INSTITUTE FOR HOMELAND SECURITY

Sam Houston State University

SOCIAL NETWORK ANALYSIS USING MACHINE LEARNING

Institute for Homeland Security

Sam Houston State University

ABM Rezbaul Islam

Ahsan Ul Islam

Social Network Analysis using Machine Learning

ABM Rezbaul Islam
Department of Computer Science
SAM Houston State University
Huntsville, Texas, USA ari014@shsu.edu
and
Ahsan Ul Islam
Department of Computer Science
SAM Houston State University
Huntsville, Texas, USA axi034@shsu.edu

Abstract

Electronic Mail (Email) has emerged as a widespread technique for exchanging messages through electronic devices, becoming an indispensable and universal communication medium. Its significance cannot be overstated, as an email address is vital for swift interactions in business, government, trade, entertainment, and various other aspacts of daily life. This mode of communication has progressively replaced traditional written methods for important correspondences, including personal and business trans- actions, where an email is given the same weight as a signed document. In social net- work analysis, a significant challenge lies in identifying essential and influential nodes within a network based on its structure. These nodes can be critical in information dissemination, decision-making processes, and network dynamics. Sentiment Analysis (SA) in text mining has emerged as an automated process to discern subjective information from textual data, such as opinions, attitudes, emotions, and feelings. While many existing approaches treat SA as a text classification problem, requiring labeled data for training machine learning models, obtaining such labeled data can be laborious and time-consuming, often requiring manual annotation efforts. Additionally, the need for transferability across different domains hinders using the same labeled data in diverse applications, necessitating the creation of unique labeled datasets for each part. Overcoming these challenges is crucial for sentiment analysis's wider adoption and effectiveness in various real-world applications. The objective of the research is to analyze the Enron email dataset by creating a directed graph that represents the email communication network. Two important graph theory metrics are used to find out the number of direct connections (emails sent) for each sender and the influence of each sender as a bridge or critical point of communication in the network. On the other hand, we will use sentiment analysis to analyze the Enron email dataset using different type of pre-trained deep learning models to find the communication type for top ten email sender which we will find using graph theory.

1 Introduction

The study of topological properties and structure of complex networks, particularly in the context of significant social networks, presents numerous challenges due to their inherent complexity. Given the extensive interconnections between nodes, tasks like evaluating centrality measurements and identifying cliques demand significant computing resources. Moreover, technology-driven social networks, such as those formed on social media platforms, introduce new dimensions of complexity with the massive volume of user-generated data. Effectively analyzing such networks requires the development of efficient algorithms, parallel processing techniques, and distributed computing to extract meaningful insights. Re- searchers continuously strive to overcome these complexities, advancing network analysis and data science methodologies to unravel the intricate dynamics of technology-based social networks.[1]

In today's digitally interconnected world, email communication plays a vital role in facilitating interactions and information exchange. The analysis of email networks offers valuable insights into communication patterns, identifying influential nodes, and understanding the flow of information within an organization or a community. In this study, initially we explore the Enron email dataset and employ graph theory metrics to conduct a comprehensive analysis of the email communication network. By constructing a directed graph that represents the relationships between email senders and recipients, we calculate two fundamental graph theory metrics: degree centrality and betweenness centrality. Degree centrality measures the number of direct connections for each sender, highlighting the most active communicators in the network. On the other hand, betweenness centrality identifies nodes that act as critical intermediaries, facilitating the efficient flow of communication. By analyzing these centrality measures, we aim to identify key players in the network and gain insights into the communication dynamics. The visualization of the email network aids in better understanding the overall structure and provides a visual representation of the communication patterns. This study contributes to the field of network analysis, offering valuable knowledge applicable to various domains, including organizational communication optimization, social network analysis, and information dissemination strategies.

The allure of online social networking and communication has captured widespread interest, leading to a recent surge in research endeavors to extract valuable insights and patterns from online documents and web content [2]. Alongside this, the Internet's vast reach and the proliferation of digital devices have resulted in electronic mail (Email) becoming an omnipresent mode of communication and social networking. Statistical data indicates a steady increase in global Email users, with an average of 1.7 Email accounts per user as of 2015 [3]. Notably, the corporate world heavily relies on Email for communication, as evidenced by the staggering volume of over 108.7 billion Emails exchanged daily [3]. Email has cemented its position as the primary means of workplace communication in the contemporary business landscape. As businesses strive to make the most of the digital age, deciphering Email con- tent and discerning communication patterns has become a pressing research focus, holding immense potential to unravel valuable insights for individuals and organizations alike.

Natural Language Processing (NLP) is the automated manipulation and understanding of human language using specialized software components. The goal is to replicate human language comprehension to some extent. NLP techniques enable the automation of human

signs and languages, such as speech, writing, and text, with reasonable accuracy. A critical challenge faced by email users is dealing with spam messages. Spam is unsolicited and irrelevant messages sent by unauthorized entities or unnecessary individuals. Given the widespread usage of emails globally, spammers exploit this platform to send unwanted content, links, and advertisements, posing a potential threat of stealing sensitive information like banking credentials. Implementing robust spam handling routines is imperative to ensure emails remain a secure and effective communication platform. Properly managed, electronic mail can become a beneficial and safe medium for communication.

In Sentiment Analysis (SA), the prevailing approach treats it as a text classification problem, requiring labeled data for model training [4]. Unfortunately, obtaining a substantial labeled dataset can be challenging in many cases. As a result, manual data labeling has become necessary, leading to significant costs and time investments. Moreover, an additional challenge lies in the need for more portability of SA models across different domains. The labeled data that works well for one part may not generalize effectively to another domain, necessitating the creation of separate labeled datasets for each domain manually. These challenges highlight the need for innovative solutions that reduce the dependency on labeled data and enhance the transferability of SA models across diverse disciplines.[5]

Sentiment analysis, a crucial natural language processing task, finds increasing applications in various domains, including customer feedback analysis and communication under- standing. In this research, we conducted a comprehensive comparative study on the Enron dataset, employing state-of-the-art language models such as BERT, RoBERTa, and XLNet for sentiment analysis. The Enron dataset, comprising real-world email communications from the Enron corporation, offers an ideal scenario to assess the efficacy of these advanced transformer-based models. Leveraging the pre-trained BERT, RoBERTa, and XLNet mod- els, we fine-tune each for sentiment classification, allowing us to explore the strengths and limitations of each approach. We address challenges related to labeled data scarcity and do- main adaptability, evaluating the performance of each model on email sentiment prediction. Our findings shed light on the benefits and trade-offs associated with using BERT, RoBERTa, and XLNet for sentiment analysis, providing valuable insights for businesses aiming to en- hance customer satisfaction and communication strategies through advanced deep learning techniques.

Finally, this study proposes an ensemble-based approach to sentiment analysis for email classification, leveraging the combined power of three cutting-edge language pre-trained models: BERT, RoBERTa, and XLNet. Each model independently predicts email sentiment types, and the ensemble method combines their outputs using majority voting and weighted voting to achieve enhanced accuracy and robustness. Integrating the Gradient Boosting Classifier introduces an additional layer of sophistication to the ensemble, optimizing the final predictions by exploiting the strengths of the individual models. The results demonstrate the efficacy of the ensemble, showcasing improved sentiment classification performance and providing valuable insights into customer feedback and communication analysis.

2 Related Works

In the field of graph theory and data mining, numerous techniques are available to evaluate the significance of nodes in a network. These approaches can be classified into two primary categories: social network analytical (SNA) methods and node deletion methods. P. Stephen and colleagues refer to these methods as the KPP-Pos (Key players problem/positive) and KPP-Neg (Key players problem/negative) problems, respectively.[15]

SNA methods aim to identify valuable characteristics that distinguish nodes within a network. Centrality-based measures, such as degree, closeness, betweenness, and sub-graph analysis, are commonly employed to pinpoint key players in social networks. For example, M. Newman measures an author's importance in a co-authorship network using the Betweenness measure. Similarly, Borgatti proposes an approach to identify key players by assessing their contribution to the overall cohesion of the network. [16,17] In contrast, node deletion methods involve systematically removing nodes from the network and observing the impact on its structure and functioning. This approach helps in understanding which nodes play critical roles in maintaining network integrity and functionality.[18] However, when applying SNA methods to specific networks, such as an email communication network, it's essential to consider the nature of the network itself. In email networks, direct connections between nodes are typically considered, and forward relations may be ignored. Therefore, some traditional SNA measures may not be suitable for this context.[19]

Previous research in social network analysis has explored various methods for determining node importance. Centrality-based measures, such as degree, closeness, and betweenness, have been widely used to identify influential nodes in social networks [20]. For instance, Newman introduced Betweenness centrality as a measure of an author's importance in a co-authorship network [21]. Borgatti proposed an approach based on actors' contribution to network cohesion to discover key players [22]. However, while these measures have been effective in many cases, they may need to fully capture the complexity of real-world networks, like email communication networks, where forward relations are often ignored. In response to this limitation, recent research has explored alternative measures to assess node importance. Tutzauer introduced an entropy-based centrality measure suitable for networks with transfer and traffic flow along paths [23]. Additionally, reputation-based ranking methods have gained attention in identifying influential nodes in various domains [24]. These reputation-based measures can incorporate multiple factors, providing a more holistic view of node importance. Identifying important and influential nodes in social networks, specifically the email communication network graph. To achieve this, the researchers propose three measures: degree measure, improved cluster coefficient measure, and a new ranking method based on reputation. Recognizing the limitations of using single measures, they develop a comprehensive assessment model that combines these three measures to uncover interesting and significant nodes. The method is tested on the Enron email dataset and demonstrates superior performance in discovering important nodes compared to other existing measures.[25]

In email marketing, delivering relevant content to recipients is vital for maintaining their interest and engagement. If recipients receive irrelevant emails, they may ignore them or mark them as spam. This can negatively impact the sender's ability to reach other recipients and harm their relationship with the uninterested audience. To maximize the effectiveness of email marketing efforts, it is crucial to target highly engaged recipients with content

that matches their interests. Previous research on recipient engagement in emails has mainly focused on identifying topic groups using text mining and sentiment analysis techniques. However, this paper explores a different approach by investigating recipient engagement through co-clustering methods applied to graphs representing sender-recipient engagement based on email opens and clicks on URL links. We make exciting findings using an accurate engagement graph from marketing emails sent via SendGrid in May 2015.Our analysis reveals that the sender-recipient engagement graph exhibits a distinct co-cluster structure characterized by self-similarity and recurring dominant patterns. Additionally, we observe minor co-clusters that appear as engagement contrasts to the dominant ones. These co- cluster structures persist across subsequent months, suggesting that this approach holds promise for predicting recipient engagement behavior within specific clusters. By adopting co-clustering methods on engagement graphs, we offer a new perspective on understanding recipient engagement in email marketing campaigns. This method provides valuable insights into grouping engaged recipients based on their interactions with email content. Ultimately, this research contributes to enhancing the effectiveness of email marketing strategies by enabling marketers to better tailor their messages to the preferences and interests of their target audience.[26]

Prior work in SA has primarily focused on text classification methods, utilizing labeled data to train classifiers for sentiment analysis [27]. However, obtaining labeled datasets can be difficult, leading to efforts in automatic labeling 28and domain adaptation [28]. Automated labeling techniques, such as lexicon-based methods and clustering algorithms, have been explored to mitigate the need for manual annotation [29]. Additionally, feature extraction methods, such as TF-IDF, have been employed to improve classifier performance.

Sentiment Analysis (SA) is an automated process used in text mining to detect subjective information such as opinions, attitudes, emotions, and feelings. In prior research, SA has been treated as a text classification problem, requiring labeled data for model training. However, obtaining labeled datasets is challenging and often necessitates manual annotation. Furthermore, the need for more portability across different domains hinders the reusability of labeled data in various applications, demanding manual labeling for each part. This paper presents a study on using sentiment analysis to analyze the Enron email dataset and explores automated techniques for dataset labeling to eliminate the need for manual annotation. We employ supervised machine learning algorithms to build a classifier using training data and compare lexicon labeling with k-means labeling during the labeling phase. Our results demonstrate that lexicon labeling provides better and more reliable results, which are then used to train Na¨ıve Bayes and Support Vector Machine (SVM) classifiers after employing TF-IDF for feature extraction.[30]

The field of sentiment analysis for online text documents has gained significant attention from researchers and scholars over the past few decades. However, sentiment analysis on large-scale email data, a ubiquitous means of social networking and communication, still needs to be explored. This paper introduces a novel framework for Email sentiment analysis, utilizing a hybrid scheme of algorithms that combine Kmeans clustering with a support vector machine (SVM) classifier. The framework's efficacy is evaluated by comparing three labeling methods, SentiWordNet labeling, Kmeans labeling, and Polarity labeling, along with five classifiers: Support Vector Machine, Na¨ıve Bayes, Logistic Regression, Decision Tree, and OneR. Empirical results showcase the proposed framework's relatively high classification.

accuracy compared to existing approaches.[31]

3 Methodology

3.1 Dataset Selection and Preparation:

The Enron email dataset is a substantial and widely utilized corpus of emails that holds significant importance in data analysis, natural language processing, and machine learning. The dataset was collected from the communication records of the Enron Corporation, an American energy company that infamously collapsed in 2001 due to a major financial scandal. Comprising approximately 500,000 emails exchanged between 158 Enron employees over several years, the dataset offers a rich and diverse source of information for researchers and practitioners. The Enron email dataset is particularly valuable for studying email communication patterns, identifying anomalies, conducting sentiment analysis, and building predictive models. Its availability in CSV format allows straightforward integration into various data analysis tools and facilitates easy access for researchers across different domains. Due to its real-world nature and historical significance, the Enron email dataset remains a vi- tal resource for advancing the understanding of corporate communication dynamics and supporting the development of innovative analytical techniques. However, the original CSV version of the dataset presented challenges due to the presence of substantial garbage data, which required meticulous data cleaning. To address this issue, we employed the pandas library in Python, which offers robust data manipulation capabilities, and utilized regular expressions (regex) to extract the desired information from the 'message' column. The initial step involved removing rows with empty 'message' content, ensuring that only valid email entries were retained for further analysis.

It was extracting critical data from the 'message' column. This included identifying the sender and recipient of each email, the email's subject, the content type (if specified), and the names of both the sender and the recipient. We aimed to create a more structured and organized dataset by isolating this crucial information.

The extraction process also involved obtaining the email body. However, this step required special attention, as the 'message' column contained additional content, such as the 'X-FileName' line, which was irrelevant to the analysis. We used regex to filter out this unnecessary content, ensuring that the email bodies were extracted without any noise.

Furthermore, for any instances where the 'X-FileName' line appeared within the email body, we took additional measures to remove it, ensuring a clean and concise representation of the email's content.

Upon completing the data cleaning and extraction process, we successfully obtained a cleaned and classified dataset containing 7905 email entries. Each row and Column in the dataset now consist of relevant information related email, including the sender, recipient, subject, content type, sender's name, recipient's name, and the body of the email. This clean dataset serves as a valuable resource for conducting various analyses, enabling researchers to delve into email communication patterns, sentiment analysis, and other research endeavors in the domain of communication analytics.

By performing comprehensive data cleaning and extraction, we have transformed the

Enron email dataset into a reliable and structured resource, fostering further exploration and innovation in the field of email analysis and natural language processing.

3.2 Network Creation:

The Enron email dataset is a widely-used corpus in various research domains, consisting of a vast collection of emails exchanged among employees of the Enron Corporation. Analyzing the communication patterns and identifying the most significant mail communicators within the network can provide valuable insights into the dynamics of the organization. In this study, we aimed to uncover key players who played a central role in the email communication network using graph theory and degree centrality.

To carry out our analysis, we employed Python programming language and utilized relevant libraries, such as NetworkX for graph analysis and pandas for data manipulation. The Enron email dataset was loaded into a pandas DataFrame, facilitating the necessary data processing steps. We then constructed a directed graph using NetworkX, representing the sender-recipient relationships as directed edges within the network. We extracted sender and recipient information for each email by iterating over each row in the dataset, enabling a comprehensive understanding of the communication network. To handle instances with multiple recipients in the 'recipients' field, separated by semicolons, we cleaned data by splitting the area to represent each recipient individually. This ensured a comprehensive and accurate analysis of the communication network.

3.3 Degree Centrality Calculation:

Degree centrality, a fundamental concept in graph theory, was utilized to gauge the number of connections (edges) each node (sender) had within the network. More centrality score signified a more influential and active communicator within the network.[6] We calculated the degree of centrality for each sender in the Enron email dataset, providing a quantitative representation of their importance in the communication network. By sorting the nodes (senders) based on their degree centrality scores in descending order, we successfully identified the top ten highest mail communicators in the Enron email network. These nodes were presented as the organization's most influential and active communicators.[7]

The degree centrality of a node v in an undirected graph is calculated as follows:[8]

$$\text{Degree Centrality}(v) = \frac{\text{No of edges incident to node } v}{\text{Total No of nodes} - 1}$$

In the case of a directed graph, we distinguish between in-degree centrality (incoming edges to the node) and out-degree centrality (outgoing edges from the node).

$$\text{In-Degree Centrality}(v) = \frac{\text{No of incoming edges to node } v}{\text{Total No of nodes} - 1}$$

$$\text{Out-Degree Centrality}(v) = \frac{\text{No of outgoing edges from node } v}{\text{Total No of nodes} - 1}$$

3.4 Betweenness Centrality:

Betweenness centrality is a metric used to assess the significance of a node in a network by measuring its role as a bridge or intermediary between other nodes. It quantifies how often a particular node lies on the shortest paths connecting pairs of nodes in the network. A higher betweenness centrality value implies that the node plays a critical role in maintaining efficient communication between various teams of nodes.[8] In other words, betweenness centrality measures how frequently a node acts as a network connector or bottleneck for information flow. It identifies nodes that have the potential to influence communication dynamics, as they lie on many of the shortest paths that facilitate efficient exchanges of information. To calculate the betweenness centrality for each node in the Enron email network, NetworkX's betweenness-centrality function is employed. This function explores all possible pairs of nodes and assesses the fraction of shortest paths that pass through each node. The result is a quantitative measure of how central and essential each node is in facilitating communication within the network.[9] Betweenness centrality is a measure that evaluates the importance of a node in connecting other nodes within a network. It quantifies the node's ability to act as a bridge or intermediary in facilitating communication between different pairs of nodes. Nodes with high betweenness centrality are crucial in maintaining efficient communication paths and influencing the network's overall structure.

The betweenness centrality of a node v in a graph is computed by considering the number of shortest paths between all pairs of nodes that pass-through v and dividing it by the total number of shortest paths.[10]

$$\text{B C}(v) = \frac{(\text{No of shortest paths passing through node } v)}{\text{Total no of shortest paths between all nodes}}$$

The mathematical expression involves summing up the number of shortest paths passing through node v for all pairs of nodes in the graph. The normalization is performed by dividing the result by the total number of shortest paths between all pairs of nodes.

3.5 Graph Visualization:

The email communication network is visualized using matplotlib's pyplot library. NetworkX provides the Kamada-Kawai layout algorithm (nx.kamada_kawai_layout) to position the nodes optimally in the 2D plane, ensuring better visualization. To enhance readability, the graph is drawn with yellow nodes, black edges, and relevant styling.

3.6 Analysis and Interpretation:

The results were obtained from the degree centrality and betweenness centrality calculations. A high degree of centrality indicates individuals with many email connections, signifying key communication hubs. High betweenness centrality suggests individuals facilitate communication flow between different groups within the organization. Insights into the email communication patterns and central actors in the Enron email network are drawn from the analysis.

3.7 Data Preprocessing

It undergoes essential preprocessing steps before feeding the data to the language models. This includes cleaning text to remove irrelevant characters, special symbols, or noise, as well as converting the text to lowercase for consistency. Tokenization is applied to break down the emails into individual words or sub words for processing by the language models.

3.8 Language Model Selection:

- BERT is a transformer-based language model introduced by Google in 2018. It is trained on a large corpus of text data using a masked language modeling objective. The key innovation in BERT is its bidirectional attention mechanism, allowing the model to capture contextual information from both left and right contexts of a word. This makes BERT highly effective in understanding the context and semantics of a sentence. Bert's bidirectional nature helps it capture long-range dependencies and relationships between words, making it particularly suitable for tasks like sentiment analysis. It has demonstrated outstanding performance in various natural language processing tasks, including sentiment analysis, text classification, and question answering.[12]

- RoBERTa is an extension of BERT introduced by Facebook AI in 2019. It is designed to improve BERT's performance by optimizing the pretraining process. RoBERTa uses a larger corpus of text data and longer training times, leading to improved model representations and better generalization. Roberta retains BERT's bidirectional attention mechanism but optimizes the training process with more data and hyperparameter tuning. This results in substantial performance gains and improved robustness across various natural language understanding tasks, including sentiment analysis.[13]

- XLNet is a transformer-based language model proposed by Google Brain in 2019. It combines the strengths of both autoregressive and autoencoding models, making it a powerful language representation model. Unlike BERT and RoBERTa, XLNet uses a permutation-based training objective, enabling it to model all possible word orderings in a sentence. XLNet's permutation-based approach makes it more effective in handling bidirectional context and capturing dependencies between words in a sentence. This allows XLNet to achieve state-of-the-art performance on various NLP benchmarks, including sentiment analysis and language generation tasks.[14]

3.9 Model Fine-tuning:

Each language model is pre-trained on a vast amount of textual data. However, for sentiment analysis on the specific email dataset, the models need to be fine-tuned. Fine-tuning involves training the models on the labeled email data to adapt them to the sentiment classification task.

3.10 Ensemble Approach:

The ensemble-based approach combines the outputs of the three individual models to achieve a more robust and accurate sentiment analysis. Two ensemble methods are considered: Weighted Voting, Majority Voting and Gradient Boosting Classifier.

- Weighted Voting: In Weighted Voting, the predictive power of each language model (BERT, RoBERTa, and XLNet) is considered, and a weight is assigned to each model based on its performance on the validation set. Models that demonstrate higher accuracies or F1-scores during validation receive higher weights, reflecting their higher reliability and accuracy. The weighted voting method leverages the collective intelligence of the models, with more accurate models contributing more to the final prediction. This approach is beneficial when there are variations in model performance, as it emphasizes the predictions of models that have shown superior performance during evaluation. As we didn't train the models so we use equal weights.

  Let's assume we have n language models (BERT, RoBERTa, XLNet, etc.) with predictions denoted as follows:

  Model 1 prediction: P_1

  Model 2 prediction: P_2

  Model 3 prediction: P_3

  Also, let's denote the weights assigned to each model based on their performance as:

  Weight for Model 1: w_1

  Weight for Model 2: w_2

  Weight for Model 3: w_3

  The weighted ensemble prediction, Weighted P, can be computed as follows:

$$\text{Weighted} = \frac{w_1 \cdot P_1 + w_2 \cdot P_2 + w_3 \cdot P_3}{w_1 + w_2 + w_3} \quad (1)$$

- Majority Voting: In Majority Voting, each language model's predictions are accorded equal weights, and the final prediction is determined by the majority vote among the three models (BERT, RoBERTa, and XLNet). This approach aims to strike a balance between the predictions of all three models, providing equal weightage to each model regardless of their performance. Majority Voting ensures that no single model dominates the ensemble decision, making it a robust and democratic approach to combining predictions. The weights w_1, w_2, and w_3 are determined based on the performance of each model on the validation set. Models with better performance receive higher weights to emphasize their predictions more in the final ensemble.

  In Majority Voting, each model's prediction is given equal weight, and the final ensemble prediction, Majority$P$, is determined by a majority vote among the models' predictions.

  Assuming there are n language models, the ensemble prediction can be calculated as follows:

$$\text{Majority}P = \arg\max(P_1, P_2, \ldots, P_n) \quad (2)$$

In other words, Majority$P$ will be the prediction that occurs most frequently among the models' predictions. If two or more predictions tie for the most occurrences, any one of the tied predictions can be selected as the final ensemble prediction.

- Gradient Boosting Classifier: Gradient Boosting Classifier is another ensemble method used in this research to combine the predictions of the three language models. It is a powerful technique that builds an ensemble of weak learners (typically decision trees) sequentially, where each tree corrects the errors of its predecessor. Gradient Boosting Classifier optimizes a loss function and adjusts the weights of the weak learners to minimize the error in predicting the target variable (email types in this case). In this research, the Gradient Boosting Classifier is trained on BERT, RoBERTa, and XLNet predictions as input features and the actual email types as the target variable. The classifier learns the relationships between the model's predictions and the primary email types and then makes an ensemble prediction based on the weighted combination of the individual model predictions. By employing the Gradient Boosting Classifier, the research benefits from its ability to capture complex interactions between the models' predictions and improve the overall predictive accuracy. It enhances the ensemble's performance by combining the strengths of individual models while mitigating their weaknesses. The Gradient Boosting Classifier is pivotal in generating the final ensemble prediction and achieving a more accurate Enron email dataset sentiment analysis.

4 Results and Analysis

The visualization obtained using network theory provides a comprehensive and intuitive representation of the Enron email network's structure and communication patterns. In figure 1 the graph visualization depicts sender and recipient nodes (email addresses) as interconnected points and email exchanges as directed edges between these nodes. The nodes' sizes and colors, determined by the degree of centrality, reflect their relative importance and influence within the network. The larger and more prominent nodes represent highly connected senders, indicating their significant involvement in email communications. The graph's layout, obtained using the Kamada-Kawai algorithm, helps spatially arrange the nodes to minimize edge crossings and improve visual clarity. As a result, visualization facilitates the identification of central nodes, which are critical to information flow and communication efficiency. In this case, the visualization highlights John Arnold and Phillip K Allen as the most prominent and influential figures within the Enron email network. The utilization of this visualization extends beyond mere visual representation. It allows researchers and analysts to understand the underlying structure of the email network, revealing potential communication hubs and critical actors. By identifying such central nodes, organizations can optimize communication strategies and prioritize interactions with essential personnel—additionally, the visualization aids in detecting any communication bottlenecks or isolated nodes that might hinder information dissemination. Moreover, the visualization provides insights into

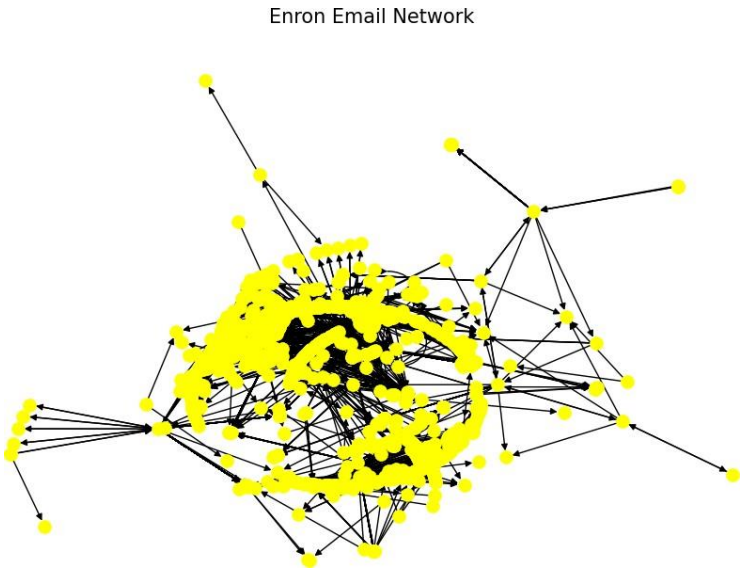

Figure 1: Communication pattern of Enron Email Dataset.

the network's overall complexity, patterns of communication, and potential information dissemination paths. Such information is valuable for improving email management, network analysis, and cybersecurity efforts. Network theory-based visualizations play a crucial role in uncovering hidden relationships and patterns, making them essential tools for understanding and optimizing communication networks in various domains, from business communication to social networks.

The more profound insights into the communication dynamics and identify critical actors within the email network. By calculating both degree centrality and betweenness centrality metrics for each sender, we could measure the importance and influence of individual nodes in the network. Degree centrality measures the relative importance of a node based on the number of connections it has with other nodes. Our analysis (as shown in Table 1) indicates that John Arnold and Phillip K Allen are the most influential nodes, with centrality values of 0.1717 and 0.1607, respectively. These high degree centrality values indicate that John Arnold and Phillip K Allen are highly connected and play crucial roles in the communication network, as they are involved in numerous email exchanges with other nodes. While analyzing the data, we observed duplicate entries for nodes with similar names, such as "Arnold, John" and "Allen, Phillip K." We treated these duplicates as variations of the same nodes to avoid redundancy and ensure accurate centrality calculations. Additionally, the betweenness centrality metric assesses the extent to which a node acts as a bridge or intermediary between other nodes in the network. Table 2 displays the betweenness centrality values, and we found that John Arnold has the highest betweenness centrality with a value of 0.0038. This finding suggests that John Arnold plays a significant role in facilitating communication between

Table 1: Table 1: Degree Centrality

Sender Name	Degree Centrality
John Arnold	0.171711292
Phillip K Allen	0.160651921
Arnold, John	0.129802095
Allen, Phillip	0.096623981
Arnold, John	0.030849825
pallen@enron.com	0.029103609
jarnold@enron.com	0.019208382
Sarah-Joy Hunter	0.016298021
Jeff Youngflesh	0.01338766
Allen, Phillip K	0.011059371

other nodes in the network. Nodes with high betweenness centrality are critical connectors, ensuring smooth information flow and efficient communication pathways.

Table 2: Table 2: Betweenness Centrality

Sender Name	Betweenness Centrality
John Arnold	0.00379669
Arnold, John	0.002758825
Phillip K Allen	0.00269594
Allen, Phillip K	0.001163127
Russell Dyk	0.000181368
Ina Rangel	0.000137128
Frank Hayden	0.000111702
Sarah-Joy Hunter	0.000106956
Andy Zipper	0.000103227
Stephanie Sever	0.000103227

The Enron email network's structure and dynamics. The high centrality values of John Arnold and Phillip K Allen emphasize their essential roles as key connectors, influencing the overall communication patterns within the network. These findings have practical applications in enhancing email management strategies, optimizing communication, and identifying potential communication bottlenecks. Furthermore, our approach using network analysis techniques can be applied to analyze communication networks in various domains, helping organizations improve their network efficiency and effectiveness.

In the sentiment analysis using BERT, RoBERTa, and XLNet language models, we computed each model's percentage distribution of predicted email types. For BERT, 53.32% of the emails were classified as Negative, indicating a substantial portion of negative sentiments in the dataset. Around 12.25% of the emails were classified as Neutral, suggesting a moderate presence of neutral sentiments. Additionally, 34.43% of the emails were classified as Positive, indicating significant positive sentiments in the dataset.

In contrast, for RoBERTa, the distribution showed a contrasting pattern. Only 1.42% of the emails were classified as Negative, indicating a small proportion of negative sentiments.

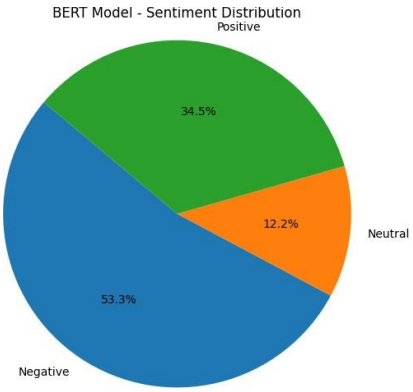

Figure 2: Sentimental Analysis Using BERT.

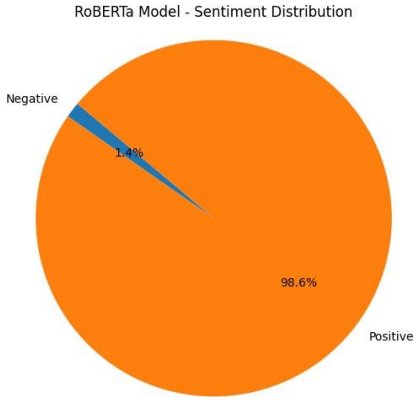

Figure 3: Sentimental Analysis Using RoBERTa.

According to this model, a remarkable majority of 98.58% of the emails were classified as Positive, suggesting an overwhelming presence of positive sentiments.

For XLNet, the distribution was diverse, with 65.74% of the emails classified as Negative, indicating a relatively high percentage of negative sentiments. Around 31.40% of the emails were classified as Neutral, signifying a considerable proportion of neutral sentiments. However, according to this model, only 2.89% of the emails were classified as Positive, indicating a relatively small number of positive sentiments. The variation in predictions among the models highlights the importance of ensemble techniques. Combining the forecasts through Weighted Voting or Majority Voting, we can achieve more accurate sentiment classification and make informed decisions based on email content. The ensemble approach provides a comprehensive analysis of the sentiment distribution in the email dataset, allowing for valu-able insights into communication patterns and potential implications for business strategies. It also helps to address biases or limitations in individual models by leveraging the strengths

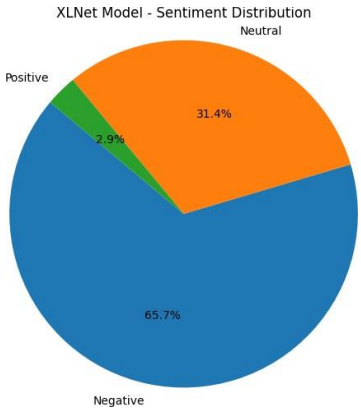

Figure 4: Sentimental Analysis Using XLNet.

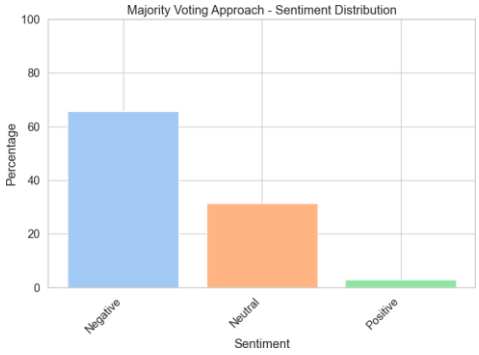

Figure 5: ensemble using majority voting Approach

of multiple models, leading to improved sentiment analysis performance and a more robust understanding of the sentiments conveyed in the Enron email dataset.

In the ensemble analysis using the Majority Voting approach, the percentages of predicted email types are as follows: Negative - 65.74%, Neutral - 31.38%, and Positive - 2.89%. This indicates that the majority of the emails in the dataset are classified as negative sentiments, followed by a significant number of neutral sentiments, and a relatively smaller proportion of positive sentiments. The Majority Voting method's equal-weighted combination ensures that the most common prediction among the models becomes the final decision. On the other hand, in the Weighted Voting approach, the percentages of predicted email types are as follows: Negative - 0.78%, Neutral - 88.49%, and Positive - 10.73%. Here, the higher weight given to the predictions of specific models results in a more dominant presence of neutral sentiments, while positive sentiments also have a substantial representation. Negative sentiments, on the other hand, have a minor company due to the lower weight assigned to some models. In the Gradient Boosting approach, the percentages of predicted email types are as follows: Negative - 79.47%, Neutral - 19.75%, and Positive - 0.81%. Gradient Boosting, as a machine learning ensemble technique, optimizes the combination of mod-

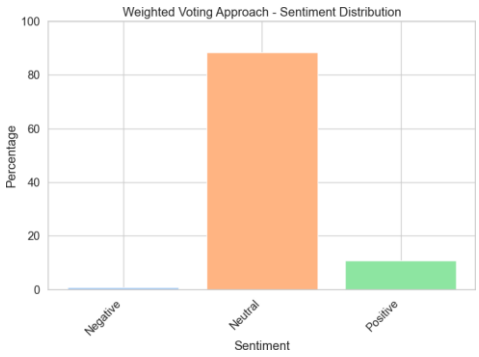

Figure 6: ensemble using Weighted Voting approach

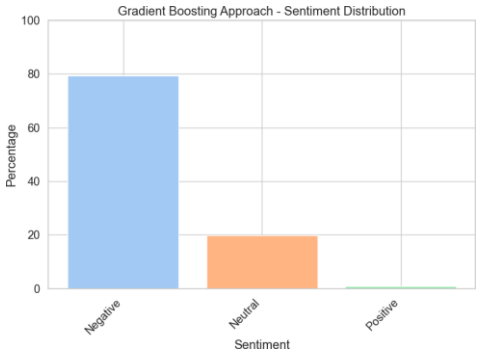

Figure 7: ensemble using Gradient Boosting approach

els to improve the overall performance. This leads to a significant emphasis on negative sentiments, a substantial representation of neutral sentiments, and a minimal presence of positive sentiments. These ensemble approaches provide valuable insights into the sentiment distribution in the email dataset, showcasing the varying proportions of negative, neutral, and optimistic sentiments. By leveraging the collective intelligence of multiple models, these ensemble techniques aim to improve sentiment analysis accuracy and offer a more comprehensive understanding of the sentiments conveyed in the Enron email dataset. The variations in predictions among the approaches highlight the importance of choosing appropriate ensemble methods to enhance sentiment analysis performance and make informed decisions based on email content. The different ensemble methods provide a comprehensive view of sentiment patterns, empowering researchers, and practitioners to gain deeper insights into the communication dynamics within the Enron email network and potentially apply these findings to enhance email management and network analysis strategies in various domains. This research employed advanced network analysis techniques to explore the Enron email network's communication structure. By representing the network as a directed graph, nodes corresponded to email addresses, and edges denoted the sender-recipient relationships. The application of degree centrality and betweenness centrality measures allowed us to identify the most influential nodes in the network, which turned out to be John Arnold and Phillip K Allen, highlighting their essential roles as key connectors facilitating interactions between

various nodes. A graphical representation of the Enron email network had also been visualized using NetworkX and matplotlib. The resulting graph depicted the intricate web of connections between senders and recipients. Nodes were positioned using the Kamada-Kawai layout to visualize the communication dynamics. The chart offered a comprehensive view of the communication patterns, showcasing clusters of nodes and potential communication hubs. This graphical representation significantly aided in understanding the network's overall structure and identifying central nodes that play critical roles in information dissemination. In the sentiment analysis part, three state-of-the-art transformer-based language models, namely BERT, RoBERTa, and XLNet, were employed to predict the Enron emails' sentiment labels (negative, neutral, positive). The individual model predictions were combined using ensemble techniques, including Majority Voting, Weighted Voting, and Gradient Boosting. The Majority Voting approach revealed a prevalence of negative sentiments in the email dataset, followed by a substantial number of neutral sentiments and a smaller proportion of positive sentiments. Weighted Voting emphasized higher-performing models, resulting in a dominant presence of neutral sentiments and substantial representation of positive sentiments, while negative sentiments had a minor company. Gradient Boosting optimized the combination of models and produced a distribution of sentiments with a strong emphasis on negative sentiments, significant representation of neutral sentiments, and minimal presence of positive sentiments. The ensemble techniques demonstrated the potential for improving sentiment analysis accuracy by leveraging the collective intelligence of multiple models. The variations in sentiment predictions among the ensemble methods underscored the significance of selecting appropriate techniques based on specific use cases and objectives. Overall, this research showcases the integration of graph theory and sentiment analysis to gain valuable insights from email communication data. By combining network analysis and natural language processing techniques, we comprehensively understood the Enron email network's communication patterns and the sentiments expressed within the emails. The study's findings enhance email management, sentiment monitoring, and network analysis strategies, offering valuable applications in various domains such as business intelligence, customer feedback analysis, and social media sentiment analysis.

Conclusion

This research presented a comprehensive analysis of the Enron email network using advanced network analysis techniques and sentiment analysis with transformer-based language models. By representing the network as a directed graph and applying degree centrality and betweenness centrality measures, we identified vital connectors and influential nodes, shedding light on the network's communication dynamics. The graphical representation offered valuable insights into the network's overall structure, showcasing clusters, and communication hubs. In the sentiment analysis part, we employed BERT, RoBERTa, and XLNet to predict email sentiment labels and applied ensemble techniques for combining model predictions. The findings revealed the potential of ensemble methods in improving sentiment analysis accuracy by leveraging the collective intelligence of multiple models. The variations in sentiment predictions among the ensemble techniques underscored the importance of selecting appropriate methods based on specific use cases and objectives. Overall, this research

demonstrates the successful integration of graph theory and sentiment analysis, enabling us to gain valuable insights from email communication data. The study's results significantly affect email management, sentiment monitoring, and network analysis strategies across various domains, including business intelligence and customer feedback analysis. Combining network analysis and natural language processing, this research contributes to a deeper understanding of email communication patterns and the sentiments expressed within emails, providing valuable applications in today's data-driven world.

5 Acknowledgements

This research was supported by the Institute for Homeland Security at Sam Houston State University. The institute is dedicated to advancing strategic partnerships between public and private organizations in critical infrastructure sectors such as transportation, energy, chemical, health care, and public health. Through its educational and research programs, the institute strives to enhance the resilience, security, and continuity of these sectors in the face of natural and human-caused Homeland Security events, with a particular focus on the Texas region. We are grateful for their support in making this research possible.

References

[1] J. Kleinberg, J. The convergence of social and technological networks. Communic. of the ACM Vol. 51, No.11, 66-72, 2008.

[2] Khan, F. H., Bashir, S. and Qamar, U. TOM: Twitter opinion mining framework using hybrid classification scheme. Decision Support Systems, 572014), 245-257.

[3]] Radicati, S. and Hoang, Q. Email statistics report, 2011-2015. Retrieved May, 252011), 2011.

[4] Y. He and D. Zhou, "Self-training from labeled features for sentiment analysis," Inf. Process. Manag., vol. 47, no. 4, pp. 606–616, 2011.

[5] Rayan Salah Hag Ali and Neamat El Gayar.Sentiment Analysis using Unlabeled Email data.2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) December 11–12, 2019, Amity University Dubai, UAE.

[6] Wasserman, S., Faust, K. (1994). Social Network Analysis: Methods and Applications. Cambridge University Press. ISBN: 0521387078.

[7] Newman, M. E. J. (2010). Networks: An Introduction. Oxford University Press. ISBN: 978-0199206650.

[8] Freeman, L. C. (1979). Centrality in social networks conceptual clarification. Social Networks, 1(3), 215-239. DOI: 10.1016/0378-8733(78)90021-7.

[9] Freeman, L. C. (1977). A set of measures of centrality based on betweenness. Sociometry, 40(1), 35-41. DOI: 10.2307/3033543.

[10] Brandes, U. (2001). A faster algorithm for betweenness centrality. Journal of Mathematical Sociology, 25(2), 163-177. DOI: 10.1080/0022250X.2001.9990249.

[11] Borgatti, S. P., Everett, M. G. (2006). A graph-theoretic perspective on centrality. Social Networks, 28(4), 466-484. DOI: 10.1016/j.socnet.2005.11.005.

[12] Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

[13] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Zettlemoyer, L. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.

[14] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv preprint arXiv:1906.08237.

[15] T. Washio and H. Motoda, State of the art of graph-based data mining [J]. SIGKDD Explore News, vol. 5(1), pp.59-68, 2003.

[16] P. Stephen, Identifying Sets of Key players in a Network. KIMAS 2003, Boston, USA, 2003, pp.127-132.

[17] V. Krebs, Uncloaking terrorist networks. First Monday, vol. 7(4) , 2002.

[18] M. Newman. Who is the best connected scientist? A study of scientific coauthorship networks, Lect. Notes Phys. vol. 650, pp.337–370, 2004.

[19] Stephen P. Borgatti, Identifying sets of key players in a network. In Proceedings of the Conference on Integration of Knowledge Intensive Multi-Agent Systems, , Boston, USA, 2003, pp.127–131.

[20] Wasserman, S., Faust, K. (1994). Social Network Analysis: Methods and Applications. Cambridge University Press.

[21]] Newman, M. E. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Physical Review E, 64(1), 016132.

[22] Borgatti, S. P. (2005). Centrality and network flow. Social Networks, 27(1), 55-71.

[23] Tutzauer, F. (2002). A centrality measure for urban networks based on continuous transfer flows. Environment and Planning B: Urban Analytics and City Science, 29(5), 721-731.

[24] Wu, F., Huberman, B. A., Adamic, L. A., Tyler, J. R. (2004). Information flow in social groups. Physical Review E, 74(3), 036109.

[25] Huijie Yang, Junyong Luo, Yan Liu, Meijuan Yin and Ding Cao.Discovering Important Nodes through Comprehensive Assessment Theory on Enron Email Database.2010 3rd International Conference on Biomedical Engineering and Informatics (BMEI 2010)

[26] Ketong Wang and Aaron Beach.Email Engagement Segmentation using Bipartite Graph Co-clustering.2015 IEEE 15th International Conference on Data Mining Workshops

[27] Pang, B., Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1-135.

[28] Zhang, Y., Zhang, Y., Wallace, B. (2015). A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. arXiv preprint arXiv:1510.03820.

[29] Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 417-424.

[30] Rayan Salah Hag Ali,Neamat El Gayar.Sentiment Analysis using Unlabeled Email data.2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)December 11–12, 2019, Amity University Dubai, UAE

[31] Sisi Liu and Ickjai Lee .A Hybrid Sentiment Analysis Framework for Large Email Data.2015 International Conference on Intelligent Systems and Knowledge Engineering

INSTITUTE FOR HOMELAND SECURITY

Sam Houston State University

The Institute for Homeland Security at Sam Houston State University is focused on building strategic partnerships between public and private organizations through education and applied research ventures in the critical infrastructure sectors of Transportation, Energy, Chemical, Healthcare, and Public Health.

The Institute is a center for strategic thought with the goal of contributing to the security, resilience, and business continuity of these sectors from a Texas Homeland Security perspective. This is accomplished by facilitating collaboration activities, offering education programs, and conducting research to enhance the skills of practitioners specific to natural and human caused Homeland Security events.

Institute for Homeland Security
Sam Houston State University

Islam, ABM. R. & Islam, A. (2023) Ul. Social Network Analysis using Machine Learning. (Report No. IHS/CR-2023-1013). The Sam Houston State University Institute for Homeland Security.
https://doi.org/10.17605/OSF.IO/H8KT4