Comparative Analysis of Existing and a Novel Approach to Topic Detection on Conversational Dialogue Data

Haider Khalid [1] and Vincent Wade [2]

[1]School of Computer Science and Statistics, Trinity College Dublin, Ireland
khalidh@tcd.ie
[2]ADAPT Centre, School of Computer Science and Statistics, Trinity College Dublin, Ireland.
vinny.wade@tcd.ie

ABSTRACT

Topic detection in dialogue datasets has become a significant challenge for unsupervised and unlabeled data to develop a cohesive and engaging dialogue system. In this paper, we proposed unsupervised and semi-supervised techniques for topic detection in the conversational dialogue dataset and compared them with existing topic detection techniques. The paper proposes a novel approach for topic detection, which takes preprocessed data as an input and performs similarity analysis with the TF-IDF scores bag of words technique (BOW) to identify higher frequency words from dialogue utterances. It then refines the higher frequency words by integrating the clustering and elbow methods and using the Parallel Latent Dirichlet Allocation (PLDA) model to detect the topics. The paper comprised a comparative analysis of the proposed approach on the Switchboard, Personachat and MultiWOZ dataset. The experimental results show that the proposed topic detection approach performs significantly better using a semi-supervised dialogue dataset. We also performed topic quantification to check how accurate extracted topics are to compare with manually annotated data. For example, extracted topics from Switchboard are 92.72%, Peronachat 87.31% and MultiWOZ 93.15% accurate with manually annotated data.

KEYWORDS

Dialogue System; Topic Detection; PLDA Model, Term-similarity Analysis; Semi-supervised Learning.

1 INTRODUCTION

Initial techniques of representing textual information for conversation were focused on keywords, which are single words or phrases that have been determined as crucial for expressing a document's content. Today's conversational systems can rely on multiple modalities such as voice [1], body motion [2], gaze movements [3]. The last five years have seen rapid growth in text-based chatbots, designed to interact via human conversation (text or speech-based) and perform specific tasks. Such task-based chatbots are usually focused on a specific domain, e.g. tourist venue or entertainment. These chatbots are mostly integrated with a software application or a web portal to ease and speed up customer support [4]. These chatbots also provide a speech interface via a dedicated device, such as a smart speaker or a mobile phone-based device. The limitations of these chatbots are typically restricted to single-turn utterances to perform some specific tasks or provide

some information requested from a user. However, such chatbots can not support continuous conversations, typically consisting of multiple utterances in conversational dialogue [5]. Human conversational dialogue is far more complex as it consists of much more than individual commands or queries but contains multiple paradigms, for example, exchanges of information across topics, discussion, argument, and storytelling [6]. Several categories for intelligent conversational systems have been identified: task-oriented, questions answering systems, (open) social conversational systems, and purposeful conversational systems [7]. In a task-oriented conversational system, the system attempts to recognise specific user intent to fill 'slots' that parameterise a query or action the system can carry out. Social conversational systems are also open-domain systems where no specific domain is defined. Instead, the system aims to establish a connection with a user to carry out a long-term conversation by satisfying the user's communication needs, social belonging and affection [8]. It can be considered a combination of task-based and open domains where the intention is to engage in the conversation with a more general goal rather than just for entertainment or specific (slot filling) purposes.

In a purposeful conversational system, the system aims to establish a connection with a user through social interaction to carry out a long-term conversation by providing some valuable information to the user rather than just chit-chat. The interaction between the user and an agent revolves around a specific topic at a particular moment, and conversation shifts accordingly when the topic has changed for the interaction [7]. During the conversation, either the user explicitly changes the topic for the interaction with an agent, or the agent switches between the topics following the Dialog Move Tree (DMT) [9]. Most of the time, people often have to talk with an agent either in first-time encounters or with some acquaintances. In this scenario, managing the conversation with a user is challenging when an agent has no prior knowledge to interact with a user. Especially when a conversational topic is not defined [10]. In a machine-oriented conversational system, the agent needs to keep engagement with the user by managing the topics because the topics can influence the relevance of the dialogue and the user's engagement in the system[11].

Thus, managing the transitions between topics and suggesting a new topic for the conversation is essential. In this scenario, either a user can propose a topic for the conversation or an agent can propose a topic and lead the conversation. This research is based on a machine-oriented conversational system, where an agent can propose a topic, steer the conversation, and switch between topics when needed. There are two possibilities in a machine-oriented conversational system. Firstly, the agent already knows the topics for the conversation if the system is trained with the labelled dataset. Secondly, the system is trained with an unlabelled dataset, and an agent does not have prior knowledge about the topics for the conversation. So, topic detection is essential in a conversational system when an agent has no prior knowledge of topics discussed in the conversation. This paper's proposed approach focuses on **"how to improve the topic detection in conversational dialogue corpora"?** The objective is to detect the topics previously held in the unsupervised dialogue corpora in the experimental phase to train the agent. **The contribution of this research is the paper presents an experiment based on two phases for topic detection from the unsupervised dialogue corpus. The experimental results show that the proposed approach performs significantly better than the traditional topic detection methods**. The experiment begins with an unsupervised, unlabeled dialogue corpus containing only dialogue utterances to run the model and detect topics. In the second phase, a partially labelled, semi-supervised dialogue corpus is used to train the model and test on unsupervised data for topic detection.

Existing approaches such as k-mean clustering and LDA model approaches are mainly used separately to detect topics from textual documents and tweets [12]. The proposed experimental approach integrates the term similarity analysis technique based on TF-IDF scores and a bag of word approach to identify higher frequency words from dialogue utterances. Secondly, the higher frequency words refine by integrating the clustering technique and elbow method for the interpretation and validation to select the optimal number of clusters. Lastly, the Parallel Latent Dirichlet Allocation (PLDA) model explains a set of cluster observations to achieve topic detection. In this approach, each dialogue utterance is considered a document. The classical bag of words approach evaluates the importance of words in each document with the TF-IDF weighting scheme. The term similarity analysis group similar terms and clusters the document-to-document and document-to-cluster combination. Also, we use the elbow method to interpret and validate consistency within-cluster analysis to select the optimal number of clusters. We use precision, recall and F-measures metrics for the evaluation to compare our results with traditional topic detection approaches. We used switchboard [1], personachat [2] and multiWOZ[3] dataset because it is based on human conversational dialogues. Moreover, these metrics we used for the evaluations are widely used to evaluate comparable systems.

The rest of the paper is structured as follows: section 2 presents background knowledge and a review of topic detection techniques. It also explains how topic detection differs in dialogue systems and the related work for topic detection. Section 3 describes the proposed experimental approach and explains the methods and techniques we follow for the experiment. Section 4 presents the experimental results to evaluate extracted topics with traditional topic detection techniques. Finally, the conclusion summarised the whole experimental proposed approach and evaluation with an explanation of future work.

2 RELATED WORK

The world has experienced a massive increase in digital data across the internet in audio, video and text. Nowadays, people are more engaged with social media, news sites and blogs to seek updated information. However, in seeking information from textual data, the topic plays a vital role in classifying, organising and identifying the nature of the document [13]. In 1996, topic detection and tracking were a DARPA (Defense Advanced Research Projects Agency) sponsored initiative to investigate state of the art in finding and following the event in a stream of broadcast news stories [14]. Topic detection is helpful in many applications such as discovering natural disasters as soon as feasible [15,16], assisting political parties in predicting election results [17], and businesses in understanding user perspectives. It is also valuable for developing marketing content to better understand client needs [18], in engaging human users with conversational machine system to provide satisfactory information needs [7]. The most common representation of topics is as a list of keywords and typically uses weights to represent the keyword's importance in the topic. The significant distinction between topic detection in textual documents/tweets and dialogue corpus is that textual documents or tweets are static data that do not change in context over time. On the other hand, dialogue conversations shift the conversations' context over time. Also, conversational dialogues are short pieces of text with distinctive writing styles, abbreviations, and synonyms.

Many techniques for topic detection have been proposed, including clustering, matrix factorization, exemplar-based method and frequent pattern mining. Unfortunately, these

[1] https://catalog.ldc.upenn.edu/LDC97S62
[2] https://www.kaggle.com/datasets/atharvjairath/personachat
[3] https://github.com/budzianowski/multiwoz/tree/master/data

techniques generate terms that may or may not be correlated. Clustering topics involves grouping similar topics into a set known as a cluster. The idea is that topics in one cluster are likely to be different compared to topics grouped under another cluster [19]. In other words,, topics in one cluster are more co-related than those in another. The centroid of each discovered cluster is used to represent this cluster, and the top t words (in terms of TF-IDF) are used as the topic's keywords. In detecting topics, each utterance in the dialogue is represented using the TF-IDF technique, and the number of topics to be discovered from dialogue corpus is used as the number of cluster (k).

Another widely used approach for topic detection is pattern mining techniques which are based on different algorithms such as Apriori algorithm [20], Rapid association rule mining (RARM) algorithm [21], ECLAT algorithm [22]. Frequent Pattern Mining (FPM) is a widely used algorithm, including a series of techniques developed to discover frequent patterns in a large dataset. The same approach can be used to detect topics proposed in [23,24]. The FPM technique has two phases. First, detect the frequent pattern and secondly, rank the pattern. The technique uses an FP-growth algorithm to detect frequent patterns and has the following steps:

- Set a threshold value and calculate the frequency of each word. Neglect the words having frequencies below the threshold value.
- Sort the pattern according to their frequencies and their co-occurrences.
- Generate association rules.

After detecting the frequent patterns, the FPM technique sorts them and returns the top k frequent patterns as the detected topics. To sort the frequent pattern, several techniques were discussed [23] such as support and lifting the patterns. FPM has also been used in conjunction with probabilistic topic models to enrich document representation before standard probabilistic topic models are processed [25]. Another variation of FPM is soft frequent pattern mining (SFPM). SFPM considers the co-occurrence between two terms and the relations between multiple terms in grouping the terms. SFPM begins with the set S, which has only one term and then extends this set greedily by measuring the similarity between the set S and each term. This process is repeated until the similarity between the set S and the next term is less than a certain threshold.

Due to the textual data length limitation, detecting the topic using FPM in the short text is more challenging than in the long text. Thus, most existing approaches are unsuitable for topic detection in the short text (the occurrence in dialogue corpora).

Another exemplar-based approach detects topics from short text and represents a topic as an exemplar. This exemplar is much easier to be interpreted by the user as it contains related terms, and it represents a topic [26]. Elbagoury [26] use exemplar-based approach to detect topics in tweets. The approach constructs the similarity matrix between every pair of tweets and categories the behaviour of the similarity distribution of each tweet into two categories, which are:

- There is a low sample variance in the similarity distribution of tweet i and therefore tweet i is similar to many tweets, or tweet i is not similar to most other tweets.
- There is a high sample variance in the similarity distribution of tweet i and therefore tweet i is similar to a set of tweets and less similar to the others, which will be a good representative of the topic it is discussing.

Matrix factorization is another type of technique and includes latent semantic indexing (LSI), which projects a data matrix X into a lower-dimensional space with latent topics. An indexing and retrieval method uses a mathematical technique called singular value decomposition (SVD) to identify the patterns in the relationship between terms and concepts

in an unstructured text collection. It is a popular text analysis technique which extracts the statistical 'contextual usage meanings of words from a large corpus of text [27]. However, there are two interpretability drawbacks to LSI. First, the factorized matrices may contain negative values without intuitive interpretation. Second, the extracted topics are latent and difficult to interpret. Non-negative matrix factorization (NMF) is another class of techniques that ensures that the factorized matrices have non-negative values. Furthermore, traditional topic detection approaches that focus on representing topics using terms are negatively affected by the length limitation and lack of contextual information.

Latent Dirichlet Allocation (LDA) is another widely used technique in natural language processing for topic detection and semantic mining from textual data [28]. LDA is a generative statistical model that explains a set of observations by breaking them down into unobserved groups, with each group explaining why some parts of the data are similar [28]. It imagines a group of words representing a predetermined set of topics and keywords. The idea behind LDA is that each document can be described by the topic distribution, and each topic can be described by a word distribution, which is the 'bag of words' assumption. In the bag of word model, the order of words is not taken into account by applying the exchangeability property of words. Topic extraction methods based on the LDA model have been widely applied in many domains, including information retrieval, text mining, social media analysis, and natural language processing. For example, topic extraction based on social media analytics improves understanding people's reactions and conversations in online communities. The limitation of the LDA model is that the extracted topics are latent and can not capture correlation. Also, the number of topics is fixed and must be known ahead of time.

Besides improving the LDA algorithm, the parallel LDA method (PLDA) [29] has also been proposed to accelerate LDA training. For example, parallel LDA divides training documents into subsets with a similar number of tokens. In parallel LDA training, the corpus-topic count and the word-topic count (keywords) are calculated based on all documents. Apart from document-wise data partition, they further divide data according to words for execution iteration. PLDA smoothes storage and computation for long distributed LDA computations and provides fault recovery.

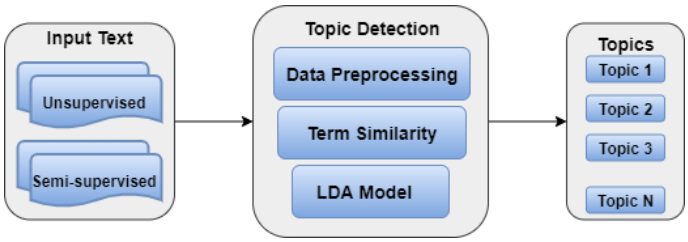
3 PROPOSED METHOD

As mentioned earlier, the standard representation of presenting topics is by multiple keywords. In the experiment, we also use the keyword extraction approach where each word in the dialogue utterance is associated with weights, representing the importance of the words. The weights of the words are also considered for the ranking. The highly ranked words are considered the keywords of the topics. Moreover, the highly-rated keywords extracted from the text can also represent the topic and a specific category like sports, music, travel, and food.

As mentioned in the introduction, the experimental approach is designed in two phases using unsupervised and semi-supervised dialogue corpora. The experimental first phase runs on three different datasets and uses unsupervised dialogue utterances to retrieve keywords. This section includes a detailed explanation of the experiment. In this phase, the PLDA model does not require training data to learn from data first and then perform topic detection. Instead, it detects the topics from an unlabelled corpus without training the model. However, the limitation is that the extracted topics are latent and not related to each other. Also, this approach does not verify "how accurate and true positive are extracted topics from the proposed approach in the experiment". To address this limitation,

in the second phase of the experiment, we use a partially annotated corpus to train the proposed model and then perform topic detection on an unlabelled corpus. The benefit of using partially labelled data for training is that the model learns from data and pre-trains itself to recognize which keyword belongs to which topic in the training phase. After training the PLDA model, run the model on unlabelled data to detect the topics. This semi-supervised learning bridges the gap between unsupervised and the semi-supervised PLDA model to discover unlabeled statistical relationships in the dialogue utterances. The partially supervised learning emphasizes the relationship between annotated topics and word features to extract topics from the dialogue utterances [30].

In order to extract topics, retrieve the keywords and obtain the semantic representation of topics from the dialogue corpus. Firstly, we preprocessed the data and then applied the term similarity analysis technique by recognising the similarity between words from dialogue utterances and grouping similar words. Secondly, we applied k-means clustering to make clusters for all high-frequency keywords. Lastly, the proposed approach used a PLDA topic model to detect the topics combined with the elbow method to select the optimal clusters. In the experimental procedure, topic detection is divided into three stages: data preprocessing, term similarity analysis with clustering and topic detection with Parallel Latent Dirichlet Analysis (PLDA) combined with elbow method; mentioned in figure **??**.



The above figure describes three blocks for topic detection. The first block is based on the unsupervised and semi-supervised dataset as an input to initialise the experiment. The second block is based on the experimental techniques, which are data preprocessing term similarity and parallel LDA model with elbow method. Finally, the third block shows the extracted topics as an output.

Fig. 1. Topic detection overview.

4 EXPERIMENTAL SETUP

4.1 Data Preprocessing

Initially, cleaning the dataset was necessary to reduce the computational power and execution time. In the experiment, we use switchboard data [31], multiWOZ dataset [32] and persona chat [33]. The data that support the findings of this study are openly available in the respiratory [4]. All these datasets are based on human conversational dialogues. In preprocessing, we removed smaller conversations, including stop words and interjections such as "uh-huh", "okay", "right", "oh", "Um-hum". These words are not keywords and are interpreted as topics [34]. We also use a markup tag filter, Stanford tagger, punctuation eraser, number filter, N character filter, stop word filter, and case conversion in the data cleaning process [35] to reduce the computational power. First, using the "Stanford tagger" node, all of the words in the documents are POS tagged. The lemma of each word is

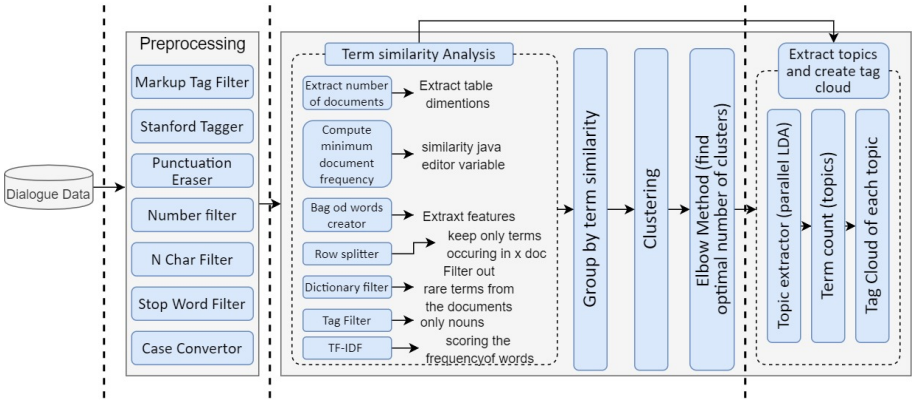
[4] https://github.com/hkmirza/Dialogue-Datasets

then extracted using a "Stanford Lemmatizer". The "Punctuation Erasure" node removes punctuation marks, filters numbers and stop words, and converts all terms to lowercase.

The following table 1 explain the data statistics used for the experimental procedure followed by the the pipeline of the experimental procedure is defined in the figure 2.

Table 1. Data Statistics

Metric	**Switchboard**	**MultiWOZ**	**Personachat**
Total Dialogues	2400	8438	10907
Total Utterances	18640	115424	162064
Average Conversation Length	12.7	13.68	7.2
Average token per utterance	21.5	13.18	10.57
Language	English	English	English



The above figure describes a detailed experimental procedure. Firstly it preprocessed the data using multiple filters. Secondly, it shows term similarity techniques on dialogue utterances to define the keywords for dialogue topics. Lastly, it shows the topic extraction with the PLDA method.

Fig. 2. The pipeline of experimental procedure[7]

4.2 Term Similarity Analysis

In term similarity, the first challenge in the topic detection experiment is finding dialogue utterances similar in content. The dialogue is composed of utterances; each utterance is considered a single document in the experiment. Documents are represented as vectors of features using the vector space model. Often these features are the terms (e.g. n-grams) that occur within the document collections. For example, if there are N terms in a document collection, each feature vector would correspondingly contain N dimensions. In this method, the feature value uses a binary value to indicate the existence of the featured term. However, the model also incorporates the frequency of a term. Suppose a term is more often used in a document. In that case, that term has greater importance in a document, which causes a problem of lending too much weight to a common term that may occur with a degree of frequency throughout the entire collection. We use the existing term frequency-inverse document frequency (TF-IDF) technique in the experiment to discount these high-frequency terms. We followed [36] formula to compute TF-IDF as:

$$\omega_i, j = tf_i, j \times log\frac{N}{n_i} \tag{1}$$

In the equation, the weight of term i in a document j for a vector is the product of its frequency in j and the log of its inverse document frequency in the document collection. The ni represents the number of documents in the collection that contain the term i, and N is representing the total number of documents in the entire collection. Considering each utterance, we utilize the frequency of a term in an utterance, discount by the log of its inverse frequency across all dialogue conversations. In the experimental approach, the bag-of-words model is used along with TF-IDF. The words that rarely occur in the short utterance may have neighbours in the feature vector space, which can identify which word belongs to which topic in the short dialogue utterance. Therefore, we enrich the bag-of-words representation by including neighbouring words in the feature vector. After detecting the high-frequency similar features, k-mean clustering involves similar grouping features into a set known as a cluster. Items in one cluster are likely to be different than those grouped under another. For each discovered cluster, its centroid is used to represent this cluster, where the highly ranked top t words are used as the keywords of the topics (in terms of TF-IDF weights). For topic detection, each utterance in the dialogue is represented using the TF-IDF technique, and the number of discovered topics is used as the number of cluster (k). In k-means clustering, the elbow method determines the optimal number of clusters. The elbow method plots the cost function value as a function of k; as k increases, the average distortion decreases, each cluster has fewer constituent instances, and the instances are closer to their respective centroids. However, as k increases, the improvements in average distortion decrease. The value of cluster k at which the improvement in distortion decreases the most is known as the elbow, and it is at this value that it should stop dividing the data into other clusters. The elbow method considers the total "within clusters sum of square (WSS) error" minimizes this to absolute value and selects the optimal number of clusters.

4.3 PLDA Model

After term similarity and refining clusters from the elbow method, we use the PLDA model for topic extraction. It is an extension of the LDA model, dividing training documents into subsets with a similar number of tokens to smoothen storage and computation and provide fault recovery. It extracts a set of topics t in documents, and each topic t represents the w set of keywords. So we can have

- $\theta td = P(t|d)$ - which is the probability distribution of topics in documents (Choose $\theta \sim Dir(\alpha)$).
- $\Phi td = P(w|t)$ - which is the probability distribution of words in topics (Choose $\Phi w \sim Dir(\beta)$
- It draws topic assignment per token by $Z_{di} \sim Multi(\theta_d)$
- It draws word assignment per token by $x_{di} \sim Multi(\Phi_{z_{di}})$

We can also say that the probability of a word given a document, $P(w|d)$, is equal to:

$$\sum_{t \in T} P(w|t,d)p(t|d) \tag{2}$$

Where T represents the total number of topics. Dir(*) refers to the Dirichlet distribution, and α and β are the hyper parameters of the model. Multi(*) denotes the multinomial distribution. Also, let us assume that we have w words in our vocabulary for all of the documents. Assuming conditional independence, we can state that:

$$P(w|t,d) = P(w|t) \tag{3}$$

Hence $P(w|d)$ is equal to:

$$\sum_{t=1}^{T} P(w|t)p(t|d) \tag{4}$$

That is the dot product of θtd and Φwt for each topic t. The approach randomly assigned weights to the probability distribution of words and topics by following three simple steps: randomly choose a topic from the distribution of topics in a document based on their assigned weights. Next, choose a word at random from the distribution of words for the chosen topic and insert it into the document. Lastly, repeat this process for the entire document.

It is an inverse process of the generative model in LDA, where known documents are models as a mixture of T latent topics. LDA training, in particular, attempts to find the posterior distribution of latent variables such as Φ, θ, and Z given word assignments, X. In PLDA, we exploit the data-level parallelism. First, we partition training documents into V subsets for parallel sampling. Initially, each document is assigned a random topic, and count matrices are calculated. Then, the corpus topic count, C_t, is duplicated for each subset. Finally, to reduce the memory requirement, the word-topic count, C_{tw}, is shared by all documents. In each iteration, k keywords, one from each subset, are fetched in parallel. After finishing parallel sampling, local copies of Ck are aggregated to the total number of topics T. We used Gibbs sampling in the PLDA. It generates Z samples by integrating Φ and θ. The conditional distribution is the topic of the i^{th} token (keywords) in the d_{th} document, z_{di}.

$$P(z_i = t|z_{-i}, w_i, d_i) \propto \frac{C^{WT}_{w_i t} + \eta}{\sum_{w=1}^{W} C^{WT}_{wt} + W\eta} \times \frac{C^{DT}_{d_i t} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha} \tag{5}$$

So we can have:

- $P(z_i = t)$ - The probability if assigning token (keyword) i to the topic t.
- z_i - Represents the topic assignments of all the other tokens (keywords).
- w_i - Word (index) for the i_{th} token (keyword).
- d_i - Document containing the i_{th} token (keyword).
- C^{WT} - Word-topic matrix, the wt matrix generated for the processing.
- $\sum_{w=1}^{W} C^{WT}_{wt}$ - The total number of tokens (keywords) in each topic.
- C^{DT} - Document-topic matrix, the dt matrix generated for the processing.
- $\sum_{t=1}^{T} C^{DT}_{d_i t}$ - The total number of tokens (keywords) in the document i.

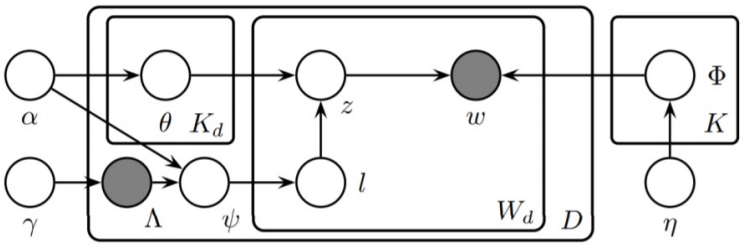
Where η determines the topic distribution for the words; the higher the number, the more evenly distributed the words will be across the specified number of topics (T). α determines the topic distribution for the documents. W is the total number of words in the total set of documents.

4.4 Semi-supervised Experimental Approach

As stated earlier in the introduction, the proposed approach significantly enhances the topic extraction and outperform with traditional method. But the limitation of this approach is the extracted topics are latent and not related to each other. Also, this method

does not verify "how accurate and true positive are extracted topics from the proposed approach in the experiment".

The second phase of the experiment aims to evaluate how accurate and true positive are extracted topics from the proposed approach in the fist phase of the experiment. We use partially manually annotated data, train the model and let the model categorize text that already knows which keywords fall under the specific topic category. We use the same switchboard, multiWOZ and Personachat dialogue datasets. In our method, we first identify the most common topics used within the dialogue utterances. The PLDA model takes as input the given conversations and detects significant words for each topic. Secondly, the trained PLDA model can determine the potential topic addressed in each conversational utterance. The utterances flow is then transformed into a sequence of potential topics within each conversation. Finally, the semi-supervised PLDA topic model is evaluated by computing its coherence over each topic's most significant words. The semi-supervised version of PLDA extends it with constraints that align some learned topics with a human-provided label. The model exploits the unsupervised learning of topic models to explore the unseen themes with each label and unlabeled themes in the large collection of data [30].



The above figure describes the graphical model of the parallel LDA model. This illustrates the distribution of words, labels, pre-defined labels in a semi-supervised approach and how the observation of keywords is being done on dialogue utterances for topic detection

Fig. 3. PLDA probabilistic graphical model: each documents word ω and label Λ are observed, with the per-doc label distribution ψ, per-doc-label topic distributions θ, and per-topic word distributions ϕ hidden variables. Because each documents label-set λ_d is observed, its sparse vector prior γ is unused; included for completeness [30].

Figure 3 shows the Bayesian graphical model for PLDA. In our approach of semi-supervised topic detection, we use a collection of documents D [5], each containing a multi-set of keywords w_d from a vocabulary V and a set of labels Λd from a space of labels L. We want to recover a set of topics Φt that fit the observed distribution of words in the multi-labelled documents, where each topic is a multinomial distribution over words from vocabulary V that co-occur with each other and some label $l \in L$.

Formally, PLDA assumes the existence of a set of labels $1..L$, each of which has been assigned to some number of topics $1..T_L$. Each topic $\Phi_{l,t}$ is represented as a multinomial distribution over all terms in the vocabulary V drawn from a symmetric Dirichlet prior η.

Each document d is generated by drawing a document-specific subset of available label classes, represented as a sparse binary vector Λd from a sparse binary vector prior. A document-specific mix $\theta_{d,l}$ over topics $1..T_l$ is drawn from a symmetric Dirichlet prior α for each label $l \in \Lambda d$ present in the document. Then, a document-specific mix of observed labels Φd is drawn as a multinomial of size $|d|$. From a Dirichlet prior $\sim \alpha L$, with each

[5] each utterance is considered as a document and collection of documents considered as a dialogue

element $\Phi_{d,l}$ corresponding to the document's probability of using label $l \sim \Lambda d$ when selecting a latent topic for each word. For derivational simplicity, we define the element at position l of $\sim \alpha L$ as αTl, so $\sim \alpha L$ is not a free parameter. Each word w in document d is drawn from some label's topic's word distribution, i.e. it is drawn by first picking a label l from Φd, a topic T from $\theta d, l$, and then a word w from $\Phi l, t$. Ultimately, this word will be picked in proportion to how much the enclosing document prefers the label l, how much that label prefers the topic T, and how much that topic prefers the word w.

5 Experimental Results

The experimental approach was tested on three different datasets and compared against traditional topic detection approaches. It defines each topic as represented by an (unknown) set of keywords. These are the keywords that dialogue utterances cover. PLDA tries to map all the (known) dialogue utterances to the (unknown) topics in a way such that those topics mainly capture the words in each dialogue utterance. For the implementation of the PLDA model, it has two hyper-parameters for training, usually called *alpha* and β.

- Alpha controls the similarity of dialogue utterances. A low value represents utterances as a mixture of a few topics. In contrast, a high value will output utterances representations of more topics – making all the utterances appear more similar to each other.
- Beta controls topic similarity for dialogue utterances. A low value of Beta represents more distinct topics and fewer in the count, but unique words belong to each topic. A high value of Beta has the opposite effect, resulting in topics containing more words in common.

The model extracts the topics with the optimal parameters alpha is 0.5, beta 0.1 and sampling iteration 1000. These are the optimal pre-defined parameters of LDA model [28]. The number of topics varies and depends on the nature of the dataset. Table 2 shows the initial extracted topics from unsupervised switchboard, personachat and multiwoz dialogue datasets.

Table 2. Detected topics from unsupervised dialogue corpus.

Switchboard			**Personachat**			**MultiWOZ**		
Topic0	**Topic1**	**Topic2**	**Topic0**	**Topic1**	**Topic2**	**Topic0**	**Topic1**	**Topic2**
call	day	car	festival	cat	travel	time	hotel	people
car	dollar	feel	dancing	animal	family	booking	food	hospital
care	house	change	movie	bacon	kid	taxi	restaurant	department
family	look	guess	theater	dog	school	parking	time	police
child	money	kid	song	hunting	nurse	address	day	phone
home	month	people	sport	fish	job	center	wifi	time
course	pay	school	play	meat	police	airport	parking	day
Job	time	sort	concert	pet	people	price	cheap	museum
kid	week	stuff	listen	catch	street	transport	phone	address
lot	yeah	talk	write	fishing	child	attraction	stay	information

The semi-supervised experimental approach performed significantly better on all three datasets and a total number of extracted topics and keywords are greater in count than the unsupervised approach. The extracted topics shows in table 3, table 4 and table 5. Tables 6, 7 and 8 presents the proposed approach's comparative analysis (unsupervised

and semi-supervised) with traditional state-of-the-art topic detection approaches. Again, higher precision, recall, and f1 scores indicate a higher level of agreement. The comparative analysis tables show that the proposed approach performed better in all evaluation metrics.

Table 3. Detected topics from semi-supervised dialogue corpus.

Topic0	**Topic1**	**Topic2**	**Topic3**	**Topic4**
elderly	dollar	car	course	feel
care	pay	people	degree	stuff
nursing	phone	engine	computer	change
family	machine	gas	college	heat
kid	change	speeding	job	holiday
children	money	diesel	talk	weather
mental	payment	checks	school	hiking
kid	card	pollution	graphics	country
story	ATM	ride	study	food
home		space	science	
		environment		

Table 4. Detected topics from semi-supervised dialogue corpus on persona chat .

Topic0	**Topic1**	**Topic2**	**Topic3**	**Topic4**	**Topic5**
dog	festival	music	visit	taste	police
hunting	halloween	dancing	england	cheese	catch
shoot	gym	play	travel	fish	family
catch	exercise	netflix	living	macaroni	child
cat	hiking	movie	michigan	food	job
animal	fishing	song	honeymoon	bake	kids
pet	swimming	concert	holiday	foodie	people
fishing	sports	carnival	school	meat	nurse
	camping	theater	asia	eat	
	ski		street	chocolate	
				bacon	

Table 5. Detected topics from semi-supervised dialogue corpus on MultiWOZ dataset .

Topic0	**Topic1**	**Topic2**	**Topic3**
taxi	restaurant	hospital	police
train	food	people	time
booking	hotel	department	phone
attraction	wifi	stay	address
price	place	town	day
time	day	admission	department
destination	price	museum	centre
parking	cheap	college	
arrival	parking	phone	
cambridge	reserve	information	
london		church	
airport			

Table 6. Comparative evaluation of different methods with proposed novel approaches on switchboard dataset.

Unsupervised and Semi-supervised Switchboard Corpus

Methods	TP Rate	FP Rate	Precision	Recall	F-Measure	Accuracy
Traditional LDA [28]	0.836	0.092	0.762	0.834	0.874	0.566
k-mean Clustering [19]	0.865	0.109	0.778	0.861	0.899	0.490
Matrix Factorization (LSI) [27]	0.613	0.082	0.881	0.873	0.876	0.545
Examplar-based Method [26]	0.704	0.097	0.878	0.857	0.867	0.684
Term Similarity + PLDA + Elbow Method(unsupervised)	0.922	0.089	0.846	0.931	0.915	0.734
Term Similarity + PLDA + Elbow Method(semi-supervised)	0.891	0.077	0.948	0.891	0.919	0.866

Table 7. Comparative evaluation of different methods with proposed novel approaches on personachat dataset.

Unsupervised and Semi-supervised PersonaChat Corpus

Methods	TP Rate	FP Rate	Precision	Recall	F-Measure	Accuracy
Traditional LDA [28]	0.684	0.423	0.617	0.658	0.637	0.551
k-mean Clustering [19]	0.601	0.398	0.605	0.615	0.605	0.457
Matrix Factorization (LSI) [27]	0.715	0.584	0.567	0.711	0.649	0.652
Examplar-based Method [26]	0.791	0.501	0.593	0.794	0.679	0.711
Term Similarity + PLDA + Elbow Method(unsupervised)	0.765	0.416	0.668	0.768	0.713	0.791
Term Similarity + PLDA + Elbow Method(semi-supervised)	0.865	0.414	0.692	0.861	0.769	0.843

Table 8. Comparative evaluation of different methods with proposed novel approaches on multiwoz dataset.

Unsupervised and Semi-supervised MultiWOZ Dialogue Dataset

Methods	TP Rate	FP Rate	Precision	Recall	F-Measure	Accuracy
Traditional LDA [28]	0.725	0.441	0.607	0.725	0.664	0.612
k-mean Clustering [19]	0.654	0.454	0.605	0.652	0.630	0.527
Matrix Factorization (LSI) [27]	0.558	0.281	0.691	0.557	0.614	0.471
Examplar-based Method [26]	0.640	0.345	0.647	0.645	0.639	0.633
Term Similarity + PLDA + Elbow Method(unsupervised)	0.731	0.354	0.691	0.731	0.712	0.722
Term Similarity + PLDA + Elbow Method(semi-supervised)	0.874	0.477	0.721	0.845	0.781	0.764

6 EVALUATION AND DISCUSSUION

The evaluation of the performance of a topic model is not an easy task. In most cases, topics need to be manually evaluated by humans, which may express different opinions and annotations. The most common quantitative way to assess a probabilistic model is to measure the log-likelihood of a held-out test set performing perplexity. However, the authors in [37] have shown that, surprisingly, perplexity and human judgment are often not correlated and may infer less semantically meaningful topics.

To evaluate our proposed approach with traditional methods, we adopt the evaluation metrics precision, recall, f-measures and accuracy: which are widely used in knowledge extraction and in information retrieval systems. We calculate the Precision, Recall, F-measure and accuracy on the whole corpora to observe the global performance of the system.

– Precision: the percentage of texts the model got right out of the total number of texts that it predicted for a given topic.

$$P = \frac{TruePositive}{TruePositive + FalsePositive}$$

$$= \frac{TruePositive}{TotalPredictedPositiveItems}$$

– Recall: the percentage of texts the model predicted for a given topic out of the total number of texts it should have predicted for that topic.

$$R = \frac{TruePositive}{TruePositive + FalseNegative}$$

$$= \frac{TruePositive}{TotalAccuratePositiveItems}$$

– F-Score: the harmonic mean of precision and recall.

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

– Accuracy: the percentage of texts that were predicted with the correct topic.

$$Accuracy = \frac{NumberofCorrectPrediction}{Totalnumberof Predictions}$$

One important note: the model does not use training data to measure performance, as the model has already seen these samples in the semi-supervised approach.

From the table 6, table 7 and table 8, we can observe that Precision, Recall and F-measure are all improved when we apply a semi-supervised approach. The reason is that using the term similarity defines similarity measures and refines the utterances with the similarity between dialogue utterances before applying the PLDA model and elbow method. As a result, the results are better with traditional approaches. We can see in switchboard F1 scores and accuracy is better than personachat and multiwoz because the switchboard dataset is small and tokens in each utterance are higher than peronachat and multiwoz dataset. So, the proposed approach performs better if the dataset size is small. Another factor is the dialogue utterance length. If we compare the F1 score of personachat and multiwoz, the result in multiwoz is slightly better than personachat because the dialogue utterances in personachat are smaller than the multiwoz. Personachat workers were instructed to keep each sentence short to a maximum of 15 words [33], which leads to more irrelevant topic/keyword extraction shown in the table 9.

To evaluate how accurate the extracted topics were from the proposed experimental approach, We used manually annotated switchboard, personachat and multiwoz dataset and performed topic quantification. As a result, we observed that the keywords in extracted topic are 92.72% similar in switchboard dataset, 87.31% in personachat and 93.15% in multiwoz shown in the table 9.

Table 9. Comparison between human annotated and machine extracted topics.

Match Type	**Switchboard**	**Personachat**	**MultiWOZ**
Exact	92.72%	87.31%	93.15%
Missing	03.41%	04.45%	02.21%
Irrelevant	03.87%	08.34%	04.64%

7 Conclusion and Future Discussion

This work proposed topic detection techniques from the unsupervised dialogue corpus by adapting term similarity analysis with Parallel Latent Dirichlet allocation (PLDA) with the elbow method. The experimental procedure was performed in two phases. Firstly, topic detection from an unsupervised dialogue dataset. Secondly, training the proposed model with a semi-supervised dialogue corpus lets the model learn to categorize text that already knows which keywords fall under the specific topic category and then perform topic detection from an unsupervised dialogue dataset. The proposed experimental approach was performed on switchboard, personachat and multiwoz dialogue datasets. The semi-supervised experimental approach performed significantly better than the unsupervised approach. The extracted keywords in each topic from semi-supervised are more cohesive and meaningful. However, in comparative analysis with traditional topic detection approaches, the semi-supervised technique performs better in all evaluation metrics. In the comparative analysis of different datasets, according to the experimental results, the proposed approach performs better if the dataset is small [6] and the dialogue utterances are not too short [7]. Topic detection alone is not useful unless the extracted topics map the dialogue dataset to find the reasoning and divide the dialogue into segments for a particular topic.

The topic detection alone cannot assist in producing a coherent response on a particular topic for human-machine conversation. Instead, it requires topic modelling, which allows for segmenting the whole unsupervised dialogue text according to a particular topic with dialogue act and contextual information of the conversation. Furthermore, dialogue topic modelling requires topic spotting to select an appropriate topic from the topic network by identifying the intent of the user query from a machine perspective. The integration of topic detection, dialogue topic modelling and topic spotting helps the dialogue manager start the conversation with humans from a machine perspective.

8 Acknowledgement

This research was conducted with the financial support of the Science Foundation Ireland under grant agreement No. (13/RC/2106) at the ADAPT SFI Research Centre at Trinity College Dublin, University of Dublin, Ireland. The ADAPT SGI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centre Programmes and is Co-funded under the European Regional Development Fund (ERDF) through grant 13/RC/2016

[6] Dataset with an average 2500 dialogues
[7] An average 20 tokens in each dialogue utterance

References

[1] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12, 2018.

[2] Ryo Ishii, Taichi Katayama, Ryuichiro Higashinaka, and Junji Tomita. Generating body motions using spoken language in dialogue. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 87–92, 2018.

[3] Edin Šabić, Daniel Henning, Hunter Myüz, Audrey Morrow, Michael C Hout, and Justin A MacDonald. Examining the role of eye movements during conversational listening in noise. *Frontiers in psychology*, 11:200, 2020.

[4] Amon Rapp, Lorenzo Curti, and Arianna Boldi. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies*, page 102630, 2021.

[5] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*, 2015.

[6] M Gašić, Dongho Kim, Pirros Tsiakoulis, Catherine Breslin, Matthew Henderson, Martin Szummer, Blaise Thomson, and Steve Young. Incremental on-line adaptation of pomdp-based dialogue managers to extended domains. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[7] Haider Khalid and Vincent Wade. Topic detection from conversational dialogue corpus with parallel dirichlet allocation model and elbow method. *arXiv preprint arXiv:2006.03353*, 2020.

[8] Kaihui Liang, Austin Chau, Yu Li, Xueyuan Lu, Dian Yu, Mingyang Zhou, Ishan Jain, Sam Davidson, Josh Arnold, Minh Nguyen, et al. Gunrock 2.0: A user adaptive social conversational system. *arXiv preprint arXiv:2011.08906*, 2020.

[9] Oliver Lemon, Er Gruenstein, Alexis Battle, and Stanley Peters. Multi-tasking and collaborative activities in dialogue systems. *Trait Autom Langues, a special issue on dialogue*, 43, 06 2004.

[10] Younhee Kim. Topic initiation in conversation-for-learning: Developmental and pedagogical perspectives. *English Teaching*, 72(1):73–103, 2017.

[11] Nadine Glas and Catherine Pelachaud. Topic management for an engaging conversational agent. *International Journal of Human-Computer Studies*, 120:107–124, 2018.

[12] Rania Ibrahim, Ahmed Elbagoury, Mohamed S Kamel, and Fakhri Karray. Tools and approaches for topic detection from twitter streams: survey. *Knowledge and Information Systems*, 54(3):511–539, 2018.

[13] Ahmed Rafea and Nada A GabAllah. Topic detection approaches in identifying topics and events from arabic corpora. *Procedia computer science*, 142:270–277, 2018.

[14] James Allan. Introduction to topic detection and tracking. In *Topic detection and tracking*, pages 1–16. Springer, 2002.

[15] Onook Oh, Kyounghee Hazel Kwon, and H Raghav Rao. An exploration of social media in extreme events: Rumor theory and twitter during the haiti earthquake 2010. In *Icis*, volume 231, pages 7332–7336, 2010.

[16] Paul S Earle, Daniel C Bowden, and Michelle Guy. Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics*, 54(6), 2012.

[17] Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Fourth international AAAI conference on weblogs and social media*, 2010.

[18] Fuji Ren and Ye Wu. Predicting user-topic opinions in twitter with social and topical context. *IEEE Transactions on affective computing*, 4(4):412–424, 2013.

[19] Yu Zhang, Kanat Tangwongsan, and Srikanta Tirthapura. Streaming algorithms for k-means clustering with fast queries. *arXiv*, 2017.

[20] Ferenc Bodon. A fast apriori implementation. In *FIMI*, volume 3, page 63, 2003.

[21] Amitabha Das, Wee-Keong Ng, and Yew-Kwong Woon. Rapid association rule mining. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 474–481, 2001.

[22] Xiaomei Yu and Hong Wang. Improvement of eclat algorithm based on support in frequent itemset mining. *J. Comput.*, 9(9):2116–2123, 2014.

[23] Luca Maria Aiello, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, Ayse Göker, Ioannis Kompatsiaris, and Alejandro Jaimes. Sensing trending topics in twitter. *IEEE Transactions on Multimedia*, 15(6):1268–1282, 2013.

[24] MD Goethals. Data mining and knowledge discovery handbook (2nd edn.) chapter frequent set mining, 2010.

[25] Hyun Duk Kim, Dae Hoon Park, Yue Lu, and ChengXiang Zhai. Enriching text representation with frequent pattern mining for probabilistic topic modeling. *Proceedings of the American Society for Information Science and Technology*, 49(1):1–10, 2012.

[26] Ahmed Elbagoury, Rania Ibrahim, Ahmed Farahat, Mohamed Kamel, and Fakhri Karray. Exemplar-based topic detection in twitter streams. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[27] Thomas K Landauer, Peter W Foltz, and Darrell Laham. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284, 1998.

[28] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[29] Yi Wang, Hongjie Bai, Matt Stanton, Wen-Yen Chen, and Edward Y Chang. Plda: Parallel latent dirichlet allocation for large-scale applications. In *International Conference on Algorithmic Applications in Management*, pages 301–314. Springer, 2009.

[30] Daniel Ramage, Christopher D Manning, and Susan Dumais. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–465, 2011.

[31] John J Godfrey, Edward C Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society, 1992.

[32] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*, 2018.

[33] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*, 2018.

[34] Anmin Mao et al. Conceptuality and context-sensitivity of emotive interjections. *Open Journal of Modern Linguistics*, 7(01):41, 2017.

[35] Suad A Alasadi and Wesam S Bhaya. Review of data preprocessing techniques in data mining. *Journal of Engineering and Applied Sciences*, 12(16):4102–4107, 2017.

[36] D Jurafsky and JH Martin. Speech and language processing 18 bt—an introduction to natural language processing, computational linguistics, and speech recognition.

An introduction to natural language processing, computational linguistics, and speech recognition, 988, 2019.

[37] Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M Blei. Reading tea leaves: How humans interpret topic models. In *Neural information processing systems*, volume 22, pages 288–296. Citeseer, 2009.

Authors

Haider Khalid is an ADAPT PhD student at Trinity College Dublin, Ireland. His current research is focused on topic management for human-machine conversational systems, particularly on topic detection, dialogue topic modelling and topic spotting with NLP techniques. He holds a Master's degree in Software Engineering from Jiangsu University China, funded by the Chinese Scholarship Council (CSC). His research interest includes Natural Language Processing (NLP - supervised and unsupervised textual data), Machine learning techniques, Data mining, Text classification, Information retrieval, and Recommendation systems.

Vincent Wade is Director of the ADAPT Centre for Digital Media Technology and holds the Professorial Chair of Computer Science (Est. 1990) in the School of Computer Science and Statistics, Trinity College Dublin as well as a personal Chair in Artificial Intelligence. His research focuses on intelligent systems, AI and Personalisation. He was awarded the Fellowship of Trinity College for his contribution to research and has published over three hundred and fifty scientific papers in peer-reviewed international journals and conferences.