Hindawi
Complexity
Volume 2024, Article ID 2775236, 24 pages
https://doi.org/10.1155/2024/2775236

WILEY | Hindawi

Research Article

Conceptualizing Discussions on the Dark Web: An Empirical Topic Modeling Approach

Randa Basheer [1] **and Bassel Alkhatib** [2]

[1]*Faculty of Information Technology and Communications, Syrian Virtual University, Damascus, Syria*
[2]*Faculty of Informatics Engineering, Damascus University, Damascus, Syria*

Correspondence should be addressed to Randa Basheer; randa_151689@svuonline.org

Received 13 October 2023; Revised 5 January 2024; Accepted 27 January 2024; Published 14 February 2024

Academic Editor: Hiroki Sayama

Social networks on the Internet have become a home that attracts all types of human thinking to exchange knowledge and ideas and share businesses. On the other hand, it has also become a source for researchers to analyze this knowledge and frame it in patterns that define types of thoughts circulating on these networks and representing the communities around them. In particular, some social networks on the Dark Web attract a special kind of thinking centered around the malicious and illegal activities disseminated on websites and marketplaces on the Dark Web. These networks involve discussions to exchange and discourse information, tips, and advice on performing such business. Studying social networks on the Dark Web is still in its infancy. In this paper, we present a methodology for analyzing the content of social networks on the Dark Web using topic modeling methods. We demonstrate the needed stages for the topic modeling process, beginning with data preprocessing and feature extraction to topic modeling algorithms. We utilize and discuss the following four topic models: LDA, CTM, PAM, and PTM. We discuss the following four topic coherence measures as evaluation metrics: UMass, UCI, CNPMI, and CV, demonstrating the selection of the best number of topics for each model according to the most coherent produced topics. Furthermore, we discuss the limitations, challenges, and future work. Our proposed approach highlights the ability to discover the latent thematic patterns in conversations and messages in the common language used in social networks on the Dark Web, constructing topics as groups of terms and their associations. This paper provides researchers with a leading methodology for analyzing thought patterns on the Dark Web.

1. Introduction

In crime analysis, studying content related to malicious and criminal activities forms a significant aspect of understanding crime and its motivations. The malpractices of technological and communication development have led to the dissemination of the malignant content and encouraged conducting of illicit businesses on various websites and social networks. This content is particularly abundant in the dark part of the Internet, or the Dark Web, which prompts researchers to analyze and study published posts and discussions on platforms on the Dark Web to enrich their knowledge about crime and the patterns of thoughts related to it.

The Dark Web is a part of the web that includes websites only accessed via encryption software, such as the onion router (TOR), and hosted on these encrypted networks. These sites cannot be reached or indexed by search engines [1, 2]. Such software provides users anonymity and traffic encryption, encouraging illicit and criminal activities to take place widely on the Dark Web without the fear of revealing identities or geographic locations and increasing their ability to avoid detection and arrest [3]. Such malicious activities include drug trade, stolen and counterfeit credit cards, fraud, terrorism and extremism, propaganda dissemination, hacking tools and tutorials, child pornography, weapon trade, and many others [1, 3, 4]. It is estimated that activities of such types form about 80% of the Dark Web [5].

Moreover, the Dark Web thrives through the rapid digital development in communication and social networking, leading to an emerged digital culture of illegal

activities on the Dark Web that is imposed on individuals to join Dark Web communities, such as the marketplaces on the Dark Web, or cryptomarkets. Such cultures are built and strengthened through social interactions, as they are crucial to ensure the operability and sustainability of cryptomarkets [6]. Therefore, some social networks on the Dark Web are essential platforms for participating in cryptomarkets, forming cybercrime communities. These social networks are mostly the forums associated with cryptomarkets, which are organized according to specific topics in which common interests and goals are shared and illicit products are presented and promoted, in addition to exchanging information, experiences, advice, and negotiations [6–8].

For these reasons, the Dark Web, in general, has gained a prominent interest from researchers. In particular, cryptomarkets and their associated forums make substantial sources for law enforcement and cybersecurity agencies to investigate and detect cybercrimes [2].

It is worth noting that the Surface Web includes communities related to criminal activities on the Dark Web that support the continuity of cryptomarkets and achieve flexibility for their systems [9]. Although the policies of platforms on the Surface Web prohibit communities related to illicit activities on the Dark Web, such communities still emerge and thrive, which makes them another target for researchers and law enforcement agencies to study and analyze them in depth [9].

Studies emphasize the importance of analyzing the content of the Dark Web communities and the related ones on the Surface Web to understand the thoughts and concepts they comprise, understand members' interests, trending topics, and crime perpetration methods, and anticipate new events [2, 10]. Therefore, the semantic analysis of the social network contents helps to discover the relationships within the semantics of messages and discussions on the social network.

Consequently, intellectual contents of Dark Web platforms have been studied and analyzed for various purposes, and different methods and tools were developed and employed. Such approaches included machine learning and data mining methodologies, such as classification, clustering, and summarization, and statistical analysis methodologies, which depend on the statistics calculated on tokens or chunks of the text. However, such techniques may suffer from missing the hidden semantic relationships in the analyzed content. Therefore, semantic-based approaches are needed to extract the shared concepts from massive textual data while considering the significant relationships within.

This paper presents an empirical methodology to analyze Dark Web social networks semantically using generative and probabilistic topic modeling. The approach utilizes several topic modeling methods and evaluation metrics, providing comprehensive insights about the semantic correlations among words and among topics and exemplifying the intellectual concepts on which the associated community is based.

The main contributions of the paper are as follows:

(1) Providing a comprehensive background of the needed stages of the followed methodology, including a thoughtful selection of data preprocessing procedures that address the particularity of the language used in Dark Web social networks, a term weighting scheme for feature extraction, and four selected topic models that serve various purposes

(2) Presenting a discussion about evaluating the generated topics to choose the optimal number of topics for each method

(3) Presenting further aspects of the generated topics by employing different visualization techniques, which help to understand the modeled content and the extracted relationships

(4) Presenting a discussion about the limitations and challenges that lead to new fields of research in the domain of Dark Web content analysis.

The remaining of the paper is structured as follows. Section 2 discusses the related work. Section 3 demonstrates the methodology basics, including topic modeling, data preprocessing, feature extraction, topic modeling algorithms, evaluation metrics, and the proposed approach. Section 4 discusses the results. Section 5 indicates limitations, challenges, and future work. Section 6 is the conclusion.

2. Related Work

Porter [2] presented a topic modeling approach to analyze a public subreddit related to the Dark Web called DarkNetMarkets. LDA was used to discover monthly-reciprocated information regarding the state of marketplaces on the Dark Web, in addition to information about security and anonymization technologies, cryptocurrency, and commercial exchange services. The research studies the recent changes and trends occurring in the Dark Web communities during a specific period. A relevance measure was used, then ranking the words for each topic according to the relevance scores, and then labeling the topics according to the ranked keywords. The results showed that during the studied period and up to the security crisis, the topics shifted from expressing normalcy and comfort to expressing a state of tension, low confidence, and an increased orientation towards a security mindset.

Similarly, Cho and Wright [11] presented a study that sheds light on the social phenomenon resulting from forum bans executed against communities of illegal products. The study included an assessment of the extent to which unexpected disruptions cause changes in the public debate. The approach included topic modeling using LDA and sentiment analysis to examine how members perceive the ban and how user participation has changed in the new system.

In Kigerl's approach [7], comments written by each user were combined into a single piece of text, making each word appears at least once in any comment. The word is represented by the number of times it is used by a particular user, using the Bag-of-Words model and a term frequency matrix,

where the row represents an individual user and the column represents a unique word containing the frequency of use of the word by each user. LDA was used to cluster texts into topics so that one user can be a member of more than one topic, with the topic probabilities totaling 1.0 for each user. The approach utilized model fit and performance measures to determine the optimal number of topics, including internal methods based on the similarity of documents assigned to the same topics and external metrics to measure the separation and distance between each topic and the others.

Kwon and Shao [12] relied on the communicative constitutive of organization (CCO) theory to analyze the human side of social networks on the Dark Web and to study the formation of knowledge in cryptomarkets' communities. The study aims to discover the characteristics that make the sociotechnical environment associated with markets on the Dark Web resilient through topic modeling. They utilized structural topic modeling (STM), an algorithm based on LDA, after NLP and text-cleaning operations, in addition to setting a minimum threshold of word frequency to ten occurrences. Quantitative and heuristic evaluations were used to determine the optimal number of topics, including exclusivity and coherence, in addition to manual reviews of thematic overlaps.

Heistracher et al. [10] highlighted the importance of determining the appropriate procedures of NLP to process the text and convert it into structured data. The research focuses on named entity recognition, relationship extraction, and event detection. The presented approach utilized topic modeling to visualize the deliberated topics and their distributions and to generate data classifications by ranking the most critical keywords of each topic. Relationship extraction was implemented using part of speech tagging, dependency tree analysis from SpaCy, neural networks, and word embeddings. DBSCAN was used to cluster the data elements according to their similarities using the cosine similarity measure.

Yang et al. [13] presented an approach to extract latent and trending topics from Dark Web forums. They suggested improving the results of the biterm topic model (BTM) by filtering words to define more coherent and interpretable topics and reduce cost and complexity. The filtering process was based on reducing the redundancy of biterms using a proposed new criterion called "generality" based on the document ratio formula. The generality measure works on identifying the least significant terms to be filtered out. They argued that if a term appears frequently and widely in the entire set of documents, the term belongs to the stopwords, while topical terms appear frequently in a few documents. The coherence measures UMass, UCI, and centroid coherence were used to assess the topics quality.

Kwon and Shao [9] presented an approach to examine the types of hidden knowledge shared in Dark Web-related communities on the Surface Web, specifically Reddit, and the extent to which the distribution of this knowledge differs in periods of the constant operation of the market and unstableness or crisis. LDA was utilized and implemented in R with the STM package. The study tried the model with different numbers of topics and applied FREX weights and semantic coherence for evaluation.

Several research studies analyzed the Dark Web content using content analysis and topic modeling techniques as a part of integrated processes. Topic modeling was used as a part of a two-step methodology to reduce the dimensionality of the feature space for better classification results [14] and with association rules mining among top words and other frequent words to describe the content of categorized Dark Web sites [15]. Some approaches involved text summarization [16] and classification techniques to classify reactions occurring in Dark Web communities in times of crises or shutdowns of the markets [17].

Topic modeling is still not widely utilized in semantic content analysis of the Dark Web communities, and shedding the light on the conceptual bases of such communities is still in its early steps. By using topic modeling methods and coherence score measures, the goal of this study is to extract the concepts forming the intellectual properties of communities around Dark Web marketplaces and represent the knowledge emerging from such communities in different forms.

3. Methodology

In this section, we begin with the basic concepts on which this study relies, including topic modeling definition in subsection 3.1, data preprocessing procedures in subsection 3.2, and feature extraction definition in subsection 3.3, and four algorithms of topic modeling are discussed in subsection 3.4, namely, latent Dirichlet allocation (LDA), correlated topic model (CTM), Pachinko allocation topic model (PAM), and pseudodocument topic model (PTM). Consequently, topic coherence measures are discussed as the evaluation metrics in subsection 3.5. Subsection 3.6 introduces the proposed approach.

3.1. Topic Modeling. In psychology, researchers define a concept as a network of correlated words. In a more generalized notion, a concept can be defined as "elements and their organization" [18].

In computational linguistics, researchers proposed several definitions of a concept, or topic. A topic can be defined as a set of words and their frequencies [19], as a set of words or phrases that represent a common temporal concept [20] or as latent topical features in given texts that correspond to contextually related words [21]. In another definition, a topic is a set of words likely to appear in the same context [22, 23].

On the other hand, topic extraction, or topic modeling, is defined as the technique used to infer conceptual topics hidden in a set of documents, or corpus [22, 24], where there is no explicit taxonomic scheme to project onto a corpus or when such projection (or labeling) is costly [25]. In another definition, topic modeling is an automated process to define the "latent thematic structure" of a corpus, summarizing the texts into topics or categorizing them into labels [26].

There are two fundamental research interests in topic modeling, topic interpretation and labeling and defining dominant topics from the generated topic models [27]. Topic modeling can be considered as a form of text mining used to extract frequent patterns of words in a corpus, where the set of words expresses a topic and thus infers the nature of the information in the document set. These interpretable topical schemes help to label each document in the corpus with an annotation. Subsequently, the annotations have usages in social computing and numerous other applications, such as information retrieval, classification, summarization, and sentiment analysis [22, 24, 28].

A topic model links documents and words using probabilistic and statistical analysis to extract latent features from the text where these features symbolize the hidden topical themes in the text [22]. Consequently, a document can contain several topics, where each topic is represented by a probability distribution over the vocabulary [29, 30].

Topic modeling can process massive volumes of textual data to extract hidden concepts, distinguish features, and define latent variables from text according to the application purpose [31]. Moreover, it can represent documents of a large corpus with concise but comprehensive commentaries [22], with the most likely ones assigned to each document [25].

Methodologically [32], it is advantageous to determine the following definitions as the basic units in the modeling process:

(i) **Word** or **Term**: the single basic unit of data

(ii) **Document**: a string of N words

(iii) **Corpus**: a collection of M documents

(iv) **Vocabulary**: the set of all the unique words in a corpus

(v) **Topic** or **Concept**: it is represented by a probability distribution over the vocabulary.

The words can be correlated through similarity, co-occurrence, proximity, and a subject-predicate structure [33, 34]. Each document is represented by a vector with dimensions corresponding to each term in the vocabulary and valued with the weights of the terms [35].

In this context, topic modeling considers that choosing the words and their positions in the text is intentional. Therefore, statistical analysis of the vocabularies and their co-occurrences with other terms in a particular text helps discover the vital concepts, premises, and intentions implied within the text [34].

In sociology, topic modeling reduces the human impact on the analysis objectivity and improves its efficiency compared to traditional methods. Therefore, advantages such as accuracy and objectivity make topic modeling a robust tool for sociologists [34, 36].

On the other hand, traditional content analysis methods suffer from unhandled polysemy. Topic modeling overcomes this limitation and considers the different meanings of the word by including the lexical context for a more accurate analysis of the word [34, 37].

Topic Modeling methods evaluate the importance of terms at several levels. Each word is weighted based on its position (i.e., the distances between words), word frequency, and the context, or semantics, of the word. Then, it follows the following steps to create a theme according to [34]:

(1) Determine which words have the highest weight (salient words)

(2) Determine the words closest to the words with the highest weights

(3) Determine the prominence of a particular set of words according to its frequency.

Depending on how high the weight of a particular set of words is, this set can be considered one of the topical themes in the corpus. The topic is thus composed of a set of significant words with sufficient proximity to each other and frequently appears as an integrated unit in the corpus [34].

Topic modeling methods can be supervised, unsupervised, or semisupervised using structured or unstructured data. Consequently, it is widely used on web resources to discover the abstract topics underlying a variety of text inputs on the web, such as short articles, chats, social media posts, user comments and reviews, blogs, emails, and other formats [31].

3.2. Data Preprocessing. The main challenge when analyzing Dark Web forums and marketplaces is the diversity of content structures and writing styles, which can contain grammatical and spelling errors, slang, and symbolized and ambiguous words intended to obscure their nature [10].

Like any text mining method, topic modeling inevitably relies on cleaning and preparing the raw text to its optimum prior analysis. For this purpose, textual data enter several preprocessing procedures to remove unimportant, irrelevant, and redundant attributes to reduce the dimensionality of the data space. Some of the most essential and widespread techniques used in text preprocessing are as follows [30, 38]:

(1) Normalization, which converts all letters to lowercase

(2) Removing common words or stopwords

(3) Removing nonalphabetic characters

(4) Removing punctuation, which includes removing all special characters and symbols (such as @ # $ %^ & * < >)

(5) Tokenization, which breaks a text into elements or attributes called tokens

(6) Lemmatization, which assembles the different morphological forms of a term in one single form (the lemma), so they can be analyzed as a single element

(7) Stemming, which returns the word to its root

(8) Parsing, which finds different dependencies between the words in a sentence and represents them in a tree structure called the parsing tree. A parsing

tree helps to find relationships among vocabularies by extracting the shortest path in the tree structure

(9) Parts of speech tagging, which determines the types of the different parts of speech that occurred in the text

(10) Removing hashtags, HTML tags, and links.

3.3. Feature Extraction. The data enter a feature extraction procedure as a further filtering process. Feature extraction is the core process of identifying word patterns and extracting topics. It aims to improve the modeling performance by reducing the dimensionality of the vocabulary space. The selected features are represented as a vector of salient words with contextual terms designated by a weighted distribution [30]. Several approaches were proposed to weight the term. The traditional and common method is the term frequency-inverse document frequency (TF-IDF) and its variants, which depend on the frequency of term occurrences within the document and through the corpus [25]. On the other hand, pointwise mutual information (PMI) is a leading weighting scheme as it weights terms regarding their dependencies and co-occurrences [25].

3.4. Topic Modeling Algorithms. Topic modeling algorithms are a multiperspective technique applied to discover the semantics in a corpus and to group extracted word patterns into topics. These algorithms generate a representation of the word meanings based on statistical and probabilistic analyses of the words [24, 28, 33]. The algebraic perspective was initially followed to reduce dimensionality, as the original matrix is decomposed into a matrix of factors. Thus, algorithms in topic modeling can be divided into the following two types: algebraic-perspective based and probabilistic-perspective based [28]. This paper concentrates on four probabilistic topic models, i.e., latent Dirichlet allocation (LDA), correlated topic model (CTM), Pachinko allocation topic model (PAM), and pseudodocument topic model (PTM). In the implementation, these algorithms follow the Bag-of-Words model to represent the documents and Gibbs sampling to identify topics. One of the most fundamental features of probabilistic models is that the topics can be extracted directly from the corpus without any predefined input from any prior knowledge [31].

The Bag-of-Words (BoWs) model represents the document as a word-document matrix regardless of the order or grammar of the words. The matrix is valued by the number of word occurrences. The word-document matrices form the input for the topic model instead of the entire document [22].

Gibbs sampler is the most common sampling method used in topic modeling. It performs conditional sampling using Markov chain Monte Carlo (MCMC), and it uses the distributions of the variables to define the posterior distribution, which is used afterwards to determine the best number of topics and identify topics [30, 39].

3.4.1. Latent Dirichlet Allocation (LDA). LDA was developed by Blei et al. [40], and it is the most popular topic modeling algorithm. It is a generative Bayesian probabilistic model that generates a distribution in document-topic and word-topic forms using the Dirichlet priors alpha and beta as hyperparameters to estimate the document-topic and the word-topic distributions, respectively [30]. The vocabulary space is transformed into a topic space due to the reduced volume of the latter. This transformation results in the following two matrices: the first represents the probability distributions of words on topics and the second represents the probability distributions of topics on documents [22]. Thus, the Bayesian model is built in a three-level hierarchy; word, document, and topic [32, 35]. Consequently, each document is a mixture of distinct topics, and each topic is composed of probabilities of words that are likely to co-occur in the topic [21]. This contrast helps to reduce ambiguity by ensuring that each document enfolds a small set of topics and that topic consists of a small cluster of words. Topic generation in LDA depends on the probability of vocabulary co-occurrences; in other words, a term may occur in different topics, but the other terms in its context determine the interpretability of the topic they represent [20]. LDA relies on the Definneti theorem to define the statistical structure of the document internally (the relationships between terms within the document) and externally (the relationships between documents). The algorithm depends on a predetermined number of topics k, where the specified number of topics is distributed over the documents with varying proportions [28]. It has proven effective in defining consistent topics, especially in datasets with sufficient vocabulary [20].

3.4.2. Correlated Topic Model (CTM). LDA suffers from a critical limitation, which is its inability to detect potential correlations between topics due to the use of the Dirichlet distribution to model the variability between topic parts. However, in real-world data, topics are interconnected [28]. To overcome this limitation, CTM was developed, by Blei and Lafferty [41], to model topics along with discovering the correlations between topics through the logistic normal distribution. Therefore, CTM has a more flexible and realistic distribution, taking into account topic parts to generate a hierarchical representation of the latent structure and correlations among components of the different topics [27, 28, 32, 42]. Studies based on the perplexity measure show that the CTM model provides a better fit than the LDA model. Moreover, for a document with a relatively small number of words, the perplexity value was much lower; thus, the certainty value was significantly higher for CTM. This advantage is because CTM uses the correlation among topics in the prediction procedure and infers that words appearing in the related topics may also occur in the document under process. Contrastly, LDA needs a higher proportion of the document to be observed and the topics to be fully generated to predict the remaining words [32].

3.4.3. Pachinko Allocation Topic Model (PAM). PAM is a topic modeling algorithm developed by Li and McCallum [43]. While CTM detects correlations between any two topics at a time, PAM creates a mixture model in the form of

a directed acyclic graph (DAG) to discover topical correlations of different types such as nested, arbitrary, and sparse correlations. The DAG is randomly constructed, where each word of the vocabulary is represented by a leaf node, and each topic is represented by an interior node; thus, the inner nodes are parent nodes, and the nodes branched from them (leaf and nonleaf) are children nodes. PAM can define correlations within the vocabulary and correlations among the topics; in other words, it explores the distributions of topics over other related topics in the form of categories of supertopics and subtopics representing a hierarchal relationships scheme [28, 32].

3.4.4. Pseudodocument Topic Model (PTM). PTM was introduced by Zuo et al. [44]. It is based on LDA in a process called self-aggregation, which implicitly groups short texts into pseudodocuments to address the problem of data sparsity, achieving higher quality with reduced training samples. PTM aggregates documents without employing supplementary information so that topic distributions are modeled on larger and fewer documents of regular sizes (the latent documents) rather than on a sparsely large amount of short texts (the observed documents). The aggregation process is done with a multinomial distribution of short texts over pseudodocuments, where each short text belongs to only one pseudodocument.

3.5. Evaluation Metrics. Evaluation metrics are used to assess the robustness of discovered topics. They evaluate the understandability (or quality) of the topics and the performance and accuracy of the modeling process. Topic evaluation can be achieved through standard measures (such as recall, precision, and *F*-score), perplexity, or semantic coherence measures [25, 28, 31]. In this paper, we choose perplexity as a prior evaluation metric to estimate the performance of the topic modeling process and topic coherence as a posterior metric to evaluate the quality of the generated topics, as it has been proven that it is well correlated with human evaluations [45].

3.5.1. Perplexity. Perplexity estimates the log-likelihood of the held-out document [28]. It is used to examine the performance of the topic model and calculated as shown in the following equation [46]:

$$\text{Perplexity}(D) = \exp\left\{-\frac{\sum_{d=1}^{M}\log P(w_d)}{\sum_{d=1}^{M}N_d}\right\}, \quad (1)$$

where D is the document set, M is the number of documents in the set, N is the number of words in document d, and $P(w_d)$ is the probability of words in document d.

3.5.2. Topic Coherence. Coherence expresses the interpretability of topics through word co-occurrences, as words that contextually and frequently co-occur in the corpus are more correlated, thus conveying a better-defined concept [25, 45]. Coherence is one of the paramount measuring techniques of topic quality. Probabilistic coherence estimates to which extent the words in a topic are correlated [28]. Coherence means that a set of statements or facts support each other contextually. In other words, they can be interpreted in a particular context that covers all or most of the facts; thus, they are coherent [47].

Several variations of coherence measures have been proposed that employ different probabilities calculations, such as PMI and NPMI. Studies proved that coherence measures based on PMI and NPMI give the highest agreement with human evaluations [47]. Pointwise mutual information (PMI) among top words of a topic assesses the amount of information gain of a word given the presence of the other word, taking into account the dependencies between words [25]. In an updated version, normalized PMI (NPMI) was developed and applied in many studies [25].

Coherence metrics consist of several components mainly divided into four dimensions, namely, segmentation (S), probability calculation (P), confirmation measure (M), and aggregation (Σ), applied to the generated topics (T) to produce the final coherence score (C), as demonstrated in Figure 1 [47].

3.5.3. Standard Coherence Measures. Several approaches have been proposed to estimate topic coherence. The measures consider the set of N top words of each topic, calculate the coherence of word pairs (w_i, w_j) based on the probabilities of word occurrences and word pair co-occurrences, and sum the resulting scores in a final coherence score of the topic. Coherence measures scores lead to a topic's keyword ranking, where the higher value indicates better topic coherence (closest to zero in case of negative values) [45, 47, 48].

In this section, we demonstrate the four customary coherence measures, namely, UMass, UCI, CNPMI, and CV (the topic coherence measures are thoroughly explained by Röder et al. [47] with an extensive comparison. We refer the interested reader to their study for further information).

UMass is calculated as shown in the following equation (ϵ is a small value added to prevent the logarithm of zero) [47]:

$$C_{UMass} = \frac{2}{N\,.\,(N-1)}\sum_{i=2}^{N}\sum_{j=1}^{i-1}\log\frac{P(w_i, w_j)+\epsilon}{P(w_j)}. \quad (2)$$

UCI relies on PMI, and it is calculated as shown in the following equation [47]:

$$C_{\text{UCI}} = \frac{2}{N\,.\,(N-1)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\text{PMI}(w_i, w_j), \quad (3)$$

where PMI is calculated as shown in the following equation [47]:

$$\text{PMI}(w_i, w_j) = \log\frac{P(w_i, w_j)+\epsilon}{P(w_i)\,.\,P(w_j)}. \quad (4)$$

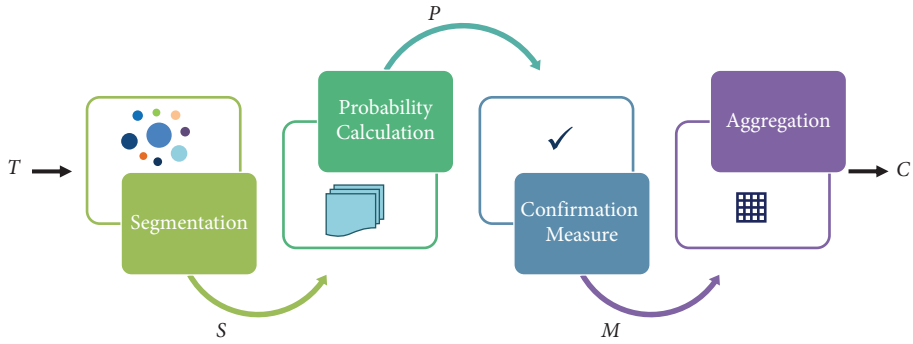


Figure 1: Dimensions of the coherence measure as inspired by [47].

The probabilities are estimated based on the number of co-occurrences of the words, and these co-occurrences are calculated from documents generated by the sliding window technique with a specified size [47].

Researchers deduced that when a word context is represented by a vector of its co-occurrences with other words within context windows (±5 words around the keyword), the topic coherence assessment is in the highest agreement with the human assessment when NPMI is used to define the elements of these vectors [47, 49]. It also achieves the highest performance when the keyword space is limited to words belonging to the same topic. Thus, for element j of the context vector $\overrightarrow{v_i}$ for the word w_i, the NPMI is calculated as shown in in the following equation (γ is a weighting factor used to give high NPMI values more weight) [47]:

$$v_{ij} = \mathrm{NPMI}\left(w_i, w_j\right)^{\gamma} = \left(\frac{\log P\left(w_i, w_j\right) + \epsilon / P\left(w_i\right).P\left(w_j\right)}{-\log\left(P\left(w_i, w_j\right) + \epsilon\right)}\right)^{\gamma}. \tag{5}$$

In a modified version of UCI, NPMI is used instead of PMI to calculate the CNPMI score [49]. Therefore, the CNPMI coherence measure is calculated as shown in the following equation:

$$\mathrm{CNPMI} = \frac{2}{N.(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathrm{NPMI}\left(w_i, w_j\right). \tag{6}$$

The final coherence measure, CV, proposed by Röder et al. [47] is based on the cosine measure with NPMI and a Boolean sliding window of size ≥50.

3.6. Proposed Approach. In this section, we demonstrate the proposed approach to model topics from Dark Web social networks. Figure 2 illustrates the methodology.

3.6.1. Dataset. For our experiments, we use a dataset of a Dark Web forum associated with the Wallstreet Cryptomarket. The dataset is retrieved from AZSecure Dark Net forums datasets (https://www.azsecure-data.org/dark-net-markets.html (accessed on 20 September 2022)) provided by Du et al. [50].

For the texts corpus creation, we selected the flat content of the post as our only interest, without revealing any identities or usernames.

3.6.2. Data Cleaning and Preprocessing. We implemented our cleaner to include several basic text-cleaning procedures and some specific ones according to the nature of the selected dataset. The data preprocessing steps are described as follows:

(1) Replace accented characters (such as à, ê, õ, and ü) with unicode characters using the unidecode Python module. The purpose of this step is to unify the character coding for a fair judgement of the words during the weighting and feature extraction phase.

(2) Normalize characters to lowercase. The same word can be found in different capitalization forms; thus, this step is necessary to bring all cases of the word into one.

(3) Remove lines, tabs, and spaces

(4) Remove hyperlinks by defining them as regular expressions using re Python module

(5) Expand contractions (such as you're = you are) by defining a list of contractions and their replacements. This step prepares for a better exclusion of stopwords performed in a later step.

(6) Remove special characters (such as ~ ! @ # $) defined as regular expressions using the re module

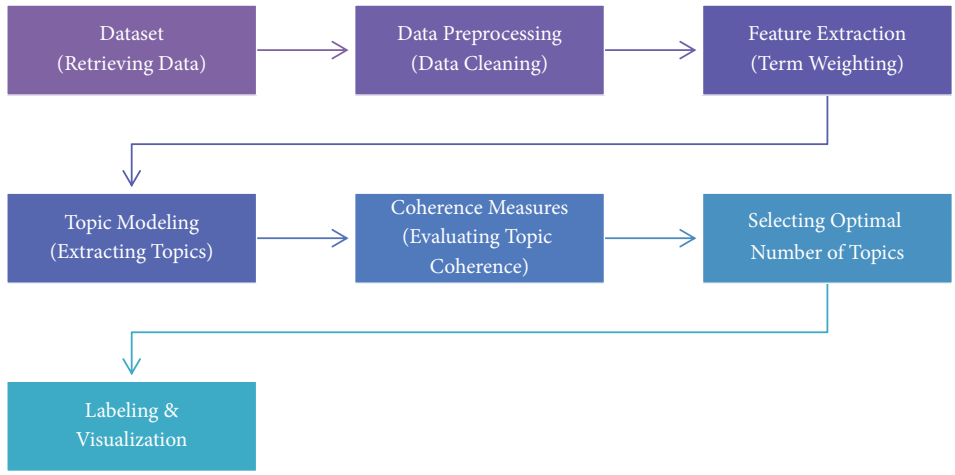


Figure 2: Methodology of the proposed approach.

(7) Lemmatization using WordNetLemmatizer package from nltk (Natural Language Toolkit (NLTK): https://www.nltk.org/ [Accessed 2 October 2022]) library

(8) Remove stopwords defined in nltk stopwords list for English

(9) Remove specific words observed through the text but add no meaning to the text, such as "wrote" and "would", and the names of the months (basically before the post declaring the posting date)

We did not include a stemming process, as we noticed that stemming transforms the word to an erroneous-spelled word or stems a word that does not need stemming (such as quality to qualiti), which may confuse the interpretation of the results. Consequently, after several cleaning trials and observations, we noticed that lemmatization produces better results than stemming. Therefore, we depend on the lemmatization process to reduce the dimensionality of the word space.

We also did not include a spelling correction process, as we discussed previously, members may use intentionally misspelled words as an obfuscation strategy, and such words may be well known among the community participants, so we keep the spelling as it is.

The following example shows a text before and after preprocessing:

Before preprocessing:

"*I believe one of the most important opportunities or factors of buying your drugs on the Dark Net is because once you get to know the markets, vendors and the people that post on these forums we can together create a community that watches out for the safety of all. The DNMs should be a place where we can be sure that we are buying safer quality drugs than what we can obtain from the streets. Therefore we need to learn about what can be the dangerous adulterants in the drugs we are buying. As for cocaine the one that is showing up most often that is particularly dangerous is LEVAMISOLE. For a better understanding of the whats and the whys there is a well written series of articles by the free weekly publication from Seattle, WA, USA "The Stranger". You can find these articles here*: https://cocaineo5z66elwy.onion/Levamisol...caine.html"

After preprocessing:

"believe one important opportunity factor buy drug dark net get know market vendor people post forum together create community watch safety dnms place sure buy safe quality drug obtain street therefore need learn dangerous adulterant drug buy cocaine one show often particularly dangerous levamisole good understand whats well series article free weekly publication seattle usa strange find article caine html"

3.6.3. Feature Extraction and Topic Modeling Algorithms Implementation. We utilized four topic modeling algorithms, which are LDA, CTM, PAM, and PTM, implemented using Tomotopy (Tomotopy: https://github.com/bab2min/tomotopy [Accessed 14 October 2022]) Python package for topic modeling. We choose Tomotopy for the variety of algorithms and utilities it provides.

As discussed in Section 3.4, we choose these four methods for their characteristics. LDA is the most common topic modeling method, which is easy and fast. CTM has an advantage over LDA that it extracts the relationships among generated topics depicted as a network. This network benefits in detecting the connections between different thoughts in the discussions, as concepts in the real world are not independent but correlated. PAM helps build a hierarchy of supertopics and subtopics, which helps to illustrate a relationship between a generalized thought to more specific ones. PTM has an advantage when utilized to extract topics from diverse sizes of posts when grouping them in pseudodocuments of regular lengths for better modeling results.

For a fair comparison between the four models, we set the basic settings and parameters equally. First, we use PMI as a word-weighing scheme to capture the salient words

Table 1: Results of the word-weighting process.

Number of documents	Total number of words	Number of used vocabularies	Entropy of words	Entropy of term-weighted words
45342	1956570	117061	7.964	9.214

according to their semantic relevance [51]. We found that the topic modeling methods performed better with PMI than TF-IDF for feature extraction, with lower entropy of term-weighted words and lower perplexity. Second, we set the iteration number of the training process where the log-likelihood stops making a significant increase. A higher log-likelihood (closer to 0) indicates a lower perplexity, and a low perplexity score means that the prior calculated probabilities define the generated topics well [7]. Thus, we set the iterations number to 30 iterations of Gibbs sampling. Each iteration implies 20 cross-sampling iterations, which makes a total of 600 iterations for each modeling process. The third setting is the number of topics k. For each algorithm, we train the model for ten candidate numbers of topics from 5 to 50 with hops of 5.

For CTM, we set the number of iterations to sample beta parameter to 5, a moderate number of iterations to regulate time cost, as CTM takes longer than the other methods. For PAM, we set the numbers of supertopics to small numbers to give it a sense of clustering, grouping the subtopics into small groups of master domains. For PTM, we set the number of pseudodocuments to ten times the number of topics ($10 * k$). Each model is run three times to estimate the best performance and select the best model for each method. Table 1 shows the result of the word-weighting process. Table 2 presents the log-likelihood per word of the last iteration for each model and each value of k.

3.6.4. Evaluating Topics Using Topic Coherence Measures. For coherence evaluations, we use the four common coherence score measures, namely, UMass, UCI, CNPMI, and CV. The evaluations were conducted for each topic model method and each value of k, and we applied the standard deviation (denoted as STDEV) to estimate the error rate of the coherence measures for the three runs. The resulting values are demonstrated for LDA, CTM, PAM, and PTM in Tables 3–6, respectively.

Figures 3–6 illustrate the coherence scores for each topic model according to the coherence measures, i.e., UMass, UCI, CNPMI, and CV, respectively.

4. Results and Discussion

From the four coherence score measures, we discuss the results from CNPMI, as NPMI proved to have the overall best performance and correlations with human evaluations [47, 49]. Thus, from Figure 5, we infer the optimal value of k for each topic modeling method. We suggest considering several high close points of coherence for a topic model for further human observation to choose the optimal number of topics that suits the aim of the research and for further better analysis and labeling. Therefore, we discuss the highest points of coherence for each topic model we conducted on the Dark Web forum. For each model, we examine the optimal value of k, demonstrate the generated topics, and label them according to the most frequent terms and most related ones in each topic. We notice that more than one topic may fall under the same label, and a topic may hold more than one label. This influence is due to the nature of the probability distributions of terms over topics and topics over documents with different proportions.

LDA shows high coherence scores at $k = 20$, 25, and 30. Two of the top three highest scores are achieved with $k = 20$ with CNPMI = 0.0509 and 0.0452 and with an error rate STDEV = 0.0076. To help demonstrate the separation of topics for each of the three topic models, we use pyLDAvis (pyLDAvis: https://github.com/bmabey/pyLDAvis [Accessed 23 October 2022]) to visualize the topic maps corresponding with the abovementioned models, for $k = 20$ at CNPMI = 0.0509, $k = 25$ at CNPMI = 0.0445, and $k = 30$ at CNPMI = 0.0390, as illustrated in Figure 7.

In Figure 7, we notice the best separation is gained from $k = 20$ (shown in a), while for 25 and 30 (shown in b and c, respectively), more topic clusters overlap. We can infer that some numbers of topics produce good coherence; however, with an increased number of topics, more topic clusters overlap. Table 7 shows the top 20 words of the topics generated by LDA for $k = 20$ at CNPMI = 0.0509.

For CTM, the top three coherence scores are achieved at $k = 5$ and $k = 10$, with the highest score of all recorded at $k = 10$ with CNPMI = 0.0173 and error rate STDEV = 0.021, with a high peak of coherence compared to the other numbers of topics. Table 8 presents the positive and negative correlations among the generated topics. We illustrate the network of the generated topics and their correlations using pyvis (pyvis: https://github.com/WestHealth/pyvis [Accessed 23 October 2022]), as shown in Figure 8. Positive correlations indicate that two words are likely to appear in the same topic, while negative correlations indicate that two words are unlikely to appear in the same topic. As mentioned earlier, correlations are calculated through the logistic normal distribution [41]. Table 9 shows the top 20 words of topics generated by CTM for $k = 10$ at CNPMI = 0.0173, with the sizes of the topics defined by the number of terms in each topic.

It is worth noting that if more number of topics with more correlations is desired for a specific analysis purpose, one may choose a higher number of topics but gain lower coherence, as CTM generates topics with dense correlations structure but less guaranteed coherence [52].

PAM achieved the top three coherence scores at $k = 10$ with CNPMI = 0.0869, 0.0542, and 0.0662 and error rate

Table 2: Log-likelihood value for each modeling algorithm and each value of k.

	$k=5$			$k=10$			$k=15$			$k=20$			$k=25$		
	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3
LDA	−8.66	−8.67	−8.67	−8.41	−8.43	−8.44	−8.24	−8.26	−8.26	−8.15	−8.14	−8.15	−8.04	−8.05	−8.05
CTM	−8.05	−8.10	−8.07	−7.64	−7.66	−7.69	−7.38	−7.42	−7.41	−7.32	−7.32	−7.32	−7.17	−7.17	−7.07
PAM	−9.93	−9.97	−10.02	−10.43	−10.40	−10.38	−10.63	−10.57	−10.58	−10.76	−10.79	−10.75	−10.85	−10.79	−10.77
PTM	−8.85	−8.84	−8.81	−8.73	−8.68	−8.69	−8.60	−8.54	−8.56	−8.50	−8.46	−8.47	−8.45	−8.44	−8.43
	$k=30$			$k=35$			$k=40$			$k=45$			$k=50$		
	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3
LDA	−7.95	−7.96	−7.97	−7.91	−7.90	−7.90	−7.86	−7.85	−7.86	−7.80	−7.79	−7.77	−7.75	−7.74	−7.76
CTM	−7.04	−7.06	−7.12	−7.03	−6.98	−7.03	−6.97	−6.80	−6.89	−7.01	−7.24	−7.11	−6.87	−6.74	−6.67
PAM	−10.89	−10.91	−10.87	−11.03	−10.97	−11.02	−11.06	−11.02	−11.05	−11.16	−11.15	−11.10	−11.12	−11.09	−11.15
PTM	−8.37	−8.36	−8.44	−8.34	−8.33	−8.33	−8.31	−8.29	−8.27	−8.27	−8.29	−8.22	−8.19	−8.22	−8.20

Table 3: Coherence scores for LDA.

	$k=5$			$k=10$			$k=15$			$k=20$			$k=25$		
	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3
CUMASS	−4.5818	−3.7039	−3.4599	−4.0605	−3.7081	−3.7773	−4.1671	−4.3727	−4.9998	−4.2603	−4.1414	−4.1743	−4.5056	−4.2672	−5.4747
STDEV		0.5900			0.1867			0.4337			0.0614			0.6396	
CUCI	−1.9514	−1.2642	−1.4639	−1.7021	−1.7921	−1.9673	−1.9565	−1.9007	−2.2416	−1.5066	−1.6862	−1.7135	−1.6519	−1.7577	−2.4349
STDEV		0.3535			0.1349			0.1829			0.1124			0.4248	
CNPMI	0.0066	0.0321	0.0196	0.0245	0.0185	0.0089	0.0287	0.0242	0.0041	0.0509	0.0452	0.0359	0.0445	0.0341	0.0133
STDEV		0.0128			0.0079			0.0131			0.0076			0.0159	
CV	0.6756	0.6281	0.6471	0.7006	0.6738	0.6943	0.7063	0.7249	0.7256	0.7126	0.7007	0.7438	0.7622	0.7332	0.7606
STDEV		0.0239			0.0140			0.0110			0.0222			0.0163	
	$k=30$			$k=35$			$k=40$			$k=45$			$k=50$		
	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3
CUMASS	−4.7084	−4.5987	−4.6954	−5.5073	−4.8649	−5.2566	−5.1379	−5.6170	−5.1214	−5.7629	−5.2985	−5.3391	−4.8740	−5.6395	−5.7052
STDEV		0.0599			0.3238			0.2815			0.2572			0.4621	
CUCI	−1.9416	−1.6969	−1.8742	−2.4134	−1.9732	−2.0141	−2.2374	−2.3828	−2.2753	−2.2347	−2.1141	−2.0447	−2.0721	−2.0365	−2.4476
STDEV		0.1264			0.2432			0.0754			0.0962			0.2278	
CNPMI	0.0338	0.0390	0.0355	0.0104	0.0311	0.0278	0.0200	0.0159	0.0195	0.0211	0.0275	0.0326	0.0229	0.0297	0.0108
STDEV		0.0026			0.0111			0.0022			0.0057			0.0096	
CV	0.7468	0.7597	0.7468	0.7602	0.7932	0.7568	0.7535	0.7673	0.7535	0.7741	0.7895	0.7737	0.7630	0.7725	0.7763
STDEV		0.0074			0.0201			0.0080			0.0090			0.0069	

Table 4: Coherence scores for CTM.

	$k=5$			$k=10$			$k=15$			$k=20$			$k=25$		
	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3
CUMASS	−3.1369	−2.9813	−2.9612	−3.1135	−3.7349	−3.8300	−4.2050	−4.2860	−5.0225	−6.0585	−5.7740	−6.3037	−6.6547	−5.2086	−6.5010
STDEV		0.0961			0.3891			0.4504			0.2650			0.7943	
CUCI	−1.4503	−1.3898	−1.9542	−1.5666	−1.7358	−2.1866	−2.6047	−2.1523	−2.6744	−3.5079	−3.3110	−3.5650	−3.5229	−2.9540	−3.4597
STDEV		0.3099			0.3204			0.2834			0.1333			0.3118	
CNPMI	0.0014	−0.0242	−0.0328	0.0173	−0.0171	−0.0208	−0.0322	−0.0195	−0.0315	−0.0623	−0.0602	−0.0607	−0.0624	−0.0446	−0.0685
STDEV		0.0178			0.0210			0.0072			0.0011			0.0124	
CV	0.5654	0.5252	0.5332	0.6343	0.6150	0.6379	0.6065	0.6441	0.6590	0.6939	0.6892	0.6993	0.7118	0.6920	0.6915
STDEV		0.0212			0.0123			0.0270			0.0050			0.0116	
	$k=30$			$k=35$			$k=40$			$k=45$			$k=50$		
	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3
CUMASS	−6.0002	−7.0800	−7.8119	−6.8520	−6.2491	−7.1897	−7.3029	−5.0659	−6.6768	−7.0007	−7.3443	−7.5679	−7.4746	−5.3560	−7.0124
STDEV		0.9114			0.4765			1.1540			0.2857			1.1140	
CUCI	−3.3906	−3.7220	−4.3293	−3.9689	−3.8324	−4.1845	−4.2177	−3.0423	−3.9720	−4.2713	−4.1629	−3.9763	−4.3125	−3.5212	−4.0024
STDEV		0.4761			0.1775			0.6200			0.1492			0.3987	
CNPMI	−0.0631	−0.0793	−0.1046	−0.0808	−0.0774	−0.0973	−0.0934	−0.0590	−0.0861	−0.0961	−0.0882	−0.0912	−0.1043	−0.0748	−0.0943
STDEV		0.0210			0.0106			0.0181			0.0040			0.0150	
CV	0.6778	0.7224	0.7135	0.7227	0.6898	0.7169	0.7182	0.6798	0.6974	0.7095	0.7145	0.7236	0.6979	0.6646	0.6135
STDEV		0.0236			0.0176			0.0192			0.0071			0.0425	

Table 5: Coherence scores for PAM.

		$k1 = 2$ $k2 = 5$			$k1 = 3$ $k2 = 10$			$k1 = 3$ $k2 = 15$			$k1 = 4$ $k2 = 20$			$k1 = 4$ $k2 = 25$	
	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3
CUMASS	−4.5682	−4.4167	−3.3346	−1.8438	−1.7641	−2.8744	−6.1847	−3.2704	−2.1185	−4.0028	−6.5858	−4.7175	−3.1601	−6.7458	−8.8022
STDEV		0.6728			0.6193			2.0958			1.3337			2.8554	
CUCI	−1.6748	−2.7993	−2.0669	−0.0846	−0.7651	−1.0505	−3.0860	−1.2615	−2.5484	−1.7979	−3.5194	−2.3722	−1.6599	−3.7740	−3.8472
STDEV		0.5708			0.4963			0.9375			0.8765			1.2422	
CNPMI	0.0452	−0.0628	−0.0077	0.0869	0.0542	0.0662	−0.0241	0.0329	0.0289	0.0479	−0.0214	0.0291	0.0450	−0.0281	−0.0144
STDEV		0.0540			0.0165			0.0318			0.0358			0.0389	
CV	0.6202	0.4941	0.5587	0.7170	0.7084	0.5713	0.5909	0.6866	0.8147	0.7042	0.6733	0.6512	0.6655	0.6023	0.7111
STDEV		0.0631			0.0818			0.1123			0.0266			0.0547	
		$k1 = 5$ $k2 = 30$			$k1 = 5$ $k2 = 35$			$k1 = 7$ $k2 = 40$			$k1 = 7$ $k2 = 45$			$k1 = 10$ $k2 = 50$	
	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3
CUMASS	−4.6596	−4.5673	−4.6108	−7.1348	−5.4745	−3.6153	−4.9706	−5.4680	−5.1653	−5.1041	−4.9456	−6.2880	−4.7598	−6.8986	−6.5116
STDEV		0.0462			1.7607			0.2506			0.7336			1.1397	
CUCI	−2.3724	−1.9703	−1.6918	−4.4737	−2.0689	−1.0260	−1.8945	−2.1215	−2.0631	−1.8972	−2.4475	−2.7510	−1.0820	−3.1024	−2.2901
STDEV		0.3422			1.7681			0.1179			0.4328			1.0167	
CNPMI	−0.0161	0.0371	0.0304	−0.0852	0.0328	0.0442	0.0236	0.0410	0.0397	0.0283	0.0063	−0.0149	0.0518	−0.0044	0.0002
STDEV		0.0290			0.0716			0.0097			0.0216			0.0312	
CV	0.6609	0.6035	0.6841	0.7005	0.6555	0.5590	0.6890	0.7498	0.7663	0.6971	0.7275	0.7128	0.6932	0.7373	0.7200
STDEV		0.0415			0.0723			0.0407			0.0152			0.0222	

Table 6: Coherence scores for PTM.

		$k = 5$ $p = 50$			$k = 10$ $p = 100$			$k = 15$ $p = 150$			$k = 20$ $p = 200$			$k = 25$ $p = 250$	
	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3
CUMASS	−2.7594	−3.2228	−4.3617	−3.2286	−3.2422	−4.1464	−4.9332	−5.6000	−5.3349	−5.7245	−6.0918	−6.2818	−5.6192	−5.8929	−5.1314
STDEV		0.8246			0.5260			0.3357			0.2833			0.3857	
CUCI	−1.6422	−1.6884	−2.8125	−1.1733	−1.2654	−1.6247	−2.4025	−2.9046	−2.3484	−2.5830	−3.1505	−3.1279	−2.5562	−2.6753	−2.0936
STDEV		0.6628			0.2385			0.3067			0.3213			0.3073	
CNPMI	0.0040	0.0159	−0.0400	0.0500	0.0510	0.0363	0.0006	−0.0212	0.0038	0.0045	−0.0281	−0.0283	−0.0013	−0.0040	0.0272
STDEV		0.0295			0.0082			0.0136			0.0189			0.0173	
CV	0.6261	0.5835	0.5630	0.6679	0.6846	0.6669	0.7305	0.7329	0.7404	0.7662	0.7396	0.7682	0.7556	0.7724	0.7662
STDEV		0.0322			0.0099			0.0051			0.0160			0.0085	
		$k = 30$ $p = 300$			$k = 35$ $p = 350$			$k = 40$ $p = 400$			$k = 45$ $p = 450$			$k = 50$ $p = 500$	
	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3
CUMASS	−6.1779	−6.5790	−6.5684	−7.0443	−7.1310	−6.0683	−6.6276	−7.4835	−6.5795	−7.0071	−7.2680	−8.1403	−7.6360	−7.6998	−7.9449
STDEV		0.2286			0.5901			0.5086			0.5935			0.1630	
CUCI	−2.7979	−2.9666	−3.0278	−3.2830	−3.4885	−3.0848	−3.2884	−3.4900	−2.8800	−3.3357	−3.4175	−3.8774	−3.4186	−3.7675	−3.8211
STDEV		0.1190			0.2019			0.3108			0.2920			0.2185	
CNPMI	−0.0098	−0.0136	−0.0267	−0.0356	−0.0386	−0.0182	−0.0316	−0.0413	−0.0172	−0.0346	−0.0312	−0.0527	−0.0309	−0.0462	−0.0486
STDEV		0.0089			0.0110			0.0121			0.0116			0.0096	
CV	0.7700	0.7713	0.7729	0.7582	0.7964	0.7906	0.7929	0.7781	0.7756	0.7987	0.8126	0.8042	0.8074	0.8173	0.7880
STDEV		0.0015			0.0206			0.0094			0.0070			0.0149	

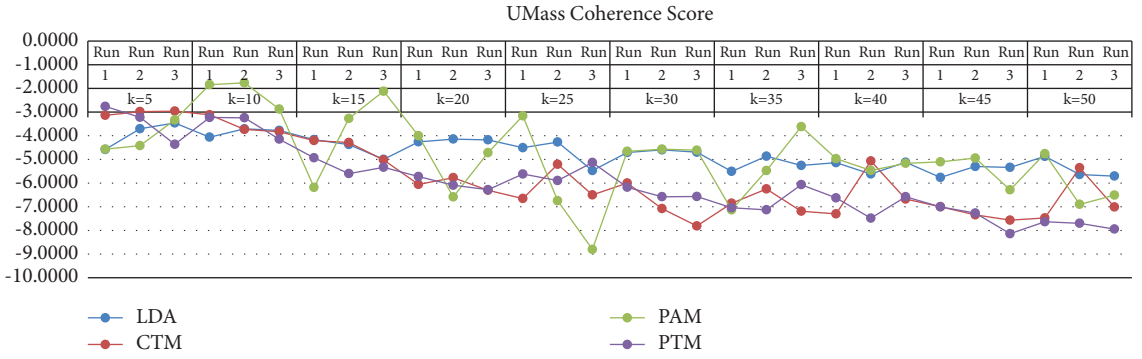


Figure 3: UMass coherence score for each topic model and each number of topics.

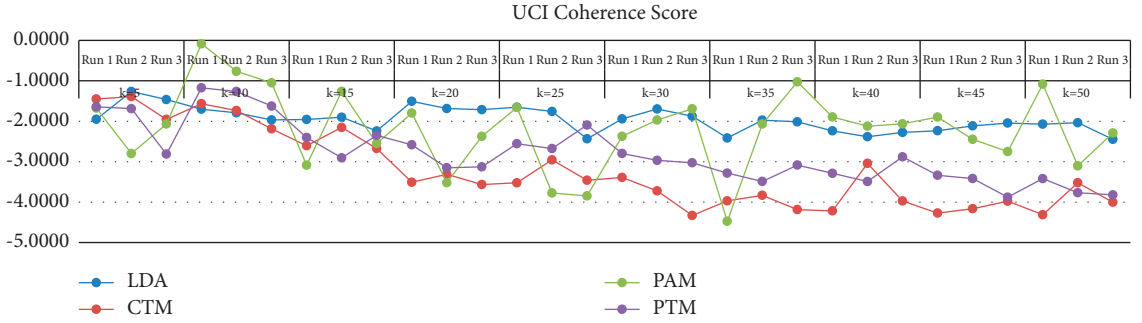


Figure 4: UCI coherence score for each topic model and each number of topics.

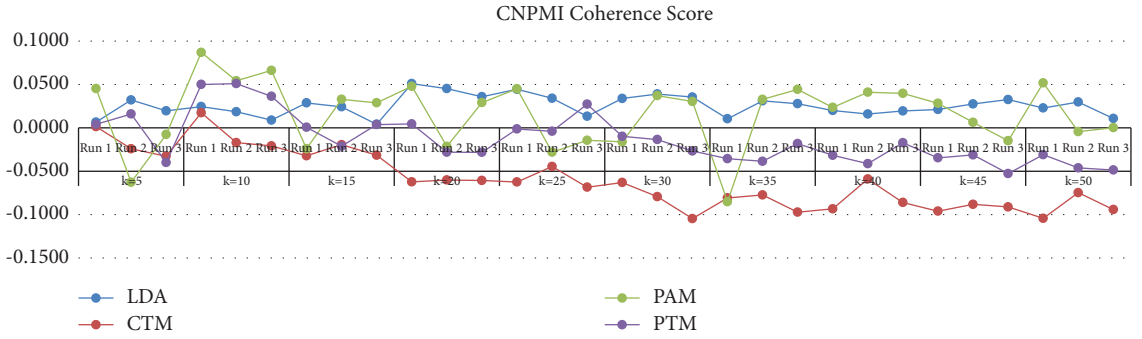


Figure 5: CNPMI coherence score for each topic model and each number of topics.

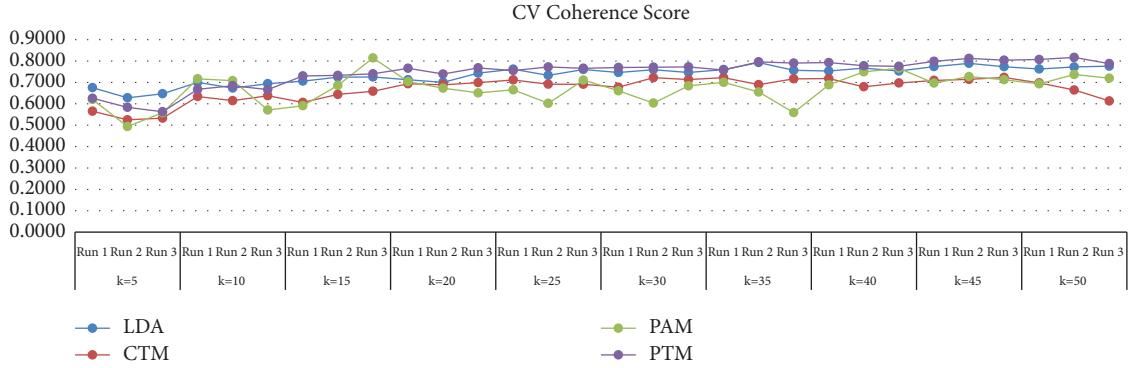
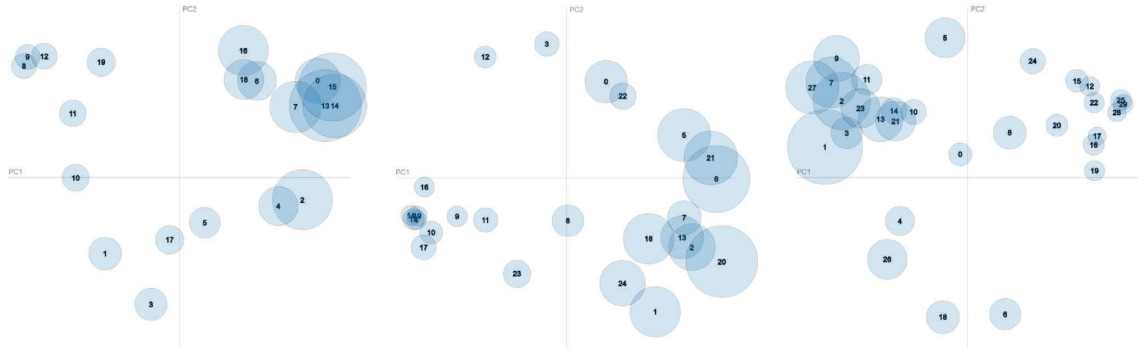


Figure 6: CV coherence score for each topic model and each number of topics.



(a) (b) (c)

Figure 7: Topic maps for the highest coherent topic models of LDA.

STDEV = 0.0165. The closest point that follows is at 0.0518 for $k = 50$. Again, a research purpose can define the best topics number to follow by determining the desired size of the hierarchy of supertopics and subtopics. For example, for 10 subtopics ($k2$), we set the number of supertopics ($k1$) to 3, while we set it to 10 for 50 subtopics. Table 10 shows the distribution of the subtopics over supertopics for $k1 = 3$ and $k2 = 10$ at CNPMI = 0.0869, with their distribution probabilities sorted descendingly. Table 11 shows the top 20 words of the subtopics.

Lastly, for PTM, the CNPMI coherence shows a high peak at $k = 10$ with the top three coherence scores recorded as CNPMI = 0.0500, 0.0510, and 0.0363. Table 12 shows the top 20 words of the generated topics for $k = 10$ at CNPMI = 0.0510. Figure 9 illustrates the topics map.

5. Limitations, Challenges, and Future Work

Our presented approach provides comprehensive discussions and solutions for a variety of issues and challenges, including a thoughtful selection of the proper data preprocessing procedures, a weighting scheme for feature extraction and suitable topic modeling methods that serve various purposes in understanding the terms' semantics and hidden topics in Dark Web discussions.

Text analysis inevitably depends on preprocessing and cleaning procedures for accurate results; thus, it is crucial to consider thoughtful decisions in selecting the appropriate ones. In social networks, texts often contain slang, misspellings, and grammatical errors, which may be intended in the case of the Dark Web in particular. Such datasets may require further manual observations of the posts to eliminate the unmeaningful words, which need training the topic model several times to detect them manually. On the other hand, we found that lemmatization produces better results than stemming for the studied dataset. However, some words are still missed by the lemmatization process and remain unlemmatized, which needs further research to enhance and extend the dictionary.

The BoW model uses a vocabulary of terms that explicitly occur in the document; thus, it ignores important correlations between the terms that do not co-occur, yet they are connected [35].

Topic models' performance and coherence are steered by the properties of the dataset. Thus, results may vary according to the dataset type, size, content, and lengths of its entries. Therefore, each dataset needs unique and thoughtful evaluation decisions to determine the best-performed coherent model. Moreover, as we observed in this study, each training process for each model with the same number of topics produced different outcomes with varied coherence scores. Thus, an analysis may need to run the experimentations several times to select the best result for each topic model and each value of k.

In our previous work [4], we discussed challenges in the field of analyzing the Dark Web content, which is mostly characterized by the language inconsistency of the discussions. This inconsistency can manifest in weak grammatical contexts and intentionally ambiguated words, such as emerged slang, intentional misspellings, abbreviated terms, and idiomatic contexts, which all are customary to the Dark Web communities but ambiguous to others outside these communities. Furthermore, unlike public or regular social networks, members of the Dark Web communities create concepts of their own that might change over time as needed and may be understood by their members only. Due to these reasons, further research is needed to analyze the intentions behind the used words and contexts, which may require the knowledge of experts to decipher the linguistic purposes. Moreover, some ethical considerations must be carefully taken when studying the Dark Web content, as the discussions may contain sensitive data.

For future work, we will add N-gram for language processing, which can help overcome the shortcomings of the BoW model. Furthermore, integrating ontologies can enhance text preprocessing and the generated topic models and labeling. Other topic modeling methods, such as the dynamic topic model (DTM), hierarchical LDA (HLDA),

Table 7: Top 20 words of the topics generated by LDA with $k = 20$.

T#0	*T*#1	*T*#2	*T*#3	*T*#4	*T*#5	*T*#6	*T*#7	*T*#8	*T*#9
110041	55138	195156	49260	80549	49287	76272	143399	31542	30021
Communication	Payment and shipping	Drug deals	Drug deals	Promoting listings	Drugs	Counterfeit items	Counterfeit items	Security	Security
Room	Utopia	Ship	Weed	Note	Bump	Passport	Guide	Pgp	Pgp
Chat	Gramo	Order	Strain	Counterfeit	Market	License	Cashout	Signature	Signature
Method	Eur	Product	Hash	Eur	Store	Transfer	Record	Begin	Begin
Cashout	Refund	Review	Kush	Bill	Wsm	Driver	Bank	Message	Badget
Fraud	Reship	Cocaine	Bud	Quality	mdma	Scan	SSN	Badget	End
Log	Kush	Gram	Utopia	Euro	Page	Card	Expectus	Sign	Sign
Everythingcc	Order	Vendor	Cannabis	Kallashnikov	Profile	Name	Pack	Sha	Sha
Change	Platinum	Quality	thc	Profile	Vendor	Phone	Extra	End	Access
Day	Eurgr	Coke	Grade	Review	Free	Com	Fullz	Hash	Message
Lack	Haze	mdma	Price	Black	Ketamine	Gold	High	Gnupg	E-mail
Chaser	Spain	Good	Haze	Lego	Tax	Porn	Card	Version	Hash
Bank	Europe	Price	Gram	Supplier	Bot	Fake	Credit	vous	Gnupg
Page	Successful	Sample	Best	Order	New	Register	Offer	Badge	Com
Dream	Wallstreet	Best	Pack	Gbp	Goznym	Account	Link	Block	Version
One	Customer	High	Smoke	Ship	Malware	Lifetime	Level	Key	Badge
Drop	Top	Lsd	ounce	pas	Cocaine	Address	Retirement	Pour	Public
Bro	Shipment	Stealth	Cart	Good	Amphetamine	Number	Lesson	Public	Key
Guide	Pgp	Day	Indoor	Sativa	Atm	Cvv	Check	sur	Block
Heist	Bubba	Come	Oil	Paper	Rebellion	Unregistered	Debit	Vendor	Vendor
Money	precio	Free	Amnesia	High	Unfree	IdSOF	EMV	Bcpg	Jevreg
T#10	*T*#11	*T*#12	*T*#13	*T*#14	*T*#15	*T*#16	*T*#17	*T*#18	*T*#19
35454	34459	33662	284319	216696	253612	117805	35751	93013	31134
General use	Community	Counterfeit cards	Dealings	Opinions	Financial matters	Networking	Prescriptions	Acquaintance	Traveling
Mirror	Supermod	visa	Order	Drug	Card	Key	Adderall	Rule	Flight
Onionshop	Shoutouts	Credit	Vendor	Consider	Account	Use	Stock	Phished	Hotel
Use	Overdue	Pgp	Dispute	Arab	Bank	tor	Pill	External	Book
Per	Thieve	Mastercard	Wsm	Sand	Drop	Tail	Suboxone	Help	Card
Eur	Honor	Signature	Get	Smash	Paypal	Pgp	Interest	Wsm	Car
Ghb	Hero	Classic	Help	Nun	Need	File	Pharmacy	Contact	Nothing
Sample	Among	Begin	Wallet	Testicle	Transfer	vpn	Xanax	Allow	Upintheair
Euro	Dabbingtimes	Debit	Send	Eskimo	Look	Click	Href	Get	Travel
Quickship	Deserve	Bank	Support	Ice	Service	Rdp	Press	Please	Rango
Liquid	Boneskoopa	Platinum	Buyer	Make	Cash	Server	Offer	Link	Service
Pill	Dont	Sign	Say	Know	Sell	Self	Cialis	Vendor	Freelance
Quality	Dude	Sha	Address	Like	Btc	Sock	Ship	Med	Change
Stealth	und	Badget	Time	Good	Interest	Encrypt	Supply	Read	Validity
Dreamwe	Code	Message	Refund	Money	Work	Window	Limit	Forum	Non
Promptly	Break	End	Ticket	People	Credit	Public	Quot	Account	Issuer
Pgp	Ich	Gold	Btc	Work	Vendor	Proxy	Brand	Close	Rental
Judge	Keep	Eurrow	Market	Forum	Get	Bridge	Medicine	Broder	Carder
Basf	Put	Ltd	Coin	Think	Contact	Computer	Legit	Welcome	Atm
vial	Wir	Plc	Monero	Dump	Buy	System	Monthly	Bro	Acquirer
Methylphenidate	mit	unite	Phished	Say	Fund	Set	Generic	Badge	National

T#~ refers to the topic ID with the number of words in each topic below it, which indicates the size of the topic.

Table 8: Correlations between topics for CTM with $k = 10$.

Positive correlations		
First topic ID	Second topic ID	Correlation
#2	#0	0.2181
#3	#0	0.1961
#3	#1	0.1806
#4	#0	0.2004
#4	#2	0.1767
#4	#1	0.1766
#5	#0	0.1483
#5	#2	0.1212
#5	#3	0.0806
#5	#1	0.0524
#6	#1	0.1893
#6	#2	0.1456
#6	#0	0.1434
#6	#3	0.1305
#6	#4	0.1233
#7	#0	0.0766
#7	#5	0.0586
#7	#3	0.0560
#7	#2	0.0518
#7	#4	0.0215
#7	#1	0.0046
#8	#5	0.0512
Negative correlations		
First topic ID	Second topic ID	Correlation
#8	#0	−0.0168
#8	#2	−0.0548
#8	#4	−0.0749
#8	#3	−0.0958
#8	#6	−0.1222
#8	#1	−0.1467
#9	#7	−0.1208
#9	#6	−0.1460
#9	#3	−0.1580
#9	#1	−0.1581
#9	#2	−0.1702
#9	#4	−0.1880
#9	#0	−0.2167
#9	#5	−0.2326

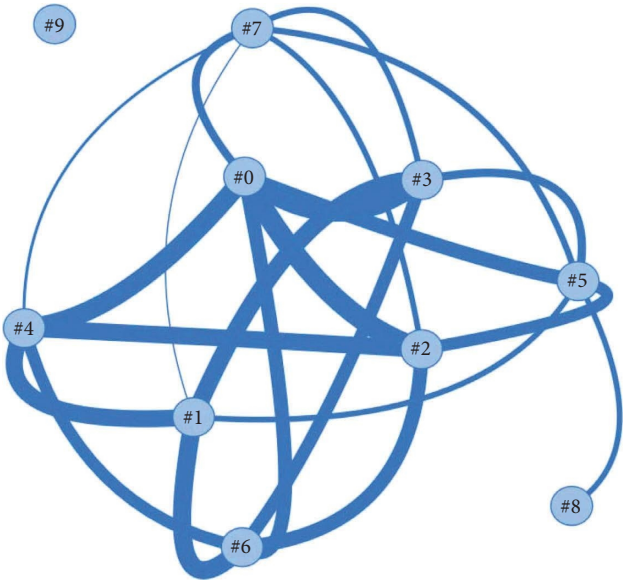


Figure 8: Correlations network for CTM with $k = 10$.

Table 9: Top 20 words of the topics generated by CTM with $k = 10$.

T#0	*T*#1	*T*#2	*T*#3	*T*#4	*T*#5	*T*#6	*T*#7	*T*#8	*T*#9
101880	83466	85822	105763	98285	211930	149347	311102	373565	435410
Communication	Drug deals	General use	Networking and security	Counterfeit items	Financial matters	Products	Shipping	Dealings and payments	Community
Drug	Utopia	Mirror	Pgp	Record	Guide	Quality	Order	Account	Vendor
Ice	Hash	Use	Key	Pack	Cashout	Ship	Time	Bank	Get
Arab	Signature	Onionshop	Wallet	Ssn	Fraud	Gram	Say	Need	Wsm
Smash	Gramo	Code	Sign	Extra	Method	Eur	Review	Card	Market
Sand	Kush	Flight	Tail	High	Chat	Sample	Like	Look	Help
Nun	Eur	Break	Begin	Retirement	Room	Product	Try	Drop	Page
Testicle	Haze	Deserve	Public	Passport	Card	Price	First	Contact	Rule
Eskimo	Sha	Among	Badget	Software	Expectus	Pill	Open	Buy	Good
Consider	Amphetamine	Honor	Encrypt	License	Day	Stealth	Really	Link	New
Phone	hcl	Hotel	Message	Net	Money	High	Buyer	Want	Offer
Name	Ketamine	Hero	tor	Ssns	Change	Black	Euro	Paypal	thank
Sock	Cocaine	Per	Use	Lesson	Bank	Cannabis	Note	Service	Phished
vpn	Heroin	Thieve	Refund	Driver	Business	Kallashnikov	Support	Sell	Send
Rdp	Weed	Supermod	Onion	Early	Lack	Weed	Take	Fullz	Work
Number	thc	Keep	File	Emv	One	track	Ask	Transfer	List
E-mail	Sign	Shoutouts	Electrum	Level	Everythingcc	mdma	thing	Btc	Please
Data	mdma	Overdue	Address	Worth	Debit	Customer	Dispute	External	Profile
Proxy	Top	Ghb	Window	Unite	Credit	Stock	Something	Log	Guy
Line	Ecstasy	Boneskoopa	Act	Visa	Dream	Package	Receive	Allow	Best
Type	Unregistered	Dude	Absolutely	Personal	Gold	Lsd	Think	Interest	Forum

T#~ refers to the topic ID with the number of vocabularies in each topic below it.

Table 10: Subtopic distributions over supertopics for PAM with $k1 = 3$ and $k2 = 10$.

Supertopic #0		Supertopic #1		Supertopic #2	
Subtopic ID	Probability	Subtopic ID	Probability	Subtopic ID	Probability
9	0.2716	9	0.2571	9	0.2802
1	0.1216	2	0.1132	1	0.1210
6	0.1158	1	0.1123	6	0.1205
2	0.1070	6	0.1117	2	0.1146
0	0.0921	0	0.1015	0	0.0873
3	0.0749	3	0.0758	3	0.0703
7	0.0596	8	0.0668	7	0.0588
8	0.0545	7	0.0579	8	0.0544
5	0.0529	4	0.0562	4	0.0543
4	0.0500	5	0.0475	5	0.0385

Table 11: Top 20 words of the topics generated by PAM with $k2 = 10$.

T#0	*T*#1	*T*#2	*T*#3	*T*#4	*T*#5	*T*#6	*T*#7	*T*#8	*T*#9
168480	282811	248213	144843	58987	53429	234668	109407	66144	589588
Drug deals	Financial matters	Communication	Community	Counterfeit items	Prescriptions	General use	Payments	Networking and security	Community
Eur	Bank	Card	Rule	Pgp	Adderall	Use	Transfer	Pgp	Order
Utopia	Guide	Room	Wsm	Signature	Gold	Mirror	Paypal	Signature	Vendor
Gram	Cashout	Chat	Phished	Begin	Quickship	Key	Card	Begin	Get
Weed	Fullz	Method	External	IdSOF	Euro	Onionshop	Account	Badget	Like
Quality	Card	Bank	Help	License	Quality	Drug	Flight	Sha	Good
Kush	Account	Cashout	Contact	Driver	Pgp	Sand	Need	Sign	Know
Ship	Record	Fraud	Get	Fake	Sample	Smash	Hotel	End	Review
Price	SSN	Log	Allow	Passport	Stock	Arab	Contact	Message	Time
Gramo	Expectus	Drop	Market	Sign	Suboxone	Nun	Service	Hash	Say
Order	Offer	Day	Vendor	Original	mdma	Testicle	Gift	Badge	Make
Haze	Credit	Credit	Please	Name	Stealth	Eskimo	Code	Gnupg	Ship
Cannabis	Drop	Visa	Bump	Message	Pill	Ice	Work	Version	Wsm
Strain	Pack	Page	Store	Badget	Begin	Consider	Fund	Unite	Dispute
Refund	High	Lack	Link	Sha	Signature	Pgp	Look	Block	Market
Reship	Extra	Change	Account	End	Hash	tor	Deal	Vendor	Look
Top	Link	Everythingcc	Page	Quality	Dreamwe	Tail	Dabbingtimes	Public	Send
Product	Check	One	Close	Apple	Pharmacy	Wallet	Btc	Key	People
Free	Need	Rdp	Profile	Cvv	Promptly	File	Med	Que	Note
Sample	Balance	Money	Shop	High	Partysquadnl	Click	Supermod	York	Try
Hash	Scan	Chaser	Forum	Hash	Interest	Encrypt	Thieve	State	Help

T#~ refers to the topic ID with the number of vocabularies in each topic below it.

Table 12: Top 20 words of the topics generated by PTM with $k = 10$.

T#0	*T*#1	*T*#2	*T*#3	*T*#4	*T*#5	*T*#6	*T*#7	*T*#8	*T*#9
216487	50034	323949	184949	97436	49958	56195	800143	53164	124255
Communication	Drug deals	Drug deals	Networking	Counterfeit items	Community	Security	Community	Promoting	Drugs
Guide	Mirror	Ship	Use	Passport	Supermod	Pgp	Vendor	Visa	Eur
Cashout	Onionshop	Like	Key	License	Thieve	Signature	Get	Credit	Utopia
Bank	Use	mdma	Address	Fake	Overdue	Begin	Wsm	Mastercard	Gram
Record	Euro	Review	Pgp	Scan	Shoutouts	Sign	Account	Self	Weed
Expectus	Ghb	Product	tor	Driver	Hero	Badget	Help	Classic	Kush
Ssn	Pill	Note	Phone	Register	Pgp	Sha	Need	und	Gramo
Card	Adderall	Sample	Tail	Com	Among	Message	Market	Debit	Price
Extra	Per	Order	Card	Quality	Honor	Hash	Bank	Bank	Order
High	Suboxone	Time	vpn	Bill	Deserve	End	Rule	Unite	Reship
Pack	Liquid	Good	Wallet	Porn	Dabbingtimes	Badge	Page	Ich	Refund
Offer	Pgp	Test	File	Original	Boneskoopa	Gnupg	Drop	Platinum	Quality
Link	Signature	Make	Server	Buy	Dude	Version	Card	Pgp	Hash
Credit	Ritalin	Cocaine	Kallashnikov	High	Dont	Vendor	Contact	Wir	Haze
Sale	Stock	Say	Name	Lifetime	Code	Key	Look	mit	Top
Many	Cialis	Take	Click	Onion	Break	Public	Please	Die	Strain
Level	url	People	Encrypt	pas	Begin	Page	Phished	Gold	Cannabis
Balance	Sample	Gold	Password	Document	Signature	Thomas	Paypal	Ltd	Bud
Worth	Begin	Come	Number	Unregistered	Put	Establishment	Order	Signature	Ship
Business	Methylphenidate	Free	Sock	Name	Keep	Jefferson	Method	State	Per
Check	Basf	Think	Public	Wallst	Long	Get	Thank	Offfers	Offer

T#~ refers to the topic ID with the number of vocabularies in each topic below it.

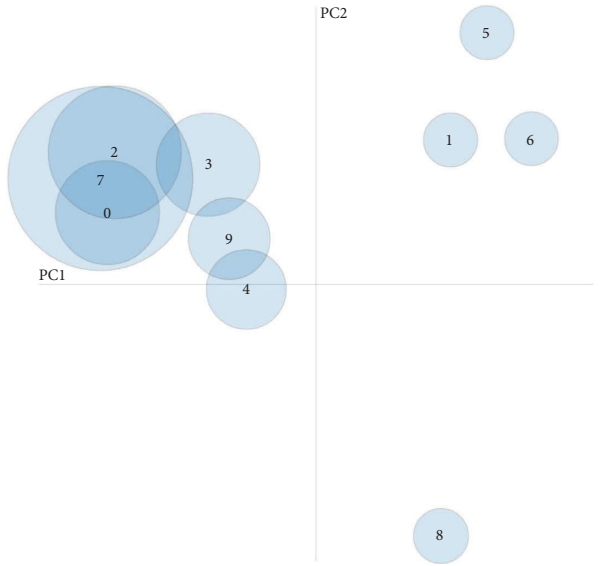


Figure 9: Topics map for PTM with $k = 10$.

and hierarchical PAM (HPAM), will be considered to extend the empirical analysis and comparison.

6. Conclusion

Topic modeling is a promising methodology to analyze the contents of social networks semantically and correlatedly. Social networks on the Dark Web are no exception. Forums associated with cryptomarkets are fraught with discussions about criminal behaviors and illegal business perpetrating. We introduced an approach to model the latent topics and their correlations from a forum of illicit and malicious activities on the Dark Web into thematic patterns. We emphasized the significance of choosing the appropriate preprocessing and cleaning procedures on which the accuracy and quality of the topic models primarily depend. We used four topic modeling algorithms: LDA, CTM, PAM, and PTM, and four coherence measures: UMass, UCI, CNPMI, and CV, and discussed their performance and outcomes for the studied dataset. According to these evaluation metrics, we examined the most coherent topics produced by each model to choose the optimal number of topics for each method. Subsequently, we visualized the results as labeled groups of semantically associated terms, including the relationships among topics for CTM and PAM. Lastly, we discussed limitations, challenges, and future work. Analyzing discussions and contents on the Dark Web can be tremendously advantageous to sociologists, criminologists, psychologists, law enforcement agencies, cybersecurity agencies, and many others. This study presents a leading start for further research in the field by providing a comprehensive approach to extracting hidden thought patterns from the Dark Web.

Data Availability

The dataset used to support the findings of this study was retrieved from an open source repository "Dark Net Forums Datasets" provided by AZSecure and can be viewed and downloaded from the URL: https://www.azsecure-data.org/dark-net-markets.html.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Supplementary Materials

A detailed list of the abbreviations and symbols used in this paper can be found in the Appendix as a ready reference. (*Supplementary Materials*)

References

[1] C. Easttom, "Dark web forensics," 2019, https://www.eccouncil.org/dark-web-forensics/.

[2] K. Porter, "Analyzing the DarkNetMarkets subreddit for evolutions of tools and trends using LDA topic modeling," *Digital Investigation*, vol. 26, pp. S87–S97, 2018.

[3] B. Akhgar, P. Bertrand, C. Chalanouli et al., "TENSOR: retrieval and analysis of heterogeneous online content for terrorist activity recognition," in *Proceedings of the Estonian Academy of Security Sciences*, Tallinn, Estonia, September 2017.

[4] R. Basheer and B. Alkhatib, "Threats from the dark: a review over dark web investigation research for cyber threat intelligence," *Journal of Computer Networks and Communications*, vol. 2021, Article ID 130299, 21 pages, 2021.

[5] Z. Csutak, "The maze of networks, the social impact and security risks of 21st century technologies," in *DIGITAL SECURITY POLICY IN THE CYBER SPACE*, T. Babos, Ed., pp. 61–81, Hungarian University of Agriculture and Life Sciences, Budapest, Hungary, 2022.

[6] M. Sawicka, I. Rafanell, and A. Bancroft, "Digital localisation in an illicit market space: interactional creation of a psychedelic assemblage in a darknet community of exchange," *International Journal of Drug Policy*, vol. 100, Article ID 103514, 2022.

[7] A. Kigerl, "Profiling cybercriminals: topic model clustering of carding forum member comment histories," *Social Science Computer Review*, vol. 36, no. 5, pp. 591–609, 2018.

[8] I. Pete, J. Hughes, Y. T. Chua, and M. Bada, "A social network analysis and comparison of six DarkWeb forums," in *Proceedings of the 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, Genoa, Italy, September 2020.

[9] K. H. Kwon and C. Shao, "Dark knowledge and platform governance: a case of an illicit E-commerce community in Reddit," *American Behavioral Scientist*, vol. 65, no. 6, pp. 779–799, 2021.

[10] C. Heistracher, S. Schlarb, and F. Ghaffar, "Information extraction from darknet market advertisements and forums," in *Proceedings of the 14th International Conference on Emerging Security Information, Systems and Technologies (SECURWARE 2020)*, Valencia, Spain, December 2020.

[11] S. Y. Cho and J. Wright, "Into the dark: a case study of banned darknet drug forums," in *Proceedings of the Social Informatics 11th International Conference, SocInfo*, pp. 18–21, Doha, Qatar, November 2019.

[12] K. H. Kwon and C. Shao, "Communicative constitution of illicit online trade collectives: an exploration of darkweb

market subreddits," in *Proceedings of the International Conference on Social Media and Society (SMSociety'20)*, Toronto, Canada, July 2020.

[13] J. Yang, H. Ye, and F. Zou, "pyDNetTopic: a framework for uncovering what darknet market users talking about," in *Proceedings of the 16th EAI International Conference on Security and Privacy in Communication Networks (SecureComm 2020)*, Washington, DC, USA, August 2020.

[14] M. Faizan and R. A. Khan, "A two-step dimensionality reduction scheme for DarkWeb text classification," in *Ambient Communications and Computer Systems RACCCS 2019*, OUCI, Ajmer, India, 2020.

[15] M. Ch A S and P. H. Rughani, "Sentiment & pattern analysis for identifying nature of the content hosted in the dark web," *Indian Journal of Computer Science and Engineering (IJCSE)*, vol. 12, no. 6, pp. 1822–1836, 2021.

[16] A. Joshi, E. Fidalgo, E. Alegre, and M. A. Nabki, "Extractive text summarization in dark web: a preliminary study," in *Proceedings of the 1st International Conference on Applications of Intelligent Systems (APPIS 2018)*, Las Palmas, Spain, January 2018.

[17] K. H. Kwon and J. Shakarian, "Black-hat hackers' crisis information processing in the darknet: a case study of cyber underground market shutdowns," *Studies in Media and Communications*, vol. 17, pp. 113–135, 2018.

[18] K. Lee and H. Jung, "Dynamic semantic network analysis for identifying the concept and scope of social sustainability," *Journal of Cleaner Production*, vol. 233, pp. 1510–1524, 2019.

[19] S. Barbon Jr., G. M. Tavares, and G. S. Kido, "Artificial and natural topic detection in online social networks," *iSys Brazilian Journal of Information Systems*, vol. 10, no. 1, pp. 80–98, 2017.

[20] H. Wandabwa, M. A. Naeem, R. Pears, and F. Mirza, "A metamodel enabled approach for discovery of coherent topics in short text microblogs," *IEEE Access*, vol. 6, pp. 65582–65593, 2018.

[21] H. Saif, T. Dickinson, L. Kastler, M. Fernandez, and H. Alani, "A semantic graph-based approachfor radicalisation detection on social media," in *Proceedings of the The Semantic Web: 14th International Conference*, Portorož, Slovenia, May 2017.

[22] B. V. Barde and A. M. Bainwad, "An overview of topic modeling methods and tools," in *Proceedings of the 2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, June 2017.

[23] R. Churchill and L. Singh, "The evolution of topic modeling," *ACM Computing Surveys*, vol. 54, no. 10s, pp. 1–35, 2022.

[24] D. Camacho, Á. Panizo-Lledot, G. Bello-Orgaz, A. Gonzalez-Pardo, and E. Cambria, "The four dimensions of social network analysis: an overview of research methods, applications, and software tools," *Information Fusion*, vol. 63, pp. 88–120, 2020.

[25] F. Viegas, S. Canuto, C. Gomes et al., "CluWords: exploiting SemanticWord clustering representation for enhanced topic modeling," in *Proceedings of the The Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19)*, Melbourne, Australia, February 2019.

[26] B. A. H. Murshed, S. Mallappa, J. Abawajy, M. A. N. Saif, H. D. E. Al-ariki, and H. M. Abdulwahab, "Short text topic modelling approaches in the context of big data: taxonomy, survey, and analysis," *Artificial Intelligence Review*, vol. 56, no. 6, pp. 5133–5260, 2022.

[27] H. H. Kim and H. Y. Rhee, "An ontology-based labeling of influential topics using topic network analysis," *Journal of Information Processing Systems*, vol. 15, no. 5, pp. 1096–1107, 2019.

[28] P. Kherwa and P. Bansal, "Topic modeling: a comprehensive review," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 7, no. 24, p. e2, 2019.

[29] S. A. Ríos, F. Aguilera, J. D. Nuñez-Gonzalez, and M. Graña, "Semantically enhanced network analysis for influencer identification in online social networks," *Neurocomputing*, vol. 326-327, pp. 71–81, 2019.

[30] S. Likhitha, B. S. Harish, and H. M. K. Kumar, "A detailed survey on topic modeling for document and short text data," *International Journal of Computer Application*, vol. 178, no. 39, pp. 1–9, 2019.

[31] R. Albalawi, T. H. Yeap, and M. Benyoucef, "Using topic modeling methods for short-text data: a comparative analysis," *Frontiers in Artificial Intelligence*, vol. 3, p. 42, 2020.

[32] I. Vayansky and S. A. Kumar, "A review of topic modeling methods," *Information Systems*, vol. 94, Article ID 101582, 2020.

[33] D. Ruan, J. Han, Y. Dang, S. Zhang, and K. Gao, "Modeling on micro-blog topic detection based on semantic dependency," in *Proceedings of the The 9th International Conference on Modelling, Identification and Control (ICMIC 2017)*, Kunming, China, July 2017.

[34] D. Valdez, A. C. Pickett, and P. Goodson, "Topic modeling: latent semantic analysis for the social sciences," *Social Science Quarterly*, vol. 99, no. 5, pp. 1665–1679, 2018.

[35] S. Bayrakdar, I. Yucedag, M. Simsek, and I. A. Dogru, "Semantic analysis on social networks: a survey," *International Journal of Communication Systems*, vol. 33, no. 11, p. e4424, 2020.

[36] M. Baranowski and P. Cichocki, "Good and bad sociology: does topic modelling make a difference?" *Society Register*, vol. 5, no. 4, pp. 7–22, 2021.

[37] A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat, and A. Hassan, "Topic modeling algorithms and applications: a survey," *Information Systems*, vol. 112, Article ID 102131, 2023.

[38] M. N. Asim, M. Wasim, M. U. G. Khan, W. Mahmood, and H. M. Abbasi, "A survey of ontology learning techniques and applications," *Database*, vol. 2018, 2018.

[39] S. Sbalchiero and M. Eder, "Topic modeling, long texts and the best number of topics. Some Problems and solutions," *Quality and Quantity*, vol. 54, no. 4, pp. 1095–1108, 2020.

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

[41] D. M. Blei and J. D. Lafferty, "Correlated topic models," *Advances in Neural Information Processing Systems*, vol. 18, p. 147, 2006.

[42] C. Bouveyron, P. Latouche, and R. Zreik, "The stochastic topic block model for the clustering of vertices in networks with textual edges," *Statistics and Computing*, vol. 28, no. 1, pp. 11–31, 2018.

[43] W. Li and A. McCallum, "Pachinko allocation: DAG-structured mixture models of topic correlations," in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, USA, June 2006.

[44] Y. Zuo, J. Wu, H. Zhang et al., "Topic modeling of short texts: a pseudo-document view," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, January 2016.

[45] R. Pandey and G. O. Mohler, "Evaluation of crime topic models: topic coherence vs spatial crime concentration," in *Proceedings of the 2018 IEEE International Conference on*

Intelligence and Security Informatics (ISI), Miami, FL, USA, November 2018.

[46] Y. Zhang, N. Ning, J. Lv, C. Song, and B. Wu, "Jointly modeling community and topic in social network," in *Proceedings of the 12th International Conference on Knowledge Science engineering and management (KSEM 2019) part of the lecture notes in computer science book series (LNAI)*, Athens, Greece, June 2019.

[47] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, Shanghai, China, February 2015.

[48] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, UK, July 2011.

[49] A. R. Hadiat, *Topic Modeling Evaluations: The Relationship between Coherency and Accuracy*, University of Groningen, Groningen, Netherlands, 2022.

[50] P.-Y. Du, N. Zhang, M. Ebrahimi et al., "Identifying, collecting, and presenting hacker community data: forums, IRC, carding shops, and DNMs," in *Proceedings of the 2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, Miami, FL, USA, July 2018.

[51] C.-X. Zhang, R. Liu, X.-Y. Gao, and B. Yu, "Graph convolutional network for word sense disambiguation," *Discrete Dynamics in Nature and Society*, vol. 2021, Article ID 2822126, 12 pages, 2021.

[52] Y. Hu, J. Boyd-Graber, B. Satinoff, and A. Smith, "Interactive topic modeling," *Machine Learning*, vol. 95, no. 3, pp. 423–469, 2014.