Comparative Analysis of Topic Extraction Methods for Early 2000s Blog Content Mining

1st Kent Nolan - SID: ?????
Auckland University of Technology
Auckland, New Zealand
email address

2nd Pedro Mroninski - SID: XTQ6111
Auckland University of Technology
Auckland, New Zealand
xtq6111@autuni.ac.nz

***Abstract*—This paper provides a brief overview of a blog mining task, focusing on the comparative analysis of two topic extraction strategies: TF-IDF and Transformer-Enhanced Latent Dirichlet Allocation (LDA). We present key findings on the top topics identified from a dataset of early 2000s blog content. Furthermore, we discuss the performance and scalability of the implemented methods, noting the challenges that would arise when applying them to larger datasets.**

***Index Terms*—topic modeling, blog mining, nlp, lda, tf-idf, demographic analysis, text mining**

I. Contributions

This section details individual contributions from each team member.

- Data preprocessing and cleaning pipeline development (Pedro)
- Implementation of topic extraction methods (Pedro + Kent)
- Demographic segmentation and analysis (Pedro + Kent)
- Results evaluation and documentation (Kent)

II. Introduction

A. Background and Motivation

This paper explores the importance of understanding demographic-specific content trends from blog data from 2001-2004 Schler et al., 2006. The value for product/service innovation based on blog analysis is discussed.

B. Research Objectives

The primary goal is to extract the two most popular topics per demographic group. Secondary goals include comparing the effectiveness of different topic extraction methods and understanding the nuanced differences in topics across demographics. This is framed within the context of an innovation company seeking to identify marketing opportunities.

C. Paper Organization

This paper is organized as follows. The next section reviews relevant literature. Then, we present our methodology, followed by the experimental setup. After that, we present and analyze the results. Finally, we discuss our findings and conclude with future work.

III. Literature Review

A. Topic Modeling Approaches

A review of traditional methods like LSA, pLSA, and basic LDA, as well as modern enhancements using transformer-based models and semantic embeddings. This includes previous work on demographic-based content analysis.

B. TF-IDF for Topic Extraction

Discussion of the theoretical foundation of TF-IDF, its strengths in identifying distinctive terms, and its limitations in capturing semantic relationships.

C. LDA and Neural Enhancements

Exploring the evolution from basic LDA to transformer-enhanced approaches, and the benefits of combining probabilistic models with embeddings, especially for recent applications in short-text analysis.

IV. Methodology

A. Data Description and Preprocessing

The dataset consists of 19,320 blog posts in XML files, with demographic metadata in the filenames. A table with demographic distribution is provided. The cleaning pipeline involved HTML removal, tokenization, lemmatization, and handling of non-ASCII characters and other noise.

B. Demographic Segmentation Strategy

Five demographic groups were created based on metadata extracted from filenames. Data validation and consistency checks were performed to ensure data quality.

C. Topic Extraction Method 1: TF-IDF Approach

This method involved document segmentation using semantic boundaries, phrase detection with Gensim, and TF-IDF vectorization with parameters such as `max_features=1000` and `ngram_range=(1,3)`. Topic quality scoring and filtering were then applied.

D. Topic Extraction Method 2: Transformer-Enhanced LDA

This approach used spaCy for semantic document creation and integrated Sentence-BERT embeddings (`all-mpnet-base-v2`). Adaptive topic number optimization via silhouette analysis was performed, followed by semantic re-ranking of topic words using centroid similarity.

E. Topic Identification and Labeling

A common post-processing pipeline for both methods was used to generate topic labels from top words and extract clauses containing dominant topics. Quality-based filtering with a determined threshold was applied.

V. Experimental Setup

A. Implementation Environment

The experiments were conducted on Google Colab, with drive mounting for data access. Required libraries include spaCy, Gensim, scikit-learn, and sentence-transformers. Computational considerations for processing 19,320 files are noted.

B. Evaluation Metrics

Topic coherence measures were used for quantitative evaluation. Manual evaluation criteria were also established, with inter-annotator agreement assessed for topic relevance.

C. Parameter Optimization

Hyperparameter tuning was performed for both methods. A cross-validation strategy was applied to demographic subsets to ensure robustness.

VI. Results and Analysis

A. Overall Topic Distribution

Presentation of the top topics identified across the entire dataset, including topic frequency and quality scores. Word clouds or other visualizations are used to illustrate major themes.

B. Demographic-Specific Results

1) Gender-Based Analysis: Analysis of topic differences between male and female bloggers, with statistical significance testing to validate the findings.

2) Age-Based Analysis: Comparison of topics for bloggers under 20 versus over 20, highlighting generational topic preferences.

3) Student Population: Identification of unique topics for the student demographic and comparison with the general population.

C. Method Comparison

A comparative analysis of TF-IDF and Transformer-Enhanced LDA performance in terms of topic quality and coherence scores. Computational efficiency is also analyzed. The agreement between methods on top topics is discussed, noting that the top topics were often the same, requiring a focus on demographic relevance.

D. Example Topic Clauses

Presentation of representative clauses for each demographic's top two topics, providing context and interpretation.

VII. Discussion

A. Key Findings

Discussion of the most significant demographic differences, any unexpected topic discoveries, and the business implications for innovation.

B. Method Strengths and Limitations

Analysis of when TF-IDF excels versus Transformer-Enhanced LDA, and the trade-offs between interpretability and semantic depth. Scalability considerations are also discussed, noting the computational intensity of the transformer-based model.

C. Challenges Encountered

Discussion of data quality issues (e.g., HTML artifacts, encoding), and the cleanup required. Computational considerations due to model sizes and challenges related to the temporal nature of the content (2001-2004 slang and language use) and handling short, sparse text are also covered.

VIII. Conclusion and Future Work

A. Summary of Contributions

A summary of how the study successfully extracted demographic-specific topics, provided a comprehensive comparison of two extraction methods, and offered practical insights for product innovation.

B. Recommendations

Recommendations on the preferred method based on the specific use case and implementation guidelines for a production environment.

C. Future Improvements

Suggestions for future work, such as real-time topic tracking, integration with modern blog platforms, exploring deep learning approaches like BERT fine-tuning, and temporal analysis of topic evolution.

References

References

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2006). Effects of age and gender on blogging. *AAAI spring symposium: Computational approaches to analyzing weblogs*, *6*, 199–205.

Appendix

The full implementation is available in a Google Colab notebook. Instructions for data placement in Google Drive and a setup guide with requirements are provided in the repository.