systems

MDPI

Article

Unveiling Gig Economy Trends via Topic Modeling and Big Data

Oya Ütük Bayılmış [1,2], Serdar Orhan [3,*] and Cüneyt Bayılmış [4]

[1] Labour Economics and Industrial Relations, Institute of Social Sciences, Sakarya University, 54050 Sakarya, Türkiye; oya.bayilmis1@ogr.sakarya.edu.tr or oyabayilmis@subu.edu.tr
[2] Sakarya Vocational School, Sakarya University of Applied Sciences, 54290 Sakarya, Türkiye
[3] Labour Economics and Industrial Relations, Faculty of Political Sciences, Sakarya University, 54050 Sakarya, Türkiye
[4] Computer Engineering, Faculty of Computer and Information Sciences, Sakarya University, 54050 Sakarya, Türkiye; cbayilmis@sakarya.edu.tr
* Correspondence: sorhan@sakarya.edu.tr

Abstract

The gig economy, driven by flexible and platform-based work, is reshaping labor markets and employment norms. Understanding public perceptions of this shift is critical for promoting social good and informing equitable policy. This study employs big data analytics and Latent Dirichlet Allocation (LDA) topic modeling to analyze 15,259 tweets collected from the X platform. Seven key themes emerged from the data, including labor precarity, flexibility, algorithmic control, platform accountability, gender disparities, and worker rights. While some users emphasized autonomy and new income opportunities, most expressed concerns about job insecurity, lack of protections, and digital exploitation. These findings offer real-time insights into how gig work is discussed and contested in public discourse. The study highlights how social media analytics can inform labor policy, guide platform regulation, and support advocacy efforts aimed at building a fairer and more resilient gig economy.

Keywords: gig economy; big data; natural language processing; Latent Dirichlet Allocation algorithm; topic modeling; social media analysis

check for updates

Academic Editor: Sitalakshmi Venkatraman

Received: 6 May 2025
Revised: 29 June 2025
Accepted: 5 July 2025
Published: 8 July 2025

Citation: Bayılmış, O.Ü.; Orhan, S.; Bayılmış, C. Unveiling Gig Economy Trends via Topic Modeling and Big Data. *Systems* **2025**, *13*, 553. https://doi.org/10.3390/systems13070553

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Digital transformation and technological advances have led to fundamental changes in the labor market, replacing traditional forms of work with more flexible and independent models. One of the most striking elements of this transformation is the gig economy, where individuals work in short-term, project-based jobs. The gig economy stands out as a system that allows individuals to offer short-term services for specific jobs through digital platforms such as Uber, Upwork, and Fiverr. Although it offers advantages such as flexibility, the opportunity to earn additional income, and geographical independence, the gig economy also brings with it significant structural problems such as a lack of social security, income instability, and a lack of job security [1–3].

In this context, understanding the social perception of the gig economy and its impacts on individuals is critical for both academic research and policymakers. Traditional survey and interview methods may fail to provide a holistic picture of perceptions of the gig economy due to methodological limitations such as limited sample sizes and response trends. In contrast, social media platforms provide large-scale data repositories where users freely and in real time share their thoughts, experiences and concerns about the gig economy. Analyzing the content collected from these platforms provides a valuable

resource for understanding individuals' perspectives, social dynamics and new business trends. This method is considered within the scope of social network analysis (SNA) and allows for the study of the collective discourse of individuals through data-driven methods [4].

The significance of understanding public perceptions of the gig economy extends beyond academic curiosity; it presents concrete opportunities to inform inclusive and evidence-based policy development. Social media platforms function not only as spaces for personal expression but also as collective arenas where structural inequalities, labor grievances, and demands for fairness in the digital workplace are voiced in real time. By analyzing these narratives, this study contributes to policymaking in key areas such as labor rights, wage protection, platform accountability, unionization pathways, and gender-inclusive employment strategies. Recognizing and systematically mapping these concerns aligns with the broader pursuit of social good by supporting the design of more equitable and resilient labor systems in the age of platform capitalism. This approach helps bridge the gap between public sentiment and regulatory intervention, enabling stakeholders to respond proactively to the evolving nature of work.

While previous research on the gig economy has offered valuable insights into its dual nature—balancing flexibility and opportunity with precarity and exploitation, these studies have predominantly relied on interviews, surveys, or theoretical analyses. This study contributes to existing literature by introducing a novel, data-driven approach to understanding how the gig economy is perceived in real time by large and diverse populations. By applying Latent Dirichlet Allocation (LDA) topic modeling to a corpus of over 15,000 social media posts, we capture organic, unprompted, and temporally situated public discourse. This computational social science perspective complements conventional labor studies by uncovering emergent themes and collective discursive patterns that may not surface through traditional methods. In doing so, our research offers a scalable and timely lens into public sentiment, thereby contributing both to academic debates and to the development of more responsive and inclusive policy frameworks.

This study aims to explore the main themes and trends discussed by social media users regarding the gig economy and to investigate how big data analytics and natural language processing (NLP) techniques can be utilized to understand its social impacts. By applying LDA topic modeling to social media data, the study identifies key narratives, sentiments, and concerns associated with the gig economy. The findings contribute to the growing intersection of big data analytics and social good by providing empirical insights into the societal perceptions of the gig economy. Furthermore, this research demonstrates the potential of social media content analysis as a scalable and real-time approach to inform policymakers, platform developers, and labor organizations seeking to promote fairer and more resilient gig work ecosystems.

In methodological terms, this study adopts an unsupervised machine learning approach to extract latent public concerns from a large, real-time dataset of 15,259 tweets. Through LDA topic modeling, the analysis identifies seven central themes that dominate social media discourse on the gig economy—ranging from job flexibility and autonomy to algorithmic control, precarity, and demands for labor rights. This approach offers a scalable and data-driven lens to understand how platform-based labor is publicly debated, thus complementing existing interview- or survey-based studies with real-time, high-volume insights.

Accordingly, the research questions used to assess possible relationships are as follows:

RQ1: What are the main themes and trends users on the X platform (formerly Twitter) are discussing in relation to the gig economy?

RQ2: How can big data analytics and natural language processing methods be utilized to explore and understand the social impacts of the gig economy based on social media discussions?

2. Literature Review

The literature reviewed in this section lays the conceptual and methodological foundation for our study, focusing on three strands: structural features of the gig economy, social media as a site for public labor discourse, and topic modeling as an analytical tool for large-scale textual data.

The gig economy has become a central focus in labor market research, emphasizing its dual character: while it offers autonomy, flexibility, and opportunities for income diversification, it also raises concerns regarding job insecurity, algorithmic management, and worker exploitation. De Stefano [5] highlights how digital labor platforms blur the boundary between employment and self-employment, complicating access to social protection. Vallas and Schor [6] interpret the gig economy as a neoliberal form of labor flexibilization, reinforcing power asymmetries between platforms and workers. Similarly, the deepening of precarity and the weakening of workers' bargaining power have been attributed to algorithmic control and rating systems [7,8]. Berg et al. [9] focus on income instability and the fragmentation of labor protections, while Meijerink and Keegan [10] argue that many gig workers experience a form of "pseudo-independence". A recent review by Pilatti et al. [3] emphasizes the structural inequalities embedded in platform labor, focusing on power dynamics, digital surveillance, and the mediating role of online social networks.

Social media platforms have increasingly become critical arenas for the formation and expression of public discourse. Platforms like X (formerly Twitter) allow users to share unfiltered, temporally grounded opinions on a variety of societal topics. Prior studies have leveraged such platforms to analyze discourse on public health [11], remote work [12], food waste [13], climate-related stress [14], urban planning [15], and utility service complaints [16]. These studies highlight the versatility of social media data in capturing public sentiment and emerging trends in real time. In the context of digital transformation, Sakas et al. [17] used big data from decentralized finance (DeFi) social media profiles to study supply chain management, while Abesinghe et al. [18] explored the perception of urban identity using geo-coded social media data. Despite the growing use of such data in other domains, social media-based analyses of the gig economy remain limited. Most existing gig work studies rely on structured interviews and surveys, often overlooking the dynamic, large-scale, and organic narratives emerging in online spaces.

Topic modeling has become a widely adopted method in computational social science for uncovering latent themes in large collections of textual data. Among these methods, Latent Dirichlet Allocation (LDA) stands out for its simplicity, scalability, and interpretability [19,20]. It has been successfully applied across domains including environmental communication, health discourse, service quality improvement, and public policy analysis [21–26]. Isoaho et al. [27] emphasize how topic modeling can complement qualitative methods by identifying thematic trends in policy texts.

Despite its popularity in various research fields [21–26], the use of LDA in gig economy studies remains limited. However, these studies often lack real-time or large-scale social media input. Our study applies LDA to over 15,000 tweets, offering a scalable and timely perspective on public discourse around gig work, thus filling a critical gap in the literature.

3. Materials and Methods

This section provides a detailed discussion of the analysis methods and research processes employed to uncover social media users' perceptions of the gig economy using big

data collected from the X platform. Topic modeling is a powerful text mining method used to automatically detect hidden themes within large text datasets and to extract meaningful insights from unstructured text. Researchers increasingly rely on topic modeling for innovation discovery, inductive classification, and a deeper understanding of online audiences and social dynamics. The LDA algorithm was used for topic modeling due to its advantages such as semantic description, generalization capability, dimensionality reduction, and mixture modeling. These procedures are schematically illustrated in Figure 1.

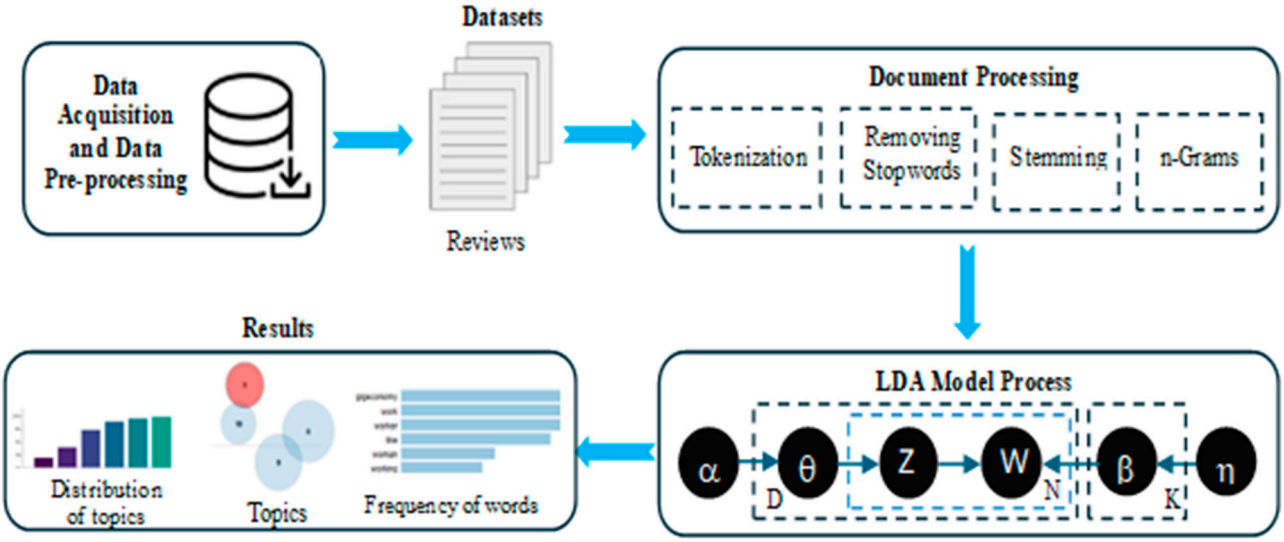


Figure 1. Methodological process followed in the research.

3.1. Data Acquisition and Data Preprocessing

In this study, a total of 15,259 tweets related to the gig economy were collected from the X platform using web scraping tools such as BeautifulSoup and the X Developer API (https://developer.x.com/en/docs/x-api, accessed on 1 April 2025). The X platform (formerly Twitter) offers a variety of features that facilitate interactions between users, and tweets often contain rich, user-generated hashtags that serve as keywords. Since the gathered information constitutes publicly available data, there were no ethical concerns regarding its use in the research process.

Both data collection and the implementation of the application were carried out using the Python (3.11.13) programming language. To enable effective analysis of the dataset obtained from the X platform, a series of data preprocessing steps were undertaken. During this stage, empty, duplicate, and spam tweets were filtered out to refine the dataset used for analysis. Additionally, unnecessary columns, as well as rows containing NaN (Not a Number) and NaT (Not a Time) values in the tweet body column, were removed. This was followed by a series of transformation procedures to prepare the data for further analysis.

The data acquisition and preprocessing steps were implemented using Python libraries such as BeautifulSoup, pandas, nltk, and re for web scraping, data cleaning, and text normalization. For topic modeling, the gensim library was used to implement the LDA algorithm. Visualization of topic distributions and keyword patterns was conducted using matplotlib, seaborn, wordcloud, and pyLDAvis. The scripts developed for this study are available from the corresponding author upon reasonable request.

3.2. Document Processing

Tweets are typically short, highly unstructured, and noisy text, which makes it difficult to extract coherent topics using topic modeling algorithms due to the low frequency of terms occurring together. It is therefore important to preprocess the collected data by

converting the unstructured text into a cleaner and more structured format suitable for topic modeling. The main goal of the preprocessing step is to improve the accuracy of topic extraction. Tweets often contain misspellings, gibberish, symbols, irrelevant characters, emoticons, and many stop words, prepositions, and punctuation marks.

The following preprocessing steps were applied to prepare the dataset for exploratory data analysis and topic modeling:

- Text normalization: All tweets were converted to lowercase to ensure consistency.
- Tokenization: Non-alphabetic characters, URLs, hyperlinks, emojis, special characters, and user mentions were removed.
- Removing stopwords: Frequently used words that do not contribute to the semantic meaning of documents, along with domain-specific stopwords, were eliminated.
- Stemming: Words were reduced to their base or root forms to consolidate different morphological variations into a single representation.
- n-Grams technique: Applied to identify meaningful word pairs and enhance the detection of contextual relationships.

These preprocessing steps resulted in a refined and structured dataset suitable for further analysis.

3.3. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation, developed by Blei et al. [19], is a generative probabilistic model designed for collections of documents. It operates on the assumption that each document is composed of a random mixture of latent topics, where each topic is characterized by a specific probability distribution over words. By modeling these distributions, LDA enables the automatic discovery of hidden thematic structures within large textual datasets. The algorithm leverages probabilistic associations between words and topics, making it particularly effective for uncovering semantic patterns and organizing unstructured text corpora [19,20].

The probabilistic generative process of the LDA algorithm is shown in Figure 1. In the LDA generative model, several key parameters define the process. The parameter $\boldsymbol{\alpha}$ controls the distribution of topics within each document, while θ represents the topic proportions for a specific document. Each word in a document is associated with a latent topic assignment, z, and the observed word itself is denoted by w. Meanwhile, β characterizes the distribution of words for each topic, and η controls the overall distribution of words across all topics. In this framework, documents are generated by first selecting topic distributions, then sampling a topic for each word, and finally choosing the actual word from the selected topic's distribution. This structure enables the identification of hidden thematic patterns within large text corpora.

Following the generative process where documents are represented as mixtures of topics and topics as distributions over words, the joint probability of all hidden and observed variables—including the topic distributions (β), document–topic mixtures (θ), topic assignments (z), and observed words (w)—is formally captured by the probabilistic model given in Equation (1). This equation establishes the probabilistic framework of LDA, which serves as the basis for the topic extraction performed in this research [26].

$$P(\beta_k, \theta_D, Z_D, W_D) = \prod_{k=1}^{K} P(\beta_k|\eta) \prod_{d=1}^{D} P(\theta_d|\alpha) \prod_{n=1}^{N} P(Z_{d,n}|\theta_d)\ P(W_{d,n}, \beta_K) \quad (1)$$

To see how each observed word w in document d is generated from the latent topics, we marginalize over z:

$$P(w \mid d) = \sum_{k=1}^{K} P(w \mid z_k) \times P(z_k \mid d) \quad (2)$$

where $P(z_k \mid d) = \theta_{d,k}$ is the proportion of topic k in document d, and $P(w \mid z_k) = \varnothing_{k,w}$ is the probability of word w under topic k.

Given its ability to uncover latent structures within large-scale textual data, LDA was deemed suitable for the objectives of this study.

To ensure theoretical validity in topic interpretation, the labeling process was guided by recurring constructs in gig economy literature. Themes such as job flexibility, precarity, platform governance, and algorithmic control were selected based on their prominence in prior scholarly works [5–9]. These dimensions represent key axes of debate in the platform labor domain. In addition, topics were grouped into positive, negative, and neutral categories following sentiment analysis conventions widely adopted in computational social science [21,27], allowing for structured interpretation of public sentiment.

The thematic labels for each topic were manually assigned by two authors with expertise in labor economics and employment research. This process involved reviewing the top keywords and representative tweets for each topic and collaboratively interpreting their semantic patterns to define titles that reflect the underlying discourse in a meaningful and context-sensitive manner.

Topic Diagnostics

To evaluate the quality of the extracted topics, we used topic coherence, exclusivity, tokens, and corpus distance diagnostics, as well as perplexity and global coherence score (c_v) metrics [28,29].

Perplexity measures the model's predictive performance by calculating the exponential of the negative log-likelihood of the test corpus, normalized by the total number of words [19]. It is computed as shown in Equation (3).

$$perplexity(D) = exp\left\{ -\frac{\sum_{d=1}^{|D|} logL(w_d|\theta_d)}{\sum_{d=1}^{|D|} N_d} \right\} \quad (3)$$

where D represents the document corpus, N_d is the number of terms in document d, and $L(w_d|\theta_d)$ is the likelihood function for document d. Lower perplexity values indicate better model performance, suggesting that the model can more accurately predict word distributions in unseen documents.

While perplexity typically decreases with more topics, this may indicate overfitting, making coherence an essential complementary metric.

Coherence (UMass) indicates how semantically similar words are within a topic; higher scores mean better consistency and easier interpretation. We measure semantic consistency among the top M words of topic k using the UMass coherence:

$$C_{UMass}(k) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} log \frac{D(w_m, w_l) + \epsilon}{D(w_l)} \quad (4)$$

where $D(w_m, w_l)$ is the number of documents containing both w_m and w_l, $D(w_l)$ is the number of documents containing w_l, and ϵ is a small smoothing constant.

Exclusivity gauges how unique a topic's terms are; higher scores mean those words rarely show up in other topics. We quantify how specific the top N words of topic k are relative to other topics:

$$Excl(k) = \sum_{w \in TopN_k} \frac{\varnothing_{k,w}}{\sum_{j \neq k} \varnothing_{j,w}} \tag{5}$$

where $TopN_k$ is the set of the top N words for topic k, $\varnothing_{k,w}$ is the probability of word w under topic k.

Tokens show the number of word instances in each topic, indicating its relative size or importance in the corpus. Formally,

$$Tokens(k) = \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbf{1}\left(z_{d,n} = k\right) \tag{6}$$

where **1** (.) is an indicator that equals 1 if the topic assignment $z_{d,n}$ for word n in document d is k and 0 otherwise.

Corpus distance indicates how much a topic's word distribution $\varnothing_k$ differs from the overall corpus distribution $\varnothing_{corpus}$, with larger values signifying greater thematic differentiation. We compute this via Jensen–Shannon divergence:

$$CorpusDist(k) = D_{JS}\left(\varnothing_k| \mid \varnothing_{corpus}\right) = \frac{1}{2} D_{KL}(\varnothing_k| \mid M) + \frac{1}{2} D_{KL}\left(\varnothing_{corpus}| \mid M\right) \tag{7}$$

where

$$M = \frac{1}{2}\left(\varnothing_k + \varnothing_{corpus}\right) \tag{8}$$

is the empirical word-frequency distribution over all documents,

$$\varnothing_{corpus} = \frac{1}{\sum_{d,n} 1} \sum_{d=1}^{D} \sum_{n=1}^{N_d} 1\left(w_{d,n} = w\right) \tag{9}$$

is the empirical word-frequency distribution over all documents, and

$$D_{KL}(P||Q) = \sum_{w} P(w)\, ln \frac{P(w)}{Q(w)} \tag{10}$$

is the Kullback–Leibler divergence. This metric helps assess how well individual topics are distinguished from the overall corpus vocabulary distribution.

In addition to these quantitative metrics, we employed LDAvis for interactive visualization of topic models. LDAvis enables users to adjust parameters and explore term–topic associations through an interactive interface that balances term frequency and topic specificity for better topic interpretation [30].

4. Results

This section presents the results of exploratory data analysis (EDA), word cloud analysis, and LDA topic modeling to understand social media user perceptions of the gig economy.

Table 1 provides a sample of randomly selected tweets that reflect diverse sentiments and viewpoints regarding the growth, impact, and opportunities associated with the gig economy.

Table 1. The randomly selected sample tweets from datasets.

Tweets
New age sectors are driving innovation, creating jobs, and boosting exports
The gig economy is booming global market projections 455 billion dollars in 2023 556 billion dollars estimated in 2024 Platforms like Work X are revolutionizing freelancing by empowering global talent to connect securely, transparently, and efficiently.
That the wages are poor in any system is a function of free markets that allow for job creation. The comparison for the workers is not if they would be happier with a cushy job in a cubicle. It is if they would prefer to not have the option of working in the gig economy
The gig economy is about to explode and its going to change everything in the next 10 years 50% of the U.S. workforce will be freelancers but here s the kicker they will actually make more than traditional employees
Remote work and telemedicine were two things that disabled people the gig economy is transforming the workforce, with platforms like uber and taskrabbit offering flexible, on-demand work opportunities

Figure 2 shows the relative frequencies of the twenty most prevalent words in the corpus. The word "economy" emerges as the most frequently occurring term. It is observed that the top 20 most frequent words are present in the randomly selected sample tweets provided in Table 1.

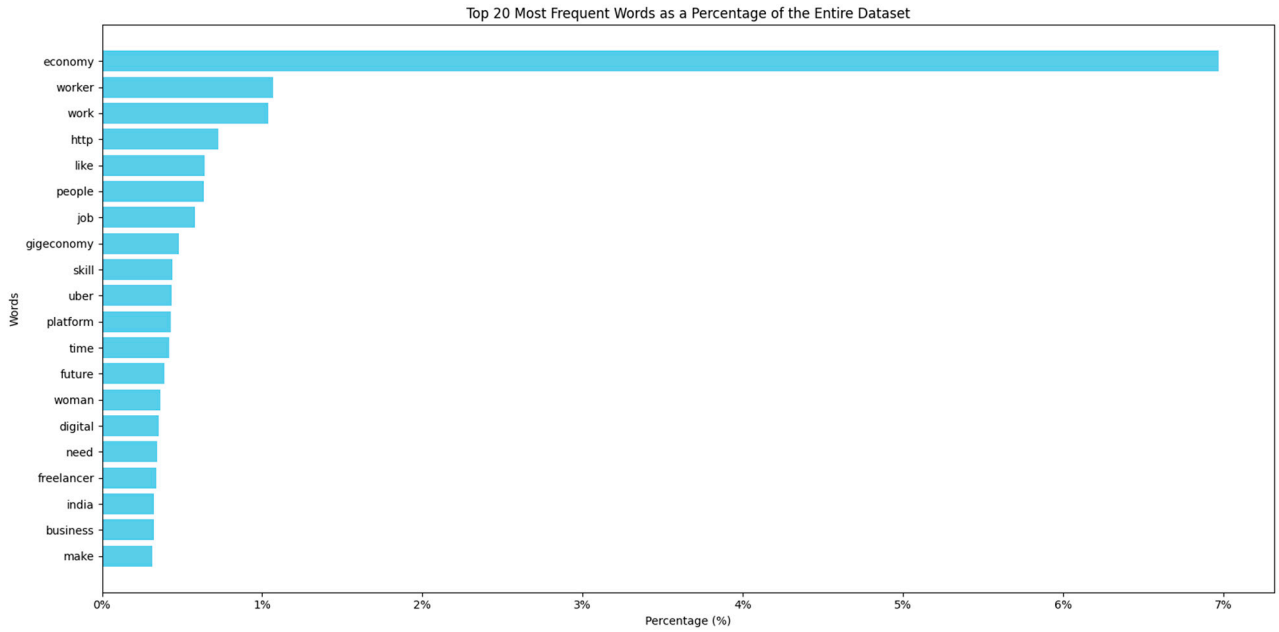


Figure 2. Top 20 most frequent words.

Word clouds provide researchers with quick and concise insights by visualizing words in text according to their frequency. Figure 3 shows a cloud plot of the most frequently used word pairs. In the figure, terms such as economy, worker, gig economy, work, people, platform, freelancer, digital, and time are prominent. This shows that the findings obtained through the word cloud visualization have a high degree of overlap with the themes analyzed.



Figure 3. The word cloud extracted from the tweet dataset.

To address RQ1, the analysis of the corpus reveals that users on the X platform predominantly discuss the gig economy in terms of economic opportunities, workforce transformation, freelancing, and flexible work models. Key terms such as "economy," "worker," "gig economy," and "platform" consistently emerge, indicating a strong emphasis on the structural and economic impacts of the gig economy.

When examining the tweets categorized into the seven different topics identified by LDA, the following explanations can be derived:

Topic 1—Workforce Dynamics and Labor Needs in the Gig Economy: This topic reflects discussions around workforce composition, emerging labor needs, and the evolving nature of work and service provision within the gig economy, particularly in the context of technological and business developments.

Topic 2—Freelancing Opportunities and Skill Development in the Gig Economy: The topic highlights discussions around freelance work opportunities, the importance of skill development, and future career prospects within the gig economy. It reflects how users perceive the gig economy as a platform for flexible employment and professional growth.

Topic 3—Platform-Based Food Delivery Services and Driver Experiences in the Gig Economy: The topic focuses on discussions related to platform-based food delivery services such as Uber and DoorDash, highlighting the experiences of drivers and workers operating within the gig economy. It also reflects the role of media coverage and reports in shaping perceptions about this segment.

Topic 4—Digitalization, Remote Work, and Online Platforms in the Gig Economy: This topic captures discussions around the digital transformation of work within the gig economy, focusing on remote work opportunities, online freelancing platforms, and payment systems that facilitate flexible employment in a digital environment.

Topic 5—Global Labor Market Shifts and Technological Disruption in the Gig Economy: The topic explores the impacts of technological advancements, decentralized platforms, and startup culture on labor markets, with a particular focus on unemployment issues and the gig economy's expansion in emerging economies such as India.

Topic 6—Women's Participation and Empowerment in the Gig Economy: This topic highlights discussions surrounding the participation of women in the gig economy, the need for supportive initiatives to enhance female workforce involvement, and global efforts

(such as those discussed at Davos and promoted through initiatives like CII for Women-Led Growth) aimed at fostering economic empowerment for women.

Topic 7—Labor Rights, Exploitation Risks, and Unionization Efforts in the Gig Economy: This topic examines concerns related to labor rights, potential exploitation in the gig economy, and efforts toward unionization and advocacy, with some discussions contextualized within specific regions such as Canada. It reflects a critical perspective on the need for greater protections for gig workers.

Table 2 presents the topic distributions derived from the topic modeling process, as well as the most salient keywords characterizing each topic.

Table 2. Distribution rates of topics within the document and frequently occurring words in the topics.

Topic Number	Keywords	Ratio
1	worker, people, like, work, need, business, tech, service, workforce, side	42.37%
2	gigeconomy, http, work, skill, future, time, freelane, opportunity, freelancer, read	18.56%
3	uber, delivery, many, part, article, driver, back, report, doordash, food	12.02%
4	digital, freelancing, platform, remote, better, work, payment, game, online, real	9.66%
5	india, data, labor, join, unemployment, technology, decentralized, publicai, like, startup	7.4%
6	woman, working, take, support, number, davos, rate, child, ciiforwomenledgrowth, ciicwl	5.88%
7	going, partner, exploitation, especially, anyone, research, canada, background, para, union	4.13%

Accounting for 42% of all tweets, Topic 1 emerges as the dominant theme, emphasizing discussions around workforce dynamics, labor requirements, and the influence of technology and service sectors within the gig economy.

Table 3 shows the percentage of tweets within each LDA-identified topic that were classified as positive, neutral, or negative. This breakdown highlights which themes (e.g., Topic 2 on skill development) are discussed most optimistically and which (e.g., Topic 7 on labor rights and exploitation) elicit more concern or negative sentiment.

Table 3. Sentiment distribution by topics.

Topic	Positive (%)	Neutral (%)	Negative (%)
1	56.3	21.7	22.0
2	65.5	25.9	8.6
3	52.4	20.7	26.9
4	61.0	22.8	16.1
5	63.5	13.4	23.1
6	50.2	10.4	39.5
7	45.2	21.2	33.6

Figure 4 presents the "Intertopic Distance Map," which provides an intuitive visualization of the similarities, differences, and relationships among the topics generated by the LDA model. The x- and y-axes represent coordinates derived from multidimensional scaling (MDS) based on Jensen–Shannon divergence between topic distributions. Each bubble represents a topic, with its size indicating the marginal topic distribution across the corpus. The bar chart on the right displays the top 30 most relevant terms associated with the selected topic. Red bars correspond to the estimated term frequency within the topic,

while blue bars reflect overall corpus frequency. The visualization was created using the LDAvis method, based on Sievert and Shirley [30].

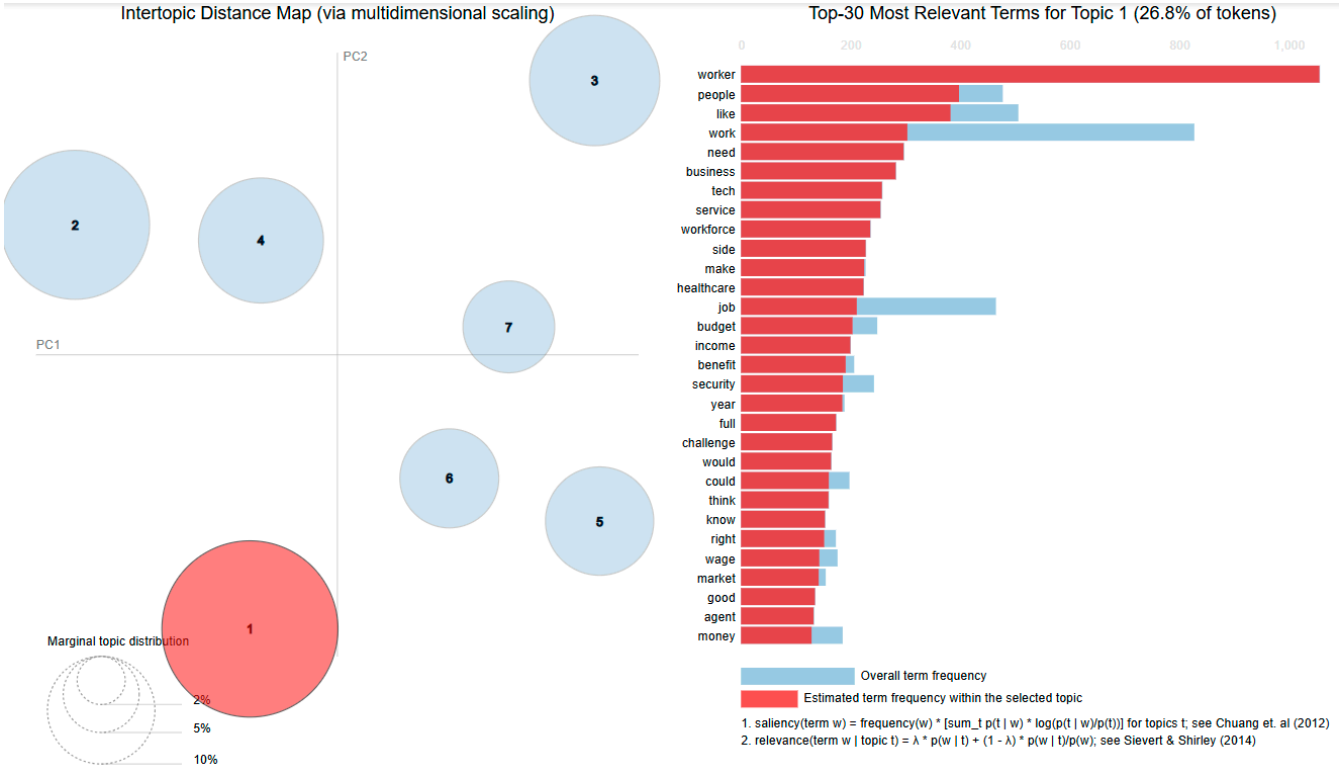


Figure 4. Intertopic distance map generated via multidimensional scaling (MDS) using Jensen–Shannon divergence between topics. Bubble size indicates the marginal topic distribution. The right-hand bar chart displays the top 30 most relevant terms for the selected topic. Red bars show term frequency within the topic, blue bars show overall corpus frequency [30,31].

To address RQ2, big data analytics and natural language processing techniques were employed to explore the social impacts of the gig economy through the analysis of Twitter discussions. After the dataset was preprocessed through cleaning, tokenization, and lemmatization, Latent Dirichlet Allocation was applied to uncover latent thematic structures. The analysis identified seven distinct topics: (1) Workforce Dynamics and Labor Needs (42.37%), (2) Freelancing Opportunities and Skill Development (18.56%), (3) Platform-Based Food Delivery Services and Driver Experiences (12.02%), (4) Digitalization, Remote Work, and Online Platforms (9.66%), (5) Global Labor Market Shifts and Technological Disruption (7.4%), (6) Women's Participation and Empowerment (5.88%), and (7) Labor Rights, Exploitation Risks, and Unionization Efforts (4.13%). Prominent keywords across these topics included worker, freelancer, platform, remote, delivery, technology, startup, woman, labor rights, and union. These findings demonstrate that big data and NLP methods enable the extraction of rich, multidimensional insights into how the gig economy is reshaping employment structures, social dynamics, and worker rights on a global scale.

Table 4 presents the coherence, exclusivity, tokens, and corpus distance values used to evaluate the quality of the topics extracted through LDA modeling in this study. Coherence reflects the degree of semantic similarity among the words within a topic, with higher values indicating greater internal consistency and interpretability. Exclusivity measures the uniqueness of the topic terms, where higher scores suggest that the words are less likely to appear in multiple topics. Tokens represent the number of word instances associated with each topic, illustrating the relative size or prominence of the topic within the corpus. Corpus distance indicates how much the word distribution of a specific topic diverges from

the overall distribution of the entire corpus, with larger values reflecting greater thematic differentiation [29].

Table 4. Diagnostic measurements for topics.

Topic	Coherence	Exclusivity	Tokens	Corpus Distance
1	0.218	0.853	16,771	1.188
2	0.306	0.881	11,987	1.545
3	0.385	0.941	9159	1.778
4	0.362	0.898	8439	1.810
5	0.400	0.927	6338	1.980
6	0.407	0.996	5290	2.027
7	0.592	0.992	4552	2.288

According to Table 4, Workforce Dynamics and Labor Needs in the Gig Economy emerges as the most dominant topic, representing the largest proportion of discussions in the dataset. A coherence value of 0.218 indicates that the words in this topic show moderate consistency with each other. An exclusivity score of 0.853, which is relatively high, suggests that the words specific to this topic are more unique compared to other topics. The token count of 16,771 is relatively high, meaning this topic occupies a significant space in the documents. A corpus distance value of 1.188 indicates that the word distribution diverges moderately from the overall dataset (corpus), but it is not an extreme outlier. While Table 4 reports per-topic semantic consistency using the UMass coherence metric, Table 5 presents a single, global c_v coherence score that evaluates the overall interpretability of the model's topic set.

Table 5. Model evaluation metrics.

Metric	Value
Perplexity	160.19
Coherence (c_v)	0.381

Table 5 reports the overall model fit and interpretability measures for our LDA model. The perplexity score of 160.19 indicates how well the model predicts held-out data (lower is better), while the c_v coherence of 0.381 reflects the semantic consistency of topics (higher is better).

5. Discussion

This study offers an exploratory analysis of public discussions about the gig economy on the X platform, leveraging big data analytics and natural language processing (NLP) techniques to identify key themes and emerging trends. By applying Latent Dirichlet Allocation (LDA) topic modeling, seven distinct themes were identified, reflecting the multifaceted nature of the gig economy as discussed by social media users. The results demonstrate that big data and NLP methods are effective tools for capturing complex and large-scale societal debates in digital environments.

The dominance of the "Workforce Dynamics and Labor Needs" topic, accounting for approximately 42% of all discussions, underscores the centrality of labor-related concerns in the gig economy discourse. Discussions frequently focused on the evolving structure of the workforce, emerging job demands, and the role of technology in reshaping employment landscapes. These findings suggest that users are critically aware of the transformations in traditional employment models and the shifting nature of work under platform capitalism.

Beyond workforce concerns, significant attention was given to freelancing opportunities, platform-based food delivery services, and remote work models. The emergence of topics highlighting freelancing and digital platforms indicates that flexibility, autonomy, and accessibility are perceived as both opportunities and challenges within the gig economy. However, while these models offer new forms of engagement, they also introduce precariousness and uncertainty, a tension that is consistently reflected across multiple topics.

The analysis further highlights critical social issues such as labor rights, exploitation risks, unionization efforts, and the gendered dimensions of gig work. Topics related to women's participation in the gig economy, as well as the risks of exploitation, showed high coherence and exclusivity values, indicating focused and significant discussion. These results point to the persistent inequalities and vulnerabilities embedded in gig work structures, emphasizing the need for inclusive policies and stronger protections for gig workers.

From a methodological perspective, the integration of big data analytics and NLP has proven valuable for analyzing large-scale, user-generated content. LDA topic modeling allowed the extraction of coherent and interpretable topics, providing a nuanced understanding of how social media users perceive and react to the gig economy. These findings illustrate the potential of computational social science approaches to contribute to broader debates on labor, technology, and society.

Overall, the findings of this study reveal the dual character of the gig economy: it is simultaneously a space of innovation and economic opportunity and a domain marked by labor precarity, inequality, and emerging risks. Future research could build upon these findings by conducting longitudinal analyses to observe how discussions evolve over time or by extending the dataset to include different social media platforms and geographic contexts. Such approaches would offer a more comprehensive view of the evolving social impacts of the gig economy.

While Twitter (X) is widely used for social commentary and public debate, we acknowledge that limiting the dataset to a single platform may constrain the generalizability of our findings. The user base of X may not fully reflect the demographics, behaviors, or discourse styles present across other social media platforms. Therefore, the conclusions presented in this study should be interpreted as indicative of public perceptions expressed specifically on the X platform. Future studies incorporating multi-platform data sources could enhance the representativeness and robustness of topic modeling results.

Existing literature on the gig economy has extensively addressed its dichotomous structure—highlighting both the flexibility it provides and the precarity it often entails. Numerous studies in labor economics and industrial relations have explored these issues through qualitative methods such as interviews, ethnographic accounts, and case studies. While our findings echo several themes already documented in the literature, this study offers a distinctive contribution by empirically capturing how these themes manifest in real-time public discourse on social media. Unlike self-reported or curated responses in traditional research, the social media corpus reflects spontaneous, large-scale perceptions from a diverse population. In this sense, our computational approach complements prior labor studies by identifying not only what issues matter in the gig economy, but also how they are collectively discussed, framed, and prioritized by the public.

6. Conclusions

This study leveraged big data analytics and natural language processing (NLP) techniques to explore how social media users perceive and discuss the gig economy. Through the application of Latent Dirichlet Allocation (LDA) topic modeling, seven major themes were identified, highlighting critical aspects such as workforce transformation, freelancing opportunities, digitalization, labor rights, and gender dynamics. The findings demonstrate

that while the gig economy offers flexibility and new opportunities, it simultaneously raises concerns regarding employment precarity, exploitation risks, and social inequalities.

The analysis further illustrates the value of computational approaches in capturing large-scale public discourse and extracting nuanced insights into complex societal phenomena. By uncovering diverse perspectives and thematic patterns within the corpus, this research contributes to a deeper understanding of the social impacts of the gig economy.

In contrast to traditional studies relying on structured data or curated interviews, this research introduces a computational perspective that captures the dynamic, organic, and collective nature of public sentiment. As gig work continues to evolve, such real-time perception analysis will be essential for informing adaptive policies that address not only labor market structures but also the lived experiences and expectations of platform workers.

Future studies could expand upon these findings by incorporating longitudinal analyses to observe changes over time or by comparing cross-platform discussions to assess variations in perceptions across different social media environments.

The findings of this study also offer practical insights for policymakers engaged in regulating platform-based labor markets. Public concerns regarding job insecurity, algorithmic control, exploitation risks, and gender disparities highlight specific areas for intervention. These themes may inform the development of regulatory frameworks aimed at improving working conditions, ensuring transparency in platform governance, and promoting equitable access and participation in the gig economy.

Author Contributions: Conceptualization, O.Ü.B., S.O. and C.B.; Methodology, O.Ü.B., S.O. and C.B.; Formal Analysis, O.Ü.B., S.O. and C.B.; Investigation, O.Ü.B., S.O. and C.B.; Validation, O.Ü.B., S.O. and C.B.; Writing—Original Draft, O.Ü.B., S.O. and C.B.; Writing—Review and Editing, O.Ü.B., S.O. and C.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study, collected from publicly available content on the X platform, is not deposited in a public repository due to the platform's terms of service and ethical considerations. However, it is available from the corresponding author upon reasonable request for academic and non-commercial purposes.

Acknowledgments: The authors thank the two anonymous referees for their suggestions that improved the work.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

EDA	Exploratory Data Analysis
LDA	Latent Dirichlet Algorithm
NaN	Not a Number
NaT	Not a Time
NLP	Natural Language Processing
SNA	Social Network Analysis
URL	Uniform Resource Loader

References

1. Woodcock, J.; Graham, M. *The Gig Economy: A Critical Introduciton*; Polity Press: London, UK, 2020; pp. 1–160.
2. Fulker, Z.; Riedl, C. Cooperation in the Gig Economy: Insights from Upwork Freelancers. *Proc. ACM Hum. Comput. Interact.* **2024**, *8*, 37. [CrossRef]
3. Pilatti, G.R.; Pinheiro, F.L.; Montini, A.A. Systematic literature review on gig economy: Power dynamics, worker autonomy, and the role of social networks. *Adm. Sci.* **2024**, *14*, 267. [CrossRef]
4. Scott, J. Social network analysis: Developments, advances, and prospects. *Soc. Netw. Anal. Min.* **2011**, *1*, 21–26. [CrossRef]
5. De Stefano, V. The rise of the "just-in-time workforce": On-demand work, crowdwork, and labor protection in the gig-economy. *Comp. Labor Law Policy J.* **2016**, *37*, 471–504. [CrossRef]
6. Vallas, S.; Schor, J. What do platforms do? Understanding the gig economy. *Annu. Rev. Sociol.* **2020**, *46*, 273–294. [CrossRef]
7. Graham, M.; Hjorth, I.; Lehdonvirta, V. Digital labour and development: Impacts of global digital labour platforms and the gig economy on worker livelihoods. *Transf. Eur. Rev. Labour Res.* **2017**, *23*, 135–162. [CrossRef]
8. Wood, A.J.; Graham, M.; Lehdonvirta, V.; Hjorth, I. Good gig, bad gig: Autonomy and algorithmic control in the global gig economy. *Work. Employ. Soc.* **2019**, *33*, 56–75. [CrossRef]
9. Berg, J.M.; Furrer, M.; Harmon, E.; Rani, U.; Silberman, M.S. *Digital Labour Platforms and the Future of Work: Towards Decent Work in the Online World*; ILO: Geneva, Switzerland, 2018; pp. 95–109.
10. Meijerink, J.; Keegan, A. Conceptualizing human resource management in the gig economy: Toward a platform ecosystem perspective. *J. Manag. Psychol.* **2019**, *34*, 214–232. [CrossRef]
11. Bogdanowicz, A.; Guan, C. Dynamic topic modeling of Twitter data during the COVID-19 pandemic. *PLoS ONE* **2022**, *17*, e0268669. [CrossRef]
12. Rojas Rincón, J.S.; Riveros Tarazona, A.R.; Mejía Martínez, A.M.; Acosta-Prado, J.C. Sentiment Analysis on Twitter-Based Teleworking in a Post-Pandemic COVID-19 Context. *Soc. Sci.* **2023**, *12*, 623. [CrossRef]
13. Jenkins, E.L.; Lukose, D.; Brennan, L.; Molenaar, A.; McCaffrey, T.A. Exploring Food Waste Conversations on Social Media: A Sentiment, Emotion, and Topic Analysis of Twitter Data. *Sustainability* **2023**, *15*, 13788. [CrossRef]
14. Bui, T.; Hannah, A.; Madria, S.; Nabaweesi, R.; Levin, E.; Wilson, M.; Nguyen, L. Emotional Health and Climate-Change-Related Stressor Extraction from Social Media: A Case Study Using Hurricane Harvey. *Mathematics* **2023**, *11*, 4910. [CrossRef]
15. Cui, N.; Malleson, N.; Houlden, V.; Yan, Y.; Comber, A. Using Twitter to understand spatial-temporal changes in urban green space topics based on structural topic modelling. *Cities* **2025**, *157*, 105601. [CrossRef]
16. Balta Kaç, S.; Eken, S. Customer Complaints-Based Water Quality Analysis. *Water* **2023**, *15*, 3171. [CrossRef]
17. Sakas, D.P.; Giannakopoulos, N.T.; Terzi, M.C.; Kanellos, N.; Liontakis, A. Digital Transformation Management of Supply Chain Firms Based on Big Data from DeFi Social Media Profiles. *Electronics* **2023**, *12*, 4219. [CrossRef]
18. Abesinghe, S.; Kankanamge, N.; Yigitcanlar, T.; Pancholi, S. Image of a City through Big Data Analytics: Colombo from the Lens of Geo-Coded Social Media Data. *Future Internet* **2023**, *15*, 32. [CrossRef]
19. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
20. Blei, D.M.; Carin, L.; Dunson, D. Probabilistic topic models. *IEEE Signal Process. Mag.* **2010**, *27*, 55–65. [CrossRef]
21. Jacobi, C.; Atteveldt, W.; Welbers, K. Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digit. J.* **2016**, *4*, 89–106. [CrossRef]
22. Calli, L.; Calli, F. Understanding airline passengers during covid-19 outbreak to improve service quality: Topic modeling approach to complaints with Latent Dirichlet Allocation algorithm. *Transp. Res. Rec. J. Transp. Res. Board* **2022**, *2677*, 656–673. [CrossRef]
23. Calli, L. Exploring mobile banking adoption and service quality features through user-generated content: The application of a topic modeling approach to Google Play Store reviews. *Int. J. Bank Mark.* **2023**, *41*, 428–454. [CrossRef]
24. Calli, L.; Alma Calli, B. Value-centric analysis of user adoption for sustainable urban micro-mobility transportation through shared e-scooter services. *Sustain. Dev.* **2024**, *32*, 6408–6433. [CrossRef]
25. Montes-Escobar, K.; De la Hoz-M, J.; Barreiro-Linzán, M.D.; Fonseca-Restrepo, C.; Lapo-Palacios, M.Á.; Verduga-Alcívar, D.A.; Salas-Macias, C.A. Trends in Agroforestry Research from 1993 to 2022: A Topic Model Using Latent Dirichlet Allocation and HJ-Biplot. *Mathematics* **2023**, *11*, 2250. [CrossRef]
26. Pilacuan-Bonete, L.; Galindo-Villardón, P.; Delgado-Álvarez, F. HJ-Biplot as a Tool to Give an Extra Analytical Boost for the Latent Dirichlet Assignment (LDA) Model: With an Application to Digital News Analysis about COVID-19. *Mathematics* **2022**, *10*, 2529. [CrossRef]
27. Isoaho, K.; Gritsenko, D.; Makela, E. Topic Modeling and Text Analysis for Qualitative Policy Research. *Policy Stud. J.* **2019**, *49*, 300–324. [CrossRef]
28. Wang, Y.; Mi, J.J.; Chen, C.; Ge, J.; Chen, Y. Evolution of China's shipping policies and attention: Evidence from LDA analysis. *Ocean. Coast. Manag.* **2025**, *267*, 107746. [CrossRef]
29. McCallum, A. Topic Model Diagnostics. UMASS. 2018. Available online: http://mallet.cs.umass.edu/diagnostics.php (accessed on 10 February 2025).

30. Sievert, C.; Shirley, K. LDAvis: A method for visualizing and interpreting topics. In Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore, MD, USA, 27 June 2014; pp. 63–70.
31. Chuang, J.; Manning, C.D.; Heer, J. Termite: Visualization techniques for assessing textual topic models. In Proceedings of the International Working Conference on Advanced Visual Interfaces, Capri Island, Italy, 21–25 May 2012; pp. 74–77. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.