Unsupervised Topic Shift Detection in Chats

Yuze Kevin Liu
McGill University
Montreal, Canada
yuze.liu@mail.mcgill.ca

Benjamin C. M. Fung
McGill University
Montreal, Canada
ben.fung@mcgill.ca

Djedjiga Mouheb
Laval University
Quebec, Canada
djedjiga.mouheb@ift.ulaval.ca

Noman Mohammed
University of Manitoba
Winnipeg, Canada
noman.mohammed@umanitoba.ca

***Abstract*—In modern digital communication, chat platforms generate vast amounts of unstructured conversational data. However, the fluid, informal, and cross-timeframe nature of these chats challenges traditional supervised topic detection approaches. Addressing this gap, we propose an entirely unsupervised framework for detecting topic shifts in chat logs. Our method first converts individual messages into contextual embeddings using *DistilBERT*, thereby capturing nuanced semantic features. Next, rather than simply segmenting messages based on fixed temporal windows, we analyze the temporal evolution of these embeddings using a novel smoothing algorithm that highlights significant changes in the semantic trajectory of the conversation. Candidate topic shift points are then identified through an unsupervised peak detection process applied to the smoothed signal. Finally, an adaptive clustering algorithm groups the segmented data to refine topic boundaries without any manual labeling, ensuring the approach remains completely unsupervised. This method not only better accommodates the dynamic characteristics of chat logs but also robustly distinguishes fine-grained topic transitions. Experimental results demonstrate that our unsupervised framework outperforms traditional lexical and statistical techniques in detecting nuanced topic changes, making it highly effective for applications in social media analysis and real-time chat monitoring.**

***Index Terms*—Topic Shift Detection, Unsupervised Learning, Group Chat, DistilBERT, Temporal Convolutional Networks.**

I. Introduction

Group chat applications, for example, Slack, Microsoft Teams, WhatsApp, etc., have become ubiquitous for personal and professional communication, generating an ever-growing volume of unstructured conversation data [1]. In large and active group chats, multiple topics are often discussed concurrently or in rapid succession, creating streams of interwoven dialogues [2]. Managing and understanding such conversations is challenging in practice: important information can be lost or overlooked when discussion threads switch or overlap without clear markers. For example, a project team's chat might shift from brainstorming ideas to discussing logistics within minutes, confusing participants and hindering efficient information retrieval [3]. This practical need to organize conversational data motivates the task of *topic shift detection* in group chats, which aims to automatically identify boundaries where the discussion topic changes. Accurate topic shift detection would greatly benefit real-world applications such as conversation summarization, context-aware virtual assistants, and chat archive management, by disentangling mixed-topic dialogues into coherent segments [4].

Detecting topic shifts in multi-party chats is a non-trivial problem due to several core challenges. First, group chats lack explicit structure: unlike well-formatted documents or threaded forums, chat messages are chronologically ordered but not hierarchically organized by topic. Participants can introduce new topics abruptly, often without explicit cues, e.g., no section headings or formal turn-taking signals, resulting in sudden context switches that are difficult to catch [5]. Second, conversations in group chats are highly dynamic and overlapping. It is common for several threads of discussion to proceed in parallel within the same channel, especially in busy chats with many participants [6]. This leads to entangled context, a single message might refer to an earlier topic while the very next message starts a different subject. Traditional dialogue segmentation methods struggle in this setting, as they typically assume a single thread of discourse [3]. Third, the language in informal chats is often noisy and context-dependent: the use of slang, emojis, abbreviations, or implicit references is prevalent, making it harder for automated approaches to judge semantic continuity.

Several related lines of research address parts of this problem, but they do not provide a comprehensive solution. Topic segmentation [7] has been extensively studied in monologues and formal texts, e.g., news articles, transcripts. Other work has applied segmentation methods to spoken meetings or two-person conversations [8]. These techniques often rely on vocabulary distribution changes or supervised classifiers to mark segment boundaries. However, they usually assume a linear conversation flow and cannot easily disentangle multiple interleaved topics [9]. Conversation disentanglement techniques attempt to separate interleaved threads in group chats or online forums [10]. More recent models incorporate neural methods or metadata to improve clustering [11]. While disentanglement is related to topic shift detection, it generally focuses on grouping messages into threads post hoc, rather than detecting the exact transition points between topics in real time. Moreover, many existing approaches, whether for segmentation or disentanglement, require large annotated datasets for training [12], which are scarce for group chats due to annotation cost and privacy concerns. Consequently, there is a clear gap in current research: no existing method adequately addresses real-time topic shift detection in unstructured multi-party chats with minimal supervision, which is crucial for practical deployment in modern chat platforms.

To address these challenges, this paper introduces a fully unsupervised framework, called *ChatSense*, designed explicitly to detect topic shifts in group chat conversations. ChatSense integrates contextual semantic embeddings using *DistilBERT*, temporal dependency modeling through *Temporal Convolutional Networks* (*TCN*), and adaptive heuristic feature extraction (e.g., time gaps, sentiment changes, and speaker role shifts). Finally, an adaptive clustering algorithm is employed to automate topic segmentation.

The contributions of this work are summarized as follows:

- *Unsupervised topic shift detection*: ChatSense does not rely on manual labeling or annotated datasets, which makes it highly practical for analyzing extensive and continuously expanding group chat datasets. This unsupervised approach is particularly advantageous for real-world applications where obtaining labeled data is costly and impractical.
- *Effective handling of topic overlap and topic entanglement*: ChatSense is specifically designed to handle complex conversation structures that are common in group chats, such as overlapping topics, rapid topic transitions, and multiple topics entanglement. Its sophisticated temporal modeling through Temporal Convolutional Networks (TCNs) enables accurate tracking of topic transitions despite the inherent complexity and asynchronous nature of the data.
- *Comprehensive feature integration*: ChatSense integrates semantic embeddings derived from DistilBERT, temporal features through TCN, and adaptive heuristic signals (e.g., message timing, sentiment shifts, and participant roles). This multi-dimensional approach significantly improves the accuracy and robustness of topic shift detection, outperforming methods that rely solely on lexical, statistical, or isolated semantic analysis.

The rest of this paper is organized as follows. Section II reviews related work on dialogue segmentation and disentanglement. Section IV details the proposed ChatSense methodology, including the chat encoding and topic shift detection algorithm. Section V presents the experimental setup and results. Section VI provides further discussion. Finally, Section VII concludes the paper and outlines the directions for future work.

II. Related Work

Topic detection in group chats and social media has become an important research area, but many existing methods still have limitations, especially when dealing with short texts, dynamic topic changes, and complex context. While various techniques have been proposed to improve topic detection, they are often optimized for traditional long-text corpora and do not consider the unique characteristics of short, informal texts.

Firstly, traditional text analysis techniques such as *TF-IDF* [13] and *LDA* [14] may not be applicable to group chat or social media data. These methods perform well in document-level topic modeling, where long texts provide sufficient context. However, short texts in group chats are fragmented and lack complete context, making it difficult for these methods to capture the temporal and fragmented nature of chat data. Furthermore, traditional word frequency analysis and lexical overlap methods may not be able to effectively handle complex linguistic phenomena such as polysemy, synonyms, and slang commonly found in group chats.

To address these challenges, *deep learning methods* have been widely explored, particularly *Recurrent Neural Networks (RNNs)* [15] and *Convolutional Neural Networks (CNNs)* [16], which are capable of capturing temporal relationships within the text. However, these models still struggle with long-range dependencies, especially in group chats where responses may be delayed by hours or even days. Furthermore, these models often assume sequential continuity, which is typically disrupted in asynchronous conversation environments, leading to performance degradation.

On the other hand, *Word2Vec* [17] and *GloVe* [18] word embeddings improve semantic understanding by capturing relationships between words, especially in longer contexts. However, these methods perform poorly on short texts because they lack effective context awareness and struggle with handling polysemy and slang. Recently, *Gupta et al.* [19] has attempted to improve word embeddings for short texts, but the variability and contextual complexity of short texts remain a significant challenge.

Another issue is the measurement of *topic coherence*. In group chats, topics can change rapidly, and methods for measuring topic coherence often fail to capture long-term consistency. Traditional word frequency-based similarity measures do not account for contextual and temporal variations, making it difficult to accurately reflect the evolution of the topic.

Moreover, many existing methods overlook the importance of *contextual awareness* in group chat data. Group chat conversations are not simply a cumulative collection of individual texts, but rather contain rich, dynamic contexts that change rapidly over time and in different situations. Traditional models often fail to effectively capture these variations, limiting their applicability in real-world scenarios. Our work introduces *DistilBERT embeddings*, which incorporate contextual word representations, enabling better handling of polysemy, synonyms, and other complex language phenomena in group chat data. DistilBERT embeddings significantly improve semantic understanding of short texts, and through an unsupervised learning approach, we can model effectively without the need for labeled data.

Furthermore, our approach introduces *Temporal Convolutional Networks (TCN)* to address the limitations of traditional RNN and CNN in handling long-range dependencies. TCN is well-suited for capturing long-range dependencies in group chat data, especially with the long time intervals between responses. Unlike sequence-based models, TCN does not assume sequential continuity, making it more suitable for asynchronous conversation environments.

In summary, while various methods have been proposed for topic detection, effective topic detection techniques for short texts, such as those found in group chats and social media, still face significant challenges. Future work should focus on improving the understanding of short texts, capturing temporal and contextual information effectively, and continuously tracking topic changes in dynamic environments. Our research addresses these gaps by introducing context-aware embeddings and methods to handle long-range dependencies, providing a more efficient and accurate solution for topic detection in group chats.

III. Problem Definition

Based on the characteristics of group chat texts and the needs of this experiment, we define three core concepts: *Chatlog*, *Utterance*, and *Dialogue*. These concepts serve as the foundation for the methods of topic shift detection and topic segmentation that will be discussed later.

$$C = [\ u_1, u_2, \ldots, u_n], \quad D_j \subseteq C, \quad u_i = \langle \text{time}_i, \text{id}_i, \text{text}_i \rangle \tag{1}$$

where:

- *Chat log (C):* A sequence of utterances representing a multi-turn conversation. Formally, we denote a chat log as $C = [u_1, u_2, \ldots, u_n]$, where u_i is the i-th utterance in the conversation.
- *Utterance (u_i):* A single message or turn in the chat. Each utterance u_i typically consists of text written by one participant at a particular time. It is the basic unit of the conversation.
- *Dialogue (D_j):* A contiguous subsequence of utterances in C that are topically coherent. In other words, $D_j = [u_p, \ldots, u_q]$ is a segment of the chat log such that all utterances in D_j pertain to the same topic. Once the topic of conversation changes, a new dialogue D_{j+1} begins.

Thus, a single chat log C may be segmented into M dialogues $D_1, D_2, \ldots, D_M$ chronologically. By definition, a *topic shift* occurs at the boundary between two consecutive dialogues in the chat log. If dialogue D_j ends with utterance u_i and the next dialogue D_{j+1} begins with utterance u_{i+1}, then we say that a topic shift occurs between u_i and u_{i+1}. The goal of *chat log topic shift detection* is: given the sequence of utterances $C = [u_1, \ldots, u_n]$, identify all indices i (with $1 \leq i < n$) such that a topic shift occurs between u_i and u_{i+1}. Similarly, the task is to predict the boundaries that divide C into dialogues $D_1, \ldots, D_M$ that correspond to distinct topics.

We can also formulate the task as a sequence labeling problem. For each position i between u_i and u_{i+1} in the chat log, we assign a binary label y_i to indicate whether a topic shift begins at that point (1 if yes, 0 if no). The sequence of labels $y_1, y_2, \ldots, y_{n-1}$ thus encodes the locations of topic changes in C. The aim of the topic shift detection model is to predict the correct topic shift label sequence for a given chat log.

IV. Methodology

Our proposed framework, *ChatSense*, is a fully unsupervised approach that combines deep semantic modeling, temporal context integration, and heuristic cues to detect topic shifts in multi-party chats. In contrast to prior methods, ChatSense does not require labeled data and is thus highly scalable. It is designed to handle the non-linear, asynchronous, and interwoven nature of group conversations by capturing both the semantic content of messages and structural indicators of topic change. An overview of the system architecture is shown in Figure 1, which illustrates the flow from raw chat messages through feature extraction and context modeling to the final clustering that identifies the boundaries of the topic change.

A. Data Preprocessing

Each raw chat log is first preprocessed to facilitate downstream analysis. Non-text elements such as HTML tags and system notifications are removed to clean message content and normalize timestamps into a standard datetime format. For every pair of consecutive messages, we compute the time interval between them as a feature (e.g., Δt_i = time gap between message u_i and u_{i-1}). We also assign each participant a unique anonymized ID for tracking speaker turns. After preprocessing, only essential columns are retained: `timestamp`, `user_id`, `message`, and `time_diff`.

B. Semantic Embedding via DistilBERT

To capture the semantic content of each utterance, we transform the message text into a high-dimensional vector using DistilBERT. Given an utterance u_i, we obtain its embedding $\mathbf{e}_i \in \mathbb{R}^{768}$ by feeding u_i into the pre-trained DistilBERT model to output token representation as the sentence vector. This 768-dimensional embedding encodes the nuanced meaning of the message in context, helping to disambiguate polysemous words and capture synonyms or slang usage common in chat conversations. The use of a transformer-based semantic representation ensures that even short, informal messages are mapped to a rich semantic space, addressing the limitations of traditional lexical similarity measures.

C. Feature Construction and Context Modeling

Each message is represented by concatenating three key features: (1) $\mathbf{e}_i \in \mathbb{R}^{768}$ DistilBERT embedding, (2) a time feature computed as $\log(1+\textit{time\_diff})$, and (3) a 100-dimensional user embedding. This results in an 869-dimensional vector for each message.

These vectors $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\}$ are input into a Temporal Convolutional Network (TCN), which uses 1D convolutions with dilation to capture sequential dependencies. The TCN is structured with several convolutional layers followed by a fully connected layer. The output of this component is a sequence of enriched feature vectors, each summarizing the content of an utterance in the context of its surrounding dialogue.

By using a TCN, our framework can learn dependencies between distant utterances without the recurrence constraints of RNN-based models, enabling more effective handling of asynchronous or interleaved dialogue flows. The TCN processes the entire sequence of message features and produces a context-aware embedding for each message. In essence,

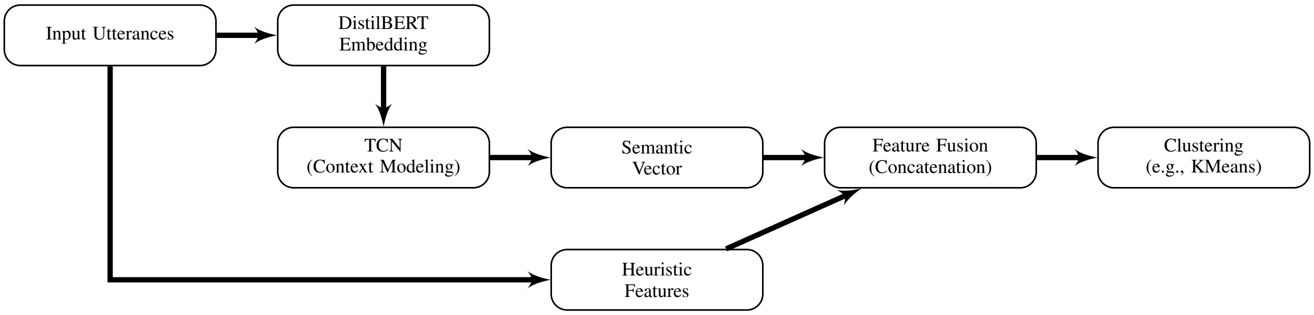

Fig. 1. Chatsense framework architecture.

it constructs higher-level features for each utterance that incorporate information from its neighboring messages within a certain temporal window defined by the TCN's receptive field. This context modeling allows the system to detect when an utterance does not fit smoothly with its predecessors – a strong indication of a topic shift. We optionally augment the TCN with a localized self-attention mechanism to let the model weigh the importance of words or features in recent messages when forming each context-aware representation, further enhancing its ability to pinpoint subtle shifts. Overall, the TCN-based context modeling enables ChatSense to track multiple overlapping topics and abrupt transitions, addressing the non-linear conversational structure challenge.

D. Heuristic Feature Extraction

In parallel with DistilBERT embeddings, we derive several heuristic features from each message and its context to capture linguistic and structural cues of topic shifts. These features include:

- *Message Length, Keyword Overlap, and Punctuation Ratio:* The number of tokens/characters in the message, lexical similarity, e.g., shared content words, and the fraction of punctuation marks, e.g., "?", "!", in a message, all of which tend to change at topic boundaries.
- *Topic Deviation, Speaker Change, and Silence Indicator:* We measure semantic deviation between messages using cosine distance between their DistilBERT embeddings. Larger distances indicate topic transitions.A binary flag for whether the speaker has changed, and a flag for prolonged silence, all of which help signal mood shifts, new subjects, or topic changes.
- *Sentiment Score and Shift:* The sentiment score of a message and its change compared to the previous message. Sentiment polarity is computed using the *SnowNLP* library for Chinese Characters and *TextBlob* for English Characters (range [0, 1]). We then calculate the sentiment shift as the absolute difference between consecutive sentiment scores, capturing mood transitions.

These heuristic features complement the semantic embeddings by providing information about conversational structure and dynamics. They are computed in an unsupervised manner from the raw data and require no domain-specific labeling. By capturing aspects like lexical continuity, discourse pauses, and speaker turns, these features help the system detect topic boundaries even in the face of interleaved and asynchronous chat threads.

E. Feature Standardization and Fusion

Since the extracted features have different scales and units, we apply standardization by using *StandardScaler* to process the data, ensuring that each feature has a mean of 0 and a standard deviation of 1, thus eliminating the dimensional differences between the features and ensuring that they are comparable on the same scale. Specifically, each continuous feature, both in the TCN embedding vector and in the heuristic feature set, is z-normalized across the chat log. Specifically, we subtract the mean and divide by the standard deviation for that feature. The fused vector $\{\mathbf{t}_1, \ldots, \mathbf{t}_n\}$ incorporates semantic information from the transformer embedding along with temporal and structural cues from the heuristics, yielding a comprehensive representation of utterance i. This multi-dimensional feature encapsulates various signals that are potentially relevant to identifying a topic shift.

F. Clustering for Topic Shift Detection

Finally, we identify topic shift points in an entirely unsupervised manner by clustering the context-enhanced utterance representations. We employ the k-means algorithm with $k = 2$ clusters to partition the set of message feature vectors. The idea is that one cluster will contain vectors corresponding to "continuation" utterances within an ongoing topic, and the other cluster will contain vectors that represent "boundary" utterances where a new topic begins. In practice, we run k-means on the set $\{\mathbf{t}_1, \ldots, \mathbf{t}_n\}$ for each chat log. Because $k = 2$ is fixed, this clustering essentially classifies each message as either a topic-boundary or a non-boundary point. Notably, this decision is made without any ground-truth labels or supervised training: clustering leverages the natural separation in the feature space created by our semantic, temporal, and heuristic integration. A message that deviates significantly in semantic context or structural pattern from its predecessors is likely to be isolated by clustering as a topic shift. After clustering, we interpret the two resulting clusters by assigning one cluster to the "no shift" ($y_i = 0$) class and the other to the "shift boundary" ($y_i = 1$) class. The sequence of labels $y_1, y_2, \ldots, y_{n-1}$ thus encodes the locations of topic

TABLE I
SUMMARY OF THE SELF-COLLECTED GROUP CHAT DATASETS.

Dataset	Domain	#Messages	Avg. Msg Length	Topic Shifts $\leq$ 10 min (%)
D5	Job Search and Employment	11,590	23.34	16.81%
D2	Student Daily Life	15,035	24.53	17.76%
D3	Renting and Housing	6,731	45.10	6.65%

changes for each u_i. We can determine which is which by examining features such as time gap or content change, which are expected to be higher in the shift cluster. The boundary detection for the chat is then given by the set of message indices that were assigned to the topic-shift cluster, i.e., if message u_i falls in the shift cluster, we mark a topic boundary before u_i.

Our clustering-based segmentation approach has the advantage of being completely data-driven and adaptive. Since no manual annotation is used at any stage, the entire method remains unsupervised. This makes ChatSense highly scalable to large or continuously streaming chat data, as it does not require retraining when new data arrives or when porting to a different domain. In summary, by combining deep semantic embeddings, temporal context modeling with TCN, and heuristic conversational features within an unsupervised clustering framework, ChatSense is able to robustly detect topic shifts in complex group chats without any labeled training data.

V. Experimental Setup and Result

This section presents the experimental results of our proposed topic shift detection framework on three datasets, comparing it with various baseline methods. The primary goal of the experiment is to validate the effectiveness and efficiency of the proposed framework, especially in terms of protecting privacy while maintaining classification accuracy when handling complex anonymization requirements. The experiment was conducted a Nvidia 2080Ti GPU with Python 3.8, Pytorch 2.0, scikit-learn 0.24, pandas 1.2.4, and NumPy 1.20. The source code of ChatSense can be obtained from GitHub[1].

A. Datasets

We conducted experiments on three self-collected social media group chat datasets, covering different discussion domains and annotated with topic shift information. Each chat log in the dataset is a multi-turn dialogue involving a group of participants. The time interval for the message ranges vary, and we specifically calculate the proportion of replies made within 10 minutes to capture rapid topic shifts. We segmented each chat log into dialogues by labeling the utterances where a new topic begins. These datasets come from group chat information on platforms, encompassing multiple topics ranging from casual discussions to task-oriented dialogues, including but not limited to job searching, student life, housing, etc. Certain platforms do not support direct extraction of group chat records. Therefore, we used specialized software to decrypt the database and extract real conversation data. Additionally, due to privacy concerns, including sensitive personal information, the dataset cannot be made public. To ensure the accuracy and integrity of the data, we filtered out incomplete or invalid messages. After this processing, the datasets provide a comprehensive view of topic segmentation in various real-world scenarios. Table I provides a summary of the datasets used in our experiments.

After filtering out incomplete messages, the datasets provide a comprehensive view of topic segmentation in various real-world scenarios.

B. Evaluation Metrics

To quantify performance, we compare the predicted topic boundaries with the ground truth annotated boundaries using standard evaluation metrics: *Accuracy and WindowDiff* [20]. Accuracy represents the proportion of correctly predicted boundaries that correspond to true topic shifts, reflecting the model's precision in detecting topic change points. WindowDiff is a standard segmentation quality measure that penalizes near-miss errors in boundary placement, meaning it penalizes small errors in boundary detection. WindowDiff provides a direct reflection of how well the overall segmentation structure of the conversation is preserved, with lower values indicating better alignment with the ground truth boundaries.

We report these metrics on the test set and compare them with the baseline models. All metrics are computed for each chat log and then averaged across the entire test set. By combining Accuracy and WindowDiff, we comprehensively assess the model's performance in topic change detection, ensuring that the model not only detects topic changes accurately but also maintains a good segmentation structure.

C. Baselines

To benchmark our proposed model, we compare it against the following baselines that mentioned in the related work under identical experimental conditions: *TF-IDF + k-means*, a traditional clustering approach using term frequency features and k-means; *LDA + k-means*, which combines Latent Dirichlet Allocation for topic modeling with k-means clustering; *Word2Vec + k-means*, using Word2Vec embeddings and k-means for topic detection. For the supervised baseline (Word2Vec), we divided the dataset into training, validation, and test sets at the chat log level. Specifically, 70% of the chat logs were used for training, 10% for validation , and the remaining 20% for testing; *DistilBERT + k-means*, an

[1] https://github.com/McGill-DMaS/ChatSense

TABLE II
PERFORMANCE COMPARISON OF TOPIC SHIFT DETECTION METHODS ACROSS DATASETS

Method	Dataset 2		Dataset 3		Dataset 5	
	Accuracy (%)	WindowDiff	Accuracy (%)	WindowDiff	Accuracy (%)	WindowDiff
LDA	67.88%	0.94	76.23%	0.91	79.93%	0.98
TF-IDF	68.76%	0.95	81.96%	0.88	94.77%	0.88
Word2Vec	84.92%	0.99	83.65%	0.88	96.74%	0.53
DistilBERT	71.64%	0.94	82.82%	0.90	61.48%	1.00
Chatsense	**72.65%**	**0.94**	**75.75%**	**0.94**	**74.95%**	**0.99**

advanced method using DistilBERT embeddings with k-means for contextualized segmentation; and *Heuristic Features + k-means*, which combines syntactic and discourse features such as sentence length and punctuation with k-means for topic boundary detection. Note that the k-means clustering method in all these baselines is analyzed using the same approach as in our proposed model.

D. Results

We evaluated ChatSense against several baseline methods on three real-world group chat datasets (D2, D3, D5). The baselines include four unsupervised approaches (TF-IDF + k-means, LDA + k-means, Word2Vec + k-means, DistilBERT + k-means), as well as a supervised baseline, which uses learned Word2Vec features and requires model training on annotated data. We report Accuracy, Precision and WindowDiff for each method on each dataset (see Table II). This comprehensive comparison highlights ChatSense's performance advantages in detecting topic shifts, as well as the trade-offs in efficiency compared to a supervised approach. Overall, ChatSense delivers competitive or superior performance, especially on multi-topic shifts and short text, on key metrics such as accuracy and WindowDiff, demonstrating its effectiveness in capturing topic boundaries in multi-topic chats.

In addition to the unsupervised baselines, we also consider a supervised baseline method: a model trained on a portion of annotated chat logs. In this case, we are using Word2Vec-based features with a learning algorithm. Supervised topic shift detection models (e.g., neural networks or classifiers) typically achieve higher raw performance metrics on test data similar to the data they were trained on. In our experiments, the supervised baseline performed slightly higher than that of ChatSense, and it showed a better segmentation error, as WindowDiff indicates after learning the specific patterns of the domain. However, ChatSense offers critical advantages in time efficiency and resource utilization, making it highly practical for real-world applications, often compensating for these small performance gaps.

VI. DISCUSSION

Our experiments confirm that ChatSense effectively addresses the challenges outlined in the introduction. By using an unsupervised strategy, ChatSense detects topic shifts in group chats without relying on manual labels, avoiding the data-dependence of supervised methods. The system integrates semantic embeddings and temporal modeling, proving robust to the asynchronous nature of multi-party conversations and capturing even subtle topic changes. The following sections discuss key aspects of our approach and outline future work in the areas of *Window Size*, *Embedding Model*, *Clustering*, *Dataset*, *Heuristic Features, Multilinguality*, and *Real-Time Tracking*. Each subsection highlights how our methodology ties back to the original goals and anticipates potential concerns.

A. Dataset, Window Size, Heuristic Features, and Clustering

The evaluation of ChatSense depends on several factors: dataset, window size, and clustering technique, all of which directly influence segmentation quality and model performance.

The dataset used for evaluation is crucial in shaping the model's effectiveness. While our current dataset contains multi-party conversations with varying topics, its scope is limited and may not fully represent the variety of online conversations. Future work will expand testing to a broader set of conversation types, including formal dialogues, e.g., meeting transcripts, and informal discussions, e.g., group chats, online forums, to assess ChatSense's robustness across different environments. We will also explore scaling ChatSense to handle larger datasets, such as millions of messages, to ensure its scalability.

In terms of window size, ChatSense avoids relying on a fixed segmentation window, thus mitigating issues like over-segmentation or under-segmentation. However, the pacing of conversations still affects the system's responsiveness. Future work could improve this by developing adaptive smoothing techniques that adjust sensitivity to conversation dynamics, enabling the system to handle both rapid and slower-paced discussions effectively.

The final component, clustering, groups semantically similar messages to detect topic boundaries. Our current approach uses k-means clustering, which requires setting the number of clusters (k) beforehand. This can lead to over-clustering or under-clustering, especially in long or multi-threaded conversations. Future work will explore hierarchical clustering or Bayesian non-parametric models to automatically determine k, as well as density-based clustering methods to better handle topics that evolve gradually or overlap.

In our experiments, heuristic features, such as message length, sentiment shifts, punctuation ratios, and speaker change indicators, played a critical complementary role alongside semantic embeddings and temporal context modeling. Specifically, heuristic features help the model detect subtle conversational cues not directly encoded in semantic embeddings, such as abrupt changes in participant behavior or prolonged conversational pauses, both of which are indicative of topic shifts. However, heuristic features alone may lack generalizability across diverse conversational domains due to their domain-specific or stylistic variations.

Future improvements in heuristic feature extraction should consider adaptive feature weighting or automatic heuristic selection methods to dynamically emphasize the most informative features in different conversation contexts. Another promising direction is the inclusion of conversational meta-features, such as reaction patterns (e.g., emoji usage), reply structure depth, or explicit user role distinctions, which may further enhance the detection accuracy of nuanced topic transitions.

In summary, the combination of dataset characteristics, window size, heuristic features, and clustering methods significantly influences ChatSense's performance. Further refinements in these areas, including dataset diversification and improved clustering techniques, will ensure better handling of various conversational contexts and scale.

B. Embedding Model and Multilinguality

Using DistilBERT embeddings, ChatSense captures the semantic context of each message and detects nuanced topic shifts. Unlike supervised classifiers, our embedding-based approach generalizes to new vocabulary and slang, enabling the system to adapt to dynamic language use. While domain-specific nuances, e.g., slang, acronyms, code-switching, etc., present challenges, ChatSense's design allows for easy substitution of different embedding models, thus enabling flexibility and better handling of evolving language and terminology.

A key area of future improvement lies in multilinguality, particularly with code-switching and cross-lingual semantics. While the current DistilBERT model is primarily focused on single language, its performance can be limited when dealing with other languages or mixed-language messages. To enhance its capabilities, we plan to integrate multilingual models. These models will allow ChatSense to understand semantic connections across multiple languages, thereby improving topic continuity detection even in conversations that switch between languages. Moreover, integrating language detection and language-specific models could address issues with low-resource languages or dialects, enabling ChatSense to handle a broader range of conversational contexts.

In future work, we aim to fine-tune the embedding model on larger multilingual chat corpora, incorporating lexical knowledge bases and context-dependent language models to further improve its sensitivity to nuances in informal and multilanguage conversations.

VII. Conclusion

The unsupervised and adaptive approach of ChatSense shows significant potential to detect topic changes in group chats. The results of our experiments demonstrate that this method can effectively identify topic changes in conversations without the need for manual annotations, making it particularly suitable for applications in information retrieval and dialogue analysis. With further refinements, including improvements to clustering techniques, multilingual support, and real-time capabilities, ChatSense will be better equipped to handle a variety of dynamic and diverse conversational contexts.

ChatSense is specifically designed for the detection of topic shifts in real-time ongoing conversations. Future work will focus on adapting the system for live deployment, emphasizing incremental processing and dynamic management of topic continuity. Incorporating features such as incremental clustering and topic reactivation mechanisms will enable ChatSense to track and update topics in real time, enhancing its applicability in areas such as live chat moderation, dynamic conversation summarization, and other real-time communication environments.

Acknowledgements

The research is supported in part by Discovery Grants (RGPIN-2024-04087) and CREATE Grants (CREATE-554764-2021) from the Natural Sciences and Engineering Research Council of Canada (NSERC), Canada Research Chairs Program (CRC-2019-00041), and R&D Spearhead Grants (2024-1301) from the National Cybersecurity Consortium (NCC).

References

[1] K. Avudaiappan, R. Ramesh, and P. Mahesh, "Analyzing multi-threaded conversations in online chat platforms," *International Journal of Information Technology*, vol. 27, no. 3, pp. 134–149, 2021.

[2] Y. Ai and J. Zhang, "Challenges in topic segmentation for multi-user conversations," *Journal of Computational Linguistics*, vol. 48, no. 1, pp. 45–68, 2022.

[3] M. Liu and X. Chen, "Automated analysis of nonlinear discussions in group chats," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 2, pp. 321–336, 2023.

[4] T. Chang and M. Zhao, "Advancements in nlp for topic modeling and summarization," *Journal of Artificial Intelligence Research*, vol. 76, pp. 233–258, 2023.

[5] L. Zhu and H. Wang, "Tracking topic evolution in multi-user online discussions," *ACM Transactions on Information Systems*, vol. 36, no. 4, pp. 29–48, 2018.

[6] J. Shen and K. Li, "Modeling multi-threaded conversations for chatbot applications," *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4210–4222, 2020.

[7] M. A. Hearst, "Texttiling: Segmenting text into multi-paragraph subtopic passages," *Computational Linguistics*, vol. 23, no. 1, pp. 33–64, 1997.

[8] G. Glavaš and J. Šnajder, "Event-centered topic segmentation," *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 415–425, 2016.

[9] S. Joty, G. Carenini, and R. Ng, "Topic segmentation and labeling in asynchronous conversations," *Journal of Artificial Intelligence Research*, vol. 47, pp. 521–573, 2013.

[10] M. Elsner and E. Charniak, "Disentangling chat," in *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP*, 2009.

[11] S. Mehri and et al., "Structured disentangled representations for dialogue," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[12] Z. Chen and T. Yang, “Transformer-based topic segmentation for multi-speaker conversations,” *Transactions of the Association for Computational Linguistics (TACL)*, vol. 9, pp. 1565–1580, 2021.

[13] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” *Information Processing and Management*, vol. 11, no. 5-6, pp. 513–523, 1975.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[15] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” vol. 9, no. 8, pp. 1735–1780, 1997.

[16] Y. Kim, “Convolutional neural networks for sentence classification,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1746–1751.

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2013, pp. 1–12.

[18] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[19] R. Gupta and Y. Lin, “Unsupervised topic shift detection in group conversations,” *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pp. 7912–7923, 2021.

[20] L. Pevzner and M. Hearst, “The WindowDiff metric for topic segmentation,” in *Proceedings of the 5th Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2002, pp. 11–18.