using a keypad, much as in emails, but Additionally, it possesses the spontaneity inherent in real-time human contact just like a phone conversation. The ability to stay in touch with colleagues, friends, and family, as well as the thrilling value of getting to know new people, trading information, and participating in township hall-type chat rooms to deliberate about nearly any subject comprehendible, propelled instant messaging to renown. Topic discovery in chat group communications becomes a crucial yet tough research task in order for users to comprehend what the group is talking about without having to read all of the messages. It is critical to assist multi-chat system users in understanding the topic of the group conversation at any given time without having to read every message sent. The goal of this research is to develop a machine-learning-powered chat application that would identify topics from posted messages in chat group conversations in real-time.

2. LITERATURE REVIEW

A topic, according to its formal definition, is a probability distribution across the terms in a lexicon. Informally, a topic implies an underlying semantic subject; a long text may simply be defined as being composed of a smaller number of topics. McAuliffe (2008). Topic modeling, as the name implies, is a technique for automatically detecting topics within a text item and generating hidden patterns from a text corpus. As a result, making decisions becomes easier. Topic modeling is distinct from rule-based text mining approaches based on regular expressions or dictionary-based keyword search strategies. It is an unsupervised method for detecting and monitoring word clusters (called "themes") inside large text clusters. Co-occurring words in a corpus are referred to as "topics." Health, doctor, patient, hospital, and farm are all examples of phrases that should be included in a decent topic model for the subject "Healthcare."

Chat logs' contents (topics) change often as a result of message sequence incoherence. Chat log messages are often brief (Rahman et al. 2013) Chat is considered a kind of natural language by us. Natural language is just the language spoken by people. (Martin and Jurafsky 2018). The means of communication have changed over time according to the situation and advancements in technology. Chat applications enable individual users to converse 'one-on-one' via 'real-time' text conversation via the Internet. Chat systems operate in a variety of ways, with some enabling communication involving two users and others allowing communication between hundreds of individuals. (Patil, 2016).

While many online apps include 'real-time' text conversation, it is typically a supplementary function, whereas, with chat systems, it is the core service. Chat offers computer-mediated communication (Dewes, Wichmann, and Feldmann, 2003). Synchronous conferencing is a word used in computer-mediated communication, collaboration, and education to refer to technologies often referred to as online chat. It is occasionally expanded to encompass audio/video conferencing and instant messaging services that include multi-user text chat.

2.1 Related Works

Bengel and colleagues (2004) created ChatTrack, a collection of programs and tools for analyzing and archiving electronic chat room talks. They collected data on conversations that took place in open chat rooms on various Internet Relay Chat networks as well as on a special server. The software keeps an XML archive of the chat room discussions' content. On-demand, subsets of this data may be categorized in order to build profiles for particular sessions or individuals. The classifier can be updated to include ideas particular to a given analyst because it was pre-trained on more than 1,500 topics. The results demonstrate that by combining automated subject recognition with chat exploration, intelligence agents may be more vigilant in the pursuit of criminals by significantly lowering the amount of time spent manually examining chat room conversations. However, it is currently not possible to automate the analysis of chat room topics over time.

Kolenda, Hansen, and Larsen (2001) investigated the use of independent component analysis (ICA) on chat room log files. The results of independent component analysis (ICA) show that it is capable of detecting relevant context structures in a chat room log file. However, the sequence of independent components cannot be identified, implying that a physical signal and the predicted independent component do not have a one-to-one relationship. Dong et al. (2006) examined the features of chat messages and proposed a strategy for detecting chat topics based on an indicative term-based categorization approach. The result indicates that the indicative term-based strategy outperforms the standard document frequency-based technique in categorizing conversation topics.

Zhang et al. (2014) suggested a novel technique for topic detection in instant messages that is suitable in the following situations: 1) irrelevant phrases continue to emerge; 2) instant messages are extremely brief, and 3) different languages are employed. They regarded each message in an online group conversation as a separate data item in a message stream and then applied Probabilistic latent semantic analysis (PLSA) to the gathered instant communications. They split multilingual messages without relying on machine translation and eliminated recurring irrelevant terms. Extensive trials using real-world Instant Messaging group data demonstrate the proposed method's efficiency. However, their topic detection algorithm was not implemented as an offline auxiliary component in instant messaging platforms which can analyze chatting history automatically and save topic detection findings to disk, so that when users wish to explore a certain online group can simply browse the topic detection results saved in advance and have a real-time experience.

3. MATERIALS AND METHOD

3.1. Dataset

For this experiment, data scraped from the Narialand Forum website (source: https://www.nairaland.com/) from 2016-2022 and the 2004-2005 BBC news dataset (source: http://mlg.ucd.ie/datasets/bbc.html) were used. The Narialand dataset contains 25785 documents, whereas the BBC news website dataset contains 26512 documents related to articles in five thematic categories.

3.2. Design of the Proposed System

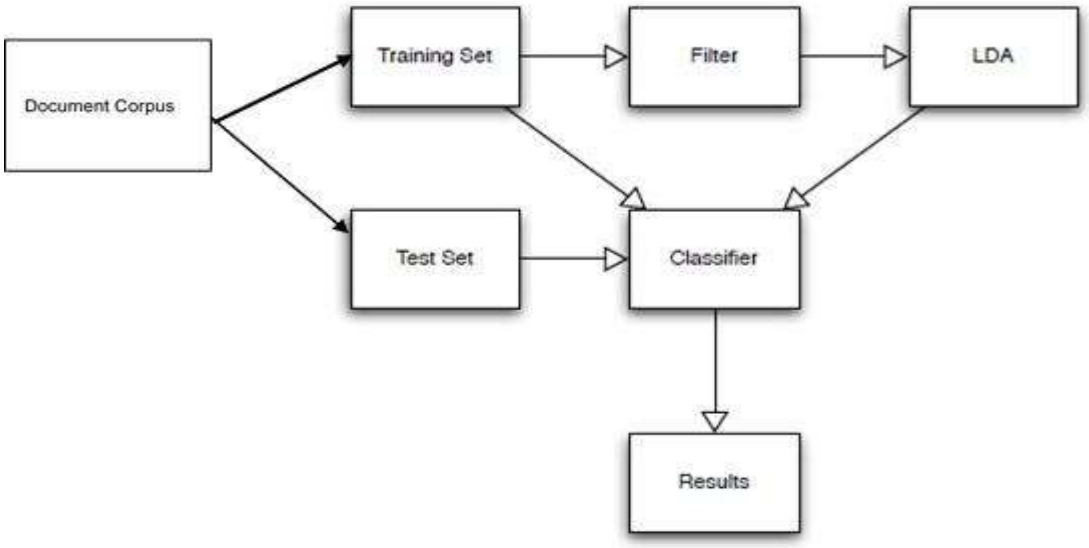


Figure 3.2 Block Diagram of the design of the Proposed system.

Figure 3.13 shows a Block Diagram of the design of the Proposed system. We divide the document corpus into two different sets: The training set and the test set. The training set would go through the pre-processing and filtering stage. After pre-processing and filtering, we subject the corpus to LDA. Latent Dirichlet Allocation (LDA) is a topic modeling technique used to assign text in a document to a specific topic. It employs the Dirichlet distribution to find topics and words for each document model and topic model. A method of unsupervised learning known as Latent Dirichlet Allocation (LDA) does not require manually labelled data. One type of probabilistic generative model is the latent Dirichlet allocation (LDA) model of a corpus (Blie, Ng, and Jordan, 2003). The development of documents with a variety of themes is a requirement for LDA. Then, in accordance with the likelihood of their spread, the topics generate words. LDA recursively tries to infer the subjects that initially gave rise to the collection of documents. (El-Habil, A. M. 2012). The output of LDA is further subjected to a Classifier such that we would be able to have labelled topics as our results.

Journal, Advances in Mathematical & Computational Sciences
Vol. 10 No. 3, September, 2022
www.isteams.net/mathematics-computationaljournal

3.3. Use case Diagram

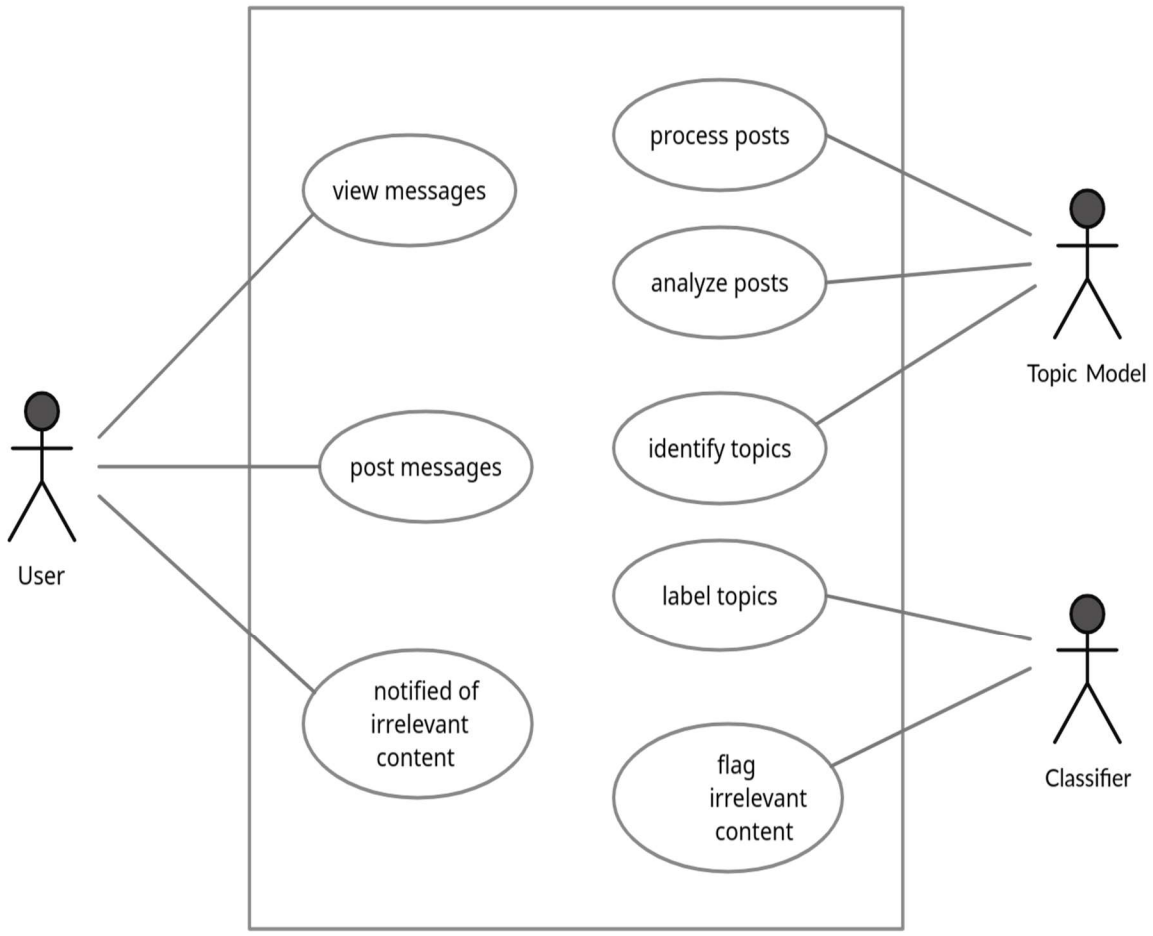


Figure 3.3 Use case diagram

The case diagram summarises the details of the proposed system users (actors) and their interactions with the system. Figure 3.3 show that users can view messages, post messages and get notifications for irrelevant content posted. Our topic model would be able to get users' posts, reprocess and analyse them and apply LDA to get topic form posts after which our classifier labels the post and flags irrelevant posts.

Journal, Advances in Mathematical & Computational Sciences
Vol. 10 No. 3, September, 2022
www.isteams.net/mathematics-computationaljournal

3.4. Use Class Diagram

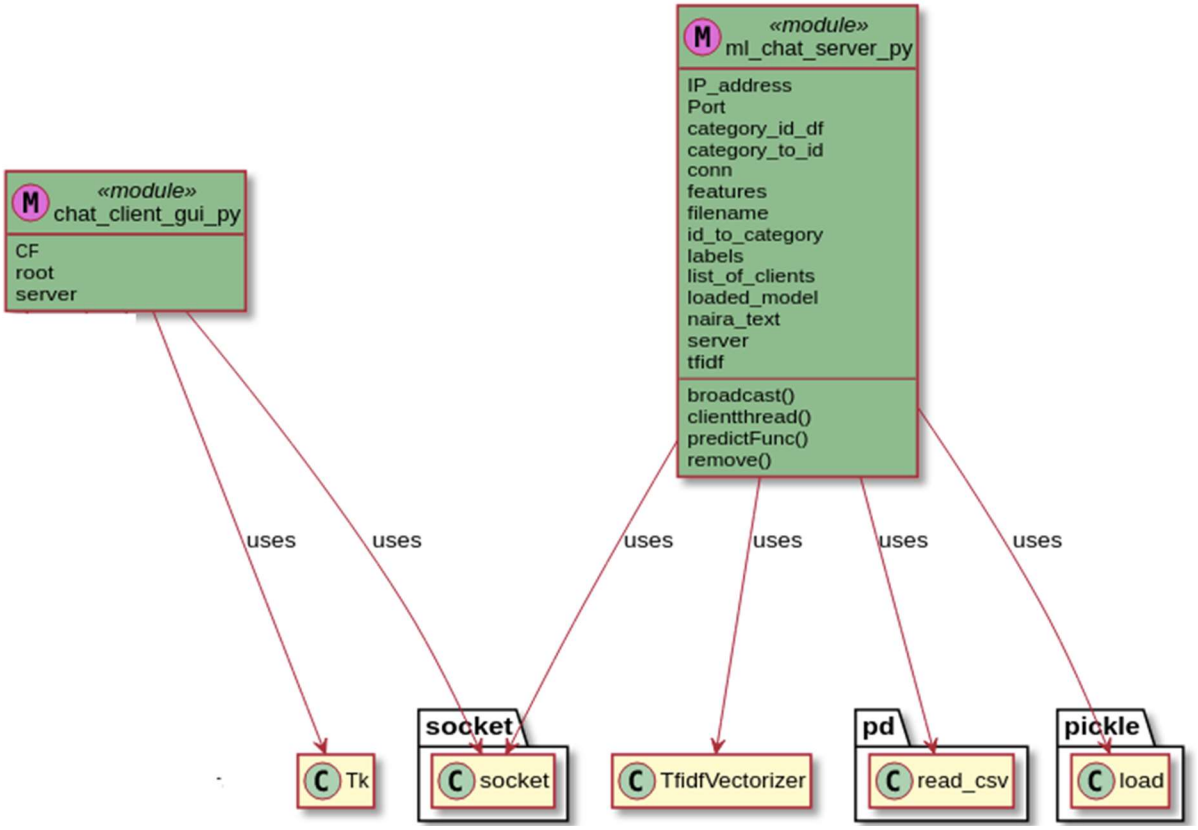


Figure 3.4 ml_chat_server and chat_client_gui module interactions with Sockets, tkinter class, Pandas package, Pickle package and TfidfVectorizer

Figure 3.4 depicts the interactions of the ml chat server and chat client gui modules with Sockets, the tkinter class, the Pandas package, the Pickle package, and the TfidfVectorizer (TermFrequency Inverse Document Frequency Vectorizer). The Tkinter interface provides a standard Python interface to the Tk GUI toolkit, which is in charge of the graphical user interface. Pickle is a Python module that defines binary protocols for serializing and de-serializing object structures. A Python object hierarchy is transformed into a byte stream in this case, while "unpickling" is the converse procedure in which a byte stream (from a binary file or bytes-like object) is converted back into an object hierarchy. The socket module makes it easier to create network servers.

Journal, Advances in Mathematical & Computational Sciences
Vol. 10 No. 3, September, 2022
www.isteams.net/mathematics-computationaljournal

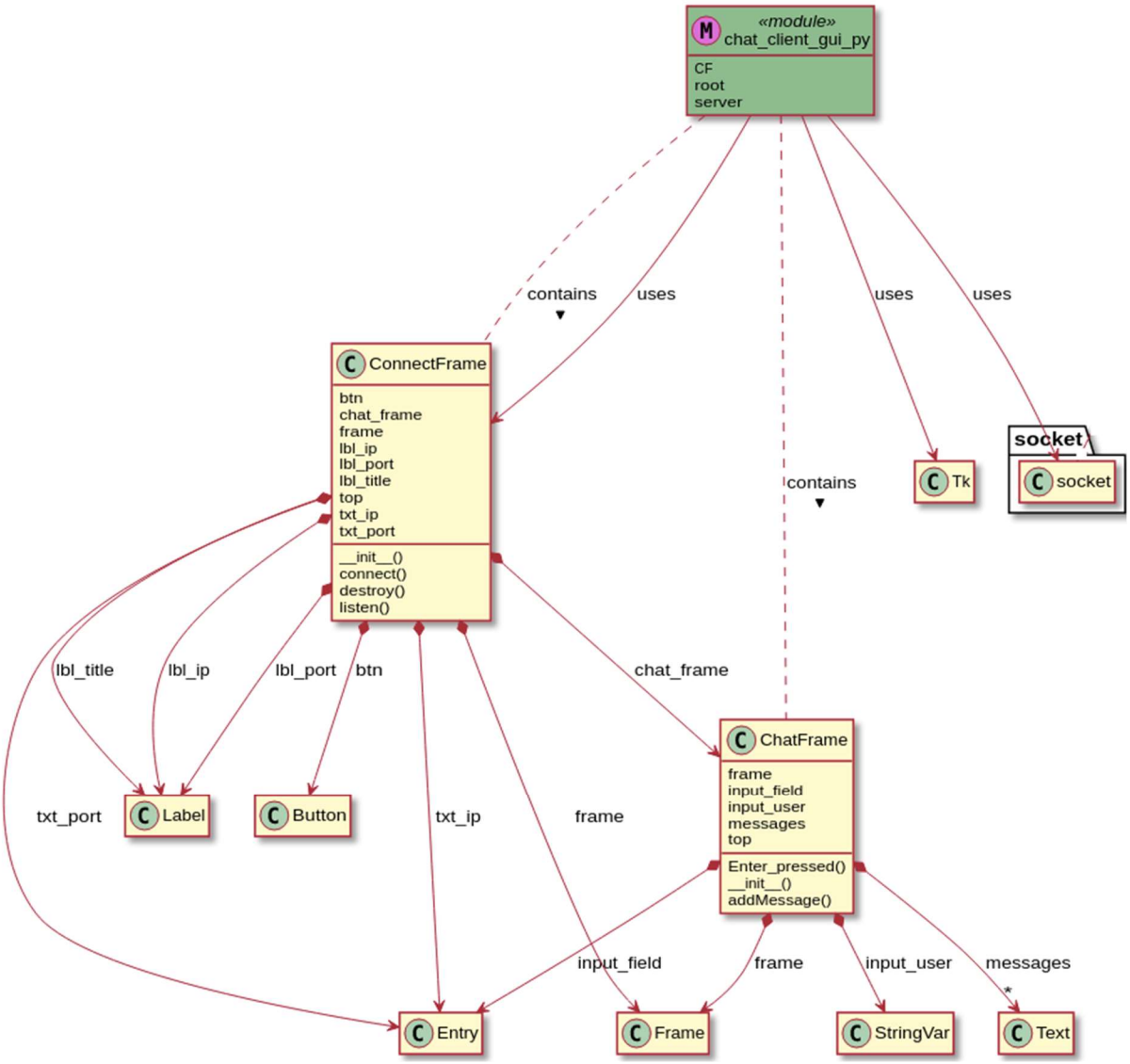


Figure 3.5 Associations and dependencies within the chat_client_gui module

Figure 3.5 Shows associations and dependencies within the chat_client_gui module. The ConnectFrame is basically a ChatFrame that organizes other components. The Label and Button class provides the mechanism to add Labels and Buttons to the Frame. The Entry class allows input from the user such as txt_ip (IP address) and txt_port (Port Number)

Journal, Advances in Mathematical & Computational Sciences
Vol. 10 No. 3, September, 2022
www.isteams.net/mathematics-computationaljournal

4. RESULTS DISCUSSION

We use multinomial logistic regression in our study because our target variable, y, spans more than two classes and we want to know the likelihood that y belongs to each possible class. A term frequency-inverse document frequency was assigned to each unigram (TF–IDF). A bit size of 15 bits was specified to extract 32,768 hashing characteristics. The top 5000 closely related traits were used in this experiment. An experiment was conducted with this model, and the model was performed with a precision of 0.979782. The generated model was used to build ML-CHAT-APP, a revolutionary multi-user chat system. ML-CHAT-APP is launched with host address 127.0.0.1. Figure 4.1 depicts the connection of two chat clients to the ml-chat server. User A has been assigned the IP address 127.0.0.1:38978, whereas User B has been assigned the IP address 127.0.0.1:38988.

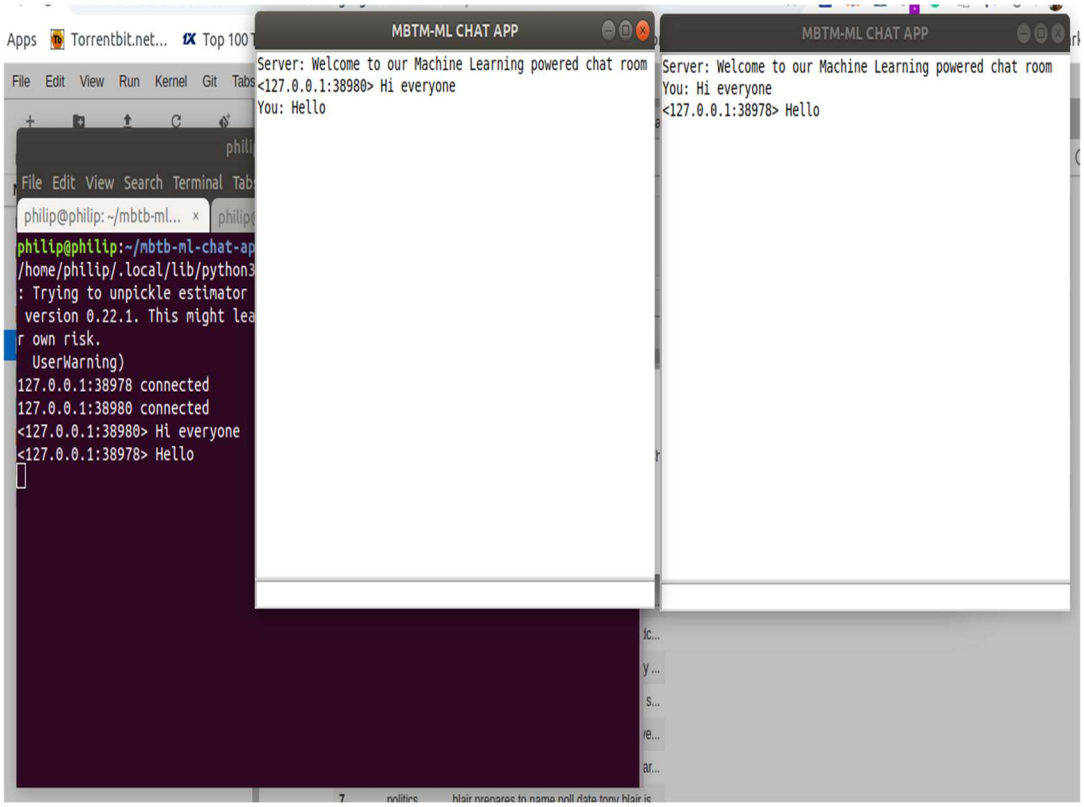


Figure 4.1– Connection of two chat clients with the ml-chat server

Journal, Advances in Mathematical & Computational Sciences
Vol. 10 No. 3, September, 2022
www.isteams.net/mathematics-computationaljournal

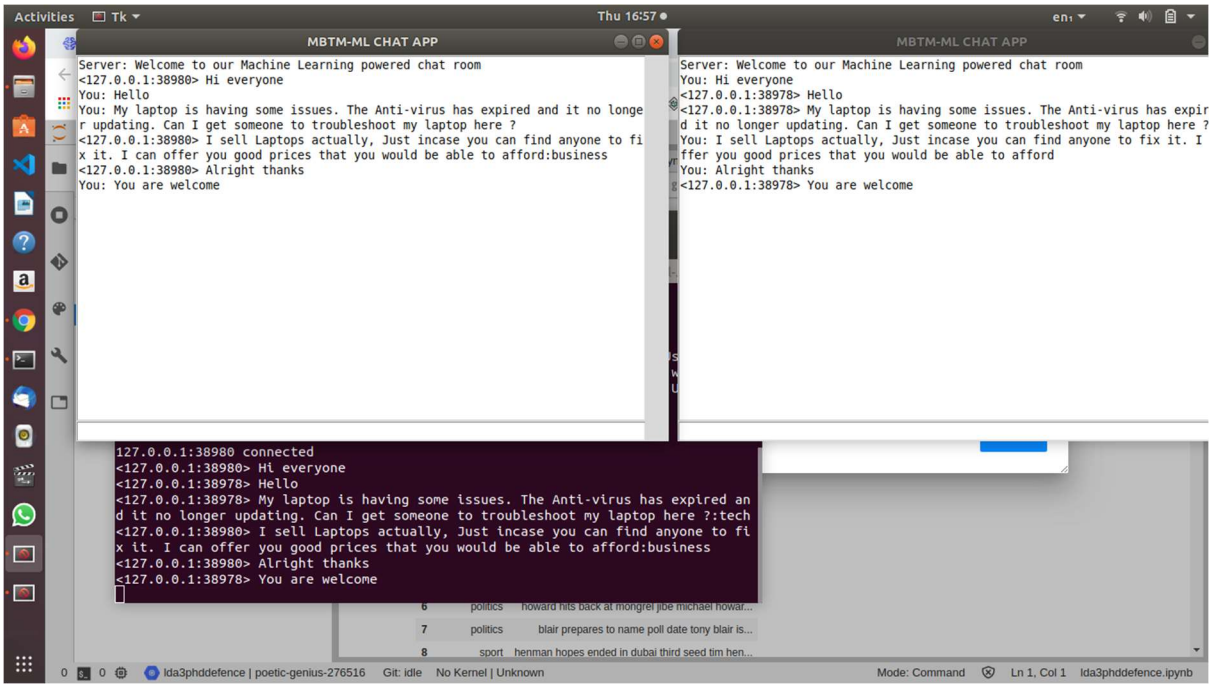


Figure 4.2 depicts the connection of two chat clients to the ml-chat server.

The discussion is taking place between two client applications with IP addresses 17.0.0.1:38980 and 127.0.0.1:38978. The prediction model is not activated unless the text is longer than 25 characters.. Based on the preceding, the integrated model correctly predicts *Technology* and appropriately detects what client user 127.0.0.1:38978 says about "laptop..having some troubles." The model also properly predicts the topic of client user 127.0.0.1:38890. "I sell laptops... I can offer you good prices..." and makes a *business* prediction.

5. CONCLUSION

We developed a machine learning-powered chat application: ML-CHAT-APP, that incorporates a combination of Latent Dirichlet Allocation and Multinomial Logistic Regression. The experiment for this model was carried out, and the model was performed with an accuracy of 0.979782. Our study can estimate the content dispersion of each sentence by counting the words in each sentence. Individual posts were treated as documents in the hopes of determining the author's or a conversation's topics. The system recognizes and categorizes topics within posts/messages as they appear. Our ML-CHAT-APP eliminates the need for a domain expert to intervene and determine whether or not a topic is related to a specific subject, In this context, we present an automated technique for labelling or categorizing the retrieved topics in instant messaging chat applications. The findings indicate that a combination of Latent Dirichlet Allocation and Multinomial Logistic Regression were complementary in terms of identifying textual topics in multi-user chat systems.

REFERENCES

1. Blei, D. M., & McAuliffe, J. D. (2008). Supervised Topic Models. New Jersey .
2. Bliei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. (J. Lafferty, Ed.) Journal of Machine Learning Research, 993-1022.
3. Bengel, J., Gauch, S., Mittur, E., & Vija, R. (2004). ChatTrack: Chat Room Topic Detection Using Classification. International Conference on Intelligence and Security Informatics (pp. 266-277). Berlin, Heidelberg: Springer.
4. Dong, H., Hui, S. C., & He, Y. (2006). Structural Analysis of Chat Messages for Topic Detection. Singapore.
5. Dewes, C., Wichmann , A., & Feldmann, A. (2003). An analysis of Internet chat systems. Florida,, U.S.A: ACM.
6. Rahman, A., Basher, M. A., & Fung, B. C. (2013). Analyzing Topics and Authors in Chat Logs for Crime Investigation. Canada.
7. Kolenda , T., Hansen, L. K., & Larsen, J. (2001). SIGNAL DETECTION USING ICA: APPLICATION TO CHAT ROOM TOPIC SPOTTING.
8. Zhang, Y., Li, Z., Gao , C., & Kaigui , B. (2017). Mobile Social Big Data: WeChat Moments Dataset, Network Applications, and Opportunities.
9. Jurafsky, D., & Martin, J. H. (2018). Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Sanford.
10. Patil, S. (2016). WhatsApp Group Data Analysis with R. International Journal of Computer Applications, 154, 31-36.
11. El-Habil, A. M. (2012). An Application on Multinomial Logistic Regression Model. Pakistan Journal of Statistics and Operation Research, 8(2), 271-291. https://doi.org/10.18187/pjsor.v8i2.234