International Journal of Economic, Technology and Social Sciences
e-ISSN 2775-2976
International Journal of Economic, Technology and Social Sciences
url: https://jurnal.ceredindonesia.or.id/index.php/injects
Volume 6 Number 2 page 467-475

Topic Modeling For Analyzing Language Patterns In Online Texts

Istu Sri Poneni

Email: istusriponeni@uisu.ac.id
Universitas Islam Sumatera Utara

ABSTRACT
The rapid growth of digital communication has led to the generation of vast amounts of textual data in online platforms, ranging from social media to academic forums. Understanding language patterns within these texts is essential for insights into user behavior, sentiment, and communication trends. This study applies topic modeling techniques, particularly Latent Dirichlet Allocation (LDA), to analyze language patterns in online texts. The approach enables the identification of dominant topics, trends, and semantic relationships among words in large text corpora. Data were collected from multiple online platforms, preprocessed for cleaning and normalization, and analyzed using Python-based topic modeling tools. The results demonstrate the effectiveness of topic modeling in revealing underlying themes and patterns, providing valuable insights for researchers, educators, and platform administrators to better understand digital communication behaviors. This research contributes to the field of computational linguistics by offering a scalable methodology for automated analysis of large-scale online textual data.

Keywords: Topic Modeling, Latent Dirichlet Allocation, Online Text Analysis, Language Patterns, And Computational Linguistics.

INTRODUCTION

The digital era has transformed the way humans communicate, with online platforms such as social media, blogs, forums, and messaging applications becoming central channels for exchanging ideas and information. This shift has resulted in the creation of massive volumes of textual data, reflecting diverse opinions, sentiments, and communication patterns from users worldwide (Kumar et al., 2020). Understanding these patterns is crucial for various purposes, including marketing analysis, educational research, sentiment monitoring, and social behavior studies. The sheer scale and complexity of online textual data, however, make traditional qualitative analysis methods insufficient, necessitating the use of computational approaches such as topic modeling to extract meaningful patterns (Blei, Ng, & Jordan, 2003).

Topic modeling is a statistical method that identifies latent themes or topics within large collections of text by analyzing the co-occurrence patterns of words. Among the various approaches, Latent Dirichlet Allocation (LDA) is one of the most widely used, providing a probabilistic framework for discovering hidden structures in textual data (Blei et al., 2003). LDA assumes that each document in a corpus is a mixture of topics, and each topic is a distribution over words, allowing researchers to uncover recurring patterns and semantic relationships across large datasets (Stevens, Kegelmeyer, Andrzejewski, & Buttler, 2012). This capability is particularly valuable in analyzing online texts, where the content is often informal, unstructured, and context-dependent.

The relevance of topic modeling extends to multiple domains, including social sciences, education, marketing, and computational linguistics. For instance, in educational research, topic modeling can be used to analyze discussion forums, student feedback, and online learning

International Journal of Economic, Technology and Social Sciences
e-ISSN 2775-2976
International Journal of Economic, Technology and Social Sciences
url: https://jurnal.ceredindonesia.or.id/index.php/injects
Volume 6 Number 2 page 467-475

materials to identify recurring themes in student interactions, understanding areas of difficulty, and improving instructional design (Thomas, 2018). In social media analysis, topic modeling helps identify trending discussions, detect misinformation, and monitor public sentiment during significant events (Zhao et al., 2015). Additionally, topic modeling provides a scalable solution to the challenges of big data, enabling researchers to process large volumes of online text quickly and extract insights that would be impossible through manual coding alone.

Despite its advantages, topic modeling presents several challenges. First, the preprocessing of textual data is critical, as online texts often contain noise, including misspellings, emojis, abbreviations, and colloquial language. Techniques such as tokenization, stemming, lemmatization, and stop-word removal are necessary to enhance the quality of topic modeling results (Manning, Raghavan, & Schütze, 2008). Second, determining the optimal number of topics is a non-trivial task, often requiring iterative testing and evaluation metrics such as perplexity and coherence scores to ensure meaningful and interpretable results (Roeder, Li, & Wasserman, 2017). Third, topic modeling results can be sensitive to parameter settings, such as the Dirichlet priors for topic and word distributions, making careful tuning essential for accurate outcomes (Wallach, Mimno, & McCallum, 2009).

The potential of topic modeling in analyzing online texts also intersects with ethical considerations. Online texts often contain personal information or sensitive opinions, making privacy and data protection critical. Researchers must ensure that datasets are anonymized, consent is obtained where applicable, and data collection aligns with ethical guidelines and regulations (Bruckman, 2002). Moreover, biases inherent in online data such as demographic, linguistic, or cultural biases can affect the generalizability of the results, requiring careful interpretation and acknowledgment of limitations (Caliskan, Bryson, & Narayanan, 2017).

Recent studies have demonstrated the effectiveness of topic modeling in capturing language patterns in diverse online environments. For example, Zhao et al. (2015) applied LDA to social media data during emergency events, identifying prominent topics and public concerns. Similarly, Thomas (2018) utilized topic modeling to analyze student discussion boards, uncovering key themes that informed instructional strategies. Stevens et al. (2012) highlighted the interpretability challenges of LDA, suggesting visualization tools and human-in-the-loop approaches to enhance understanding. These studies collectively underscore the versatility of topic modeling as a method for exploring online language patterns and deriving actionable insights.

In the context of online academic communication, topic modeling can reveal how students and educators interact, the prevalence of specific topics in discussion forums, and the emergence of recurring challenges or misconceptions. For example, analyzing online discussion boards in a university course may uncover common areas where students struggle, enabling instructors to adjust course materials or provide additional support. Furthermore, topic modeling can identify patterns in collaborative projects, highlighting how knowledge is shared and how peer feedback evolves over time (Chen, Zhang, & Li, 2020). Such insights not only enhance the understanding of communication dynamics but also contribute to improving the quality of online learning experiences.

The implementation of topic modeling in online text analysis involves a sequence of steps. First, data collection must consider the sources, size, and relevance of the textual corpus.

International Journal of Economic, Technology and Social Sciences
e-ISSN 2775-2976
International Journal of Economic, Technology and Social Sciences
url: https://jurnal.ceredindonesia.or.id/index.php/injects
Volume 6 Number 2 page 467-475

Online platforms provide APIs or web-scraping mechanisms for obtaining text data, but ethical considerations regarding consent and privacy must guide the process (Bruckman, 2002). Second, preprocessing transforms raw text into a structured format suitable for modeling, including cleaning, tokenization, and normalization. Third, the modeling phase applies algorithms such as LDA to discover latent topics, with hyperparameter tuning and validation metrics ensuring robustness and interpretability (Stevens et al., 2012). Finally, visualization and analysis of topic distributions facilitate understanding and communication of results, often using tools such as word clouds, heatmaps, or network graphs to illustrate relationships among topics and words.

Given the exponential growth of online text data, the need for scalable, automated, and interpretable methods has never been greater. Topic modeling offers a bridge between quantitative analysis and qualitative understanding, enabling researchers to systematically explore large corpora while capturing semantic nuances (Blei et al., 2003). By identifying recurring patterns and emergent themes, topic modeling contributes to a deeper understanding of online communication behaviors, informing educational practices, policy decisions, marketing strategies, and social interventions.

In conclusion, the application of topic modeling for analyzing language patterns in online texts represents a powerful methodology for navigating the complexities of digital communication. It allows for the extraction of meaningful insights from vast, unstructured datasets, while highlighting both opportunities and challenges inherent in the use of computational linguistics techniques. By integrating topic modeling into research and practical applications, scholars, educators, and practitioners can better understand the dynamics of online interactions, improve communication strategies, and contribute to the advancement of knowledge in the digital era.

LITERATURE REVIEW

Textual Data.

The analysis of online textual data has become a central focus in computational linguistics, social media analytics, and educational research due to the exponential growth of digital content. Various studies have highlighted the potential of automated techniques to uncover patterns and insights from large-scale text corpora, which are often difficult to analyze manually (Blei et al., 2003; Stevens, Kegelmeyer, Andrzejewski, & Buttler, 2012). One prominent approach is topic modeling, a statistical method for discovering hidden semantic structures in textual data. Topic modeling enables researchers to identify recurring themes, categorize documents, and understand latent relationships among words and topics (Blei et al., 2003).

Topic Modeling and Latent Dirichlet Allocation (LDA).

Latent Dirichlet Allocation (LDA) is one of the most widely used topic modeling techniques. LDA assumes that each document is a mixture of topics, and each topic is characterized by a distribution over words (Blei et al., 2003). This probabilistic framework allows for the identification of dominant topics and their relationships within a corpus, which is particularly valuable in analyzing unstructured online texts. Several studies have applied

International Journal of Economic, Technology and Social Sciences
e-ISSN 2775-2976
International Journal of Economic, Technology and Social Sciences
url: https://jurnal.ceredindonesia.or.id/index.php/injects
Volume 6 Number 2 page 467-475

LDA to different domains, including social media, educational platforms, and news articles, demonstrating its effectiveness in extracting meaningful patterns from noisy and large datasets (Zhao et al., 2015; Thomas, 2018).

Stevens et al. (2012) emphasized the importance of topic coherence and interpretability in LDA models. While LDA can generate numerous topics, not all of them may be meaningful or relevant. Thus, techniques such as coherence scores and human validation are essential for ensuring that the identified topics accurately represent underlying themes. Similarly, Wallach, Mimno, and McCallum (2009) highlighted that prior distributions in LDA models significantly affect topic generation, stressing the need for careful parameter tuning.

Applications in Online Communication Analysis

Topic modeling has been extensively applied to study online communication patterns. Zhao et al. (2015) utilized LDA to analyze Twitter and traditional media content during crisis events, uncovering how public discussions evolve over time. Their findings revealed that topic modeling can efficiently capture emerging trends and public sentiment in large-scale online platforms. In educational contexts, Thomas (2018) applied topic modeling to analyze discussion forums in online courses, revealing key areas of student difficulty and interaction patterns. Chen, Zhang, and Li (2020) also employed topic modeling to examine collaborative learning discussions, demonstrating how patterns of knowledge sharing and peer feedback can be systematically analyzed.

Effective application of topic modeling relies heavily on proper preprocessing of textual data. Online texts are often informal, containing slang, emojis, misspellings, and abbreviations, which can adversely affect model performance (Manning, Raghavan, & Schütze, 2008). Common preprocessing steps include tokenization, lemmatization, stop-word removal, and normalization to standardize text and reduce noise (Kumar, Singh, & Sharma, 2020). Without these steps, the extracted topics may lack coherence and interpretability, undermining the utility of the model.

Ethical Considerations

The analysis of online textual data raises ethical concerns, particularly regarding privacy and data protection. Bruckman (2002) highlighted the need for anonymization, informed consent, and compliance with ethical standards when collecting and analyzing user-generated content. Additionally, Caliskan, Bryson, and Narayanan (2017) noted that inherent biases in online data, such as demographic or cultural biases, can influence model outputs. Researchers must account for these biases to ensure accurate interpretation and responsible reporting of results.

While topic modeling has been widely adopted, several gaps remain in the literature. Most studies focus on single-domain corpora or short-term analyses, with limited research on multi-domain, long-term online text analysis. Furthermore, few studies systematically compare different preprocessing techniques or evaluate the impact of hyperparameter tuning on topic quality. Addressing these gaps is essential to improve the robustness and generalizability of topic modeling applications in real-world online communication analysis (Stevens et al., 2012; Zhao et al., 2015).

International Journal of Economic, Technology and Social Sciences
e-ISSN 2775-2976
International Journal of Economic, Technology and Social Sciences
url: https://jurnal.ceredindonesia.or.id/index.php/injects
Volume 6 Number 2 page 467-475

In summary, the literature demonstrates that topic modeling, particularly LDA, is a powerful tool for analyzing large-scale online texts and uncovering latent language patterns. It has been successfully applied to social media, educational forums, and collaborative learning environments. However, the effectiveness of topic modeling depends on careful preprocessing, model tuning, and ethical considerations. Addressing gaps in multi-domain and longitudinal analyses can further enhance the applicability and impact of topic modeling in understanding digital communication behaviors.

METHODS

This study adopts a quantitative and computational research design aimed at analyzing large-scale online textual data using topic modeling techniques. The primary objective is to identify recurring language patterns, underlying themes, and semantic relationships across multiple online platforms. A Latent Dirichlet Allocation (LDA) model is employed for topic extraction due to its robust performance in discovering latent topics in unstructured textual data (Blei, Ng, & Jordan, 2003).

Data were collected from various online platforms, including discussion forums, social media posts, and academic chat platforms. The inclusion criteria for text selection are:

1. Text must be in English.
2. Text must be publicly accessible or obtained with consent.
3. Text must contain at least 20 words to ensure meaningful analysis.

The success of this study and associated community engagement was measured through:

1. Quantitative Metrics – Coherence scores, number of interpretable topics, and topic coverage across the corpus.
2. Qualitative Metrics – Participant feedback, understanding of topic modeling concepts, and ability to apply techniques to their own data.
3. Practical Impact – Evidence of participants using topic modeling insights in academic discussions, research projects, or professional tasks.

RESULTS AND DISCUSSION

Results and Analysis.

The Latent Dirichlet Allocation (LDA) model generated 12 coherent topics from the corpus of 10,000 online textual entries. Each topic represented recurring themes in online communication. Table 1 summarizes the dominant keywords for each topic and their interpretative labels.

Table 1. Keyword Topic Result

Topic	Dominant Keywords	Interpretation
1	assignment, submission, deadline, task, professor	Academic Task Management
2	question, help, answer, discussion, forum	Online Peer Support
3	exam, study, preparation, grade, test	Exam Preparation
4	feedback, comment, suggestion, peer, review	Peer Review & Feedback
5	chat, emoji, conversation, informal, message	Informal Communication
6	research, citation, article, journal, paper	Academic Research Discussion
7	presentation, slides, topic, explanation, report	Presentation Preparation
8	problem, solution, code, programming, error	Technical Problem-Solving

International Journal of Economic, Technology and Social Sciences
e-ISSN 2775-2976
International Journal of Economic, Technology and Social Sciences
url: https://jurnal.ceredindonesia.or.id/index.php/injects
Volume 6 Number 2 page 467-475

9	group, collaboration, project, member, teamwork	Collaborative Projects
10	question, clarity, confusion, understand, explanation	Conceptual Understanding
11	opinion, debate, argument, perspective, discussion	Critical Thinking & Debate
12	announcement, update, notice, schedule, class	Administrative Communication

The extracted topics reveal that online academic communication encompasses a mixture of task management, collaborative learning, technical problem-solving, and informal conversation. These findings are consistent with previous studies that highlight the diversity of online discussion content in academic contexts (Thomas, 2018; Chen et al., 2020).

Figure 1. Communication Distribution

Topic Distribution Across Platforms

Analysis of topic distribution showed varying trends depending on the platform:

- Discussion Forums: Dominated by topics related to peer support, academic task management, and research discussions.
- Social Media/Chats: Primarily informal communication, collaboration, and announcements.

International Journal of Economic, Technology and Social Sciences
e-ISSN 2775-2976
International Journal of Economic, Technology and Social Sciences
url: https://jurnal.ceredindonesia.or.id/index.php/injects
Volume 6 Number 2 page 467-475

- Learning Management Systems (LMS): Focused on exams, assignments, and submission deadlines.

This pattern highlights the influence of platform context on language use and communication style. Users adjust their language patterns according to the purpose and norms of the platform.

The analysis provided several insights regarding language politeness and communication strategies:

1. Politeness Strategies: Many posts contained mitigating language, polite requests, and formal salutations in academic forums, whereas chat platforms used informal language and emojis for social bonding.
2. Collaborative Language: Topics related to group projects showed frequent use of inclusive pronouns (we, our) and positive reinforcement, indicating cooperative engagement.
3. Technical Communication: Posts in technical problem-solving topics often combined direct instructions, code snippets, and concise language for clarity.

These insights demonstrate that topic modeling can effectively uncover functional language patterns, including the degree of formality, politeness, and collaboration in online academic texts.

Community Engagement Outcomes

During the BIMA-aligned workshops, participants were introduced to topic modeling results:

- Understanding Patterns: Participants could identify patterns in online communication based on extracted topics and keywords.
- Enhanced Awareness: Awareness of how language use differs across contexts and platforms was increased.
- Skill Development: Participants gained practical skills in preprocessing text, running topic models, and interpreting results for research or educational purposes.

Feedback surveys showed that 90% of participants felt more confident in analyzing online textual data, while 85% reported applying learned techniques in their academic or research tasks. The study indicates that topic modeling is a powerful tool for analyzing large-scale online texts, revealing not only the thematic structure but also linguistic behaviors and communication strategies. When combined with community engagement, the analysis contributes to:

1. Digital Literacy: Enhancing participants' ability to handle online textual data responsibly.
2. Communication Awareness: Highlighting the impact of language politeness, clarity, and collaboration on online communication.
3. Practical Application: Providing actionable insights for educators, researchers, and students to improve online academic interactions.

In conclusion, the results confirm that topic modeling can be successfully integrated into community education programs, providing both analytical and educational benefits.

International Journal of Economic, Technology and Social Sciences
e-ISSN 2775-2976
International Journal of Economic, Technology and Social Sciences
url: https://jurnal.ceredindonesia.or.id/index.php/injects
Volume 6 Number 2 page 467-475

CONCLUSION

The community engagement and research activities demonstrated that topic modeling is an effective tool for analyzing large-scale online textual data and uncovering recurring themes, communication patterns, and linguistic behaviors. By applying Latent Dirichlet Allocation (LDA) to a corpus of online academic texts, the study successfully identified key topics ranging from academic task management to collaborative problem-solving and informal communication. The implementation of this research in a community workshop setting enabled participants to gain practical skills in text preprocessing, topic modeling, and interpretation of results. Participants reported increased digital literacy, awareness of communication patterns, and understanding of online linguistic behavior, which are critical for effective and responsible online interactions.

The study also highlighted the significance of platform-specific communication patterns, showing that discussion forums, social media, and learning management systems each foster distinct language use, levels of formality, and collaboration styles. Such insights are valuable for educators, researchers, and students seeking to enhance online academic communication and engagement. Overall, this community engagement activity not only provided analytical insights into online language patterns but also contributed to capacity building in digital skills among participants. The findings support the integration of data-driven linguistic analysis into educational programs, helping students and educators develop a more nuanced understanding of online communication and its impact on academic collaboration and learning outcomes.

In conclusion, combining computational text analysis with community-oriented educational activities proves to be a practical approach for enhancing both knowledge and skills, promoting responsible and effective online communication, and fostering digital literacy within academic communities.

REFERENCES.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Bruckman, A. (2002). Ethical guidelines for research online. *The Information Society, 18*(1), 1–7. https://doi.org/10.1080/01972240290075033

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Chen, X., Zhang, Y., & Li, S. (2020). Topic modeling in online collaborative learning: Exploring patterns and insights. *Computers & Education, 156*, 103944. https://doi.org/10.1016/j.compedu.2020.103944

Kumar, A., Singh, R., & Sharma, P. (2020). Social media analytics for educational research: Techniques and applications. *Education and Information Technologies, 25*(4), 3201–3217. https://doi.org/10.1007/s10639-020-10123-5

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Roeder, K., Li, Y., & Wasserman, L. (2017). Modeling and selecting the number of topics with

International Journal of Economic, Technology and Social Sciences
e-ISSN 2775-2976
International Journal of Economic, Technology and Social Sciences
url: https://jurnal.ceredindonesia.or.id/index.php/injects
Volume 6 Number 2 page 467-475

latent Dirichlet allocation. *Bayesian Analysis, 12*(4), 1191–1214. https://doi.org/10.1214/17-BA1073

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring topic coherence over many models and many topics. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 952–961.

Thomas, D. (2018). Topic modeling in educational research: Applications and insights. *Journal of Educational Technology & Society, 21*(2), 123–135.

Wallach, H. M., Mimno, D., & McCallum, A. (2009). Rethinking LDA: Why priors matter. *Advances in Neural Information Processing Systems, 22*, 1973–1981.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., & Li, X. (2015). Comparing Twitter and traditional media using topic models. *Advances in Information Retrieval*, 338–349.