Identifying the Influences Behind the LinkedIn Posts using Topic Modeling and Sentiment Analysis

R Nagaraj[1], Rohith Adithya C R[2], Sakalabathula Sri Chakra Teja[3], Dr.Deepika T[4]

rnagaraj3004@gmail.com[1], rohi.adi3115@gmail.com[2] , Srichakratejasakalabathula@gmail.com[3]
t_deepika@cb.amrita.edu [4]

Department of Computer Science Engineering, Amrita School of Computing, Coimbatore,
Amrita Vishwa Vidyapeetham, India.[1-4]

Abstract. The ubiquity of social networking has transformed daily life. People constantly share opinions, rate products, and conduct business on these platforms. Social media is vital in business, helping establish connections with prospects and clients. LinkedIn, specifically, excels in building trust through success stories, promotions, and recommendations. It connects professionals worldwide with the general public. However, only a few posts gain significant influence, and the factors behind this remain unknown. To address this, we propose a solution involving topic modeling and sentiment analysis. Our project aims to uncover the influences behind LinkedIn posts by scrutinizing media content, utilizing natural language processing techniques, and identifying sub-topics and aspects through topic modeling and sentiment analysis

Keywords: Social Networking, LinkedIn, Topic modeling, Sentiment analysis.

1 Introduction

Social media platforms, originally channels for self-expression and communication, have transformed into battlegrounds involving users, employers, and platform owners, with these conflicts evident at the interface level. LinkedIn, a significant business network site, provides valuable data for micro and macro network analysis. Despite the wealth of LinkedIn data, effective mining necessitates unconventional database management methods. This research paper concentrates on post influences, utilizing hypothetical testing to evaluate media content's importance, topic modeling to uncover top subjects, and fine-grained sentiment analysis. By leveraging LinkedIn data, the paper divulges the most reacted topics and sentiments for each post, providing users with valuable insights. The paper comprises six sections, including a discussion of related works in Section 2, an architecture diagram in Section 3, a comprehensive methodology analysis in Section 4, experimental findings in Section 5, the conclusions and future work outlines in Section 6.

IACIDS 2023, November 23-25, Lavasa, India

DOI 10.4108/eai.23-11-2023.2343215

2 Related works

Our research drew inspiration from several studies. Poonguzhali et al. [1] inspired our sentiment analysis work on LinkedIn. Dr. Anbazhagan et al.'s SELDAP model [2] influenced our topic modeling and sentiment analysis. Ravikrishna B et al. [3] introduced methods to identify hot topics in social media, which guided our approach. We also incorporated techniques from Abuzayed et al.'s work [5], which favored BERTopic in topic modeling. Chandra R et al. [6] informed us about the sentiment analysis, while Ramya G R et al. [7] influenced our strategy for identifying influential nodes on Twitter. Our approach to sentiment analysis in Hadoop Distributed File System-related tweets was influenced by Parveen H et al. [8] and Naveenkumar et al. [12]. Moghadasi M. N et al. [9] shaped our analysis of user engagement and opinions in Facebook comments.

3 Architecture Diagram

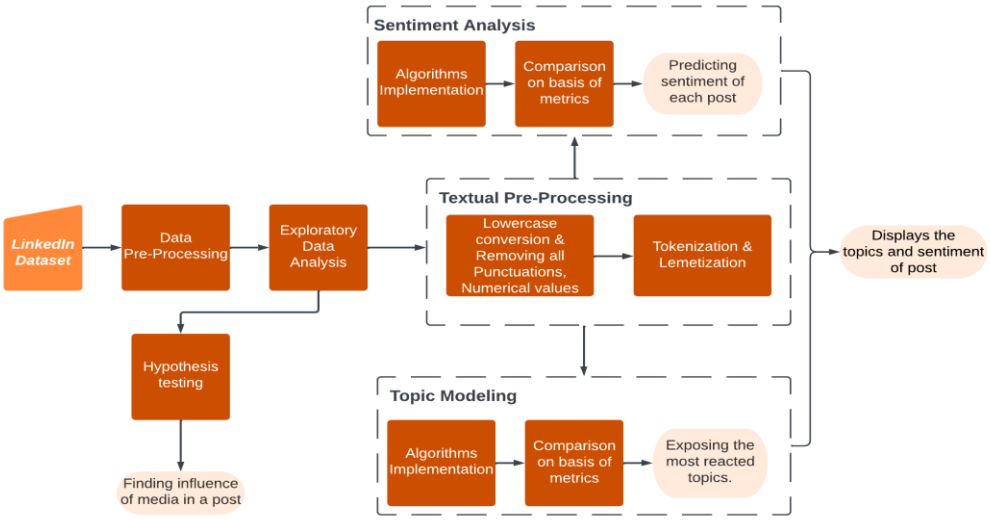

Fig. 1. Architecture diagram.

4 Methodology

To tackle our problem, we aligned and analyzed the dataset with live LinkedIn columns, eliminating null values and extraneous columns. This yielded a focused exploratory dataset. Our analysis unveiled a clear link between followers and reactions. Moreover, we assessed the influence of media in posts by scrutinizing media type counts and author attractiveness. Through Hypothesis testing, we quantified media's significance, confirming a 16% enhancement in a post's impact, see equation 1.

$$Z = (\bar{x} - \mu) / (s / \sqrt{n}) \quad (1)$$

We used NLP techniques, including Stemming, Lemmatization, Parts of Speech Tagging, typo handling, punctuation, and Spacy tokenization, for Topic Modeling and Sentiment Analysis. Phrase modeling captured meaningful multi-word combinations, ensuring consistent analysis with standardized post content. LDA in combination with BERT were chosen for Topic Modeling after evaluating LDA, BERT, RoBERTa, and NMF for efficiency, validated by Coherence and Perplexity metrics. VADER proved the most effective in Sentiment Analysis among options such as SVM, RoBERTa, and Naive Bayes based on multiple metrics. An intuitive Streamlit interface was developed for user interaction, allowing the input of post content and the display of topics and sentiments for each post.

5 Results and Analysis

5.1 Importance of media

Our primary goal was to assess media's impact on LinkedIn posts. After refining the data, we conducted exploratory analysis, identifying outliers like LinkedIn user, who garnered more reactions and followers. We observed a positive correlation between followers and reactions, indicating that users with larger followings tend to receive more reactions. To gauge what makes a LinkedIn post successful, we computed an "attractivity" score for each author. In our analysis of media's impact on attractivity, we found an average increase of 38.8% with a substantial variance of 476.62%. As the variance value is high, we couldn’t come to a conclusion that adding media will increase the attractivity. So, we did Hypothesis Testing and confirmed our assumption, validating a 16% increase with 95% confidence. This underscores the significance of incorporating media to boost post attractivity.

5.2 Topic modeling

We employed Topic Modeling to identify sub-topics in posts, using NLP techniques and data cleaning through the MVP approach. This included text cleaning, lowercasing, and removing elements like "..see more," punctuation, hashtags, white spaces, and emojis. Cleaned data formed a corpus, and phrase modeling generated multi-word tokens. LDA algorithm was applied using Gensim to uncover document structures, with topics manually reviewed. PyLDAvis visualized topic distribution, and we shortlisted topics based on LinkedIn trends, confirming their presence in posts. The model processed raw posts, revealing associated topics and values shown in Figure 2.

Fig. 2. Topics present in raw post is displaye

We employed other language models , BERT and RoBERTa, for masked language modeling. BERT predicts original tokens from masked ones, while RoBERTa, an enhanced version, excels in topic identification due to its larger batches, longer training, dynamic masking, and omission

of the next sentence prediction task. We implemented both models using sentence transformers ("distilbert-base-nli-mean-tokens"), with support from UMAP and HDBSCAN for embeddings and clustering. We also utilized TF-IDF techniques to reveal the top twenty words per topic, with higher scores determining clusters and topic words. NMF was employed as a third algorithm for topic identification, creating a document-term matrix and decomposing it into weight (W) and coefficient (H) matrices. To expedite topic generation, we considered combining algorithms, specifically LDA and BERT, which were integrated, then data were processed by auto-encoding it. Separate data vectorization for LDA and BERT resulted in grouped topics which was not in other models and this combination perfomed well with our dataset, visually represented as a word cloud in Figure 3.

Fig. 3. WorldCloud- LDA – BERT

After implementing and analyzing the algorithms separately, we assessed their performance using crucial metrics like coherence and perplexity. In Table. 1, you can observe the performance values, clearly indicating that the combination of LDA and BERT outperforms other approaches in our dataset.

Table 1. Topic Modeling Metrics.

Metric/Models	LDA	BERT	RoBERTa	NMF	LDA + BERT
Coherence	-3.55	0.018	-2.45	0.38	0.59
Perplexity	-7.61	0.98	0.1	0.0055	0.14

5.3 Sentiment Analysis

Our final analysis aims to identify sentiments in posts to enhance user comprehension. We used VADER, a well-known rule-based sentiment analysis method, offering compound scores from -1 (highly negative) to +1 (highly positive), along with separate scores for positivity, negativity, and neutrality. VADER relies on a lexicon to associate words with sentiment scores, providing intensity scores for each word. We processed posts by breaking sentences into words, and calculated compound scores and these scores for each post were saved in a new dataframe, including the post, scores, polarity, and sentiment, ensuring sentiments matched post content, validating our predictions. We also examined word polarity and frequency in the cleaned data, exploring terms like "Apprenticeship vs. Company" and "Employee vs. Productivity." SVM is excellent at handling linear and non-linear data, suitable for high-dimensional feature spaces, and resistant to overfitting and imbalanced datasets. In our analysis, we used a linear-kernel SVM, splitting the data into test and train sets to predict sentiment and scores, highlighting sentiment differences. RoBERTa, a potent transformer-based model, efficiently extracts context

from text, predicting sentiments based on labels. Naive Bayes, a simple yet robust probabilistic algorithm, assumes conditional feature independence given the class label. We divided the data into test and train sets, used MultinomialNB to predict scores for post features, and displayed the predicted sentiment. After predicting sentiment using these methods, we compared their performance based on metrics like accuracy, precision, recall, F1 score, and confusion matrices. Table 2 highlights that VADER and SVM excel in predicting sentiments. A closer examination confirms VADER as the best-performing algorithm for sentiment prediction.

Table 2. Sentiment Analysis Metrics.

Metric/Models	VADER	SVM	RoBERTa	Naive Bayes
Accuracy	0.922	0.91591	0.660	0.5994
F1 Score	0.922	0.91374	0.706	0.6144
Precision	0.923	0.91523	0.660	0.7494
Recall	0.922	0.91591	0.624	0.5994

5.3 Interface

We developed a Streamlit interface that combines Topic Modeling and Sentiment Analysis to provide users with valuable insights into post content. By using the LDA-BERT combination and VADER, the interface offers a detailed analysis, including line-by-line examination and displays the topics and sentiments which will certainly help the LinkedIn users to precisely understand whether they conveyed the same as per they needed , as shown in Figure. 4.

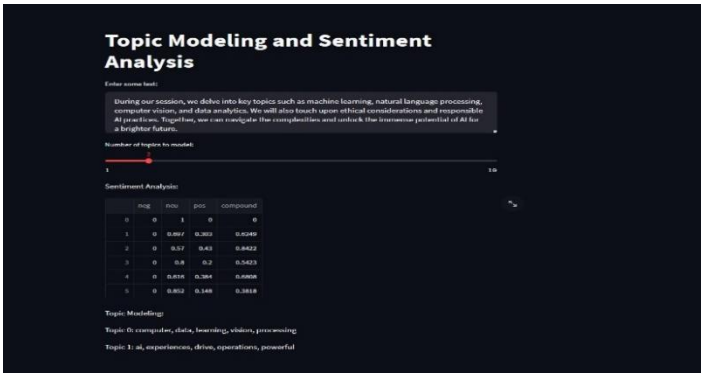

Fig. 4. Streamlit interface

6 Conclusion and Future works

Our research highlights the importance of topic modeling and sentiment analysis in understanding the factors that impacts the reach of LinkedIn posts. Our analysis reveals that integrating media content, focusing on popular topics and sub-topics, and maintaining a good sentiment can enhance engagement and reach. These insights can guide users in optimizing their content for better performance on the platform. By applying these techniques, users can create more effective and engaging content that reaches a broader audience. The future work invloves the Natural Language Generation techniques to provide personalized recommendations and

actionable insights based on the analysis results. In addition exploring network dynamics beyond individual posts to identify influencers, understand interactions between posts, and analyze how engagement impacts a user's overall content reach.

References

[1] Poonguzhali, R., Vinothini, S., Waldiya, V., & Livisha, K. (2018). Sentiment analysis on linkedin comments.International Journal of Engineering Research & Technology IJERT (ICONNECT), 6(7), 415

[2] Dr. Anbazhagan M and Arock, M., "Integrated topic modeling and sentiment analysis: a review rating prediction approach for recommender systems", Turkish Journal of Electrical Engineering and Computer Sciences, vol. 28, pp. 107-123, 2020.

[3] B.Ravi Krishna, P.Akhila, S.Sowjanya and B.Keerthana, "Prediction of Hot Topic in Social Media Based on User Participation Behavior in Social Hotspots," 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2021, (pp. 1545-1548)

[4] Rohani, V.A., Shayaa, S., & Babanejaddehaki, G. (2016, August). Topic modeling for social media content: A practical approach. In 2016 3rd International Conference on Computer and Information Sciences (ICCOINS) (pp. 397-402) IEEE.

[5] Abuzayed, A., & Al-Khalifa, H. (2021). BERT for Arabic topic modeling: an experimental study on BERTopic technique. Procedia Computer Science, 189, 191-194.

[6] Chandra, R., & Saini, R. (2021). Biden vs trump: Modeling US general elections using BERT language model. IEEE Access, 9, 128494-128505.

[7] Ramya, G. R., & Bagavathi Sivakumar, P. (2021). An incremental learning temporal influence model for identifying topical influencers on Twitter dataset. Social Network Analysis and Mining, 11(1), 1-16.

[8] Parveen, H., & Pandey, S. (2016, July). Sentiment analysis on Twitter Data-set using Naive Bayes algorithm. In 2016 2nd international conference on applied and theoretical computing and communication technology (iCATccT) (pp. 416-419) IEEE.

[9] Moghadasi, M. N., Safari, Z., & Zhuang, Y. (2020, December). A sentimental and semantical analysis on facebook comments to detect latent patterns. In 2020 IEEE International Conference on Big Data (Big Data) (pp. 4665-4671) IEEE.

[10] Ostrowski, D. A. (2015, February). Using latent dirichlet allocation for topic modelling in twitter. In Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015) (pp. 493-497) IEEE.

[11] Kumar, S., Kumar, M. & Soman, K. (2019). Deep Learning Based Part-of-Speech Tagging for Malayalam Twitter Data (Special Issue: Deep Learning Techniques for Natural Language Processing). Journal of Intelligent Systems, 28(3), 423-435. https://doi.org/10.1515/jisys-2017-0520

[12] Naveenkumar, K. S., R. Vinayakumar, and K. P. Soman. "Amrita-cen-sentidb 1: Improved twitter dataset for sentimental analysis and application of deep learning." In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1-5. IEEE, 2019.

[13] T. Rajasundari, Subathra P., and Dr. (Col.) Kumar P. N., "Performance Analysis of Topic Modeling Algorithms for News Articles", Journal of Advanced Research in Dynamical and Control Systems, vol. 2017, pp. 175-183, 2017.