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

R.Nagaraj, Rohith Adithya C.R, Sakalabathula Sri Chakra Teja, Deepika T
Department of Computer Science and Engineering,
Amrita School of Computing, Coimbatore,
Amrita Vishwa Vidyapeetham, India.
Mail Id: cb.en.u4cse19640@cb.students.amrita.edu, cb.en.u4cse19645@cb.students.amrita.edu,
cb.en.u4cse19649@cb.students.amrita.edu, t_deepika@cb.amrita.edu.

July 6, 2023

***Abstract*— The proliferation of social networking in the modern world has made it a ubiquitous presence in people's lives. On a daily basis, individuals post and share their opinions, rating products, and engage in business-related activities. Social media platforms play a crucial role in the business field, facilitating the establishment of relationships with prospects and clients. As such, LinkedIn has emerged as the ideal online platform for building trust by sharing success stories, business promotions, and recommendations through posts. This professional website serves as a means of connecting the world's professionals with the general public. At the same time, many people post opinions and professional content regularly, and only a select few post become influenced by reaching larger community. The factors responsible for this influence remain unknown, prompting us to propose a solution based on topic modeling and sentiment analysis. Our project aims to identify the influences behind LinkedIn posts by examining the role of media content in posts, and by utilizing natural language processing techniques to analyze the underlying aspects of posts which then examined using topic modeling and sentiment analysis to identify subtopics and aspects present in the posts.**

Keywords: Topic modeling, Media, Natural language processing, Sentiment analysis, Topics.

I. Acknowledgments

We express our gratitude to Mr. Jeremy Aranico for generously providing us with the dataset, which was instrumental in enabling us to achieve our research objectives and obtain the desired outcomes.

II. Introduction

Social media platforms have become highly popular venues for self-expression, communication, and self-promotion. However, rather than solely facilitating online identity formation, these platforms have become sites of contention among users, employers, and platform owners over control of online identities. These struggles are played out at the interface level.

LinkedIn, one of the world's largest social networking sites, is a valuable source of business network information. It can be used to count both micro and macro-scale networks of people. The vast amount of data available on LinkedIn represents an example of "big data" that cannot be effectively mined using conventional relational database management methods. However, smaller subsets of this data can be analyzed to generate potentially useful results. By making use of profile information, linkages offered through individual users' connections, and group functions, it is possible to map a person's interests and network of contacts, as well as to list entire organizations.

This research paper aims to analyze and determine the influences of a post by checking the importance of media content through hypothetical testing, most reacted topics using Topic modeling and analysing the sentiments using fine grained Sentiment Analysis approach by utilizing necessary data available from LinkedIn about user's posts, connections, media type and other relevant information. At last we will check each post and display the most reacted topics and sentiment behind that particular post which would give an idea to the user about the topics conveyed and the sentiment present in it before getting posted.

A. Motivation

As LinkedIn is widely used, it is crucial to understand what constitutes a high-quality post on the platform. Such understanding is necessary for expanding one's reach, fostering meaningful interactions with prospects, and ultimately increasing business revenue.

III. Related works

The study conducted by Poonguzhali et al. [1] which focused on sentiment analysis to analyze human text and determine whether the user's comment on LinkedIn was positive or negative. This helped and paved us a way to solve our problem statement. Here, authors employed the NLTK to analyze user comments, which helped in splitting the words and identified the polarity of the comments. This approach proved to be useful in determining the sentiment of user comments on LinkedIn and made us to work on identifying the influence behind the post.

As it was mentioned we decided to use topic modeling and sentiment analysis and to get assured of combining it we referred to the research work done by Dr. Anbazhagan et.al. [2] where they used topic modeling and sentiment for a review rating which was a recommending system. In this they have used SELDAP model which is basically the combination of sentiment(VADER and LDA based Prediction. Initially they found the topics and sentiment score of each review using the proposed model and then they used to train a supervised machine learning regression model which predicted the rating for corresponding reviews. This made us to go forward and implement it to our dataset.

Ravikrishna B et al. [3] proposed a prediction model for identifying hot topics based on user participation behavior in social hotspots. The authors first employed a Back Propagation neural network to analyze user participation behavior to accomplish this. Next, they identified hot social topics using the term frequency-inverse document frequency(TF.IDF) Model and calculated the weights of all topics to determine their significance within the data collection. Overall, their approach enabled accurate identification and prediction of hot topics in social media settings.

Rohani et al. [4] aimed to develop a practical topic model utilizing LDA for extracting topics from a social media corpus. The authors employed the LDA algorithm to preprocess the data, which involved removing punctuation, extra spaces, and other unnecessary patterns to achieve this. Subsequently, the LDA algorithm was used to identify topics within the collection of documents. Overall, their approach enabled efficient topic extraction from social media data. In our research study, we implemented Bidirectional Encoder Representations from Transformers(BERT) as a novel technique and referred to Abuzayed et al. [5] research to guide our approach. In their study, the authors compared the performance of the BERTopic model with that of the well-known LDA and Non-negative Matrix Factorization(NMF) models. To measure coherence between the high-scoring words in the dataset and the embedding models, the authors utilized Normalized Pointwise Mutual Information (NPMI). They employed AraVec2.0 and Pre-trained Language Model(PLM) to embed words into tokens and passed these tokens to BERTopic. Hierarchical Density-Based Spatial Clustering of Applications with Noise(HDBSCAN) and Uniform Manifold Approximation and Projection(UMAP) were then used to reduce dimensionality, and transformer models were employed to evaluate the model. The study findings revealed that the BERTopic model outperformed the LDA and NMF models, achieving a high positive NPMI score.

Chandra R et.al. [6] aimed to analyze tweets related to the US 2020 elections using Long Short-Term Memory (LSTM) and BERT to gain insights into crowd behavior and viewpoints during the election period. The study also utilized sentiment analysis to predict the election outcome. The results revealed that based on the tweets during the electoral campaigns, Biden had a higher likelihood of winning the election. Overall, the authors' approach provided valuable insights into crowd behavior and sentiment during the US 2020 elections. We referred to the work of Ramya G R et al. [7], where they demonstrated that DC-FNN classification outperforms fixed clustering and NLP-based sentiment analysis in identifying the influential node in the Twitter dataset based on categories such as pricing, service, and timeliness. They tested these approaches on the likelihood function using iterative logistic regression analysis and the Temporal Influential Model (TIM), and the results show that they are more accurate than other methods in terms of Precision, Recall, and F-measure. We gained an understanding of the topic's role in a node's influence level as a result of this research.

Parveen H et al. [8] utilized a Twitter dataset to analyze the emotions expressed in tweets related to Hadoop Distributed File System(HDFS). The authors employed Hadoop Distributed File System for analysis and conducted preprocessing on the data. Two types of preprocessed data were considered, one with emoticons and one without emoticons. The Naive Bayes algorithm was employed for sentiment analysis, and the results showed that the dataset with emoticons yielded better performance than the one without emoticons. The study analyzed tweets from various perspectives, including positive, negative, and neutral sentiments, and demonstrated the usefulness of tweets in predicting product sales, evaluating service quality, and gathering user feedback.

Moghadasi M. N et al. [9] conducted a study on the Facebook comment section to analyze user engagement through sentiment and semantic analysis. With the abundance of opinions and feedback present in the Facebook comment section, this study provides valuable insights into understanding user engagement across various topics. The authors utilized VADER, semantic analyzer, and cosine similarity functions to analyze the data, and also calculated Average Response Time (ART) and Average Comment Length (ACL) to gain a better understanding of the opinions and emotions conveyed. Through this approach, the authors were able to effectively analyze the comments and responses in the Facebook comment section and provide meaningful insights into user engagement in social media.

IV. Architecture Diagram

In this section, we present the proposed method's architecture diagram, which depicts the flow of processes involved. Each process is highlighted in a box to facilitate understanding and we have splitted into two phases for better clarity. The input to the proposed method is the LinkedIn dataset, which is subjected to pre-processing techniques such as data exploration and cleaning. Further, exploratory data analysis is performed on the dataset, and insights from important columns such as Reactions and Comments are obtained. The output is then divided into two stages: (i) to check the media's importance in the post using Hypothesis testing, and (ii) to perform textual processing using prominent NLP libraries for tasks such as case conversion, punctuation removal, etc., shown in the Figure 1. The processed text is then given as input to the Language Processing block, where tasks such as

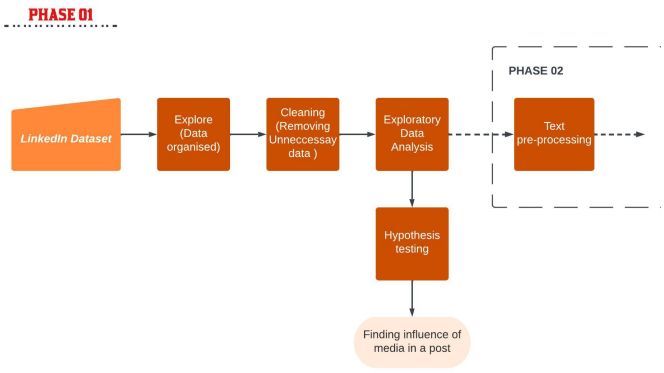

Fig. 1: Phase: 01 Architecture Diagram

tokenization, lemmatization,typo correction, phrase modeling and bi-grams are performed. This output is then passed to the Topic modeling and Sentiment Analysis block where we implemented many algorithm and obtained the best performing algorithm for our dataset as shown in Figure 2. The finalized algorithm's are then fine-tuned which then passed to display the topics and sentiments of each post.

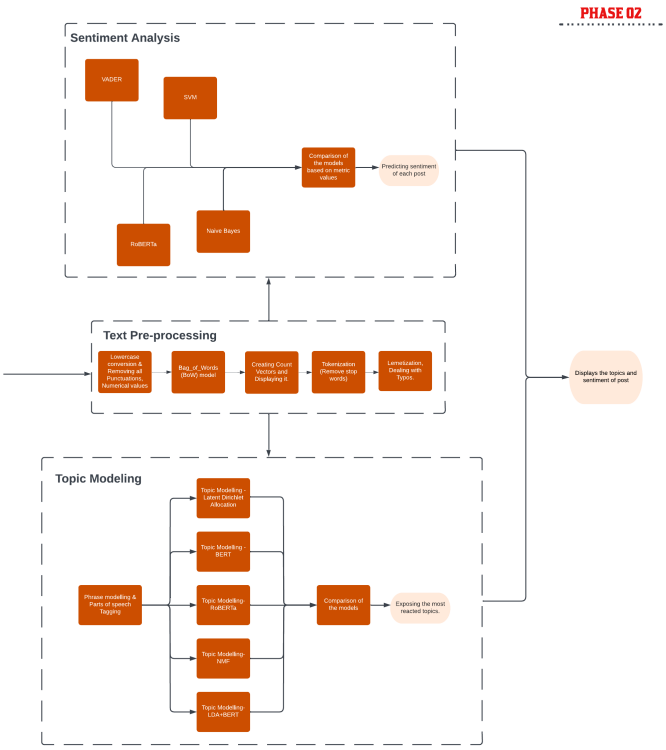

Fig. 2: Phase: 02 Architecture Diagram

V. Methodology

This section will cover the approaches and methods used by us in detail to solve our problem statement. Initially we analyzed our dataset and identified the columns are same as in real time LinkedIn such as name, connections, comments, time spent, followers, locations, media type, media etc., as shown in the Figure 3. But we were able to see some null values and unwanted columns so we dropped them and obtained the data which can be explored efficiently. So, we dropped those data and we obtained the dataset with clean values.

```
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Unnamed: 0        34012 non-null  int64
 1   name              34012 non-null  object
 2   headline          34012 non-null  object
 3   location          31740 non-null  object
 4   followers         33970 non-null  float64
 5   connections       25713 non-null  object
 6   about             34012 non-null  object
 7   time_spent        34011 non-null  object
 8   content           31996 non-null  object
 9   content_links     34012 non-null  object
 10  media_type        26779 non-null  object
 11  media_url         34012 non-null  object
 12  num_hashtags      34012 non-null  int64
 13  hashtag_followers 34012 non-null  int64
 14  hashtags          34012 non-null  object
 15  reactions         34012 non-null  int64
 16  comments          34012 non-null  int64
 17  views             0 non-null      float64
 18  votes             86 non-null     object
dtypes: float64(2), int64(5), object(12)
memory usage: 4.9+ MB
```

Fig. 3: Overall dataset information

Now we did exploratory analysis to get more insights of data dependencies and role of each values. For this, we did the author's (LinkedIn user) analysis on basis of reaction and followers of every author's. By doing this we were able to understand if the number of followers increases then number of reactions increases as shown in the Equation 1.

.

$$n_{followers} = kn^{\mu}_{react} \tag{1}$$

Also, we focused in identifying the influence of media in the post by analyzing with respect to media types counts, media attractivity of each author's. By doing this we got some variations in attractivity and variance so we used statistical method Hypothesis testing, which helps to estimate the relationship between two parameter. We found the exact value of media importance as per the Equation 2 and from this z-score test we got to known that adding media will influence the post by 16%

$$Z = \frac{\bar{x} - \mu}{s/\sqrt{n}} \tag{2}$$

As our next move of identifying the influences behind the posts, we used Topic modeling, an algorithm for extracting topics from a collection of documents and Sentiment Analysis, an opinion mining technique to determine the sentiment present in the post. As a part of it we applied some prominent NLP techniques by following Minimum Viable Product approach such as Stemming, Lemmatization, Parts of speech tagging, dealing with typos, Punctuation & Tokenization by using Spacy, a prominent and fast industrial-strength natural language library. We also did the Phrase modeling to combine tokens and represent meaningful multi-words. To clean the data, we converted all post content to lowercase and removed hashtags to ensure consistency while analysing. Now we applied the Topic modeling algorithms such as Latent Dirichlet Allocation(LDA),Bi- Directional Encoder Representation Transformer(BERT),Robustly Optimized Bi- Directional Encoder Representation Transformer(RoBERTa) & Non-Negative

Matrix Factorization(NMF) as we wanted to test which model works well with our dataset and provide the efficient topics after fine tuning. Here, all the model performed well but we wanted to check whether combination of the above would give us much more efficient results or not. For that, we researched for prominent model combination and found that LDA and BERT performs well. Based on the results obtained from all these model, we compared it using Topic modeling metrics such as Coherence and Perplexity where we obtained good values for LDA & BERT combination and decided to finalize it as our Topic modeling algorithm for predicting the topics. The same approach we followed for the Sentiment Analysis where we applied the algorithm such as Valence Aware Dictionary for Sentiment Reasoning(VADER), Support Vector Machine(SVM), RoBERTa and Naive bayes. Here, we were able to predict the sentiment present in the each posts and compared all the models using the metrics such as Accuracy, F1 score, Precision, Recall and also found the confusion matrix. From this comparison we decided to use VADER for sentiment prediction as it performed well. As we implemented our objective individually we wanted to merge topics and sentiment results for each post. For that, we used a Streamlit as a interface where we combined all the above procedure and made a single input space where we pass the post content which then displays the topics and sentiment present in that post.

VI. Results and Analysis

A. Importance of Media

As our first module is to find the importance of media and we now pass the real time dataset of LinkedIn for pre-processing where we removed the unwanted and null values and obtained the clean dataset. After that we did exploratory analysis where we analyzed the author's based on followers and reaction count as shown in the Equation 1. From this analysis we found that author's such as Simon and Richard are outliers as they have more number of reaction and followers respectively as shown in Figure 4 and those both author have more than other author's. So to get much more idea of spread we employed log scale visualization to entire dataset as shown in Figure 5. From this we observed

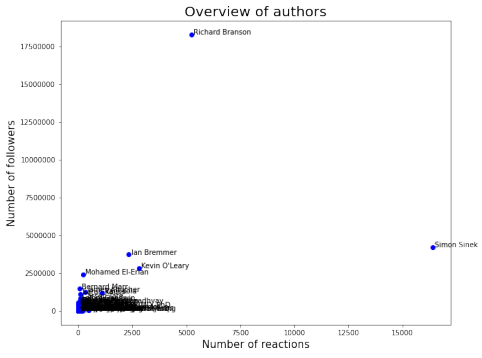

Fig. 4: Overview of authors

that higher the number of followers, higher the number of reactions which made us to understand that if a author

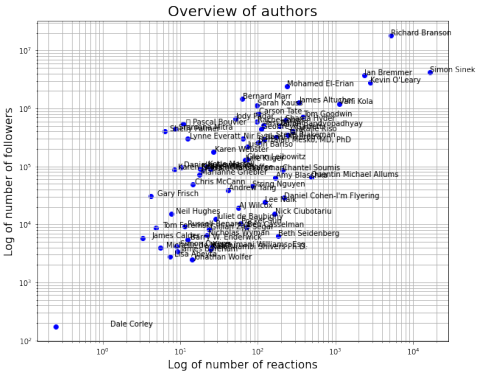

Fig. 5: Entire dataset

has good number of followers then they can have a good quantity of reactions. Therefore, in our analysis of identifying factors that makes LinkedIn post to be good and for that we decided to analyze one author's post at a time by calculating attractivity score by using the Equation 3.

$$attractivity = \frac{\#reactions + \#comments}{\#followers} \tag{3}$$

The attractitvity was calculated to the entire dataset by removing the outlier author's and were able to infer that there are some other outliers so we decided to drop the attractivity factor. But we thought of finding the media attractivity of a each post and for this we found the types of media present in the dataset and found articles and images are used majorly. Now we checked the influence of media on terms of attractivity and for that we created a separate column where if a posts do not have a media then it will have valuea as 0 and if media is present then it will have value as 1. By doing this we were able to find for specific author and to get much more confidence we checked the media attractivity for the entire dataset and found that some authors are outliers considering the difference between a post with or without media. So, we considered the 95% distribution for finding the average and variance of media attractivity by removing all the outliers. From this we inferred on an average adding a media increases the post attractivity by 38.8% and the variance by 476.62% . As the variance value is high we cannot come to conclusion that adding media will increase the attractivity.
So, to get a clear idea we implemented the well known statistical method Hypothesis Testing where we considered the assumption as H_0 = Media doesn't improve attractivity, H_1 = Media improves attractivity and significance level as 0.5 where the Z-score is 3.57. Based on this score we checked with Z-table and got the values as 0.9978 which is greater than our assumption of 95% data. By this we can understand clearly that our assumption H_0 gets rejected and get confirmed that adding media will help increasing the attractivity of the post. But still to get much more confident we did another Hypothesis Testing where we consider hypothesis mean as 30 and based on that Z-score we finalized our hypothesis mean as 16 and calculated the Z-score where we got 2.05 which confirmed us that adding a media in post increases the attractivity by 16% with a

confidence of 95%. By this we found that adding media to the post is very important as it increases the attractivity of the post.

B. Topic Modeling

As the next part of the analysis, Topic modeling was implemented to identify the sub-topics present in the posts. Here NLP techniques were employed and data cleaning was carried out to obtain efficient results. The textual pre-processing was done using the Minimum Viable Product approach, where unwanted text was identified, and letters were converted to lowercase. The "..see more" text at the end of the post was removed and punctuation, hashtags, white spaces, emojis were also eliminated. Lemmatization was done, and stop words were removed to ensure that the sentence's meaning was not affected. Spacy, a prominent NLP library, was used for this process, and the results are shown in Figure 6. The processed

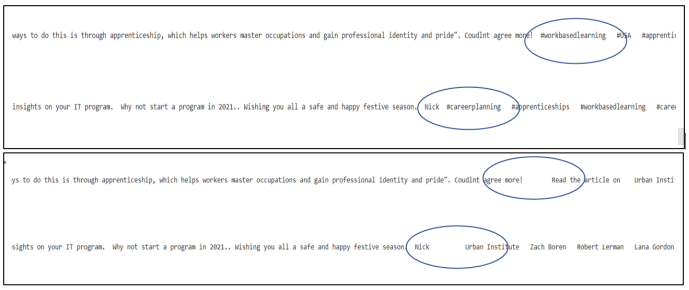

Fig. 6: Removal of hashtags

data was made into a cleaned corpus and it was stored as a dataframe, which then used for creating the phrases. The phrase modeling was done to form the combinations of tokens which represent meaningful multi-word concepts.

1) LDA: As we got the cleaned data now we implemented LDA method to identify the most reacted topics. LDA is fully unsupervised and discovers topics automatically by maximizing the likelihood of observing the document in a corpus based on assumptions.It captures some latent structure and organization within the document, meaningfully interpreting the subject material. The Gensim library was used to implement LDA due to its high-performance parallelized implementation and also by assuming a bag of words where a document is represented by the counts of distinct terms occurring within it. After training the model we obtained the topics, which were inspected manually with the token that has been grouped. The topic spread and the relationship between the topics were visualized in an interactive format using PyLDAvis in Figure 7 , where the top 30 most relevant words of a particular topic were inferred.

As in LDA topic need to be specified and for that we shortlisted the topics based on most LinkedIn trends and checked whether the those topics are present in each post by passing through the function. The results of the same is shown in the Figure 8 where the example_post1 is the raw post present in the dataset and after passing it to model we get the topics present along with the values.

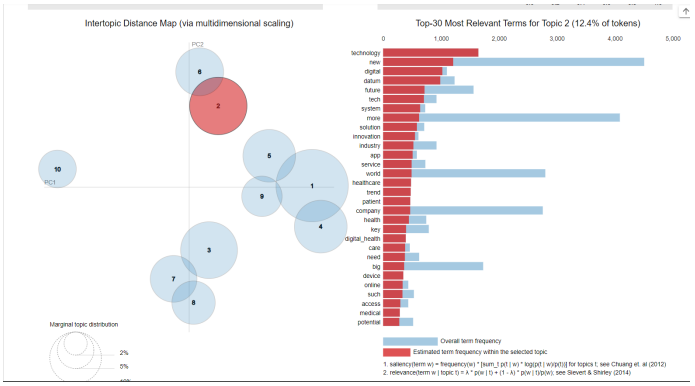

Fig. 7: Left side is the topic and Right side is top relevant words of that topic

Fig. 8: Topics present in raw post is displayed

2) BERT & RoBERTa: BERT is trained on large amounts of text data using an unsupervised learning approach called masked language modeling, where a percentage of the input tokens are randomly masked, and the model is trained to predict the original tokens from the masked tokens and RoBERTa is a pre-trained language model Like BERT, where it is trained using a large amount of text data and a masked language modeling objective whereas RoBERTa improves upon BERT in several ways, such as using larger batches, longer training time, dynamic masking, and removing the next sentence prediction task. These improvements result in better performance on various NLP tasks. Here both algorithm was used to have the most efficient way of topic identification, as it is designed to pre-train deep bidirectional representations from the unlabelled text by joint conditioning on both the left and right context. The pre-trained BERT model was fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of NLP tasks. The sentence transformers were used for implementing both the algorithms by methods of "distilbert-base-nli-mean-tokens". For the embeddings and clustering we used UMAP and HDBSCAN on the basis of metrics such as cosine and Euclidean, respectively. By doing this, clusters were formed, and a dataframe was created to store the values. To identify the topics, methods of TF-IDF were followed, and the top twenty words per topic were extracted, which were then stored in a dataframe. Based on the high score, the corresponding cluster and the words of the topic were identified. In addition to differentiate the formed clusters we visualized on different neighbour values. The topics generated from this algorithm is shown in the Figure 9

```
top_n_words[145][:10]
```

```
[('study', 0.6009197489380554),
 ('analysis', 0.4645219286360231),
 ('artist', 0.2776103017299387),
 ('imagine', 0.23051924831041584),
 ('test', 0.21973424644507877),
 ('piie', 0.21511597763130594),
 ('topical', 0.21511597763130594),
 ('appear', 0.21511597763130594),
 ('provocation', 0.21511597763130594),
 ('signage', 0.19971270695219603)]
```

Fig. 9: Top words belonging to the selected topic

3) NMF: For the fourth model we have applied NMF for finding the topics we will use the cleaned content column as the input for the vectorizer which will be used to create a document term matrix. In NMF the document term matrix will be decomposed into two matrices W and H, where W represents matrix of weights and H represents matrix of coefficients. The features names will be used to interpret the topics by identifying the most important words for each topic. After, that we have found the topic clusters which will be compared with each row and if the topics are relevant to the text of that particular row then the topic weight is given which is shown in Figure 10

```
Row 33933:
Topic 0:
cheating butter criticalinfrastructure cashier davos buildingproducts brk dazzle blinded bin
Topic weight: 0.0
Topic 1:
blasting creativewriting blast bestowed assessing cameron assessed crispy came charter
Topic weight: 0.0
Topic 2:
bureaucratic blur bestteam businesspeople ageist benioff breed apoorva congratulate classmate
Topic weight: 0.0010893837144578888
Topic 3:
cab auspicious correction brandingstrategy agricultural agnext bbc brandmanagement backpack alska
Topic weight: 0.0011429862704509016
Topic 4:
coaster apprentice abia craze accommodative apprenticeship aided abide abhishek creativewriting
Topic weight: 0.001275941635441206
Topic 5:
cornell babson compositional cancelling beneficiary assurance buckled benioff acknowledgement damiani
Topic weight: 0.007846445319645918
Topic 6:
california brandmanagement abbacchi bend darker dana congrats congratulate alipay charge
Topic weight: 0.0
Topic 7:
buffett armored decently blackness deepest bunk crunch alwayslearning decor burdensome
Topic weight: 0.002002998547301766
Topic 8:
darin daring blur dark admits bethany decides datascientist bloodstream celebratesuccess
Topic weight: 0.0
Topic 9:
```

Fig. 10: Topics clusters and weights

4) Combination of LDA & BERT: We considered using a combination of algorithms to create topics much more quickly and effectively after implementing and analysing the outcomes of individual algorithms. So for this combination we took the cleaned data did Auto encoder and imported model where raw texts were processed. In addition, we vectorized the data separately for LDA and BERT after which the topic were creation function was defined. At last we got the topics in terms of cluster and for better understanding we displayed the topics in world cloud as shown in the Figure 11 and Figure 12. From the world cloud we can infer that the combination has created the topic efficiently.

5) Comparison of all Topic Modeling Algorithms: As we implemented the algorithm individually and obtained the desired results and now we wanted to compare the model performance on basis of metric such as coherence and perplexity and the values for each algorithm is shown in the Table I where we can see that combination of LDA and BERT performs well with our dataset.

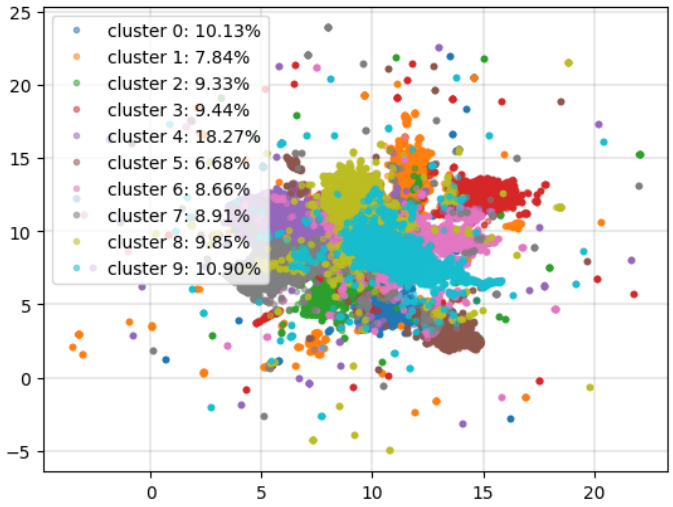

Fig. 11: Cluster- LDA - BERT

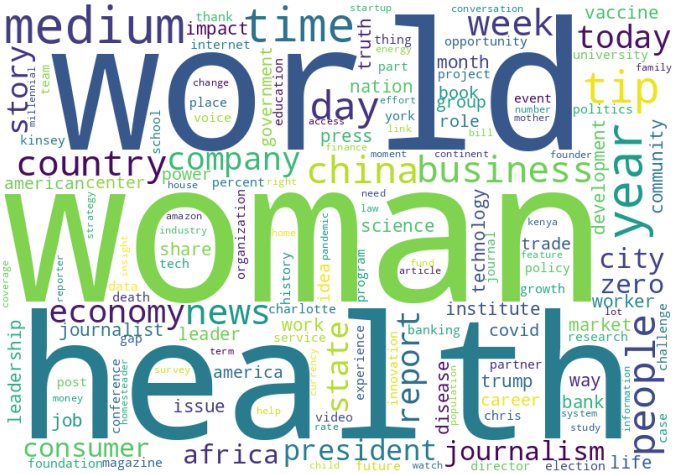

Fig. 12: WorldCloud- LDA - BERT

Metrics/models	LDA	BERT	ROBERTA	NMF	LDA & BERT
Coherence	-3.58	0.018	-2.45	0.38	0.59
Perplexity	-7.61	0.98	0.1	0.0055	0.14

TABLE I: Topic Modeling Metrics

C. Sentiment Analysis

The last part of our analysis is to identify the sentiments present in the posts. Identifying sentiments will really help the user to understand what type sentiment is conveyed through the post to other users. As explained earlier the data after textual pre-processing was used and we implemented different prominent algorithm to predict sentiments.

1) VADER: Initially we implemented VADER which is popular rule based sentiment analysis methods designed specifically for analyzing sentiment present in the texts. VADER provides sentiment analysis results in terms of a compound score, which represents the overall sentiment of a text on a continuous scale between -1 (extremely negative) and +1 (extremely positive). It also provides separate scores for positive, negative, and neutral sentiment. There are many approaches and we used Lexicon based approach because it makes use of a pre-made lexicon (dictionary) that has words or expressions connected to sentiment scores. A sentiment intensity score is given to each word in the lexicon, indicating the level of positivity, negativity or neutral connected to the word. The basic required packages and lexicon library was imported after which streamed post was generated where the sentences were splitted into words. Then compound score was found on basis of values as shown in the Figure13. The

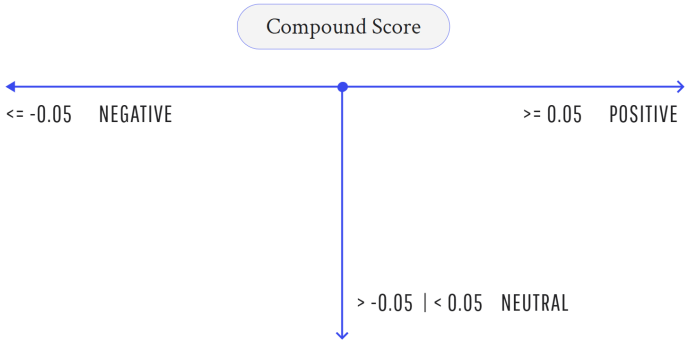

Fig. 13: Compound score

compound score was found to each posts by creating a new dataframe comprising the post, scores, polarity and sentiment as shown in the Figure 14. From this figure we could able to see that the sentiment of a post is almost closer to the post as we read it and this assured us that we have good sentiment prediction. In addtion we found the polarity and density of some frequent words present in the cleaned data such as Apprenticeship vs Company, Employee vs Productivity.

	Posts	Sentiment	Score	Polarity
	healthy future work employee skill productive ...	positive	{'neg': 0.0, 'neu': 0.594, 'pos': 0.406, 'comp...	0.8910
	national disability advocate able act	neutral	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	0.0000
	none	neutral	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	0.0000
	month company most modern apprenticeship_progr...	positive	{'neg': 0.0, 'neu': 0.447, 'pos': 0.553, 'comp...	0.9074
	fortunate time administration insight innovati...	positive	{'neg': 0.0, 'neu': 0.617, 'pos': 0.383, 'comp...	0.9313
...	...	...	...	...
	quick post small_college easy final nail coffi...	positive	{'neg': 0.084, 'neu': 0.642, 'pos': 0.274, 'co...	0.8807
	jeering right many pathway success lawyer cont...	positive	{'neg': 0.116, 'neu': 0.473, 'pos': 0.412, 'co...	0.9509
	interested strange account mine wary falsehood...	positive	{'neg': 0.129, 'neu': 0.432, 'pos': 0.439, 'co...	0.5106
	birth thousand nature course nature beauty abo...	positive	{'neg': 0.165, 'neu': 0.627, 'pos': 0.208, 'co...	0.3612
	concerned small_college massive admission frau...	negative	{'neg': 0.322, 'neu': 0.678, 'pos': 0.0, 'comp...	-0.5859

Fig. 14: Dataframe with sentiment and scores

2) SVM: Due to SVM's proficiency in handling both linearly separable and non-linearly separable data, and it also has the capacity to handle high-dimensional feature spaces. Additionally, they are resilient against over-fitting and can handle imbalanced dataset. We implemented the model by using the linear kernel where we were able to split data into test and train. By doing this we calculated the predicted sentiment and its score for test data and then compared with existing sentiment. The dataset was formed with these values which gave us a brief difference's between the sentiment of each posts as shown in the Figure 15

test_data

	Posts	Sentiment	Score	Polarity	predicted_sentiment	predicted_sentiment_score
27209	time premium payment thing other big streamer ...	neutral	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	0.0000	neutral	neutral
27210	brand_playbook vision statement success brand ...	positive	{'neg': 0.0, 'neu': 0.398, 'pos': 0.602, 'comp...	0.8625	positive	positive
27211	proud debut short online_course brand_playbook...	positive	{'neg': 0.0, 'neu': 0.721, 'pos': 0.279, 'comp...	0.4767	positive	positive
27212	ready digital	positive	{'neg': 0.0, 'neu': 0.286, 'pos': 0.714, 'comp...	0.3612	positive	positive
27213	professional negotiator poetic win_win outcome...	negative	{'neg': 0.351, 'neu': 0.649, 'pos': 0.0, 'comp...	-0.6597	neutral	neutral
...	...	...	...	...	...	...
34007	igniter optimist idealist world different one ...	positive	{'neg': 0.0, 'neu': 0.408, 'pos': 0.592, 'comp...	0.9808	positive	positive
34008	executive shareholder coach desire fair weathe...	positive	{'neg': 0.0, 'neu': 0.49, 'pos': 0.51, 'compou...	0.7430	positive	positive
34009	many end year year origin journey world vast_m...	positive	{'neg': 0.0, 'neu': 0.885, 'pos': 0.115, 'comp...	0.2500	positive	positive
34010	customer employee least second people	neutral	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	0.0000	neutral	neutral
34011	small work big way big work world small way	neutral	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	0.0000	neutral	neutral

Fig. 15: Sentiment predicted using SVM

3) RoBERTa: RoBERTa can efficiently extract dependencies and contextual information from the text because it is a potent transformer-based model. We imported the model libraries and implemented it to the entire dataset where we predicted the sentiment for each post based on labels such as Label 0: Neutral, Label 1: Negative and Label 2: Positive. These label were displayed with each post and score for the same was calculated.

4) Naive Bayes: The probabilistic algorithm Naive Bayes is straightforward but powerful for sentiment analysis. Based on the Bayes theorem, it makes the assumption that the features whether they be words or other attributes are conditionally independent of one another given the class label. Here the data was splitted into test, train where vectorizer and features were found for posts. But using MultinomialNB scores were predicted to the post features and the predicted sentiment of each post were also displayed.

5) Comparison of all SA algorithms: As we were able to predict the sentiment to the cleaned data using four prominent algorithms and we had also compared the models on basis of metrics such as Accuracy, precision, recall and F1 score and confusion matrix. From this Table II we can clearly

Metrics/models	VADER	SVM	Naive Bayes	RoBERTa
Accuracy	0.922	0.91591	0.660	0.5994
F1 Score	0.9221	0.91374	0.706	0.6174
Precision	0.923	0.91523	0.660	0.7494
Recall	0.922	0.91591	0.624	0.5994

TABLE II: Sentiment Analysis Metrics

understand the VADER and SVM performs well in predicting the sentiments for our dataset and to be much closely we found VADER is best performing algorithm to our dataset in predicting the sentiment.

D. Interface

We developed an interface using Streamlit by combining both Topic Modeling and Sentiment Analysis to provide valuable insights of the post. By leveraging LDA - BERT

Combination and VADER, the interface offers a comprehensive analysis of the post content including line by line analysis and providing the topics and sentiments as shown in the Figure 18

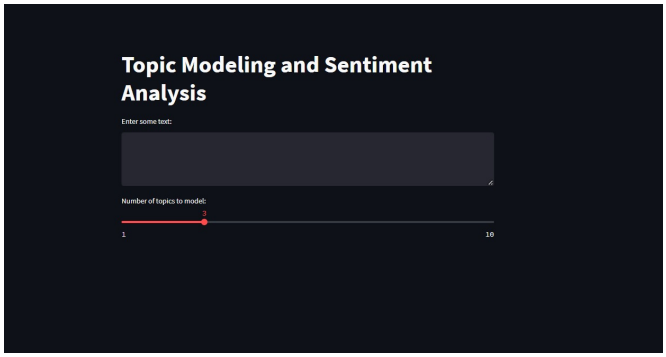

Fig. 16: Streamlit interface

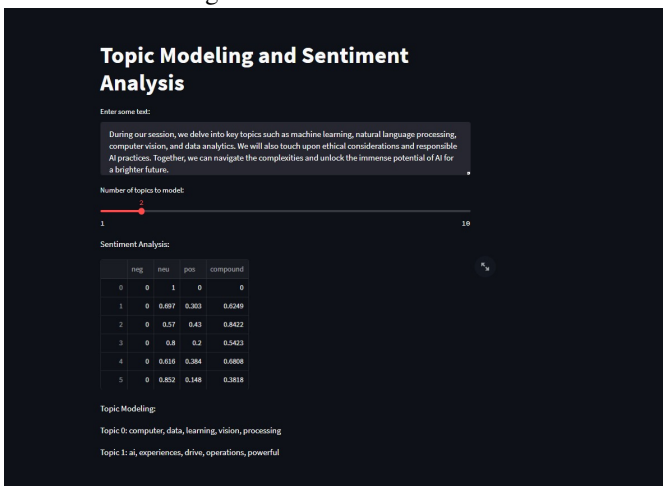

Fig. 17: Streamlit with text

Fig. 18: Streamlit

VII. Conclusion and Future work

Our research demonstrates that topic modeling and sentiment analysis can provide valuable insights into the factors that influence the reach of LinkedIn posts. Our analysis shows that incorporating media content, focusing on popular or trending topics, and using positive sentiment can increase engagement and reach. These findings can serve as a guide for users looking to improve their content and achieve their goals on the platform. By utilizing these techniques and insights, users can create more effective and engaging content that reaches a wider audience.

The major works need to be done in future are:
Natural Language Generation: Implement natural language generation techniques to automatically generate personalized recommendations and actionable insights based on the analysis results. This can provide users with concise and actionable suggestions to improve their post reach.
Social Network Analysis: Extend the analysis beyond individual posts by considering the network effects. Explore how posts interact with each other, identify influencer, and analyze the impact of engagement on the overall reach of a user's content.

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] 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.

[11] E.S.Negara, D.Triadi and R.Andryani, "Topic Modelling Twitter Data with Latent Dirichlet Allocation Method," 2019 International Conference on Electrical Engineering and Computer Science (ICECOS), 2019,(pp. 386-390).

[12] R. Prasanna Kumar, "A Comprehensive Survey on Topic Modeling in Text Summarization", 5th International Conference on Micro-Electronics and Telecommunication Engineering, Springer book series on "Lecture Notes in Networks and Systems". 2021.

[13] 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.

[14] 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