A Semantics-enhanced Topic Modelling Technique: Semantic-LDA

DAKSHI KAPUGAMA GEEGANAGE, YUE XU, and YUEFENG LI, Queensland University of Technology, Australia

Topic modelling is a beneficial technique used to discover latent topics in text collections. But to correctly understand the text content and generate a meaningful topic list, semantics are important. By ignoring semantics, that is, not attempting to grasp the meaning of the words, most of the existing topic modelling approaches can generate some meaningless topic words. Even existing semantic-based approaches usually interpret the meanings of words without considering the context and related words. In this article, we introduce a semantic-based topic model called semantic-LDA that captures the semantics of words in a text collection using concepts from an external ontology. A new method is introduced to identify and quantify the concept–word relationships based on matching words from the input text collection with concepts from an ontology without using pre-calculated values from the ontology that quantify the relationships between the words and concepts. These pre-calculated values may not reflect the actual relationships between words and concepts for the input collection, because they are derived from datasets used to build the ontology rather than from the input collection itself. Instead, quantifying the relationship based on the word distribution in the input collection is more realistic and beneficial in the semantic capture process. Furthermore, an ambiguity handling mechanism is introduced to interpret the unmatched words, that is, words for which there are no matching concepts in the ontology. Thus, this article makes a significant contribution by introducing a semantic-based topic model that calculates the word–concept relationships directly from the input text collection. The proposed semantic-based topic model and an enhanced version with the disambiguation mechanism were evaluated against a set of state-of-the-art systems, and our approaches outperformed the baseline systems in both topic quality and information filtering evaluations.

CCS Concepts: • **Applied computing** → **Document management and text processing**; • **Computing methodologies** → **Information extraction;**

Additional Key Words and Phrases: Topic modelling, semantics, concepts, disambiguation

ACM Reference Format:
Dakshi Kapugama Geeganage, Yue Xu, and Yuefeng Li. 2024. A Semantics-enhanced Topic Modelling Technique: Semantic-LDA. *ACM Trans. Knowl. Discov. Data.* 18, 4, Article 93 (February 2024), 27 pages. https://doi.org/10.1145/3639409

1 INTRODUCTION

Immense volumes of text are generated around the world via social media, academic and research activities, news, digital libraries, and e-commerce activities such as customer reviews and product

Authors' address: D. Kapugama Geeganage, Y. Xu, and Y. Li, Queensland University of Technology, 2 George Street, Brisbane, Queensland, Australia, 4000; e-mails: {dakshi.kapugamageeganage, yue.xu, y2.li}@qut.edu.au.
Author's current address: D. Kapugama Geeganage, Queensland Government Procurement, Department of Energy and Climate, Brisbane, Queensland, Australia, 4001.

ACM 1556-4681/2024/02-ART93
https://doi.org/10.1145/3639409

promotions. Understanding, analyzing and summarizing textual content have become key tasks in our daily routines. Topic modelling is one of the state-of-the-art techniques used to understand, analyze, and summarize textual content. Topic models with different capabilities are widely applied to discover hidden themes in large text collections. Most of the prevailing topic models [4, 6–8, 17] apply probabilistic techniques such as word frequency and word co-occurrences to identify hidden topics in text collections. **Latent Dirichlet Allocation (LDA)** [8] is one of the most extensively used probabilistic topic models. However, statistical features alone may be insufficient to correctly understand the content of text collections. Thus, topic modelling approaches are evolving from purely statistical [4, 5, 8, 17, 30] to semantic-based approaches [10, 12, 18, 21, 34, 37–40] with the increasing awareness of the importance of semantics in the topic generation process. Although LDA is widely applied in many text-mining applications, limitations exist in the LDA topic model due to semantics not being captured and considered in the topic generation process.

It is essential to discover the accurate meaning of words in the semantic capture process. Existing semantic-based approaches incorporate different techniques such as external knowledge bases/ontologies, embeddings, knowledge graphs, and so on, with most of these approaches directly interpreting the meaning of words by mapping them with concepts from an external ontology. The ontology should contain a set of concepts and a list of words that match each concept. Hence, word–concept relationships are considered when describing the meanings of words. Some existing approaches [34, 39], however, directly use the word–concept relationships provided by the ontology in capturing the meaning of words without considering whether these word–concept relationships are relevant to the specific document collection in question. In the external ontology, the word–concept relationships are defined based on different corpora, therefore, the actual strength of the word–concept relationships in the collection may not be reflected by the pre-defined strength value in the ontology. A review of the literature regarding semantic-based topic models revealed that the existing semantic-based topic models [18, 34, 39, 40] do not always generate meaningful topics.

In this article, we introduce a semantic-based topic model called semantic-LDA that integrates the meanings of the words into the topic-generation process. A collection of documents, words in documents, and topics in the collection are the main building blocks in conventional topic modelling. We used an external ontology to conceptualize the meanings of words by matching them with concepts in the ontology. The Probase knowledge base was used [36] to conceptualize the words from the documents.

A human reader understands the content of a document through the context of the words in the document, so a single isolated word from a document may not be sufficient to reveal the meaning of this word in the document. This is because the exact meaning of a word in a document is related to the context of the word in the document. The meaning of the word depends on the context of the word [15, 32], where related words help to understand the context. For example, if the word "*bank*" comes with "*river*," "*water*," and "*boat*," then "*bank*" indicates a river bank or "the land along the edge of a river." But if the related words are "*loan*," "*account*," and "*customer*," then "*bank*" is related to a financial institution. Hence, in our proposed approach, we generate semantic patterns [14] to represent the topics in the documents. A semantic pattern contains a set of similar or relevant words that match common concepts, and semantic patterns can better represent the meaning of documents because they are derived from commonly shared concepts.

Essentially, in our approach, we propose a novel method to quantify the relationships between words and concepts based on the probabilistic distribution of words in the documents in the collection. Quantifying the word–concept relationships by considering the word distribution in the text collection reflects the actual relationships in the particular collection, which is much more accurate than using the pre-defined relationships from the external ontology.

A word, phrase, or sentence can be considered to be ambiguous if it is not defined with a clear meaning or if it has multiple potential meanings. Thus, the inability to interpret the meaning of a word using the concepts available in the ontology can be defined as a form of ambiguity. To handle the issue of word ambiguity, we propose a new method to interpret the meaning of the words that cannot be matched with any concept in the Probase ontology. We conceptualize those unmatched words with the concepts of related or similar words. Hence, in the topic modelling process, our semantic-based topic model (semantic-LDA) can interpret the meaning of each of the words in the collection using concepts.

The main contributions of this research include:

— A novel semantic-based topic modelling technique to capture the semantics of documents and generate a topic model to represent the main themes in a collection of documents.
— A new method to quantify the relationships between words and concepts based on a collection of documents instead of using pre-defined relationships from an external ontology.
— A new method to handle ambiguous words that cannot be matched with any concept in the ontology and incorporate them into the topic modelling process.

Section 2 explains the research efforts related to semantic-based topic modelling, and Section 3 presents the word conceptualization process. Section 4 introduces the semantic-based topic model. Experimental results, evaluation, and discussion are presented in Section 5. The conclusion is found in Section 6.

2 RELATED WORKS

Topic modelling techniques evolved from statistical to semantic-based approaches as a result of recognizing the importance of the meaning of the content rather than simply considering the frequency and co-occurrence of words. Semantic-based topic modelling approaches were introduced to capture and explain the meaning of words in the topic-generation process. How to interpret the meaning of words or text was the key question posed by the semantic topic models. Ontologies or knowledge taxonomies [23, 34, 38, 39], embeddings [18, 29, 40], knowledge graphs [41], and hybrid approaches have been incorporated in semantic-based topic-modelling approaches. Furthermore, semantic-based topic models have been developed for various applications such as opinion mining, sentiment analysis, and recommendation in social media [1, 26, 27, 33].

Ontologies or knowledge taxonomies [19] are employed to explain the meaning of words. Commonly used ontologies and lexical knowledge bases include Probase [36], Freebase [9], Yago [31], Wikipedia, DBpedia [3], Open Directory Project, BabelNet [25], and WordNet [22].

Some researchers have combined probabilistic models with ontology-based techniques to enhance the quality of their topic models. For example, Yao et al. [39] introduced a semantic-based topic model called Probase-LDA that combines LDA with the Probase ontology, with the Probase ontology being used to determine the semantics. Probase-LDA [39] derives the asymmetric Dirichlet priors based on Probase concepts rather than generating topics. The performance of Probase-LDA exceeded that of DF-LDA [2] and GK-LDA [11]. Subsequently, Yao et al. [38] introduced a probabilistic topic model by combining Wikipedia knowledge. Utilizing the Wikify service [23] to capture the textual content, they used a set of Wikipedia articles and link probabilities. The conceptualization topic model (CLDA) [34] is another semantic-based topic model that consolidates Probase with LDA. CLDA [34] is a four-layered word-based topic model that generates a concept layer based on the conditional probability of concepts, given words provided by Probase.

Word embeddings have also been employed in topic modelling to capture semantics. Zhang et al. [40] presented an embedding enhanced topic model that incorporated word and topic embeddings with LDA. They introduced an integrated framework that maps the topic–word structure

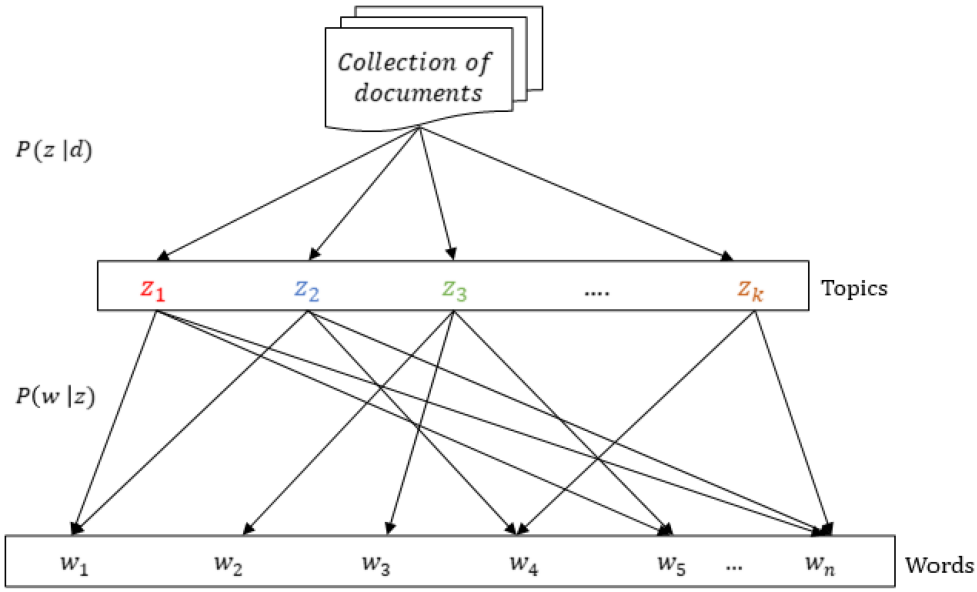

Fig. 1. Hierarchical structure of topic models.

using different word-embedding techniques such as GloVe, CBOW, and Word2Vec with Skip-Gram. In addition, Hong et al. [18] proposed two knowledge-embedded LDA versions called EK-LDA and EIK-LDA. These models use entity linking (word meanings are extracted using word links to Wikipedia [16]) to gain knowledge, and semantics are grasped via a pre-trained word vector.

Although capturing the semantics in the topic generation process is an important contribution, we identified the needs of effectively interpreting the semantics and incorporate the meanings of ambiguous words as research gaps. To address these research gaps, we introduced semantic-LDA+, and in our experiments, we evaluated Probase ontology-based semantic topic models [34, 39]. In the evaluation, we observed that the topic words generated by our approach are more focused and meaningful than topic words generated by most of the other recent semantic-based topic models [18, 40].

3 WORD CONCEPTUALIZATION

Traditional topic modelling approaches group words based on their co-occurrences without explicitly considering the meaning of the words, and each group represents a topic. Figure 1 shows the traditional topic modelling approach, which consists of three tiers: documents, topics, and words.

According to Figure 1, each document is represented by a probability distribution of a set of topics, and each topic is represented by a probability distribution of words, which is called a topic representation. But the words in a topic may be not related to each other in terms of meaning. They are grouped as a topic due to their co-occurrence in the documents. Simply, the words that often occur together are grouped to represent a topic without considering their semantics.

To improve the meaningfulness of topic representations, inspired by the work in Reference [34], in our semantic-LDA model, we include a concept tier between words and topics to conceptualize the meanings of the words before generating the topics. Figure 2 depicts the four-tier structure of semantic-LDA. In Figure 2, each topic is represented by a probability distribution of concepts, and each concept is represented as a probability distribution of words. Thus, the words are described

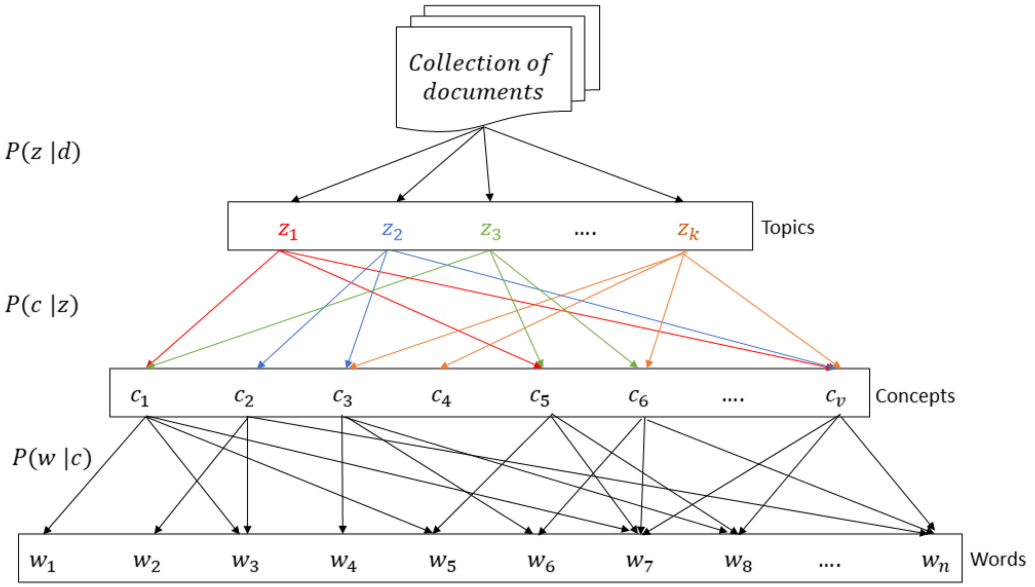

Fig. 2. Topic modelling process of concept-embedded topic modelling.

in terms of concepts, and concepts are then used to represent topics that cover the themes of the document collection. Hence, in our approach, the semantics of words are interpreted by the concepts. The word–concept relationships are generated and quantified in a separate step to that of generating the topic model. In contrast to the work in Reference [34], we generate and quantify the word–concept relationships from the documents in the document collection rather than using the relationships provided by the Probase ontology.

In this article, we quantified the word–concept relationships based on word–concept matchings. An individual word can have different meanings due to the polysemia of words. This means that a word can be matched with multiple concepts in the Probase ontology. Given a collection of documents covering different topics, it is, therefore, challenging to correctly match words and concepts.

Our solution to deal with this challenging problem is to first group semantically similar words to form what we refer to as semantic cliques [13, 14], then match the words in each clique with concepts from the Probase ontology. We assume that the cliques can be considered to be potential topics to represent the collection. Therefore, we can represent the collection in terms of semantic cliques, then interpret each clique by matching the words in the clique with Probase concepts.

Our experimental results demonstrated that the semantic cliques can represent the topics of a collection [14] (refer to Section 3.1). We further propose a method to quantify the word–concept relationships based on the probabilistic distribution of the word–concept matches in the collection. The process of generating semantic cliques, word–concept matching, and quantifying the relationships is called word conceptualization, as shown in Figure 3.

Figure 3 depicts semantic-LDA's topic modelling process. As shown in Figure 3, the document collection is the input and there are two main stages in the semantic topic modelling. Stage 1, word conceptualization, consists of three tasks, i.e., generating word-based semantic cliques from the collection of documents, generating semantic patterns for each semantic clique based on matching concepts, and quantifying the word–concept relationships from the collection based on the

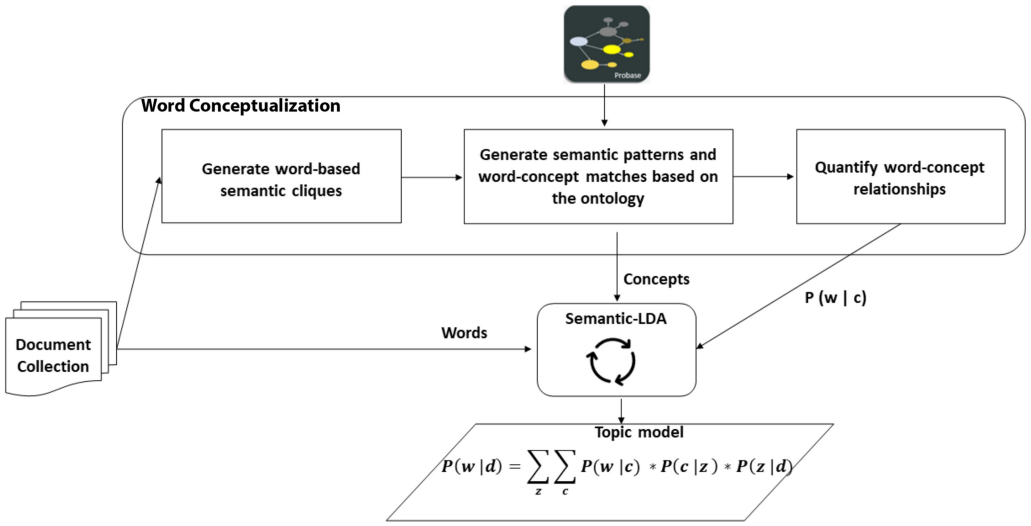

Fig. 3. Topic modelling process of semantic-LDA.

probabilistic distribution of the word–concept matches in the collection. The semantic topic model is then generated in stage 2, which consists of a set of topics described by the probability distributions of topics, concepts, and words.

3.1 Semantic Cliques

The foundation of our semantic-LDA model is semantic cliques, each of which consist of a set of semantically similar words from the collection. To generate the semantic cliques, we need to represent words in terms of their meaning. Recently, word embeddings have been popularly used to represent the meaning of words. In this article, the word embedding method GloVe [28] is used to generate an embedding for each word in the collection. Let D be the text collection and $W_D = \{w_1, w_2, w_3, \ldots, w_n\}$ be the set of unique words after pre-processing the content. In the pre-processing stage, we lemmatized the words and selected the words with a document frequency above 0.1. For each word $w_i \in W_D$, a vector $vec(w_i)$ is generated for word w_i. Then, semantic cliques are generated by clustering the words based on their embeddings. We applied soft clustering to generate semantic cliques where the same word can occur in multiple clusters based on context similarity. A semantic clique is a set of words that are semantically similar or related with each other. The words in each clique $g_i \in G$ are semantically similar because they are grouped based on word similarity. A semantic clique represents a potential topic in the collection.

3.2 Word–concept Relationships

We propose a method to generate semantic patterns based on word and concept matches and then quantify the relationships among the words and concepts.

3.2.1 Semantic Patterns. To interpret the meaning of each clique, we matched the words in each semantic clique with concepts in the Probase ontology [36]. If a word occurs in a concept, then we call it a match. Each word in a semantic clique can be matched with several concepts. Similarly, a concept can match more than one word within a semantic clique. A set of words that match a set of common concepts are grouped together to form a semantic pattern [14]. The meaning of a clique can be interpreted by the semantic patterns in the clique.

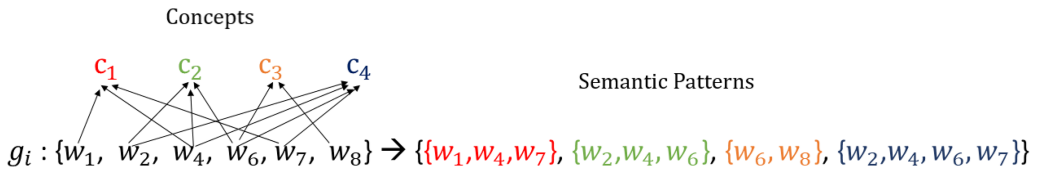

Fig. 4. Example of word–concept matching and semantic patterns.

Formally, let $G = \{g_1, g_2, \ldots, g_n\}$ be a set of cliques and C be a set of concepts in Probase, then for each word $w \in W_D$, a set of concepts can be retrieved from Probase that are related to w. For each related concept $c \in C$, Probase provides a conditional probability $P(c|w)$ to measure the relevancy between the word w and the concept c. For a word w and a concept c, if $P(c|w) > 0$, then c is considered to be a concept related to w. In other words, w is matched with c. Let $Concept(w)$ denote a set of related concepts of word w in semantic clique $g \in G$. $Concept$ is a mapping defined as follows:

$$Concept : W_D \rightarrow 2^C,\ \forall(w) \in W_D,\ Concept(w) = \{c \in C \mid P(c|w) > 0\}. \tag{1}$$

For each word $w \in W_D$, the top k related concepts are defined as $Concept^k(w)$, which contains the related concepts with the highest probability value $P(c|w)$. $Concept^k(w)$ can be defined as follows:

$$Concept^k(w) = argMax^k_{c \in Concept(w)} \{P(c|w)\}. \tag{2}$$

The related concepts of the words in a clique can be defined as follows:

$$Matched\_concepts(g_i) = \bigcup_{w \in g} Concept(w).$$

For Figure 4, $Matched\_concepts(g_i) = \{c_1, c_2, c_3, c_4\}$. The set of concepts in $Matched\_concepts(g_i)$ represents the meanings of words in clique g_i. Each concept in $Matched\_concepts(g_i)$ is related to a set of words in g_i. For example, in Figure 4, concept c_1 is related to w_1, w_4, and w_7, because c_1 is a related concept of these words, i.e., $w_1, w_4, w_7 \in Concept(c_1)$. The set of related words of a concept can be considered to be a pattern that represents the meaning of the concept, e.g., $\{w_1, w_4, w_7\}$ is a pattern that represents the meaning of c_1.

Definition (Semantic Patterns). For each matched concept $c \in Matched\_concept(g)$, where g is a clique, a semantic pattern sp in g, relating to the concept c, satisfies the following conditions:

(1) $sp(c, g) = \{w \in g | c \in Concept(w)\}$,
(2) $|sp(c, g)| > 1$.

For each matched concept $c \in Matched\_clique(g)$, $sp(c, g)$ is a maximum set of words in g, which are matched with this concept, and the number of matched words is larger than 1. We call this set of words a semantic pattern.

The above definition indicates that the words in a semantic pattern share some commonly matched concepts. The motivation for generating semantic patterns in a clique is to use the patterns to represent the matched concepts in this clique.

Figure 4 provides an example of the semantic pattern generation process.

In Figure 4, g_i is a semantic clique that contains a set of related words w_1, w_2, w_4, w_6, w_7, and w_8. The matching concepts from Probase for all of these related words are given below:

$Concept(w_1) = \{c_1\}$
$Concept(w_2) = \{c_2, c_4\}$
$Concept(w_4) = \{c_1, c_2, c_4\}$

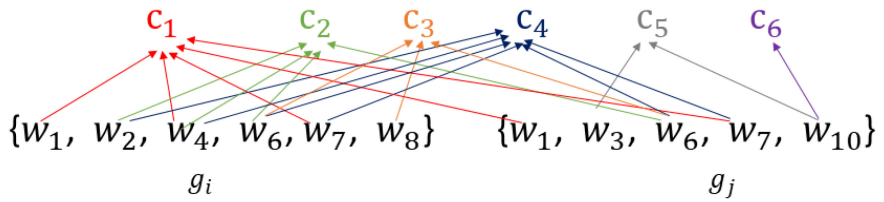

Fig. 5. Example of word–concept matching in two semantic cliques.

$Concept(w_6) = \{c_2, c_3, c_4\}$
$Concept(w_7) = \{c_1, c_4\}$
$Concept(w_8) = \{c_3\}$.

According to the common matching concepts, w_1, w_4, and w_7 match concept c_1; w_2, w_4, and w_6 match concept c_2 and c_4; w_6 and w_8 match concept c_3; and w_2, w_4, w_6, and w_7 match with concept c_4. The following semantic patterns are then generated based on the commonly matched concepts:

$\{w_1, w_4, w_7\} \Rightarrow c_1$
$\{w_2, w_4, w_6\} \Rightarrow c_2$
$\{w_6, w_8\} \Rightarrow c_3$
$\{w_2, w_4, w_6, w_7\} \Rightarrow c_4$.

Figure 5 depicts another example of how we match concepts with words in two semantic cliques.

Different cliques can share some common words and also contain different words. By matching the words to concepts, the generated semantic patterns and the matched concepts can differentiate between different cliques. Figure 5 shows two semantic cliques g_i and g_j. The words in g_i are the same as g_i in Figure 4. Clique g_j contains a few words (w_3 and w_{10}) that do not occur in g_i.

Word–concept matches in semantic cliques g_i and g_j are as follows:

$Concept(w_1) = \{c_1\}$
$Concept(w_2) = \{c_2, c_4\}$
$Concept(w_3) = \{c_5\}$
$Concept(w_4) = \{c_1, c_2, c_4\}$
$Concept(w_6) = \{c_2, c_3, c_4\}$
$Concept(w_7) = \{c_1, c_4\}$
$Concept(w_8) = \{c_3\}$
$Concept(w_{10}) = \{c_5, c_6\}$.

From these matches, we can generate semantic patterns for each clique. For clique g_i, we have the same semantic patterns as in Figure 4, which are $\{w_1, w_4, w_7\}$, $\{w_2, w_4, w_6\}$, $\{w_2, w_4, w_6, w_7\}$, and $\{w_6, w_8\}$, matched with concepts c_1, c_2, c_4, and c_3, respectively. For clique g_j, we generate three semantic patterns, which are $\{w_1, w_7\}$, $\{w_3, w_{10}\}$, and $\{w_6, w_7\}$, matched with concepts c_1, c_5, and c_4, respectively. The semantic patterns and the matched concepts indicate that the two cliques share some common meaning defined by concepts c_1 and c_4, and also contain some difference in meaning defined by concepts $\{c_2, c_3\}$ for clique g_i and concept c_5 for clique g_j.

3.2.2 Quantify the Relationships among Concepts and Words. The basic idea is to derive conditional probability $P(w|c)$ of word w given concept c based on frequency of the word–concept matching across all cliques.

According to Bayes theorem, we have the following equation, where w, c, g is a word, concept, and clique, respectively:

$$P(g|w,c) = \frac{P(w,c|g)*P(g)}{P(w,c)} = \frac{P(w|c,g)*P(c,g)}{P(w|c)*P(c)}.$$

From $\sum_{g\in G} P(g|w,c) = \frac{\sum_{g\in G} P(w|c,g)*P(c,g)}{P(w|c)*P(c)} = 1$, we have the following equation to calculate $P(w|c)$:

$$P(w|c) = \frac{\sum_{g\in G} P(w|c,g) * P(c,g)}{P(c)}. \tag{3}$$

Let $N_{c,g}$ denotes the number of concept–word matches between c and the words in g, $N_{C,g}$ denotes the total number of concept–word matches for the words in g, N_c denotes the number of matches between c and words in all cliques, N_C denotes the total number of concept–word matches for all cliques, and $N_{w,c,g}$ denotes the number of matches between w and c in g.

For the example in Figure 5, $N_{c_1,g_i} = 3$, where c_1 matches with w_1, w_4, and w_7 in g_i, $N_{C,g_i} = 11$, which contains all the word–concept matches in g_i (number of arrows going from g_i) exactly shows the number of word–concept matches in g_i. $N_{c_1} = 5$ where c_1 matches with w_1, w_4, w_7 in g_i and w_1, w_7 in g_j. We count all the concept–word matches in both cliques g_i and g_j, then $N_C = 20$ (number of arrows shows the number of word–concept matches where 11 matches in g_i and 9 matches in g_j). As an example for $N_{w,c,g}$, $N_{w_1,c_1,g_i} = 1$ where there is a single occurrence of a word matching with a concept within one clique.

We define the prior probability $P(c)$, the joint probabilities $P(w,c,g)$, and $P(c,g)$ as follows using the notations defined above:

$$P(c) = \frac{N_c}{N_C}$$

$$P(w,c,g) = \frac{N_{w,c,g}}{N_C}$$

$$P(c,g) = \frac{N_{c,g}}{N_C}.$$

Based on the probabilities defined above, we can define the following conditional probabilities:

$$P(w|c,g) = \frac{P(w,c,g)}{P(c,g)} = \frac{\frac{N_{w,c,g}}{N_C}}{\frac{N_{c,g}}{N_C}} = \frac{N_{w,c,g}}{N_{c,g}}.$$

Substitute $P(w|c,g)$, $P(c,g)$, and $P(c)$ in Equation (3), we have

$$P(w|c) = \frac{\sum_{g\in G} \frac{N_{w,c,g}}{N_{c,g}} * \frac{N_{c,g}}{N_C}}{\frac{N_c}{N_C}}.$$

Thus, the conditional probability $P(w|c)$ can be calculated as follows:

$$P(w|c) = \frac{\sum_g N_{w,c,g}}{N_c}. \tag{4}$$

A word w and a concept c match exactly one time in one clique, so, $N_{w,c,g}$ is defined as below:

$$N_{w,c,g} = \left\{ \begin{array}{ll} 1, & \text{if } w \in g \text{ and } c \in Concept(w). \\ 0, & \text{otherwise.} \end{array} \right\}$$

Since w, c matches only once in one g, $N_{c,g}$ is the number of words in g that match with c, i.e., $N_{c,g} = \sum_{w\in g} N_{w,c,g}$.

Table 1. Example of Word-concept Matching

Semantic Clique	Semantic Pattern	Matched concept
g_i	$\{w_1, w_4, w_7\}$	c_1
g_i	$\{w_2, w_4, w_6\}$	c_2, c_4
g_i	$\{w_6, w_8\}$	c_3
g_i	$\{w_2, w_4, w_6, w_7\}$	c_4
g_j	$\{w_1, w_7\}$	c_1
g_j	$\{w_6\}$	c_2, c_3, c_4
g_j	$\{w_6, w_7\}$	c_4
g_j	$\{w_3, w_{10}\}$	c_5
g_j	$\{w_{10}\}$	c_5, c_6

Therefore, $N_c = \sum_{g \in G} N_{w,c,g}$.
Finally, we can derive $P(w|c)$ as follows:

$$P(w|c) = \frac{\sum_{w \in g} N_{w,c,g}}{\sum_{g \in G} N_{c,g}}. \tag{5}$$

According to the Figure 5, Table 1 shows the semantic patterns and matched concepts in each semantic clique.

As defined above, N_C is the total number of concept–word matches, $P(c)$ is the percentage of matches involving c out of all the matches. Based on the word–concept matching in the two semantic cliques in Figure 5, the prior probability of each concept can be calculated as follows:

All the word–concept matches are counted to calculate N_C. Each concept–word match is counted, e.g., (w_1, c_1), (w_4, c_1), (w_6, c_1) in g_i are counted as 3 concept word matches. Similarly, after counting all the word–concept matches (in Figure 5, number of arrows indicates the word concept matches), $N_C = 21$.

$$P(c_1) = 5/21, \quad P(c_2) = 4/21, \quad P(c_3) = 3/21$$
$$P(c_4) = 6/21, \quad P(c_5) = 2/21, \quad P(c_6) = 1/21$$
$$P(c_1) + P(c_2) + P(c_3) + P(c_4) + P(c_5) + P(c_6) = 1.$$

As defined above, $N_{w,c,g} = 1$ if w matches c in clique g. $N_{c,g}$ counts the number of words that concept c matches in each clique g and $\sum_{g \in G} N_{c,g}$ is the total number of words matching with concept c over all the cliques. Finally, according to Equation (5); $P(w|c)$ can be calculated for each word w and concept c as follows:

$$P(w_1 \mid c_1) = 2/5, \quad P(w_2 \mid c_2) = 1/4, \quad P(w_2 \mid c_4) = 1/6$$
$$P(w_4 \mid c_1) = 1/5, \quad P(w_4 \mid c_2) = 1/4, \quad P(w_4 \mid c_4) = 1/6$$
$$P(w_6 \mid c_4) = 2/6, \quad P(w_6 \mid c_2) = 2/4, \quad P(w_8 \mid c_3) = 1/3$$
$$P(w_7 \mid c_1) = 2/5, \quad P(w_7 \mid c_4) = 2/6, \quad P(w_3 \mid c_5) = 1/2$$
$$P(w_{10} \mid c_5) = 1/2, \quad P(w_{10} \mid c_6) = 1/1, \quad P(w_6 \mid c_3) = 2/3.$$

3.3 Handle Unmatched Words

We matched the words in semantic cliques with the concepts in Probase and generated semantic patterns. It is unrealistic to find a knowledge base that contains all the words in the world. There can be missing words even in a sound knowledge base. Most of the existing semantic-based topic modelling approaches ignore the unmatched words in the topic modelling process or do not consider the word–concept relationship when the word is not available in the ontology. In our

research, we introduce a novel method based on word embeddings to handle unmatched words based on related and similar meaning words.

The similar semantic patterns of an unmatched word are identified based on the similarity of word embeddings of the unmatched word and the words in semantic patterns. For a given word w and a pattern p, the similarity between w and p is defined below:

$$sim(w,p) = \frac{1}{|p|}\sum_{t\in p} embedding\_sim(w,t). \tag{6}$$

$embedding\_sim(w,t)$ is the similarity between the embeddings of w and t, $sim(w,p)$ is the average similarity of w with the words in p.

For a given semantic clique $g_i \in G$, a set of patterns can be generated, $P_i = \{p_1, p_2, \ldots, p_m\}$, and the words in $mg_i = \bigcup_{p\in P_i} p$ can be matched with some concepts in Probase. mg_i is a set of matched words for clique g_i. However, if mg_i is a true subset of g_i, then there must be some words in g_i that are unmatched words. Let $ug_i = (g_i - mg_i)$ and $ug_i \neq \phi$, for each $w \in ug_i$, the top-k most similar patterns are calculated using the equation below:

$$simPattern(w) = argMax^{k}_{p\in P_i}\{sim(w,p) > l\}, \tag{7}$$

where l is a threshold. Since w is similar to the patterns in $simPattern(w)$, it is reasonable to consider that w can be semantically interpreted by the concepts of the patterns. That means it can share the same concepts of some words in the patterns. Therefore, we can update these patterns by adding the word w into the patterns, i.e., for each $p \in simPattern(w), p = p \bigcup \{w\}$.

If $simPattern(w) = \emptyset$, it means that no pattern is similar to w. This indicates that w is irrelevant to the semantic content of this collection. In this case, $P(w|c)$ is set to 0 for $c \in C$, meaning that w will never be chosen as a topic word.

As per Equation (7), we selected the most similar semantic patterns for each unmatched word, then add the unmatched word into those patterns. Algorithm 1 depicts the process of handling ambiguity for one semantic clique.

ALGORITHM 1: The process of handling ambiguity for unmatched words

Input: A set of semantic patterns $SP_i = \{sp_{i1}, \ldots, sp_{i|SP_i|}\}$ and a set of unmatched words UW_i in clique g_i.
Output: A set of modified semantic patterns generated for clique g_i.

1: **for** each unmatched word $uw \in UW_i$ **do**
2: $simPattern(uw) := argMax^{k}_{sp\in SP_i}\{sim(uw, sp) > l\}$
3: **if** $simPattern(uw) > 0$ **then**
4: **for** each pattern $p \in simPattern(uw)$ **do**
5: $SP_i := SP_i - \{p\}$
6: $p := p \bigcup \{uw\}$
7: $SP_i := \{p\} \bigcup SP_i$
8: **if** $simPattern(uw) = 0$ **then**
9: **for** each concept $c \in C$ **do**
10: $P(w|c) = 0$

A set of semantic patterns $SP_i = \{sp_{i1} \ldots, sp_{i|SP_i|}\}$ and the set of unmatched words UW_i of semantic clique $g_i \in G$ will be the input for Algorithm 1. First, we get each unmatched word $uw \in UW_i$ for the semantic clique (line 1). A set of semantic patterns with the highest similarity to the

unmatched word, $simPattern(uw)$, is selected (line 2). The unmatched word uw will subsequently be added to the most similar patterns (lines 4–6) resulting in modified semantic patterns.

After adding the unmatched words into the similar patterns, then we will derive the probability using the method in Section 3.2.2.

4 SEMANTIC-BASED LDA

In this section, we introduce our proposed Semantic-LDA that incorporates concepts into LDA to enhance the meaningfulness of the topic model. The traditional LDA topic model is a three-tier topic model with words, topics, and documents. Our semantic-LDA interprets the meaning of words with "concepts" by incorporating a concept tier in the topic model to build a four-tier topic model.

4.1 Generative Process of Semantic-LDA

Let D be the collection of document $D = \{d_1, d_2, \ldots, d_M\}$, which contains M documents, and each document consists of a set of words from a vocabulary W_D. Same as all probabilistic topic models, our Semantic-LDA is a generative probabilistic model of a document collection. The documents in the collection are represented as mixtures over a set of latent topics. Different from most of topic models where each topic is characterized as a distribution over words, in Semantic-LDA, each topic is characterized as a distribution over concepts, and each concept is characterized as a distribution over words. The concepts in Semantic-LDA are acquired from the external knowledge base Probase. Let $C = \{c_1, \ldots, c_V\}$ be the set of concepts obtained from Probase.

Similar to all topic models, each document is represented by a multinomial distribution over a set of K topics, $\theta_d = (\vartheta_{d,1}, \vartheta_{d,2}, \ldots, \vartheta_{d,K})$, $\vartheta_{d,j} = P(z_j|d)$, $\sum_{j=1}^{K} \vartheta_{d,j} = 1$, z_j denotes the jth topic. In Semantic-LDA, each topic is represented by a probability distribution over concepts. For the jth topic, we have $\Phi_j = (\varphi_{j,1}, \varphi_{j,2}, \ldots, \varphi_{j,V})$, V is the number of concepts, $\varphi_{j,i} = P(c_i|z_j)$ and $\sum_{i=1}^{V} \varphi_{j,i} = 1$. For each concept $c_i \in C$, c_i is represented by a probability distribution over words, $\Psi_i = (\Psi_{i,1}, \ldots, \Psi_{i,n})$, $\Psi_{i,l} = P(w_l|c_i)$, $\sum_{l=1}^{n} \Psi_{i,l} = 1$, and $P(w_l|c_i)$ was calculated using Equation (3) introduced in Section 3.2.2. α is Dirichlet prior parameter for document-topic distribution, Dirichlet distribution of concept prior is denoted by β.

4.1.1 Generative Process of Semantic-LDA without Ambiguity Handling. In this section, we introduce our Semantic-LDA model based on the words that match with the Probase concepts and do not consider the unmatched words. These unmatched words will be considered as atomic concepts, and $\Psi_{w,w} = P(w|w) = 1$ will be generated for an unmatched word w.

The generative process of the Semantic-LDA that does not handle the unmatched words is given in Algorithm 2, where the number of unmatched words is m. The graphical model of Semantic-LDA is shown in Figure 6.

Lines 1 and 2 in the generative process show that, the distribution of each concept including atom concepts is drawn from Dirichlet distribution for each topic k. Then, topic distribution for each document is drawn from Dirichlet distribution in lines 3 and 4. Lines 5 and 6 generate latent topic for each word and derive ϱ from Bernouli (line 7 in the generative process). If $\varrho = 1$, which indicates that a word can be matched by one or more concepts in Probase, then a concept c will be generated from $Mult(.|\Phi_z)$ (line 11). Then, a word w is selected based on the conditional probability $P(w|c)$ as given in line 12 in the generative process. If $\varrho = 0$, which indicates that no word can be matched by any concept in Probase, then a word will be generated from $Mult(.|\Phi_z)$, as given in line 9.

The graphical model of Semantic-LDA is given below.

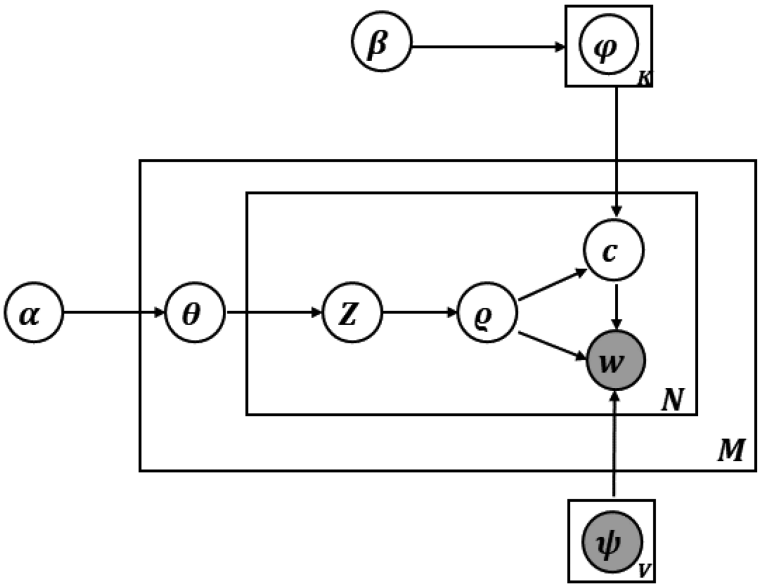

Fig. 6. Graphical representation of Semantic-LDA without ambiguity handling.

ALGORITHM 2: Generative process for Semantic-LDA

1: **for** each topic k where $k \in \{1, \ldots, K\}$ **do**
2: Generate $\Phi_k = (\phi_{k,1}, \ldots, \phi_{k,V}, \phi_{k,V+1}, \ldots, \phi_{k,V+m})^T \sim Dir(.|\beta)$
3: **for** each document $d \in D$ **do**
4: Generate $\theta_d = (\theta_{d,1}, \ldots, \theta_{d,K})^T \sim Dir(.|\alpha)$
5: **for** each i where $i = \{1, \ldots, N_d\}$ **do**
6: Generate $z_{i,d} \in \{1, \ldots, K\} \sim Mult(.|\theta_d)$
7: Derive $\varrho \sim$ from Bernoulli, if $\varrho = 1$, the word to be generated is matched with a concept in Probase, $\varrho = 0$ indicates the word to be generated is not matched with any concept.
8: **if** $\varrho = 0$ **then**
9: Generate $w_{d_i} \sim Mult(.|\Phi_z)$
10: **else**
11: Generate $c_i \in C \sim Mult(.|\Phi_z)$
12: Select a word w_{d_i} based on Ψ_i, i.e., $P(w_{d_i}|c_i)$

4.1.2 Generative Process of Semantic-LDA with Ambiguity Handling. We extended Semantic-LDA by handling unmatched words, as explained in Section 3.3. The extended model is denoted as Semantic-LDA+.

Accordingly, all the words can be matched with one or more concepts, and we directly derive the matching concepts and calculate the $P(w|c)$ probability based on our approach. The generative process of the Semantic-LDA+ is given in Algorithm 3.

In lines 1 and 2 in the generative process for Semantic-LDA+, concept distribution is drawn from Dirichlet distribution for each topic k. Then, topic distribution for each document is drawn from Dirichlet distribution in lines 3 and 4. Line 6 generates a latent topic for each word and line 7 generates a concept from $Mult(.|\Phi_z)$. Then, draw a word based on the conditional probability $P(w|c)$, as given in line 8 in the generative process.

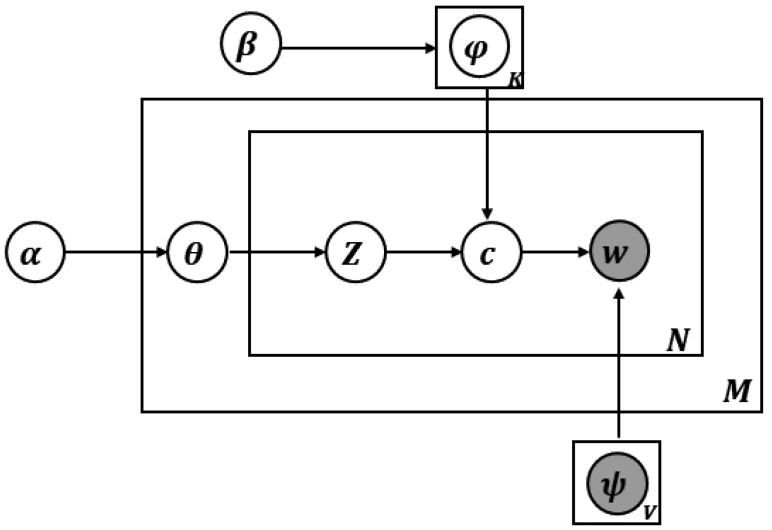

Fig. 7. Graphical representation of Semantic-LDA+.

ALGORITHM 3: Generative process for Semantic-LDA+

for each topic k where $k \in \{1, \ldots, K\}$ **do**
  Generate $\Phi_k = (\phi_{k,1}, ...\phi_{k,V})^T \sim Dir(.|\beta)$
for each document $d \in D$ **do**
  Generate $\theta_d = (\theta_{d,1}, \ldots, \theta_{d,K})^T \sim Dir(.|\alpha)$
  for each i where $i = \{1, \ldots, N_d\}$ **do**
    Generate $z_{i,d} \in \{1, \ldots, K\} \sim Mult(.|\theta_d)$
    Generate $c_i \in C \sim Mult(.|\Phi_z)$
    Select a word w_{d_i} based on Ψ_i, i.e., $P(w_{d_i}|c_i)$

The graphical model of Semantic-LDA+ is shown in Figure 7.

4.2 Inference Process and Estimate Parameters

Collapsed Gibbs sampling was used in the inference process. According to our approach, all the words are interpreted by at least one concept.

If the word is matching with a concept in Probase, then the sampling probability of a topic $z_{\_d,i}$ and a concept $c_{\_d,i}$ for the ith word in document d is calculated as below, where the topic assignment and the concept assignment to this word is k and l, relatively.

$$P(z_{\_d,i} = k, c_{\_d,i} = l|w, z_{\_d,i}, c_{\_d,i}; \alpha, \beta) \propto \frac{\beta + n^{c}_{.,k}}{\sum V * \beta + n^{(.)}_{.,k}} \cdot \frac{\alpha + n^{d}_{.,k}}{\sum K * \alpha + n^{(d)}_{.,.}} \cdot P(w|c) \tag{8}$$

$z_{\_d,i}, c_{\_d,i}$ denote the topic assignments of all other concepts and concept assignment of all other word tokens. V is the number of concepts. $n^{(.)}_{.,k}$ indicates the number of concepts in C for the kth topic without counting this concept assignment $z_{\_d,i}$. $n^{d}_{.,k}$ denotes the count of the kth topic assigned to some concepts in the dth document without counting the concepts for z. $n^{(d)}_{.,.}$ represents the summation of the distribution, and $n^{c}_{.,k}$ denotes the the number of times concept $c_{\_d,i}$ assigned

to the kth topic without counting this assignment $z_{\_d,i}$. $P(w|c)$ value is derived according to our novel concept–word probability calculation mechanism.

Equations (9) and (10) explain $\hat{\Phi}_{c,k}$ and $\hat{\theta}_k$ as below:

$$\hat{\Phi}_{c,k} = \frac{\beta + n^{c}_{.,k}}{\sum V * \beta + n^{(.)}_{.,k}}, \tag{9}$$

$$\hat{\theta}_{d,k} = \frac{\alpha + n^{d}_{.,k}}{\sum K * \alpha + n^{(d)}_{.,.}}. \tag{10}$$

With our ambiguation handling mechanism, we could interpret all the words with the related concepts. Unmatched words are interpreted with similar meaning and related words, which is similar to human understanding. In the generative process, lines 6–8 will be executed to generate topics, concepts, and words, since our ambiguation handling mechanism describes all the words with related concepts. Accordingly, Equation (8) is applied to calculate the probability of topic z and concept c for the i_th word in document d. Equations (9) and (10) can be used to estimate the parameters.

5 EVALUATION AND RESULTS

We evaluated the effectiveness of semantic-LDA from different perspectives with a set of standard evaluation techniques. There are three parts to the evaluation. In the first part, we conducted a human-based evaluation of the quality of the semantic cliques produced, as semantic cliques are the foundation of semantic-LDA. In the second part, we evaluated the quality of the topics generated by our semantic-LDA topic model in terms of the perplexity, coherence of the topics, and observations of the meaningfulness of the topic words. An information filtering-based evaluation was conducted in the third part where the topic words were considered to be features representing the user's multiple topic interests for filtering relevant documents. We also conducted the same experiments on the ambiguity handling method and compared the topic quality and information filtering results against methods that ignore unmatched words.

5.1 Human-based Evaluation

A human-based evaluation was conducted to assess the quality of our proposed semantic cliques, which are the foundation of our semantic-LDA model. We consider the semantic cliques to be word-based topic representations of the latent topics in the collection. The semantic cliques are generated based on word similarity derived from word embeddings that capture the meaning of the words. Human beings understand the content of text documents based on their understanding of the meaning of words in the text. Inspired by how humans understand content, we proposed the semantic-based approach to generate semantic cliques from a text collection, and furthermore, a semantic-based topic model. Hence, we wanted to gauge human opinions of the topics generated by several topic modelling techniques, including the semantic cliques generated by our model. An online survey was conducted to evaluate the semantic cliques from a human perspective.

In particular, we wanted to evaluate whether the topic words generated by our semantic cliques are related to each other and more meaningful to humans than the topic words generated by other topic modelling approaches.

We invited undergraduate students, postgraduate students, and participants with at least a diploma-level educational qualification who have English comprehension and analytical skills. News stories were extracted from the RCV1 dataset to create 30 collections, each containing two documents of news stories. Thirty sample comprehension questions were created for the human

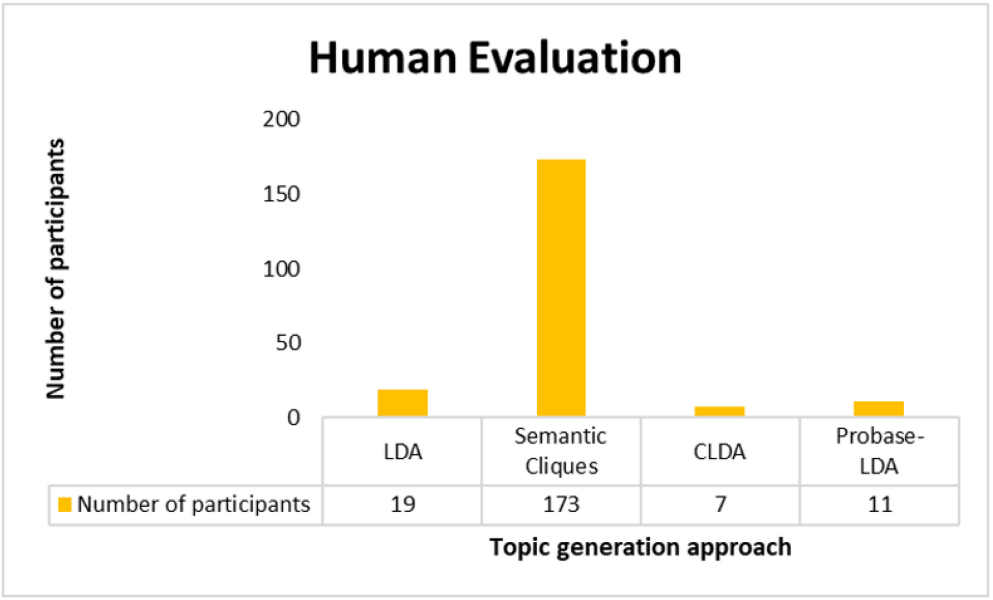

Fig. 8. Outcome of the human evaluation.

evaluation, with each question consisting of two stories. The participants were asked to carefully read the two stories and select the topic list that best describes the given stories from four prospective topic lists. Each participant was also asked to attempt to answer a randomly prompted comprehension question. Topic representations were generated using several topic modelling techniques including LDA [8], Probase-LDA [39], CLDA [34], and our semantic cliques. Five topics were generated, each with 10 topic words.

The participants were instructed to read the documents carefully and select the most meaningful and appropriate topic list for the given document collection. Classmarker.com online quiz maker[1] was used to create the online questionnaire. As shown in Figure 8, of the 210 participants who completed the questionnaire and contributed to the final results, 173 participants (82.4%) selected the topic words generated from our semantic cliques as the most meaningful and appropriate topic lists. Therefore, it is evident that the semantic cliques generated from our approach closely match human perception and denote a meaningful set of topics.

We randomly interviewed a few of the participants and queried them about the reasons for their choices. Most of the participants who selected the results of our approach stated they had made their choices due to the similarity and relevancy of words in each topic. However, 19 participants selected topics produced by the LDA model; most of these participants selected LDA due to finding the frequently occurring words in the documents also in the same topic. The 18 participants who selected the other two approaches indicated there was no specific reason for their selections and said they selected those approaches due to their perception. Hence, it is clearly proven that our semantic cliques are closer to human understanding than topic words generated by other approaches.

[1] https://www.classmarker.com/

5.2 Evaluation on Topic Quality

Several evaluation measures, such as perplexity, coherence, and word distribution in topic representations, have been widely used in some state-of-the-art systems [8, 34, 39] to evaluate the quality of generated topics.

The three datasets used in our experiments are **Reuters Corpus Volume I (RCV1)** [20], R8, and 20Newsgroup[2] datasets, all of which contain news stories. Semantic cliques can be considered to be word-based topic representations where each clique represents one topic. In this evaluation, we assessed the quality of the topics generated by the semantic cliques and our two proposed topic models, semantic-LDA and semantic-LDA+, by comparing them with the topic quality of topics generated by three word-based state-of-the-art systems (i.e., LDA, Probase-LDA, and CLDA).

Both the proposed models (i.e., semantic-LDA and semantic-LDA+) and the state-of-the-art systems (i.e., CLDA and Probase-LDA) are extensions of the LDA model. The evaluation parameters were set with the same values for all of the models, i.e., $\alpha = 1, \beta = 0.1$, and 2,000 iterations of the topic generation process were run for each model.

(1) Perplexity

Perplexity is a statistical-based evaluation mechanism widely used to measure the generalization of topic models. Perplexity calculates the alignment of new data (test dataset) with the learned data (training dataset) by calculating the log-likelihood of the test dataset. A higher likelihood is reflected by a lower perplexity value, and approaches that generate lower perplexity scores are considered to be good quality topic models.

The evaluation was conducted by varying the number of topics, and the perplexity score was calculated for each combination. The perplexity score can be calculated as follows:

$$Perplexity(D_{test}) = exp\left\{-\frac{\sum_{d=1}^{M} log\, P(W_d)}{\sum_{d=1}^{M} N_d}\right\}, \tag{11}$$

where M denotes the number of documents in the test dataset, and $P(W_d)$ and N_d denote the probability of document d and size of d, respectively.

Figure 9 shows the perplexity score by number of topics for the semantic cliques, semantic-LDA and semantic-LDA+ approaches, and state-of-the-art systems for the RCV1, R8, and 20Newsgroups datasets. Semantic-LDA+ and semantic-LDA show the lowest perplexity scores for all three datasets for all topic number variations, which indicates topic models of high quality.

(2) Coherence

Topic coherence evaluates the topics by looking at how often topic words in each topic occur together in documents. Hence, coherence indicates the semantics of generated topics. Coherence measures the closeness between the topic words within a topic. The quality of the generated topics can be measured by evaluating the closeness of the topic words. As semantic-LDA is a semantic-based topic model, coherence is an important measure to evaluate the quality of the semantics of the topic words. The coherence is calculated using the equation proposed in Reference [24], as shown in Equation (12).

$$Coherence(t) = \sum_{m=2}^{M_t} \sum_{l=1}^{m-1} log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})}, \tag{12}$$

[2] http://qwone.com/~jason/20Newsgroups/

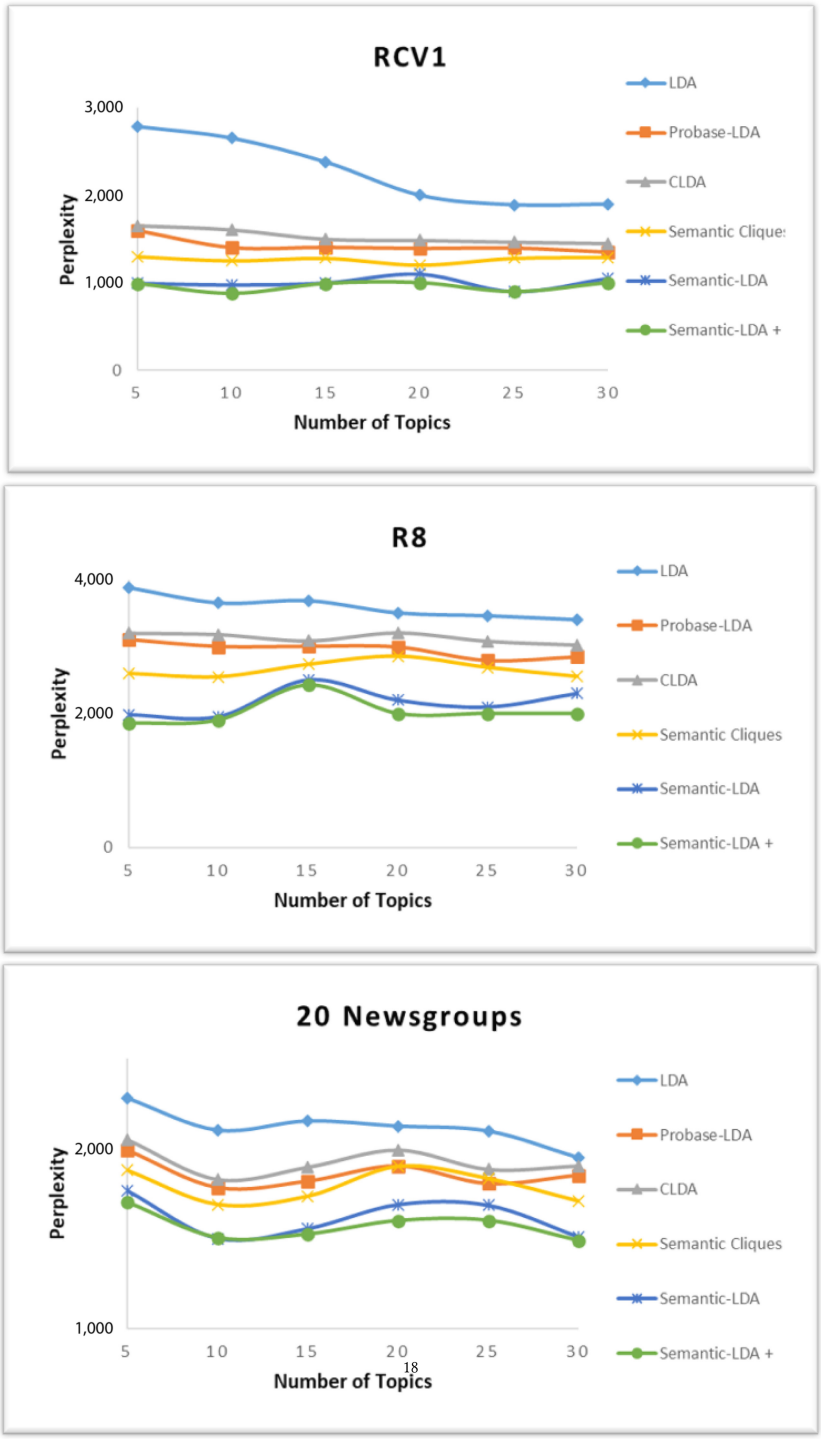

Fig. 9. Perplexity score by number of topics.

where $v_m^{(t)}$ and $v_l^{(t)}$ are the mth and lth words within topic t, M_t is the number of topic words in topic t, $D(v_m^{(t)}, v_l^{(t)})$ is the co-document frequency of the two words, and $D(v_l^{(t)})$ is the document frequency of the word $v_l^{(t)}$.

We evaluated the coherence of the topic models for the three datasets. A higher coherence score indicates a better quality topic model with highly relevant topic words in each topic. Figure 10 shows the coherence scores for different numbers of topics for our semantic-based topic models and the state-of-the-art systems for RCV1, R8, and 20 Newsgroups datasets. As shown in Figure 10, Semantic-LDA+ shows the highest and semantic-LDA shows the second-highest coherence score for all three datasets for all topic number variations. A higher coherence score indicates high-quality topics and implies semantic similarity among the topic words.

(3) Meaningfulness of topic words

An evaluation was performed of the quality of the topic words generated by the semantic-based topic models, i.e., semantic-LDA+, Probase-LDA, and CLDA. For this evaluation, five topics were generated by each topic model using collection 114 in the RCV1 dataset. Table 2 shows the topic distribution $P(z)$ and topic words in each topic (z). Topics are ordered in decreasing order of topic distribution.

For semantic-LDA+, the words in T_1 are related to gas emissions and energy consumption, T_2 has words related to climate change, T_3 describes global pandemics and recycling, words about cows are found in T_4, and finally, T_5 contains words related to the global economy and countries. The topic words generated by the other two approaches are not very focused compared to our approach. As shown in Table 2, the Probase-LDA topic words are scattered and it is difficult to group the topic words into categories or domains compared to topic words generated by our approach. CLDA topic words are better than those of Probase-LDA but some topic words are not focused on a specific topic. For example, T_5 in CLDA has words such as *thursday, past, clear,* which are not focused on the last topic. Some topic words from the CLDA and Probase-LDA models are apparently meaningless or irrelevant, such as *clear*, *thursday*, *exist*, *level*, and so on. In contrast, the topic words generated by our topic model appear to be more meaningful and related to each other.

In addition, we compared the quality of topics generated by the semantic-LDA+ model with topics from two other semantic-based topic models, i.e., embedding-based topic model EETM, which uses three different word embedding methods (i.e., CBOW, Skip, GloVe) [40], and knowledge-based topic models EK-LDA and EIK-LDA [18]. A separate evaluation was conducted for each of the methods, and they were evaluated using the results provided in each paper [18, 40], respectively. We compared the topic words provided in each paper [18, 40] with the topic words generated by our semantic topic model (for comparison, we used the same datasets specified in each paper [18, 40]).

We generated topics from the 20Newsgroups and Reuters datasets and mapped them with the topics generated by the EETM model. Table 3 and Table 4 present four topics generated by semantic-LDA+ and EETM with the three different embedding techniques using the 20NG and Reuters datasets. For each topic, a set of topic words generated for the topic is listed.

Zhang et al. [40] mainly categorized the 20Newsgroups dataset into the following list of topics: *technology*, *education*, *politics*, and *medicine*. The topics generated by semantic-LDA+ are more focused on *technology*, *politics*, and *medicine* domains. (Topic 1 is related to *politics*, Topic 2 is *graphics*, Topic 3 is *security and cryptography*, overall Topics 2 and 3 are related to *technology*, and Topic 4 is related to *medicine* in 20NG.) For example, as shown in Table 3, *image*, *jpeg*, *gif*, *graphic*, and *color* in Topic 2 are strongly related to *graphics* and can be categorized under *technology*. There are no meaningless or irrelevant words such as *going* (in EETM (Skip)), *will* (in EETM (GloVe)) among the top topic words. Furthermore, the 20Newsgroups dataset contains a set of emails and

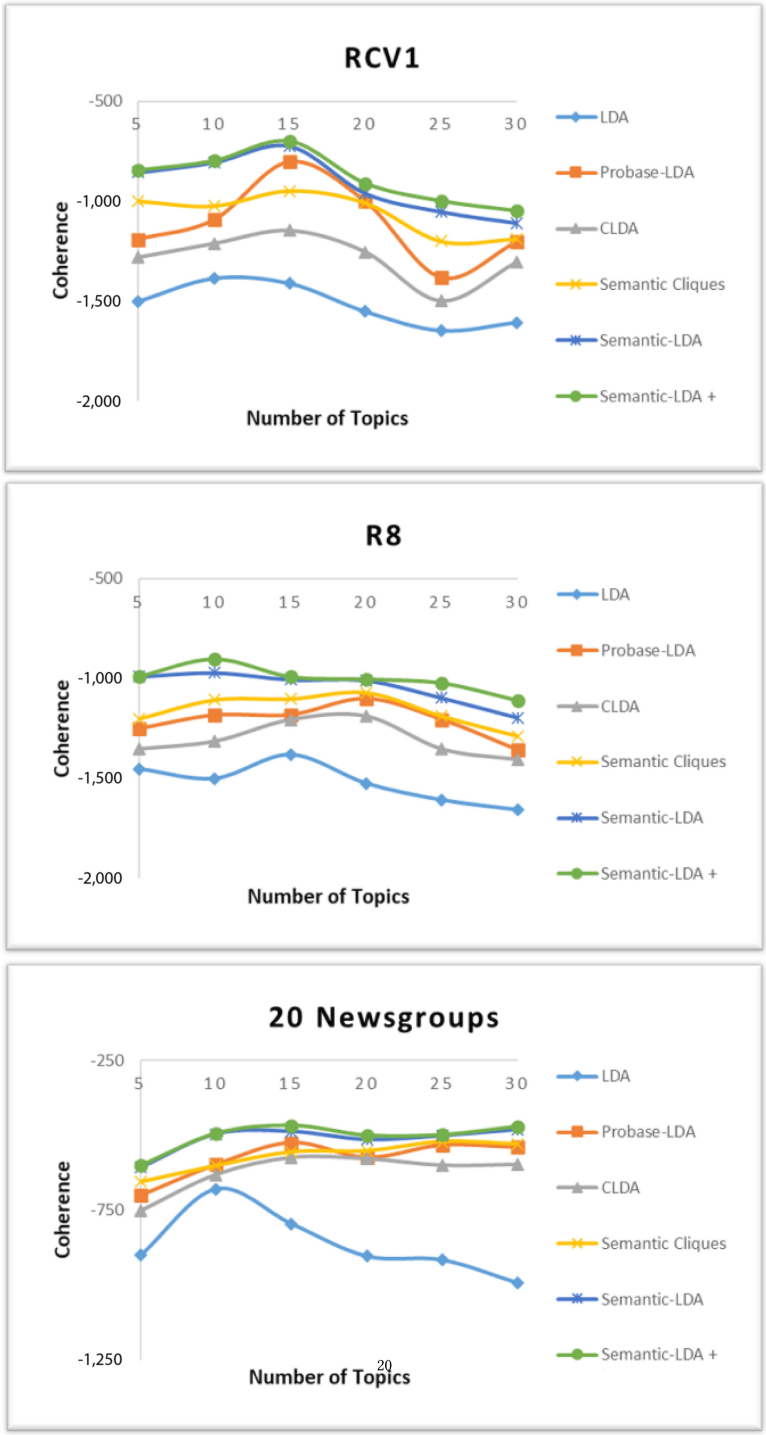

Fig. 10. Coherence score by number of topics.

Table 2. Topic Distribution and Topic Words Generated by the Probase-LDA, CLDA, and Semantic-LDA+ Models

	Semantic-LDA+		CLDA		Probase-LDA	
Topic z	Topic P(z)	Topic words	Topic P(z)	Topic words	Topic P(z)	Topic words
T_1	0.476	*coal*, *petroleum*, *fossil*, *consumption*, *energy*, *gas*, *dioxide*, *technology*, *emission*, *natural*	0.473	*coal*, *country*, *emission*, *year*, *world*, *carbon*, *fossil*, *nation*, *technology*, *asia*	0.438	*atmosphere*, *gas*, *university*, *current*, *burning*, *average*, *net*, *clear*, *indication*, *result*
T_2	0.192	*methane*, *carbon*, *snow*, *battle*, *geochemist*, *air*, *region*, *dioxide*, *cycle*, *plant*	0.258	*dioxide*, *plant*, *carbon*, *air*, *considerable*, *gas*, *snow*, *battle*, *balance*, *rate*	0.170	*scientist*, *country*, *world*, *measurement*, *forecast*, *insurance*, *butterflies*, *reduce*, *level*, *stabilisation*
T_3	0.130	*dioxide*, *carbon*, *fuel*, *develop*, *nation*, *curb*, *emission*, *epidemic*, *growth*, *recycle*	0.113	*cow*, *mad*, *british*, *herd*, *uk*, *disease*, *cut*, *epidemic*, *warming*, *cull*	0.139	*carbon*, *warmer*, *emission*, *fuel*, *effect*, *dying*, *crop*, *epidemic*, *equipment*, *land*
T_4	0.103	*species*, *flee*, *herd*, *cow*, *extinct*, *ecologist*, *disease*, *warming*, *mad*, *cull*	0.079	*group*, *enforce*, *warming*, *export*, *temperature*, *uk*, *equipment*, *trade*, *great*, *domestic*	0.130	*fossil*, *increase*, *flee*, *population*, *making*, *provoke*, *cut*, *euphydryas*, *exist*, *mexico*
T_5	0.099	*grow*, *economic*, *united*, *british*, *studies*, *china*, *uk*, *population*, *scientific*, *trade*	0.077	*population*, *agency*, *clear*, *warming*, *extinct*, *climate*, *thursday*, *past*, *find*, *record*	0.123	*warming*, *report*, *scientific*, *fear*, *influence*, *cooler*, *parmesan*, *insect*, *reduction*, *produce*

in EETM the word *edu* is extracted as a topic word for the second topic for each of the three EETM methods. However, the word *edu* may not be an important topic word. The topics generated by semantic-LDA+ correspond with the *politics*, *graphics*, *crypto*, and *medical* collections. There is no specific collection for *education* in the dataset, and EETM mainly categorized the words extracted from email domains (*edu*, *nntp*, *host*, etc.) as the *education* category, as shown in the second topic. Similarly, for the Reuters dataset, as shown in Table 4, Zhang et al. [40] generated topics that cover *economics*, *trade*, and *financial policy*. Topics generated by semantic-LDA+ also cover the topic ECAT (Economics) [20], and topics are more focused on sub areas of the economy such as *oil*, *export materials*, *stock market*, and *financial institutions*, which can be easily assigned to topics.

Then, we evaluated the topic quality of semantic-LDA+ compared with EK-LDA and EIK-LDA [18]. In their paper, the authors emphasized the topics related to guns and crimes generated from the 20Newsgroup dataset by the EK-LDA and EIK-LDA models. Table 5 shows the topic words generated from semantic-LDA+, EK-LDA, and EIK-LDA related to topics "guns and crimes."

Table 3. Topic Words Generated by the EETM and Semantic-LDA+ Models from the 20NG Dataset

Semantic-LDA+				EETM (CBOW)			
Topic 1	Topic 2	Topic 3	Topic 4	Topic 1	Topic 2	Topic 3	Topic 4
president	image	key	health	president	graphics	key	medical
stephanopoulos	jpeg	clipper	cancer	clinton	edu	clipper	harvard
myers	gif	pgp	medical	stephanopoulos	3d	chip	medicine
political	graphic	encryption	disease	myers	ray	encryption	health
government	color	chip	hiv	george	bbs	keys	treatment
EETM (Skip)				EETM (GloVe)			
Topic 1	Topic 2	Topic 3	Topic 4	Topic 1	Topic 2	Topic 3	Topic 4
president	graphics	key	medical	president	graphics	key	medical
clinton	edu	des	pitt	will	edu	number	health
stephanopoulos	3d	pgp	disease	clinton	ray	bit	disease
going	ray	public	gordon	stephanopoulos	3d	chip	cancer
myers	pub	bit	cancer	press	art	bits	patient

Table 4. Topic Words Generated by the EETM and Semantic-LDA+ Models from the Reuters Dataset

Semantic-LDA+				EETM (CBOW)			
Topic 1	Topic 2	Topic 3	Topic 4	Topic 1	Topic 2	Topic 3	Topic 4
oil	wheat	stock	bank	oil	foreign	coffee	dollar
gas	coffee	market	interest	gas	government	export	west
gasoline	export	trade	rate	tax	exchange	brazil	exchange
crude	sugar	share	economic	pct	debt	quotas	rates
fuel	grain	investor	bond	production	banks	meeting	policy
EETM (Skip)				EETM (GloVe)			
Topic 1	Topic 2	Topic 3	Topic 4	Topic 1	Topic 2	Topic 3	Topic 4
oil	bank	coffee	west	oil	bank	coffee	exchange
gas	banks	export	exchange	crude	banks	export	west
crude	loan	quotas	dollar	gas	loans	quotas	dollar
texas	loans	brazil	paris	barrels	interest	ico	baker
barrel	interest	ico	baker	ecuador	credit	brazil	paris

Even though all three approaches generated topic words related to guns and crimes, the topic words generated by semantic-LDA+ are more focused on weapons and there are no general words such as *people*, *carry*, or *control* as found in the topic words generated by EK-LDA and EIK-LDA.

Additionally, we observed the significance of the ambiguity handling approach, Semantic-LDA+ for collections contain news about crimes, politics, disasters, and accidents where many words are not conceptualized by Probase. Table 5 shows the topic-words that Semantic-LDA and Semantic-LDA+ generated for the RCV1-102 collection, which contains news related to crimes for three topics (similar topics generated by the two approaches were compared).

As shown in Table 6, Semantic-LDA+ generated topics with crime-related words that were not present in Probase, such as *kill, abuse, porn, pornography, crime, criminal* in the topic-word lists. The topics generated with those words are meaningful and more related to the context.

Table 5. Topic Words Generated for a Topic by the EK-LDA, EIK-LDA, and Semantic-LDA+ Models

Semantic-LDA+	EK-LDA	EIK-LDA
gun	guns	guns
firearm	police	crime
weapon	carry	police
crime	killed	kill
death	crime	killed
pistol	weapon	weapon
explosive	death	death
military	people	control
handgun	firearms	deaths
shotgun	control	firearms

Table 6. Topic Words Generated for a Topic from Semantic-LDA+ and Semantic-LDA

Topic	Semantic-LDA+	Semantic-LDA
T1	dutroux, nihoul, paedophile, abuse, kidnap, scandal, murdering, kill, crime, rapist	scandal, kidnap, arrest, teenage, offender, murdering, bury, body, death, victim
T2	police, cop, interpol, crime, investigation detective, officer, criminal, arrest, coverup	police, arrest, interpol, investigator, official, case, detective, government, coverup, search
T3	girl, child, sexually, women, porn, seized, abduction, kill, pornography, prostitution	teenage, young, girl, child, death, victim, abduction, body, aged, scandal

5.3 Evaluation of Information Filtering

The semantic-LDA's topic words are generated from a user's document collection and can be used to represent the user's topic interest for an information filtering system. The topic words generated from the semantic-LDA model are used as features to represent the information needs of the user. The relevance of a new document d is evaluated by matching words in the new document with the topic words in the topic model and is calculated based on the topic distribution and concept–word distribution as defined below.

$$Score(d) = \sum_{w \in d} \sum_{z \in Z} \sum_{c \in C} P(w|c) * P(c|z) * P(z|d) \tag{13}$$

In our semantic topic model w, c, and z indicate word, concept, and topic in the user interest model, respectively. $P(z|d)$ denotes the topic distribution of the semantic topic model. $P(c|z)$ denotes the concept distribution for a topic in the semantic topic model. $P(c|z)$ and $P(z|d)$ are derived using Equation (9) and Equation (10), respectively, during the topic model inference process. We derive $P(w|c)$ using our method proposed in Section 3.2.2. The information filtering-based evaluation was conducted using the RCV1 dataset. Table 7 shows the information filtering results compared with other state-of-the-art systems.

Top-20, precision, recall, F1-measure, and **Mean Average Precision (MAP)** values were evaluated in the experiment. Top-20 was calculated based on the relevance score calculation of the highest-ranking documents. The average precision score of the highest 20 documents was considered as Top-20. F1-measure was calculated based on the precision and recall where $F1 = \frac{2*precision*recall}{precision+recall}$. The mean value of the average precision for each collection was calculated as MAP. Top20 and F1 $Improvement\%$ indicate the improvement of semantic-LDA+ compared to the

Table 7. Result Comparison of All Models for the Information Filtering Semantic-topic Model

Approach	Top 20	F1	MAP
Semantic-LDA+	**0.472**	**0.468**	**0.470**
Semantic-LDA	**0.463**	**0.445**	**0.457**
Semantic cliques	**0.460**	**0.434**	**0.449**
Probase-LDA [39]	0.413	0.327	0.429
CLDA [34]	0.380	0.376	0.401
LDA-words	0.458	0.426	0.421
Improvement baselines %	1.52%	5.87%	7.5%

LDA-words and MAP *Improvement%* the improvement of semantic-LDA+ compared to Probase-LDA. LDA-words showed the highest F1-score, while Probase-LDA showed the highest MAP score.

According to the results in Table 7, it is clearly evident that our models outperformed other approaches in the information filtering evaluation. Among the three proposed approaches, the semantic-LDA+ approach (which is an enhancement of the semantic-LDA approach) scored the highest results.

5.4 Discussion

- Semantic topic model

We conducted several experiments to evaluate the generated topics. Our semantic-LDA+ approach has the lowest perplexity score for all three datasets, followed by semantic-LDA. Semantic cliques has the third-lowest perplexity scores for most occurrences. In contrast, LDA has the highest perplexity score. Semantics have played a significant role in the topic quality, and the semantic-based topic models (Probase-LDA and CLDA) have comparatively low perplexity scores compared to the LDA model.

Semantic-LDA+ has the highest coherence score, which demonstrates a high level of closeness between the topic words within each topic. Semantic-LDA has the second-highest coherence score with only a slight difference to that of semantic-LDA+. For fewer topic words (e.g., 10–15), the performance of semantic-LDA and semantic-LDA+ is similar due to fewer ambiguous words being drawn into the topic range. But for the RCV1 and R8 datasets, semantic-LDA+ has been improved for topic range 20–30, as more ambiguous words can be meaningfully handled by semantic-LDA+.

Semantic cliques and Probase-LDA have shown the next-highest coherence score in different topic number variations. However, all of the semantic-based topic models have a similar coherence score, and it is clearly evident that all of the semantic-based approaches have a higher coherence score than that of the LDA approach. Hence, there is a strong correlation between semantics and the coherence of topics.

Semantic-LDA+ has produced a set of words in each of its topics that are more closely related and more meaningful compared to the other two semantic-based topic models, CLDA and Probase-LDA. Considering that the semantic cliques and semantic patterns are the foundation of the proposed semantic-LDA models, we believe that generating the word–concept relationships from the collection itself might contribute to generating a set of meaningful topic words. In the comparison to recent semantic-based topic models [18, 40], it is clearly evident that the topic words generated by semantic-LDA+ are more focused and there are no meaningless or irrelevant topic words generated, while meaningless words do occur in the topic word lists generated by semantic topics models in References [18, 40].

Semantic-LDA+ outperformed the other approaches in the information filtering evaluation. The second-highest Top20 and F1-measure values were generated by LDA and the second-highest MAP value by Probase-LDA. Since LDA has frequently occurring words in the topic lists, it might be a reason for performing well in information filtering evaluation.

- Complexity of the semantic topic model

The semantic-LDA topic model consists of two stages: concept–word relationship calculation based on semantic patterns, and topic modelling. For the topic modelling stage, we have applied the LDA, as the inference process of LDA applies the Gibb sampling. There is a linear complexity for each iteration of Gibbs sampling of LDA with the number of topics (k) and the number of documents (M) as $O(k * M)$ [35].

Then, the concept–word relationship is calculated based on the semantic-pattern generation process. The complexity of semantic pattern generation is calculated based on the number of cliques and the number of words in each clique. The number of cliques is similar to the number of topics in the topic representation. Hence, the number of cliques can be considered to be k and the number of words M with the assumption that the number of words in each clique is not greater than the number of documents. In the semantic pattern generation process, a set of matching concepts are retrieved from Probase for each word in the clique. Let N be the number of concepts in Probase. The complexity can be derived as $O(k * M * logN)$.

Hence, the overall complexity of the semantic-LDA approach is $O(k * M * logN)$.

6 CONCLUSION

Understanding the semantics is essential in topic modelling, and we proposed a semantic-based topic model called semantic-LDA, which incorporates semantics into the topic modelling process. We have incorporated a two-phase semantic capturing mechanism (semantic cliques and semantic patterns) to conceptualize the meanings of words before applying the topic model. The human evaluation proved that our semantic cliques are closer to human understanding than topic lists produced by other approaches, and we applied semantic cliques as the foundation for semantic-LDA. Existing semantic-based approaches quantify the relationship between words and concepts based on a pre-calculated value. Our approach has introduced a novel mechanism to derive the word–concept relationships from the collection itself. We quantify the word–concept relationships based on the word–concept matchings in our collection. Some important words may be neglected by the prevailing approaches due to their unavailability in the knowledge bases. Therefore, we introduced an approach to handle ambiguity by interpreting the unmatched words into relevant words using word embeddings. Hence, all of the important words in the collection are conceptualized and meaningfully included in the final topic list. Our semantic-LDA model highlights the contributions of integrating semantics into topic modelling, quantifying the word–concept relationships based on the word–concept matchings in the collection, and handling ambiguity. A topic quality evaluation and an information filtering evaluation were conducted, and it is evident that our semantic-based approach is more effective than other state-of the-art systems.

REFERENCES

[1] Siamak Abdi, Jamshid Bagherzadeh, Gholamhossein Gholami, and Mir Saman Tajbakhsh. 2021. Using an auxiliary dataset to improve emotion estimation in users' opinions. *J. Intell. Inf. Syst.* 56, 3 (Apr. 2021), 581–603. DOI: https://doi.org/10.1007/s10844-021-00643-y

[2] David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09)*. ACM Press. DOI: https://doi.org/10.1145/1553374.1553378

[3] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A nucleus for a web of open data. In *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Conference on Asian Semantic Web Conference (ISWC'07/ASWC'07)*. Springer-Verlag, Berlin, 722–735.

[4] David M. Blei. 2012. Probabilistic topic models. *Commun. ACM* 55, 4 (Apr. 2012), 77. DOI: https://doi.org/10.1145/2133806.2133826

[5] David M. Blei. 2012. Topic modeling and digital humanities. *J. Digit. Human.* 2, 1 (2012).

[6] David M. Blei and John D. Lafferty. 2005. Correlated topic models. In *Proceedings of the 18th International Conference on Neural Information Processing Systems (NIPS'05)*. MIT Press, Cambridge, MA, 147–154. Retrieved from http://dl.acm.org/citation.cfm?id=2976248.2976267

[7] David M. Blei and John D. Lafferty. 2007. A correlated topic model of science. *Annals Appl. Stat.* 1, 1 (June 2007), 17–35. DOI: https://doi.org/10.1214/07-aoas114

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3 (Mar. 2003), 993–1022. Retrieved from http://dl.acm.org/citation.cfm?id=944919.944937

[9] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'08)*. ACM Press. DOI: https://doi.org/10.1145/1376616.1376746

[10] Chaitanya Chemudugunta, America Holloway, Padhraic Smyth, and Mark Steyvers. 2008. Modeling documents by combining semantic concepts with unsupervised statistical learning. In *Lecture Notes in Computer Science*, Vol. 5318, Springer Berlin, 229–244. DOI: https://doi.org/10.1007/978-3-540-88564-1_15

[11] Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013. Discovering coherent topics using general knowledge. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM'13)*. ACM Press.

[12] Yang Gao, Yue Xu, and Yuefeng Li. 2015. Pattern-based topics for document modelling in information filtering. *IEEE Trans. Knowl. Data Eng.* 27, 6 (June 2015), 1629–1642. DOI: https://doi.org/10.1109/tkde.2014.2384497

[13] Dakshi Kapugama Geeganage, Yue Xu, and Yuefeng Li. 2019. Topic representation using semantic-based patterns. In *Proceedings of the 17th Australasian Conference: Data Mining (AusDM'19). Communications in Computer and Information Science*, Thuc D. Le, Kok-Leong Ong, Yanchang Zhao, Warren H. Jin, Sebastien Wong, Lin Liu, and Graham Williams (Eds.). Springer Singapore, 28–40. DOI: https://doi.org/10.1007/978-981-15-1699-3_3

[14] Dakshi T. Kapugama Geeganage, Yue Xu, and Yuefeng Li. 2021. Semantic-based topic representation using frequent semantic patterns. *Knowl.-based Syst.* 216 (Mar. 2021), 106808. DOI: https://doi.org/10.1016/j.knosys.2021.106808

[15] Silvia P. Gennari, Maryellen C. MacDonald, Bradley R. Postle, and Mark S. Seidenberg. 2007. Context-dependent interpretation of words: Evidence for interactive neural processes. *NeuroImage* 35, 3 (Apr. 2007), 1278–1286. DOI: https://doi.org/10.1016/j.neuroimage.2007.01.015

[16] Nitish Gupta, Sameer Singh, and Dan Roth. 2017. Entity linking via joint encoding of types, descriptions, and context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/d17-1284

[17] Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*. ACM Press. DOI: https://doi.org/10.1145/312624.312649

[18] Yang Hong, Xinhuai Tang, Tiancheng Tang, Yunlong Hu, and Jintai Tian. 2020. Enhancing topic models by incorporating explicit and implicit external knowledge. In *Proceedings of the 12th Asian Conference on Machine Learning (Proceedings of Machine Learning Research*, Vol. 129*)*, Sinno Jialin Pan and Masashi Sugiyama (Eds.). PMLR, Bangkok, Thailand, 353–368. Retrieved from http://proceedings.mlr.press/v129/hong20a.html

[19] Bayzid Ashik Hossain, Abdus Salam, and Rolf Schwitter. 2020. A survey on automatically constructed universal knowledge bases. *J. Inf. Sci.* (June 2020), 016555152092134. DOI: https://doi.org/10.1177/0165551520921342

[20] David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. RCV1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.* 5 (Dec. 2004), 361–397. Retrieved from http://dl.acm.org/citation.cfm?id=1005332.1005345

[21] Krys Kochut, Mehdi Allahyari, Seyedamin Pouriyeh, and Hamid R. Arabnia. 2015. OntoLDA: An ontology-based topic model for automatic topic LabelingOntoLDA: An ontology-based topic model for automatic topic labeling. In *Proceedings of the IEEE 14th International Conference on Machine Learning and Applications*.

[22] George A. Miller. 1995. WordNet: A lexical database for english. *Commun. ACM* 38, 11 (Nov. 1995), 39–41. DOI: https://doi.org/10.1145/219717.219748

[23] David Milne and Ian H. Witten. 2013. An open-source toolkit for mining Wikipedia. *Artif. Intell.* 194 (Jan. 2013), 222–239. DOI: https://doi.org/10.1016/j.artint.2012.06.007

[24] David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*. 262–272. Retrieved from http://dl.acm.org/citation.cfm?id=2145432.2145462

[25] Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. 216–225.
[26] Amjad Osmani and Jamshid Bagherzadeh Mohasefi. 2022. Weighted joint sentiment-topic model for sentiment analysis compared to ALGA: Adaptive lexicon learning using genetic algorithm. *Computat. Intell. Neurosci.* 2022 (July 2022), 1–35. DOI: https://doi.org/10.1155/2022/7612276
[27] Amjad Osmani and Jamshid Bagherzadeh Mohasefi. 2022. Opinion mining using enriched joint sentiment-topic model. *Int. J. Inf. Technol. Decis. Mak.* 22, 01 (Sept. 2022), 313–375. DOI: https://doi.org/10.1142/s0219622022500584
[28] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 1532–1543. Retrieved from http://www.aclweb.org/anthology/D14-1162
[29] Min Shi, Yufei Tang, Xingquan Zhu, Jianxun Liu, and Haibo He. 2019. Topical network embedding. *Data Min. Knowl. Discov.* 34, 1 (Oct. 2019), 75–100. DOI: https://doi.org/10.1007/s10618-019-00659-7
[30] Mark Steyvers and Tom Griffiths. 2007. Probabilistic topic models. *Handb. Latent Semant. Anal.* 427, 7 (2007), 424–440.
[31] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago. In *Proceedings of the 16th International Conference on World Wide Web (WWW'07)*. ACM Press. DOI: https://doi.org/10.1145/1242572.1242667
[32] Patrizia Tabossi. 1991. Understanding words in context. In *Advances in Psychology*. Elsevier, 1–22. DOI: https://doi.org/10.1016/s0166-4115(08)61527-7
[33] Mir Saman Tajbakhsh and Jamshid Bagherzadeh. 2019. Semantic knowledge LDA with topic vector for recommending hashtags: Twitter use case. *Intell. Data Anal.* 23, 3 (Apr. 2019), 609–622. DOI: https://doi.org/10.3233/ida-183998
[34] Yi-Kun Tang, Xian-Ling Mao, Heyan Huang, Xuewen Shi, and Guihua Wen. 2017. Conceptualization topic modeling. *Multim. Tools Applic.* 77, 3 (Sep. 2017), 3455–3471. DOI: https://doi.org/10.1007/s11042-017-5145-4
[35] Xing Wei and W. Bruce Croft. 2006. LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06)*. Association for Computing Machinery, New York, NY, 178–185. DOI: https://doi.org/10.1145/1148170.1148204
[36] Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q. Zhu. 2012. Probase. In *Proceedings of the International Conference on Management of Data (SIGMOD'12)*. ACM Press. DOI: https://doi.org/10.1145/2213836.2213891
[37] Yue Xu, Hanh Nguyen, and Yuefeng Li. 2020. A semantic based approach for topic evaluation in information filtering. *IEEE Access* 8 (2020), 66977–66988. DOI: https://doi.org/10.1109/ACCESS.2020.2985079
[38] Liang Yao, Yin Zhang, Baogang Wei, Lei Li, Fei Wu, Peng Zhang, and Yali Bian. 2016. Concept over time: The combination of probabilistic topic model with Wikipedia knowledge. *Expert Syst. Applic.* 60 (Oct. 2016), 27–38. DOI: https://doi.org/10.1016/j.eswa.2016.04.014
[39] Liang Yao, Yin Zhang, Baogang Wei, Hongze Qian, and Yibing Wang. 2015. Incorporating probabilistic knowledge into topic models. In *Advances in Knowledge Discovery and Data Mining*. Springer International Publishing, 586–597.
[40] Peng Zhang, Suge Wang, Deyu Li, Xiaoli Li, and Zhikang Xu. 2020. Combine topic modeling with semantic embedding: Embedding enhanced topic model. *IEEE Trans. Knowl. Data Eng.* 32, 12 (Dec. 2020), 2322–2335. DOI: https://doi.org/10.1109/tkde.2019.2922179
[41] Jia Zhu, Zetao Zheng, Min Yang, Gabriel Pui Cheong Fung, and Yong Tang. 2019. A semi-supervised model for knowledge graph embedding. *Data Min. Knowl. Discov.* 34, 1 (Sep. 2019), 1–20. DOI: https://doi.org/10.1007/s10618-019-00653-z

Received 9 June 2022; revised 16 January 2023; accepted 15 December 2023