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

Anonymous ACL submission

Abstract

Topic detection in dialogue corpora has become a major challenge for a conversational systems, with efficient conversational topic prediction being a critical part of constructing cohesive and engaging dialogue systems (Sun et al., 2019). This paper proposed unsupervised and semi-supervised techniques for topic detection in conversational dialogue corpora and compared them with existing techniques. However, these existing topic detection techniques are widely applied to textual tweets, blogs, documents, textual data on the web. Therefore, we applied these existing techniques to dialogue corpora to detect the topics and compared them with the proposed approach because textual dialogues typically are irregular and short sentences. The paper proposes a novel approach for topic detection, which combines the clustering of known similar words, TF-IDF scores and 'bag of words' techniques (BOW) with the Parallel Latent Dirichlet Allocation (PLDA) Model to achieve topic detection. The approach also integrates the elbow method for interpretation and validation to select the optimal number of clusters. The paper comprises a comparative analysis of traditional LDA and clustering approaches across both unlabelled (unsupervised) and partially labelled (semi-supervised) switchboard corpus with a proposed novel approach. The evaluation results shows that proposed approach performs best using partially labelled topic dialogue corpora and out performed traditional and unsupervised methods.

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 (Porcheron et al., 2018), body motion (Ishii et al., 2018), gaze movements (Šabić et al., 2020), etc. The last five years has seen a rapid growth in text-based chatbots which are designed interact via human conversation (text or speech based) and to perform specific tasks. Such task based chatbots are usually focused on a specific domain e.g. tourist venue, entertainmnent etc. These chatbots are mostly integrated with a software application or a web to ease and speed up customer support (Rapp et al., 2021) or are offered using a speech interface via a dedicated device e.g. smart speaker, or via mobile phone based device. However, such chatbots are typically restricted to single turn utterances to perform some specific tasks or information request e.g. Smart devices such as Alexa, OK Google etc. However such conversational systems can not support continuous conversational dialogue (Lowe et al., 2015). Human conversational dialogue is far more complex as it consists of much more that individual commands or queries, but contains multiple paradigms for example, exchanges of information across topics, discussion, argument and story telling. (Gašić et al., 2014). Several categories for intelligent conversational systems have been identified: task-oriented, questions answering systems, (open) social conversational system, and purposeful conversational systems (Khalid and Wade, 2020). In task oriented conversational agents, the agent is attempting to recognise specific user intent to fill 'slots' that parameterise a query or action which the agent is able to carried out. Social conversational systems are also known as open-domain systems where no specific domain is defined. Instead, the system aims to establish a connection with a user to carry out long term conversation by satisfying user needs of communication, social belonging and affection (Liang et al., 2020). It can be thought of as combination of task based and open domain where the intention is to engage in the conversation with a more general goal rather than just for entertainment or for specific (slot filling) purposes.

In a purposeful conversational system, the system aims is to establish a connection with a user with social interaction to carry out 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 (Khalid and Wade, 2020). During the conversation, either the user explicitly changes the topic for the interaction with a machine, or the machine switches between the topics following the Dialog Move Tree (DMT) (Lemon et al., 2004). People often find themselves in a situation where they have to talk with agent, either in first-time encounters or with some acquaintances. In this scenario, when an agent has no prior knowledge to interact with a user, managing the conversation with a user is challenging, especially when a conversational topic is not defined (Kim, 2017). In a machine-oriented conversational agents, the agent needs to keep their human users' interaction continuous by managing the topics because the topics can influence the relevancy of the dialogue and the user's engagement in the system(Glas and Pelachaud, 2018).

Thus in order to manage the transitions between topics and suggest a new topic for the conversation, we need to be able to detect the topics that were previously held in the dialogue corpora. In this paper, the experiment is based on two phases for topic detection from the dialogue corpus. In the first phase of the experiment, we use an unsupervised, unlabelled dialogue corpus that contains only dialogue utterances. In the second phase of the experiment, we used a partially labelled, semi-supervised dialogue corpus to train the model for topic detection.

Existing approaches such as k-mean and LDA model approaches, are mostly used separately to detect topics from textual documents and tweets (Ibrahim et al., 2018). The topics extracted from the textual data and tweets can typically only be used to determine the category of the text. In the proposed approach, we combine existing techniques for topic detection to improve the accuracy of topics. The experiment integrates the k-mean clustering technique based on TF-IDF scored and a bag of word approach with parallel latent dirichlet allocation (PLDA) model and elbow method. In the experiment, we applied these approaches to the dialogue dataset to detect topics to initiate conversation between humans and machines. In this approach, each dialogue is converted as a document in the pre-processing data phase. Then, using the classical bag of words approach with the TF-IDF weighting scheme, dialogues are represented. The similarity measure is used for clustering the combination of document-to-document and document-to-cluster. Also, we use the elbow method to interpret and validate consistency within-cluster analysis to select the optimal number of clusters. To study the performance of semantic similarity between similar words, noise is removed from the data pre-processing phase; we use precision, recall and F-measures for the evaluation and compare our results with traditional LDA and clustering techniques (Khalid and Wade, 2020).

To evaluate of topic detection approach we designed a two phase experiment based on using unlabeled and partially labeled corpora. In the first phase of the experiment, unlabeled dialogue utterances are used from the switchboard corpus. The approach is purely unsupervised to extract topics from the dialogue corpus. The proposed approach significantly enhances the topic extraction and outperforms with traditional LDA model and K-mean clustering (Jelodar et al., 2019). However a limitation of this approach is that 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 these limitations, in the second phase of the experiment we use a partially annotated dataset.

Thus a second phase the experiments was carried out in which we evaluated the accuracy of the extracted topics when using such a partially annotated dataset. This type of semi-supervised learning bridges the gap between unsupervised learning 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 (Ramage et al., 2011).

This paper is organised as follow: section background knowledge and related work describe how topic detection is different in dialogue system and explains the related work for topic detection. Then, section, proposed approach briefly explains the methods and techniques we follow for the experimental procedure. The next section, describes the experimental results produced from using the proposed approach and evaluation of 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 etc. Nowadays, people are more engaged with social media, news sites and blogs to seek updated information. However, seeking information from textual data, topic detection plays a vital role in classifying and organising and identifying the nature of the document (Rafea and GabAllah, 2018). In 1996, topic detection and tracking was a DARPA (Defense Advanced Research Projects Agency) sponsored initiative to investigate state of art in finding and following the event in a stream of broadcast news stories (Allan, 2002). Topic detection is useful in many applications such as discovering natural disasters as soon as feasible (Oh et al., 2010; Earle et al., 2012), assisting political parties in predicting election results (Tumasjan et al., 2010), and businesses in understanding user perspectives. It is also useful in developing marketing content to better understand client needs (Ren and Wu, 2013), in engaging human users with machine conversational system to provide satisfy information needs (Khalid and Wade, 2020). 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 major distinction between topic detection in textual documents/tweets and dialogue corpa 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 (Khalid and Wade, 2020). Also, the conversational dialogues are short pieces of text with irregular writing styles, abbreviations, and synonyms.

Many techniques for topic detection have been proposed, including clustering and frequent pattern mining. Unfortunately, these techniques generate terms that may or may not be correlated to one another. Clustering topics involves the grouping of similar topics into a set known as a cluster. The idea being topics in one cluster are likely to be different when compared to topics grouped under another cluster (Zhang et al., 2017). In other words topics in one cluster are more co-related to each other than topics in another cluster. For each discovered cluster, its centroid is used to represent this cluster, where the top t words (in terms of TF-IDF) are used as the keywords of this topic. To detect topics, each utterance in the dialogue is represented using the TF-IDF scheme, and the number of topics to be discovered is used as the number of cluster (k).

Another approach widely used for topic detection are pattern mining techniques which are based on different algorithms. Frequent Pattern Mining (FPM) is widely used algorithm which includes a series of techniques developed to discover frequent patterns in a large database of transactions. The same approach can be used to detect topics as proposed in (Aiello et al., 2013; Goethals, 2010). The FPM technique has two phases. First, detect the frequent pattern and secondly, rank the pattern. The technique uses FP-growth algorithm to detect frequent patterns has following steps:

- Set a threshold value and calculate the frequency of each word. Neglect the words having frequencies below than 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 (Aiello et al., 2013) such as support and lift the patterns. FPM has also being used in conjunction with probabilistic topic models to enrich document representation before standard probabilistic topic models are processed (Kim et al., 2012). Another variation of FPM is soft frequent pattern mining (SFPM). SFPM considers both 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 limitation of the length of the text, detecting the topic using FPM in the short text is more challenging than in the long text. Thus, most of the existing approaches are not suitable for topic detection in short text (which is the occurence in dialog corpora).

Another approach, termed an 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 (Elbagoury et al., 2015). Elbagoury (Elbagoury et al., 2015) 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 for 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 patterns in the relationship between terms and concepts in an unstructured collection of text. It is a popular text analysis technique which extracts the statistical 'contextual usage meanings of words from a large corpus of text (Landauer et al., 1998). However, LSI has two interpretability drawbacks: for instance, the factorized matrices may contain negative values with no intuitive interpretation. Second, the extracted topics are latent and difficult to interpret. Non-negative matrix factorization (NMF) is another class of techniques that guarantees that the factorized matrices contain non-negative values. Furthermore, some 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 (Blei et al., 2003). 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. Topic extraction based on social media analytics improves understanding of people's reactions and conversations in online communities. In addition to extracting useful patterns and understandable patterns from their interactions, as well as what they share on social media websites such as Twitter and Facebook. The limitation of the LDA model is the extracted topics are latent and can not capture correlation. Also, the number of topics are fixed and must be known ahead of time. The table **??** shows the different existing techniques and their advantages and drawbacks.

3 Proposed Method

As mentioned earlier, the common representation of presenting topics is by a set of multiple keywords. In the experiment, we also use the keyword extraction approach where each keyword is associated with weights, and weight represents the importance of the keyword and are also considered for ranking the most appropriate keyword. Moreover, the highly ranked keywords extracted from the text can also be used to represent the topic and a specific category like sports, music, travel, food.

In order to extract these keywords from the dialogue textual data and obtain semantic representation of topics, firstly, we combined term similarity analysis by analysing frequent pattern in the dataset to detect topics and k-means clustering to make clusters for all high-frequency words in topics. Secondly, the proposed approach used an LDA topic model combined with elbow method to select the optimal number of clusters. In the experimental procedure, topic detection is divided into three stages: namely data pre-processing; term similarity analysis with clustering and elbow method; and topic detection with Parallel Latent Dirichlet Analysis (PLDA) mentioned in figure 1.

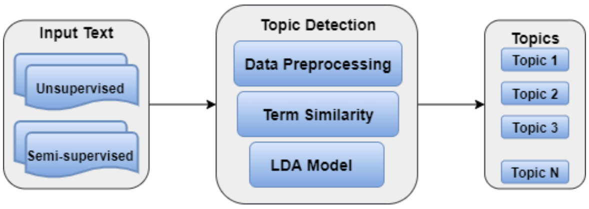

Figure 1: Topic detection from textual dialogue corpus

3.1 Experimental Setup

Data Preprocessing: Initially, cleaning the dataset was necessary for the experimental purpose to reduce the computation. In the experiment we use subset of switchboard corpus (Godfrey et al., 1992), which consisted of approximately 2,400 two-sided telephone conversations between 543 speakers (302 male, 241 female) from all over the United States was used. We used total 2145 conversation and removed smaller conversations such as "uh-huh", "okay", "right", "oh", "um-hum" etc. We 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.

In term similarity, the first challenge in the experiment for topic detection is to find the dialogue utterances that are similar in content under the term similarity analysis. The dialogue is composed of utterances, and in the experiment, each utterance is considered a single document. The vector space model is used to represent documents as vectors of features. Often these features are the terms (e.g. n-grams) that occur within the document collections. If there are N terms in a document collection, each feature vector would correspondingly contain N dimensions. In this method, the feature value use binary value to indicate the existence of the featured term. The model also incorporate the frequency of a term, the more often a term is used, the greater the importance of that term in a document. This also has a problem of lending too much weight to a common term that may occur with degree of frequeny throughout the entire collection. In the experiment we use term frequency-inverse document frequency (TF-IDF) to discount these high frequency terms. We followed (Jurafsky and Martin, 2019) formula to compute TF-IDF as:

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

Where the weight of a term i in the document vector for j is the product of its frequency in j and the log of its inverse document frequency in the collection, with ni representing the number of documents in the collection that contain term i and N representing the total number of documents in the 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 able identifying which word belongs to which topic in the short dialogue utterance. Therefore, we enrich the bag-of-words representation by including neighbouring words found in the feature vector. After detecting the high frequency similar features, k-mean clustering involves the grouping of similar features into a set known as cluster. Objects in one cluster are likely to be different when compared to objects grouped under another cluster. For each discovered cluster, its centroid is used as a representative of this cluster where the top t words (in terms of TF-IDF weights) are used as the keywords of this topic. For the purpose of detecting topics, each utterance in the dialog is represented using TF-IDF scheme and the number of topics to be discovered is used as the number of cluster (k). In k-means clustering, the elbow method is used to determine the optimal number of clusters. The elbow method plots the value of the cost function 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 decreases. The value of k at which the improvement in distortion decreases the most is known as the elbow, and it is at this value that we should stop dividing the data into further clusters. The elbow method consider the total "within clusters sum of square (WSS) error" andminimizes this to absolute value and selects the optimal number of cluters. After term similarity and refining clusters from elbow method we use PLDA model for topic extraction. It imagines a fixed set of topics and each topic represents s set of keywords. The pipeline of the experimental procedure is defined in the figure 2.

As stated earlier in the introduction, the proposed approach significantly enhances the topic extraction and outperform with traditional LDA model and K-mean clustering with elbow 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. A possible way to take par-

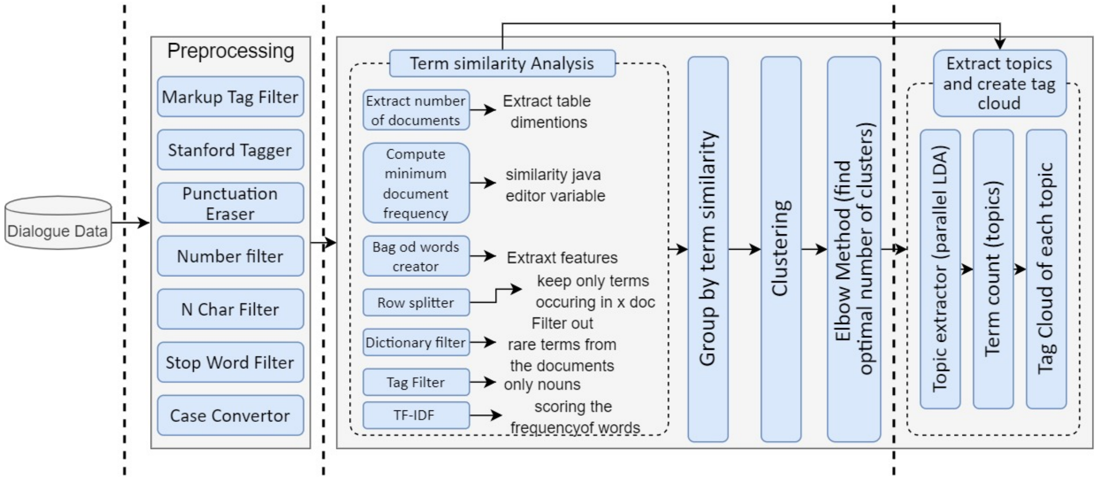

Figure 2: The pipeline of experimental procedure (Khalid and Wade, 2020)

tial manually annotated data, train the model and let the model categorize text that already knows which keywords fall under the specific topic category. We will use the same switchboard dialogue corpus we used in the topic detection experiment for this experiment. In our method, we first identify the most common topics used within the switchboard dialogue corpus. 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 (Ramage et al., 2011).

4 Experimental Results

The approach used PLDA model with elbow method to detect fixed set of topics. It defines each topic as represented by an (unknown) set of words. These are the topics that dialog utterances cover. PLDA tries to map all the (known) dialog utterances to the (unknown) topics in a way such that the words in each dialog utterance are mostly captured by those topics. The implementation of PLDA has two hyperparameters for training, usually called $\alpha and \beta$.

- Alpha controls the similarity of dialog utterances. A low value represents utterances as a mixture of 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 is the same but for topics, it controls topic similarity. A low value will represent topics as more distinct by making fewer, more unique words belong to each topic. A high value will have 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. The number of topics is 3, with ten words in each topic in the first phase of the experiment with unsupervised unlabeled dialogue corpus, shows in table 1.

Topic0	Topic1	Topic2
call	day	car
car	dollar	feel
care	house	change
family	look	guess
child	money	kid
home	month	people
course	pay	school
Job	time	sort
kid	week	stuff
lot	yeah	talk

Table 1: Topic detection from unsupervised dialogue corpus.

Methods	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**Accuracy**
Traditional LDA	0.836	0.092	0.762	0.834	0.874	0.566
Clustering	0.865	0.109	0.778	0.861	0.899	0.490
FPM+PLDA+Elbow Method(unsupervised)	0.922	0.089	0.846	0.931	0.915	0.734
FPM+PLDA+Elbow Method(semi-supervised)	0.891	0.077	0.948	0.891	0.919	0.866

Table 2: Comparative evaluation of different methods with proposed novel approaches.

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 3: Topic detection from semi-supervised dialogue corpus.

In the second phase of the experiment, the model performs significantly better with semi-supervised training data and extract 5 topics with a random number of co-related keywords, shows in the table 3.

The comparative analysis of traditional techniques and the proposed method is shown in the table 2. One important note: the model shouldn't use training data to measure performance, as the model has already seen these samples. The one possible way to measure would be to take manually annotated data, don't use it to train, and then use it to test when the model is trained. For the evaluation of different topic detection methods, precision, recall, F-scores and accuracy were used.

Evaluation Metrics:

- Accuracy: the percentage of texts that were predicted with the correct topic
- Precision: the percentage of texts the model got right out of the total number of texts that it predicted for a given topic.
- 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
- F-Score: the harmonic mean of precision and recall.

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 (Chang et al., 2009) have shown that, surprisingly, perplexity and human judgment are often not correlated and may infer less semantically meaningful topics. A potential solution to this problem is provided by the topic coherence, a typical way to assess qualitatively topic models by examining the most likely words in each topic.

5 Conclusion

This work proposed topic detection techniques from the dialogue corpus by adapting pre-existing techniques includes clustering, LDA model and elbow method. The experimental procedure performed two phases. Firstly, topic detection from unsupervised dialogue corpus by using the cross-validation technique. secondly, training the proposed model with semi-supervised dialogue corpus to let the model learn to categorize text that already knows which keywords fall under the specific topic category and extract the topic from unlabelled dialogue corpus. The topic detection experimental procedure was performed in different steps. In the first step, different data pre-processing existing techniques were used to remove noise from the dialogue corpus. Then the data is being used to extract similar features and transform them into clusters. Next, the Elbow method was used for interpretation and validation to select the optimal number of clusters. Finally, the PLDA model performs topic detection based on BOW and TF-IDF and computing topics. We compared our approach with the traditional clustering and LDA model, and the semi-supervised approach performed well in the evaluation.

6 Ethical Consideration

In the proposed experimental techniques, we ensure no ethical problems. The experiment is designed and inspire all computing professionals in the dialogue domain, including individuals, students, instructors and anyone who belongs to the natural language processing community. The dataset; switchboard corpus is publicly available for experimental purposes. According to the University of Pennsylvania, the switchboard data was assembled and published by the LDC, and all identified issues with the original publication of speech files have been addressed.

Before the experimental procedure, we ensure that there are no abusive and negative statements in the data to avoid physical or mental harm and no discrimination based on age, color, handicap, ethnicity, family status, gender identity, labor union participation, military status, nationality, race, religion or belief, sex, sexual orientation, or any other unsuitable consideration.

References

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

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

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

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

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

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

Rui Máximo Esteves, Thomas Hacker, and Chunming Rong. 2013. Competitive k-means, a new accurate and distributed k-means algorithm for large datasets. In *2013 IEEE 5th International Conference on Cloud Computing Technology and Science*, volume 1, pages 17–24. IEEE.

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

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

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. 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.

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

Xiangnan He, Hanwang Zhang, Min-Yen Kan, and Tat-Seng Chua. 2016. Fast matrix factorization for online recommendation with implicit feedback. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 549–558.

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

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

Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2019. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211.

D Jurafsky and JH Martin. 2019. 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.

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

Hyun Duk Kim, Dae Hoon Park, Yue Lu, and ChengXiang Zhai. 2012. 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.

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

A King. 2012. Online k-means clustering of nonstationary data. course project report.

Mikhail Krivenko and Vitaly Vasilyev. 2009. Sequential latent semantic indexing. In *Proceedings of the 2nd Workshop on Data Mining using Matrices and Tensors*, pages 1–9.

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

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

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

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

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

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

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

Daniel Ramage, Christopher D Manning, and Susan Dumais. 2011. 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.

Amon Rapp, Lorenzo Curti, and Arianna Boldi. 2021. 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.

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

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

Jian Sun, Wu Guo, Zhi Chen, and Yan Song. 2019. Topic detection in conversational telephone speech using cnn with multi-stream inputs. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7285–7289. IEEE.

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

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

7 Appendix

Algorithms	Advantages	Disadvantage
Sequential K-means (King, 2012)	It allows updating the model as new data is received with different time. Sequential k-means should be used when user expect the data to be received one by one or in chunks.	It doesn't allow development of an optimal set of clusters and for effective results. It is dependent on the order in which the data is received. More computational cost because each time when the model receives a new data point, it computes the whole data to find a similar object.
Distributed K-means (Esteves et al., 2013)	It allows parallel approach for clustering large datasets distributed across several machines. It decrease the complexity and computational time.	Symmetric encryption scheme compromised data privacy and can easily decrypt by other owners over the network. If the datasets are encrypted under the cloud's public key, data owners cannot decrypt their uploaded data due to not knowing the private key.
FP-Growth FPM (Aiello et al., 2013)	FP-Tree is expensive to build consumes more memory. Less time as compared to Apriori algorithm.	Dataset is scanned only two times.
LDA (Blei et al., 2003)	It is completely unsupervised and can learn the topics without the need for annotated training data. In the model, documents are distributed over topics and can classify documents by high probability topics.	The topics are uncorrelated (Dirichlet topic distribution cannot capture correlations). Fixed K (the number of topics is fixed and must be known ahead of time).
LSI (Krivenko and Vasilyev, 2009)	It provides low computational complexity of dimensionality reduction and low computation for the feature selection from feature space. The dimension of space is not fixed and dynamically changes to ensure a given level of relative approximation error of a matrix of observations.	The algorithm is not able to compute the emergence of a new relationship in the data.
ALS (He et al., 2016)	It works without a learning rate by an exact optimization in each parameter update bypassing the well-known difficulty for tuning gradient descent methods. It also minimizes the entire loss function at once by decomposes a large matrix into products of matrices and then alternately update each block of parameters.	Optimization of different matrices is non-convex and hard to solve at the same time.
NMF (He et al., 2016)	It makes low-rank approximate factorizations and minimizes least-squared error over input data. It can take many iterations from scratch to reach a suitable solution. Still, it allows us to stop during the iteration process and make a solution, regardless of the chosen rank of the factorization.	NMF is not convex and can give spurious, non-optimal results.

Table 4: Advantages and disadvantages of different existing techniques for topic detection.