Journal of
Information Systems
Applied Research

Volume 17, No. 1
April 2024
ISSN: 1946-1836

In this issue:

iscap
Information Systems
& Computing
Academic Professionals

The **Journal of Information Systems Applied Research** (JISAR) is a double-blind peer reviewed academic journal published by ISCAP, Information Systems and Computing Academic Professionals. Publishing frequency is three issues a year. The first date of publication was December 1, 2008.

JISAR is published online (https://jisar.org) in connection with the ISCAP (Information Systems and Computing Academic Professionals) Conference, where submissions are also double-blind peer reviewed. Our sister publication, the Proceedings of the ISCAP Conference, features all papers, teaching cases and abstracts from the conference. (https://iscap.us/proceedings)

The journal acceptance review process involves a minimum of three double-blind peer reviews, where both the reviewer is not aware of the identities of the authors and the authors are not aware of the identities of the reviewers. The initial reviews happen before the conference. At that point papers are divided into award papers (top 15%) and other submitted works. The non-award winning papers are subjected to a second round of blind peer review to establish whether they will be accepted to the journal or not. Those papers that are deemed of sufficient quality are accepted for publication in JISAR. Currently the acceptance rate for the journal is approximately 35%.

Questions should be addressed to the editor at editor@jisar.org or the publisher at publisher@jisar.org. Special thanks to members of ISCAP who perform the editorial and review processes for JISAR.

2024 ISCAP Board of Directors

Journal of
Information Systems Applied Research

Using Topic Modeling to Identify Factors Influencing Job Satisfaction in the IT Industry

Muge Kosar
mkosar1@student.gsu.edu
Andrew Young School of Policy Studies

Frank Lee
flee@gsu.edu
J. Mack Robinson College of Business

Georgia State University
Atlanta, Georgia, US

Abstract

This study aims to identify factors influencing job satisfaction in the IT industry. Using the Latent Dirichlet Allocation (LDA) machine learning technique, over 5,000 employee reviews from 17 leading digital companies are analyzed. The analysis reveals nine key topics IT employees value: management skills and responsibilities, workplace culture and environment, and job experience and flexibility. The comparison of LDA topics with those of a human rater shows moderate overlap with topics identified by independent researchers in work culture and environment, management, career advancement, work-life balance, and compensation and benefits. The sentiment analysis reveals that most reviews are positive.

Keywords: IT Industry, Job Satisfaction, Latent Dirichlet Allocation, Employee Reviews, Sentiment Analysis

Recommended Citation: Kosar, M., Lee, F. (2024). Using Topic Modeling to Identify Factors Influencing Job Satisfaction in the IT Industry. *Journal of Information Systems Applied Research.* 17(1), pp. 4-20, https://doi.org/10.62273/TVTU9122

Using Topic Modeling to Identify Factors Influencing Job Satisfaction in the IT Industry

Muge Kosar and Frank Lee

1. INTRODUCTION

The significance of employee satisfaction in organizations cannot be overstated, as it directly impacts a company's success and competitiveness (Oshagbemi, 2003). Given the growing significance of Information and Communication Technologies (ICT) in today's environment, it has become crucial for organizations to understand the factors that contribute to employee satisfaction and align them with their corporate strategy (Holland & Bardoel, 2016). Effective human resource management is critical in motivating employees and reducing turnover (Sainju et al., 2021).

Job satisfaction studies have explored a wide range of factors that impact job satisfaction, including employee motivation, corporate performance, absenteeism, turnover, and the financial performance of companies. Personal attributes and organizational factors such as age, gender, level of education, company size, and industry have also been examined concerning job satisfaction.

This study uses topic modeling algorithms to determine factors in IT industry employee satisfaction and utilizes a multi-step process for text analysis. The first step involves extensive text pre-processing, including removing stop words, lemmatization, extracting nouns, and generating bigrams. The second step employs Latent Dirichlet Allocation (LDA) topic modeling to identify critical topics from a dataset of 5,000 employee reviews from 17 leading digital companies. The third step involves comparison with human rater results to ensure the accuracy and consistency of the analysis. Finally, sentiment analysis is conducted to classify each review's tone, providing insight into employees' positive and negative sentiments. Through this comprehensive process, the study provides valuable insights for human resource management, helping them to improve employee satisfaction and, thus, increase the company's competitiveness.

2. LITERATURE REVIEW

Employee satisfaction is an examination of existing research on how satisfied employees are with their current employment in their field. Some studies identify essential employee satisfaction factors, such as job satisfaction, work-life balance, organizational culture, and employee retention. Additionally, other studies examine the impact of these factors on employee engagement, productivity, and turnover rate. Appendix A, Table 1 summarizes the literature review.

This study examines job satisfaction factors among IT industry employees through the topic modeling algorithm LDA. Text mining has been identified as a practical approach for analyzing employee reviews and identifying job satisfaction factors. Studies such as Jung & Suh (2019) and Luo et al. (2016) have found that text mining can extract information from employee reviews that may be difficult to identify through other methods.

Trends in IT Industry

Since the pandemic in the year 2020, there has been a significant rise in layoffs. Barnett and Li (2023) estimate that there has been an increase in tech layoffs since the pandemic began in 2020. The estimates consist of prominent corporations such as Meta Platforms, the parent company of Facebook, and Amazon, along with smaller enterprises within the United States and abroad(Barnett & Li, 2023). Deagon (2023) reported that this trend is due to companies hiring excessively during the pandemic and laying off workers due to decreased demand for tech products. Furthermore, according to a report by Challenger Gray and Christmas Inc. (2023), the number of job layoffs reported in February 2023 was the most for the month since 2009, indicating the trend of layoffs in the IT business is still ongoing.

Layoffs in the IT sector are expected to affect employee satisfaction and well-being, emphasizing the importance of effective employee support strategies. Businesses may improve employee retention and satisfaction and boost long-term success and competitiveness by doing so.

3. DATA & METHODOLOGY

This study collected over 20,000 United States top digital company reviews from Indeed.com using scraping. The analysis focuses on a subset of more than 5,000 reviews from January 2018 through October 2018 related to 17 selected IT companies. An employee review contains the textual description and a star rating in which 1 means a negative experience, and 5 indicates a positive experience. Appendix B, Figure 1 summarizes all the processes for this paper.

Text Pre-processing

Latent Dirichlet Allocation (LDA), a popular machine learning approach commonly used for topic modeling, is employed to analyze the impact of work-life balance on job satisfaction. After the employee reviews have been collected, the first step is to pre-process the data to prepare it for analysis. The initial analysis stage involves data pre-processing, which includes cleaning the employee reviews by removing HTML tags, web links, punctuation marks, non-alphanumeric characters, special symbols, and white spaces. All the text data is converted to lowercase, and duplicated rows are removed. Tokenization is then performed, where the data is broken down into individual words or phrases. Stop words, which do not hold analytical value (e.g., "a," "and" "the"), are removed. Lemmatization reduces words to their root form, and nouns are extracted. N-grams are generated to capture the relationship between words and the context in which they appear, specifically bigrams which are contiguous sequences of two words. Incorporating N-grams aims to enhance the analysis and gain a more nuanced understanding of the textual data.

Topic Modeling

Topic modeling is a technique for uncovering hidden topics in a large text corpus. The Latent Dirichlet Allocation (LDA) algorithm is one of the most widely used topic modeling methods. It represents documents arising from multiple topics, where a topic is defined as a distribution over a fixed vocabulary of terms by Blei and Lafferty (2009). LDA asserts that probabilistically distributed topics can be represented by words, as described by Blei et al. (2003). The text data must be pre-processed and expressed numerically to apply LDA for topic modeling.

Creating a bag-of-words (BoW) text representation is essential for the LDA model. The bag-of-words model is a statistical framework for representing text data as a numerical vector, where each dimension corresponds to a unique word in the text corpus's vocabulary, and its focus is solely on the frequency of occurrence of each word in the document. (Zhang et al., 2010). The first step in this process is to split pre-processed data into training and test sets with a 0.4 test and 0.6 training split, a commonly used ratio. This split is done to evaluate the performance of the model on unseen data. Next, a dictionary is created to represent the vocabulary of the text corpus using tokenized data. An individual integer ID is given to each distinct word in the corpus. As LDA uses numerical data rather than text data, this is important. Once the dictionary is created, a document-term matrix is generated. Each row indicates a document, and each column shows a word in this numerical representation of the text data. The matrix shows word frequency in the document. This matrix is referred to as a "bag-of-words" (BoW) representation since it only considers the frequency of each expression and ignores the order in which the words appear in the document.

The underlying topics in the corpus can be determined using LDA by numerically expressing the text data using a document-term matrix. The objective is to pinpoint the subjects most relevant to the data and use them to comprehend the text corpus.

The LDA model can be used after creating the BoW representation. The model's hyperparameters must first be optimized before it can be used. This process is important since they control the model's behavior and can significantly affect the quality of the topic model provided. A grid search approach is used to find the alpha and beta values that maximize the coherence score.

When the alpha and beta hyperparameters have been optimized, the LDA model can be used on the training and test set. After using the LDA model on the training and test sets, assessing its performance is crucial to ensure the topics it generates are meaningful and understandable. To evaluate the performance of the LDA model, perplexity and coherence scores, commonly used metrics, are calculated for both the training and test data sets to assess the performance of the LDA model.

Perplexity measures how well the model predicts new data, and lower values indicate better performance (Blei, 2003). Coherence refers to a metric used to evaluate the level of semantic similarity among the most relevant or high-scoring words in a topic. (Röder et al., 2015). It measures how interpretable and meaningful the model's output is. Higher coherence values indicate better performance.

By evaluating both the perplexity and coherence scores, the quality of the topic model produced by the LDA algorithm can be determined. Low perplexity and high coherence scores show that the LDA model has made meaningful topics that accurately represent the text corpus's fundamental themes and recurring patterns.

The final step in the LDA topic process involves assigning topics to all the data and calculating the probability that each text belongs to each topic. This step is essential as it enables us to recognize the underlying themes and patterns existing in the corpus of texts, which can be helpful for later tasks like sentiment analysis.

In conclusion, the alpha and beta hyperparameters of the LDA model are tuned before being applied to the text corpus. The model's performance is assessed, topics are allocated to all data, and the probability of the topic that each text belongs to each subject is calculated.

Comparison with Human Rater

To compare the topics generated by the LDA model with those of human raters, four independent researchers who understand natural language processing and textual analysis were requested to examine employee reviews gathered and identify the topics mentioned. The four researchers belonged to the same university and individually picked 33 reviews at random for this task. The objective is to compare their results to the LDA's to identify similarities or differences.

The Jaccard similarity score measures the similarity between two sets by dividing the intersection size by the union size. This method was also employed by Guo et al. (2017) to compare the dimensions of previous studies and their analysis. This study uses the Jaccard similarity score to compare the topics identified by an LDA model to those identified by independent human researchers. A comparison is made between the topic assignments of the LDA model and those of four independent human researchers (A, B, C, and D). The score ranges from 0 to 1, where 0 suggests no overlap between the two sets and 1 indicates a perfect match.

This approach is helpful for several reasons. It first provides a benchmark against which to compare the performance of the LDA model. We may assess the model's topic assignment accuracy by comparing the topics the model identified with those identified by human raters. Second, this process can give insight into how different people perceive and identify the topics in the same set of texts. The LDA model and human raters can be compared to reveal their strengths and weaknesses. Lastly, the LDA model's topics may be validated using this approach. The topics are significant and relevant to the text corpus if the topics identified by the model and human raters overlap.

Overall, comparing the LDA model's topic assignments with the outcomes of human raters helps assess the model's performance, understand how various individuals perceive topics, and validate the topics the model identified.

Sentiment Analysis

Sentiment analysis analyzes written or spoken material, such as a social media post or review, to assess its emotional tone. (Jung & Suh, 2019) The procedure uses computational and natural language processing to identify and extract text sentiment. Numerous techniques and tools are available for sentiment analysis, with the Natural Language Toolkit (NLTK) being one of the most popular. The NLTK Python library provides plenty of tools and pre-trained models for sentiment analysis. This library allows the text to be analyzed and classified as positive, negative, or neutral. (Yao, 2019) This study uses the VADER (Valence Aware Dictionary for Sentiment Reasoning) lexicon of the NLTK library to generate sentiment analysis for each topic.

VADER measures the text's overall sentiment score by aggregating the sentiment scores of individual words from a dictionary of terms labeled with their corresponding sentiment scores (positive, negative, or neutral). (Elbagir & Yang, 2019) With this technique, it is possible to understand the sentiment expressed in the text in more depth and nuance.
Consequently, by using the VADER lexicon of the NLTK library, we can conduct sentiment analysis on the topics generated by the LDA model and understand the emotional tone and sentiment expressed in the text.

4. RESULTS

After pre-processing, text data only includes nouns and bigrams. According to Figure 2 and

Figure 3, the most frequent word is management, and the most frequent bigram is life balance.

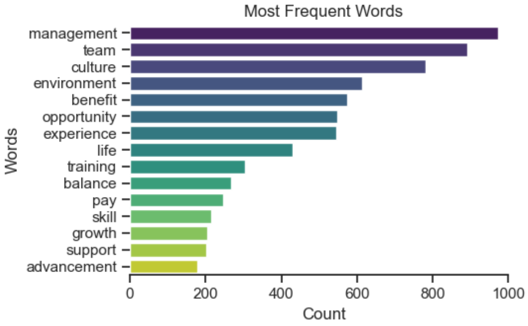


Figure 2 The Most Frequent Words

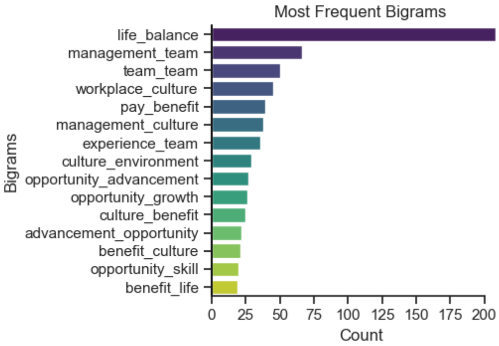


Figure 3 The Most Frequent Bigrams

The final alpha and beta values were found to be 0.31 through the LDA tuning and grid search processes. It is important to emphasize that the entire (100%) corpus, not just a subset (75%), was used for the tuning and grid search process. The whole corpus is used for LDA tuning and grid search to ensure that the resulting topic model is representative of the entire corpus. Suppose only a subset of the corpus was used. In that case, likely, the resulting model might not fully account for the complexity of the text corpus, which could lead to biased or unreliable conclusions.

Alpha and beta are critical parameters in the LDA model that affect the final topic model. As Blei (2012) explains, alpha determines each document's topic distribution, with higher alpha values indicating that each document is more likely to contain diverse topics. In contrast, beta is a parameter that controls the distribution of words within each topic, with higher beta values indicating that each topic is more likely to include multiple words. Appendix C, Table 2 shows alpha and beta values with the highest coherence score. The topic modeling metric with the highest coherence score is crucial since it shows that the results are more understandable and coherent. High coherence scores make each topic's top words more semantically similar, making them easier for human interpreters to comprehend. Therefore, the alpha and beta values that give the highest coherence score were found to be 0.31.

The LDA model is trained using these values after determining the optimal alpha and beta values. The final model has nine topics. Appendix D, Table 3 shows job satisfaction factors and their associated keywords identified by the LDA model. This information helps us understand the key topics and trends in the text corpus linked to job satisfaction.

By identifying these factors and keywords, we can determine which aspects of employee job satisfaction are most important. This information can assist firms in enhancing job satisfaction by helping them understand the factors that affect it.

Additionally, the LDA model is used to compute the frequency of each topic in the text corpus. The frequency of each topic in the text corpus can give helpful insight into which employees frequently mention topics and, as a result, may be particularly significant or influential for their overall work experience and job satisfaction.

Rank	Topic	Frequency	Percentages
1	1	1348 docs	23.19
2	2	923 docs	15.86
3	6	650 docs	11.18
4	9	621 docs	10.68
5	8	576 docs	9.91
6	7	576 docs	9.91
7	3	538 docs	9.25
8	5	344 docs	5.92
9	4	238 docs	4.09

Table 4: Dominant Topics and Their Ranking

Table 4 displays a detailed ranking of topics from the LDA model based on frequency.

Frequencies indicate each topic's importance to study respondents; one such frequency, 23.19% for Management Skills and Responsibilities, topped this list, showing its prominence among employees regarding job satisfaction. By ranking topics this way, we aim to illustrate which factors most contribute to job satisfaction among participants.

Table 5 presents the performance scores for the LDA model, including coherence scores and perplexity scores for both the training and test datasets. Higher coherence scores indicate more coherent topics. Conversely, the perplexity score measures the model's prediction ability, with lower numbers suggesting better ability.

Set	Perplexity Score	Coherence Score
Training	-6.79	0.58
Test	-7.25	0.77

Table 5: Performance Scores

The training dataset's coherence score is 0.58, which indicates that the LDA model's generated topics are moderately coherent. The test dataset's coherence score is more excellent (0.77), demonstrating the LDA model's ability to generalize effectively to new data and provide more coherent topics. This coherence score is crucial since it suggests that the LDA model can accurately capture text corpus patterns and themes even with new data.

The training dataset's perplexity score is -6.79, which shows that the LDA model can predict the text corpus relatively accurately. The test dataset's perplexity score is slightly higher, at -7.25, indicating that the LDA model can still accurately predict it.

According to the scores, the LDA model can produce accurate and coherent topics and generalize effectively to new data. The scores suggest that the LDA model may be able to reveal themes and patterns in large text collections.

To assess the LDA model's accuracy and reliability, topic assignments from the LDA model were compared to the ratings provided by human raters. The comparison explicitly considers whether the LDA model can reliably identify the themes mentioned in the text corpus and whether its conclusions are compatible with human assessors' reports.

The comparison's findings revealed that both the LDA model and human raters identified similar themes, particularly related to work culture, management, work-life balance, and benefits.

Appendix E, Table 6 provides a comprehensive summary of how topic identification by the LDA model and human evaluation is similar. Comparing these results helps us assess the LDA model's accuracy, reliability, and ability to detect text corpus topics.

According to Jaccard scores in Table 6, The Jaccard similarity scores obtained are as follows:

- LDA vs. Independent Researcher A: 0.42
- LDA vs. Independent Researcher B: 0.6
- LDA vs. Independent Researcher C: 0.42
- LDA vs. Independent Researcher D: 0.5

In this study, the Jaccard similarity scores show moderate agreement between the LDA model and the independent researchers. The LDA model and Independent Researcher B had the highest agreement (0.6), followed by Independent Researcher D (0.5), and the LDA model and Independent Researchers A and C had the lowest agreement (0.42).

In this study's final step, employee reviews were analyzed using sentiment analysis to determine how individuals felt about various job satisfaction-related topics. Figure 4 shows that most reviews were positive, organizational strategy and commitments were the least positive, and compensation and benefits were the most positive.

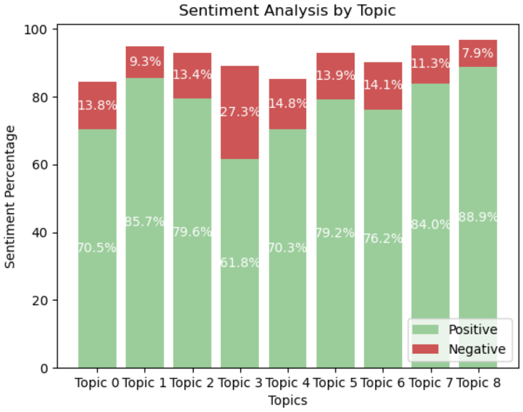


Figure 4 Sentiment Analysis by Topics

Word clouds were created for both positive and negative sentiment words related to all factors to understand better the sentiment conveyed towards specific job satisfaction factors, as seen in Figures 5 and 6. The word clouds highlight the words that refer to job satisfaction factors most frequently in positive and negative ways. According to Figures 4 and 6, benefit, management, and culture are positive words, while complaint, training, and pressure are negative.

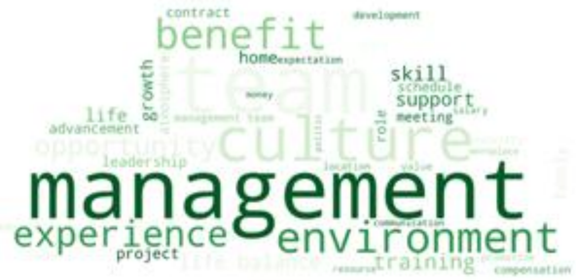


Figure 5 Positive Sentiment Words

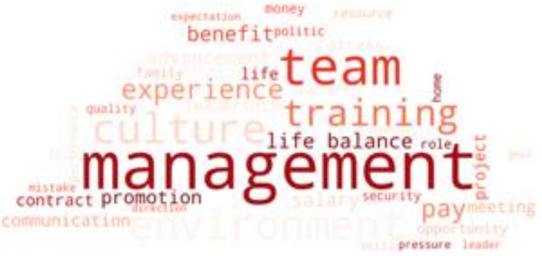


Figure 6 Negative Sentiment Words

Findings

Based on the analysis conducted in this study, several key findings can be drawn about the factors affecting employee job satisfaction.

This study used the LDA model to find job satisfaction factors in the IT industry and began with text pre-processing as the first step of the analysis. After pre-processing the text data to include only nouns and bigrams, the study of the most frequent words and bigrams identified management as the most frequent word and life balance as the most frequent bigram. This finding supports the importance of management-related factors and work-life balance in employee job satisfaction, as identified in the subsequent analysis using the LDA model and sentiment analysis techniques.

The LDA model identified nine critical topics related to job satisfaction, including management skills and responsibilities, compensation and benefits, work-life balance, and organizational culture and environment. Frequency analysis revealed management skills and responsibilities as the most frequently discussed topic, ranking first overall and underscoring their significance to employee job satisfaction. The coherence and perplexity scores for the LDA model were also evaluated to determine its performance. The results showed that the LDA model could generate coherent and accurate topics and generalize well to new data.

The results of the Jaccard similarity score indicate that the LDA model's topic assignments show some similarity to those made by independent researchers. Comparing the LDA model's topic assignments with the results of human raters also showed similarities between the topics identified by the two groups, with the work culture and environment, management, and work-life balance. The results indicate moderate agreement between the LDA model and the human researchers, with some variation observed across different comparisons.

The highest Jaccard similarity score was observed between the LDA model and Independent Researcher B (0.6), suggesting that the model's topic assignments, in this case, were in closer agreement with the human researcher. This could be attributed to the specific nature of the text samples analyzed, the parameters used in the LDA model, or the topic categorization criteria employed by Researcher B, which may have been more consistent with the model's output.

On the other hand, the lowest agreement was observed between the LDA model and Independent Researchers A and C (0.42). The differences between the LDA model and these researchers could be due to various factors, such as differences in the researchers' domain expertise, subjective interpretation of the text, or the inherent limitations of the LDA model in capturing subtle distinctions between topics. This finding suggests that there may be specific topic assignments or nuances in the text data that the LDA model struggled to capture accurately.

The average Jaccard similarity score between the LDA model and Independent Researcher D (0.5) further supports the notion that the model's performance exhibits some similarity to human annotators, but there is still room for improvement. The discrepancies observed could result from the LDA model's assumptions, parameter settings, or the quality of the input data.

Finally, sentiment analysis was used to determine the overall sentiment toward job satisfaction factors. The study showed that most employee reviews were positive, with compensation and benefits receiving the highest positive sentiment. Organizational strategy and commitments received the highest percentage of negative sentiment. The word clouds generated for positive and negative words related to job satisfaction factors highlighted the most frequently used positive and negative words related to all elements.

Discussion

In the IT industry, employee satisfaction has become increasingly important due to the impacts of the COVID-19 pandemic on layoffs. As companies compete to attract potential employees with valuable offers, there is a need for additional research on employee satisfaction and feasible solutions.

Several studies examined how vital employee satisfaction is in the IT industry, and the results show that certain factors are crucial. Ganga's (2022) research found that a positive work environment and an excellent work-life balance were the two most important determinants of job satisfaction across industries. This study also emphasizes the importance of the workplace in the IT sector, suggesting that organizations should prioritize a comfortable and productive workplace.

Moreover, the research by Sainju et al. (2021) demonstrates that management is also a significant factor in the IT sector's job satisfaction. Employees in this industry value effective management practices and need a supportive, empowering work environment to thrive. The study results highlight the significance of companies implementing management strategies prioritizing employee satisfaction.

Moro et al. (2021) identified work exhaustion as the primary cause of job dissatisfaction in the IT industry. However, this current study found that job experience and flexibility are the most dissatisfying factors. These results suggest that companies must create a flexible work environment that allows employees to enhance their job experience.

Finally, Jung and Suh's (2019) study found that project planning is a new factor for job satisfaction. Similarly, this study highlights the importance of project planning in job satisfaction. These findings suggest that companies must prioritize project planning as a critical factor in creating a conducive work environment for their employees.

One of the unique characteristics of this study is its in-depth investigation of factors not addressed prominently in previous research. This research focuses on IT industry specifics, unlike previous studies that focused on generic issues like workplace environments and management. We highlight the often under-emphasized significance of job experience and flexibility as key determinants of job satisfaction. The findings show that project planning is crucial to IT job satisfaction. Using powerful NLP techniques and an in-depth dataset, this research reveals what IT employees genuinely value, enabling focused human resource practices to improve job satisfaction.

The Jaccard similarity scores obtained in this study demonstrate that the LDA model's topic assignments show some similarity to those made by human researchers. However, the varying levels of agreement across different comparisons suggest that the model's performance could be further optimized to better align with human assessments. Future research could focus on refining the LDA model's parameters, exploring alternative topic modeling approaches, or incorporating additional domain knowledge to improve the model's performance and achieve better agreement with human annotators.

Overall, this research offers valuable insight into the factors contributing to job satisfaction in the IT industry. By acknowledging the significance of work environment, management practices, work flexibility, and project planning, businesses can create a better workplace for employees and improve their job satisfaction.

Challenges

The results of this study highlight the importance of utilizing human rater and LDA topic modeling in analyzing employee reviews for IT companies. Both methodologies observed work culture and environment, management, and work-life balance and benefits, which were similar. The result indicates that using multiple methods can provide a more comprehensive understanding of the topics in the data.

However, the human rater exercise also revealed challenges in analyzing text data. One of the challenges was the variability in the number of topics identified by each independent researcher. Although the goal was to obtain nine topics from each researcher, some identified only 7-8 topics. The various approaches and findings from each group highlight the subjectivity of human coding and the possibility of inconsistency in the coding process. Therefore, it is essential to establish clear coding guidelines and procedures for ensuring analysis consistency and accuracy. In addition, the complexity of the data requires expertise in both natural language processing and the industry being analyzed. Interdisciplinary collaboration and expertise in conducting text analysis research are needed.

Practical Implications and Implementation for IT Companies

Practicality is paramount in our research findings since theoretical understandings do not lead to real-world changes alone. This analysis diagnoses and informs IT firms on job satisfaction determinants. IT firms seeking actual benefits from our study should consider how it may be used.

By understanding and addressing key elements such as management effectiveness, work-life balance, and organizational culture, companies can improve employee satisfaction rates and enhance retention rates - ultimately decreasing turnover costs while creating a cohesive and productive work environment. Satisfied workers can boost innovation, productivity, and brand image but may also create unique challenges.

Translating research into practice can present unique challenges, with companies facing initial resistance when trying to alter long-standing practices or introduce new ones. Allocating appropriate time and finances can be demanding; furthermore, after implementation, a constant feedback process is needed to guarantee that new strategies fulfill employee demands and adapt to changing dynamics. Despite such difficulties, however, IT companies could stand to gain immensely by making this transition systematically and with dedication; their potential benefits could be immense.

6. CONCLUSIONS

This paper explores IT job satisfaction factors by analyzing five thousand employee reviews from 17 prominent digital companies using LDA. Data pre-processing includes stop word removal, lemmatization, noun extraction, and bigram generation for nuanced text data analysis. The study identifies nine key topics IT employees value, with management being the most frequent word and life balance being the most frequent bigram. Comparison with the topics identified by independent researchers reveals moderate overlap, particularly in work culture and environment, management, and work-life balance and benefits. The Jaccard similarity scores obtained in the study highlight that the LDA model's topic assignments show some similarity to those made by independent researchers.

The sentiment analysis is also conducted to classify each review's tone, finding that most reviews are positive. Notably, compensation and benefits have the highest percentage of positive sentiment, while organizational strategy and commitments have the highest rate of negative sentiment.

Finally, the analysis visualizes positive and negative sentiment words related to all factors, with benefit, management, and culture as positive words, while complaint, training, and pressure are negative.

However, it is essential to note that this study only analyzed employee reviews from 17 leading digital companies. Thus, the results may not represent the broader IT industry, as smaller or less prominent companies were not included.

In summary, this study provides valuable insights into what IT employees value in their work environment and highlights the importance of factors to improve employee satisfaction and retention. Organizations can use these findings to devise more effective support strategies that ensure success for IT employees. Future studies may expand by studying diverse companies or including more variables for a comprehensive understanding of job satisfaction within IT firms. Still, nonetheless, this work serves as a basis for companies looking to enhance employee well-being.

7. REFERENCES

Barnett, A., & Li, M. (2023, January 3). Tech layoffs are happening faster than at any time during the pandemic*. The Wall Street Journal.* Retrieved March 6, 2023, from https://www.wsj.com/articles/tech-layoffs-are-happening-faster-than-at-any-time-during-the-pandemic-11672705089

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

Blei, D. M., & Lafferty, J. D. (2009). Topic models. In A. Srivastava, & M. Sahami

(Eds.), *Text Mining: Classification, Clustering and Applications* (pp. 71-93). Cambridge: Chapman and Hall/CRC.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Challenger Gray & Christmas Inc. (2023). Job cuts hit 77,770 in February 2023; highest ytd since 2009. *Challenger Gray & Christmas Inc.* Retrieved March 13, 2023, from https://omscgcinc.wpenginepowered.com/wp-content/uploads/2023/03/The-Challenger-Report-February23-1.pdf

Dai, W., & Wu, N. (2017). Profiling essential professional skills of chief data officers. In *Proceedings of Twenty-Third Americas Conference on Information Systems.*

Deagon, B. (2023). Tech industry layoffs show no signs of abating as businesses undo overhiring. *Investor's Business Daily.* Retrieved March 10, 2023, from https://www.investors.com/news/technology/tech-layoffs-show-no-signs-of-slowing/

Edmans, A. (2011). Does the stock market fully value intangibles? Employee satisfaction and equity prices. *Journal of Financial Economics*, *101*(3), 621–640. https://doi.org/10.1016/j.jfineco.2011.03.021

Elbagir, S., & Yang, J. (2019). Twitter sentiment analysis using natural language toolkit and VADER sentiment. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, *122*.

Ganga, A. (2022). Employees satisfaction in different industries: an exploratory review of the literature. *International Journal of Economics and Management Systems*, *7*. http://www.iaras.org/iaras/journals/ijems

Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management*, *59*, 467–483. https://doi.org/10.1016/j.tourman.2016.09.009

Holland, P., & Bardoel, A. (2016). The impact of technology on work in the twenty-first century: exploring the smart and dark side. *International Journal of Human Resource Management*, *27*(21), 2579–2581. https://doi.org/10.1080/09585192.2016.1238126

Huselid, M. A. (1995). The impact of human resource management practices on turnover, productivity, and corporate financial performance. *Academy of Management Journal*, *38*(3), 635-872.

Jung, Y., & Suh, Y. (2019). Mining the voice of employees: a text mining approach to identifying and analyzing job satisfaction factors from online employee reviews. *Decision Support Systems*, *123*. https://doi.org/10.1016/j.dss.2019.113074

Kalra, V., & Aggarwal, R. (2017). Importance of text data pre-processing & implementation in rapidminer. In *Proceedings of the First International Conference on Information Technology and Knowledge Management*, *14,* 71–75. https://doi.org/10.15439/2018KM46

Luo, N., Zhou, Y., & Shon, J. J. (2016). Employee satisfaction and corporate performance: mining employee reviews on Glassdoor.com. In *Proceedings of the 37th International Conference on Information Systems.*

Moro, S., Ramos, R. F., & Rita, P. (2021). What drives job satisfaction in IT companies? *International Journal of Productivity and Performance Management, 70*(2), 391–407. https://doi.org/10.1108/IJPPM-03-2019-0124

Oshagbemi, T. (2003). Personal correlates of job satisfaction: empirical evidence from UK universities. *International Journal of Social Economics, 30*(11–12), 1210–1232. https://doi.org/10.1108/03068290310500634

Rast, S., & Tourani, A. (2012). Evaluation of employees' job satisfaction and role of gender difference: an empirical study at airline industry in Iran. *International Journal of Business and Social Science, 3*(7), 91–100.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, 399–408. https://doi.org/10.1145/2684822.2685324

Sainju, B., Hartwell, C., & Edwards, J. (2021). Job satisfaction and employee turnover determinants in Fortune 50 companies: insights from employee reviews from Indeed.com. *Decision Support Systems,* *148*. https://doi.org/10.1016/j.dss.2021.113582

Yang, S., & Zhang, H. (2018). Text mining o Twitter data using a latent dirichlet allocation topic model and sentiment analysis. *International Journal of Computer and Information Engineering, 12*(7).

Yao, J. (2019). Automated sentiment analysis of text data with NLTK. *Journal of Physics: Conference Series, 1187*(5). https://doi.org/10.1088/1742-6596/1187/5/052020

Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics, 1*(1–4), 43–52. https://doi.org/10.1007/s13042-010-0001-0

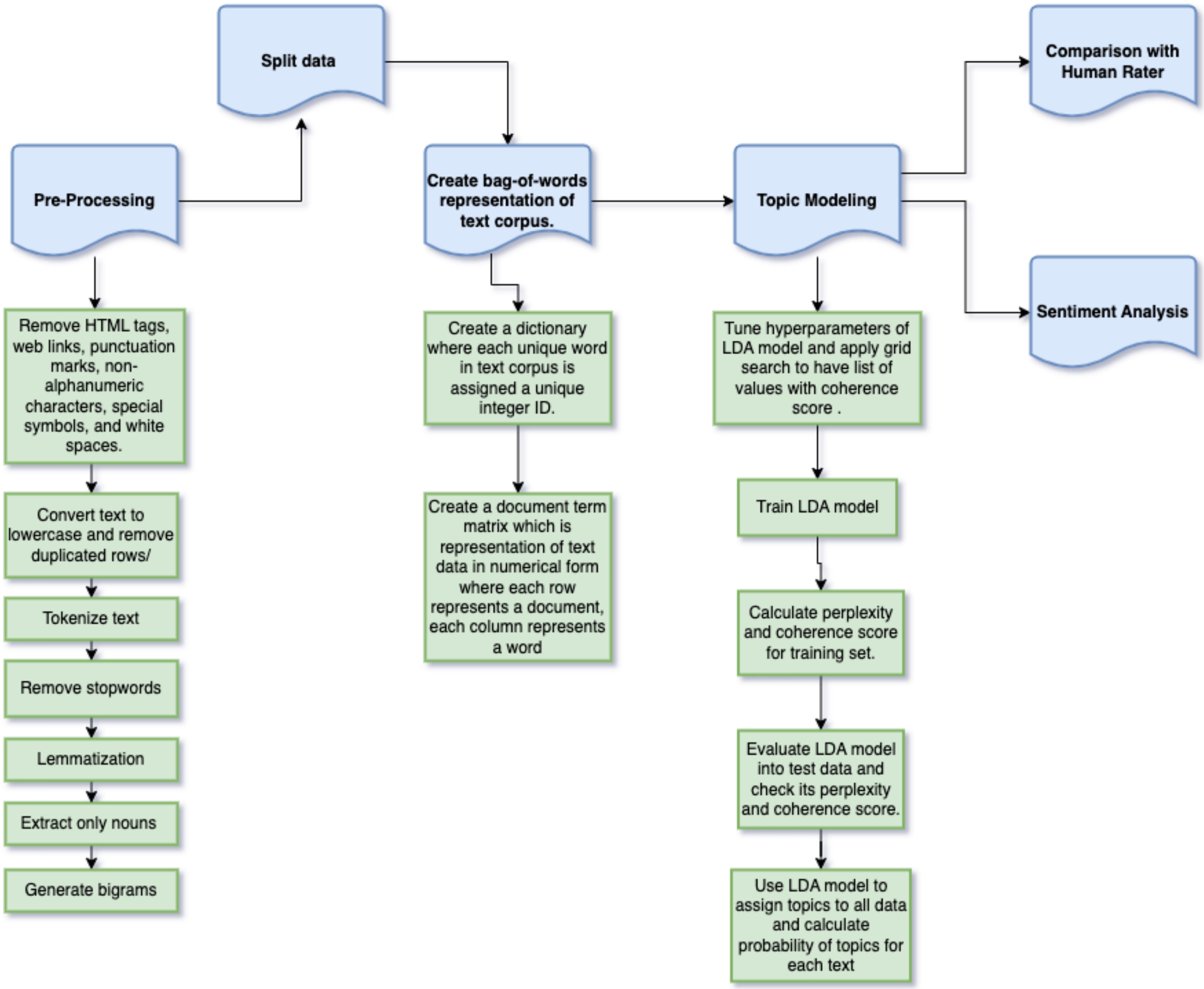
APPENDIX A
Table 1: Literature Review

Source	Purpose	Data & Method	Findings
Dai and Wu (2017)	The study examines current chief data officers' professional skills and education (CDOs) by analyzing their LinkedIn resumes using topic modeling techniques.	621 members of the CDO group on LinkedIn Topic Modelling: NMF and Latent Dirichlet Allocation (LDA)	This study finds that CDOs typically have diverse skills and educational backgrounds, including expertise in business strategy, data governance, and data architecture. Additionally, more specific and coherent skills are captured by NMF than LDA.
Edmans (2011)	This study investigates the link between long-run stock returns and employee satisfaction.	Great Place to Work Institute created 57 question survey. ("100 Best Companies to Work for in America." is the data source.) Regression analysis	According to this study, shareholder returns and employee satisfaction are positively associated.
Ganga (2022)	This study analyzes the employee review literature in different industries and finds the reasons for job satisfaction or dissatisfaction.	Twelve research papers from 2010 to 2020 across various industries and many employee review websites. Topic Modelling: LDA	This study finds that work environment and work-life balance play an essential role in most industries.
Guo et al. (2017)	This study examines how tourists express their satisfaction in reviews from TripAdvisor.	Two hundred sixty-six thousand five hundred forty-four hotel reviews from the TripAdvisor website. Topic Modelling: LDA	This study concludes that online reviews could provide valuable insights into tourist satisfaction and that LDA is an effective tool for analyzing these reviews.

Jung and Suh (2019)	This study analyzes online employee reviews to identify factors contributing to job satisfaction.	204,659 online employee reviews from 4,347 firms in 10 industries on jobplanet.co.kr (35,063 reviews about 844 firms in the IT industry in South Korea used for topic modeling) Topic Modelling: LDA	This study finds five new job satisfaction factors that had not been considered in the literature: Project, Software development, Inter-firm relationship, Marketing, and Overseas business. This study concludes that employee reviews provide insight into specific aspects of the job and the organization that affects job satisfaction, such as communication, management, and work-life balance.
Kalra and Aggarwal (2017)	This study highlights the importance of text data pre-processing and how it can be implemented using the RapidMiner tool.	The Times of India news site contains the 44th President Obama's letter. Various Text Mining Algorithms	This study concludes that data preparation is crucial in machine learning (ML), and pre-processing text data is necessary for cleaning, transforming, and formatting it for use in ML algorithms.
Luo et al. (2016)	This study analyzes employee satisfaction and its relation to corporate performance from anonymous employee reviews on Glassdoor.com.	Two hundred fifty-seven thousand four hundred fifty-four reviews from 425 companies representing 21 industry sectors on Glassdoor.com from 2008 to 2014. Text mining, descriptive data analysis, and regression analysis.	This study shows the top 5 industries that received the most reviews: Technology, Retailing, Financials, Telecommunications, and Health Care. According to findings, innovation plays a significant role in the technology industry, while quality is the driving force in the retailing and financial sectors. Overall, employee satisfaction and corporate performance have a significant correlation. The study identifies safety, communication, and integrity as negatively correlated with performance.
Moro et al. (2021)	This study examines the factors that contribute to job satisfaction among employees in IT companies	Fifteen thousand reviews from the top 15 US technology companies from Glassdoor.com. Support vector machine (SVM).	This study finds that communication, management, and work-life balance are essential drivers of job satisfaction. This study also finds positive attitudes of coworkers, contributing to job satisfaction. However, the main reason for job dissatisfaction is work exhaustion.

Oshagbemi (2003)	This study examines the research on the associations between job satisfaction and factors including age, gender, rank, and length of service.	Questionnaire - university teachers in the UK, 23 institutions. Job Descriptive Index, Regression analysis	This study identifies two significant factors that determine an individual's level of job satisfaction in higher education: their academic rank and length of service. It also shows an individual's gender, age, and length of service at their current university do not directly affect their overall job satisfaction. Still, gender and academic rank are statistically significant predictors, as well as age and length of service in higher education. However, a significant positive association exists between job satisfaction and academic rank, while length of service is negatively related.
Rast and Tourani (2012)	This study assesses the degree of job satisfaction among employees and investigates how gender impacts their job satisfaction.	Survey and questionnaires - employees from 3 private airlines in Iran. Descriptive analysis, Independent-sample t-test	This study identifies several key factors that contribute to job satisfaction, including supervision, relationships with coworkers, current salary, work type, and career advancement opportunities. According to the findings, there is no notable disparity in job satisfaction levels between male and female employees, and, in general, employee satisfaction with their jobs is moderate.
Sainju et al. (2021)	This study examines hidden aspects of employee satisfaction.	Six hundred eighty-two thousand one hundred seventy-six employee reviews of Fortune 50 companies from Indeed.com. Structural Topic Modeling (STM).	This study finds that employees in the retail industry prioritize Pay & Benefits and Length of Breaks when it comes to job satisfaction. In contrast, employees in the technology sector place greater emphasis on achieving a healthy Work-Life Balance. This study also suggests that management is a crucial job satisfaction factor.
Yang and Zhang (2018)	This study analyzes audience reviews of Thor movie on the released day from tweets.	One hundred eighty-five thousand one hundred eighty-five retrieved tweets from Twitter. Topic Modeling: LDA Sentiment analysis	This study finds that LDA is an effective tool for identifying the topics discussed on Twitter related to a particular event or topic, and sentiment analysis can be used to determine the overall sentiment of tweets related to that topic.

APPENDIX B
Figure 1: Proposed Methodology


Split data
Pre-Processing
Create bag-of-words representation of text corpus.
Topic Modeling
Comparison with Human Rater
Sentiment Analysis
Remove HTML tags, web links, punctuation marks, non-alphanumeric characters, special symbols, and white spaces.
Convert text to lowercase and remove duplicated rows/
Tokenize text
Remove stopwords
Lemmatization
Extract only nouns
Generate bigrams
Create a dictionary where each unique word in text corpus is assigned a unique integer ID.
Create a document term matrix which is representation of text data in numerical form where each row represents a document, each column represents a word
Tune hyperparameters of LDA model and apply grid search to have list of values with coherence score .
Train LDA model
Calculate perplexity and coherence score for training set.
Evaluate LDA model into test data and check its perplexity and coherence score.
Use LDA model to assign topics to all data and calculate probability of topics for each text

APPENDIX C
Table 2: LDA Tuning Results

Validation Set	Topics	Alpha	Beta	Coherence
75% Corpus	9	0.31	0.31	0.525523
100% Corpus	9	0.31	0.31	0.51706855
75% Corpus	10	0.31	0.31	0.51509178
75% Corpus	10	0.31	0.01	0.51375766

APPENDIX D
Table 3: Job Satisfaction Factors

No	Factors	Keywords
1	Management skills and responsibilities	management, skill, meeting, value, security
2	Workplace culture and environment	culture, environment, support, workplace, compensation
3	Job experience and flexibility	experience, home, role, money, direction
4	Organizational strategy and commitments	politic, mission, strategy, budget, commitment
5	Project planning	project, expectation, stress, goal, user
6	Teamwork and collaboration	team, family, quality, feedback, ability
7	Work-life balance and well-being	life, training, balance, schedule, hr
8	Career advancement and leadership development	opportunity, advancement, leadership, contract, promotion
9	Compensation and benefits	benefit, pay, growth, atmosphere, location

APPENDIX E
Table 6: A Comparison of Topics Between LDA and Human Evaluation (Researcher)

Factors	LDA	A	B	C	D
Management skills and responsibilities	✓	✓	✓	✓	✓
Workplace culture and environment	✓	✓	✓	✓	✓
Job experience and flexibility	✓	✓	✓	×	×
Organizational strategy and commitments	✓	×	×	×	×
Project planning	✓	×	×	×	×
Teamwork and collaboration	✓	×	×	×	✓
Work-life balance and well-being	✓	×	✓	✓	✓
Career advancement and leadership	✓	✓	✓	✓	✓
Compensation and benefits	✓	✓	✓	✓	✓
Coworkers	×	✓	×	×	×
Vacation	×	✓	×	×	×
Diversity	×	✓	×	✓	×
People	×	×	✓	×	×
Education	×	×	×	✓	×
Challenges	×	×	×	✓	×
Networking	×	×	×	×	✓

Notes: ✓= included × = not included. The Jaccard coefficient is 0.42 between LDA analysis and researcher A. The Jaccard coefficient is 0.6 between LDA analysis and researcher B. The Jaccard coefficient is 0.42 between LDA analysis and researcher C. The Jaccard coefficient is 0.5 between LDA analysis and researcher D.

West Nile Virus in Colorado: Analytic and Geospatial Models of the Virus in Colorado

Johnny Snyder
josnyder@coloradomesa.edu
Davis School of Business
Colorado Mesa University
Grand Junction, Colorado 81506

Abstract

West Nile Virus found its way to North America in 1999, starting with the diagnosis of two cases of encephalitis in the Queens borough of New York City. WNV had found its way to Colorado by 2002. The main vector for West Nile Virus is the mosquito, primarily the *Culex* species. This research shows, from historical data along with qualitative, quantitative, and geospatial methods, that the primary variables behind West Nile Virus cases by county in Colorado are the county's urban/rural classification, water area in the county (in square miles), and if it is an El Niño year or not. Other variables, including population density in the county, and the average precipitation and temperature over the period July to October, are discussed and their merit in a model presented. Mapping tools are used to illustrate the presence of West Nile Virus as well as its spread, over time, through the counties in Colorado. The data set for this study covers 2005 to 2021 for the 64 counties of Colorado.

Keywords: West Nile Virus, analytic model, virus model, geospatial analysis

Recommended Citation: Synder, J., (2024). West Nile Virus in Colorado: Analytic and Geospatial Models of the Virus in Colorado. *Journal of Information Systems Applied Research*. 17(1), pp 21-32, https://doi.org/10.62273/FOWT9206

West Nile Virus in Colorado: Analytic and Geospatial Models of the Virus in Colorado

Johnny Snyder

1. INTRODUCTION

West Nile Virus (WNV), a mosquito transmitted virus, was first identified in 1937 in the West Nile province of Uganda (Kramer, Li, & Shi, 2007). WNV emerged into North America, and in particular, New York in 1999 and has since spread throughout the USA, Canada, and Mexico (Petersen, 2019; Kramer, Li, & Shi, 2007). Colorado recorded its first case of WNV in 2002 (CDPHE, 2022) and saw a large number of cases in 2003 due, it is believed, to the initial large-scale testing for WNV, and the lack of immunity in the population (Marzec, 2022). Starting in 2005, the pattern of cases of WNV stabilized, to include high and low years, but nothing as severe as the initial highs of 2003 and 2004.

WNV can infect birds, which also serve as a reservoir and diffuser for the virus, horses, and humans, with the latter two being the most negatively affected by the virus (Peterson, 2019; Kramer, Li, & Shi, 2007). WNV is spread, primarily, by the enzootic cycle between birds and mosquitoes. There are over three hundred species of birds that have been identified as viable WNV hosts and generally the mosquito genus *Culex* is the transmission vector to humans (Ciota, 2017). Most humans show no symptoms upon being infected with WNV (80%), while approximately 20% exhibit flu-like symptoms, and less than 1% become seriously ill with arboviral encephalitis (Ciota, 2017; Wildlife Futures Team, 2021). In Colorado, there have been over 5800 cases of WNV identified in humans since WNV arrived to the state in 2002 (CDPHE, 2022). Extrapolating, this implies that there could have been over 23,000 undetected cases with up to 290 cases of encephalitis, making this a public health issue involving hospital stays, transplantation medicine, and transfusion concerns (Petersen, 2019; Ronca, Ruff, & Murray, 2021). While there are vaccines against WNV for horses, first approved in 2005, there are no human vaccines that have gone beyond Phase I/II trials, indicating a need for other mitigation strategies for the virus and its vectors (Ronca, Ruff, & Murray, 2021).

Prevention of WNV through target mitigation strategies can assist the state of Colorado in preserving the health of the population, as well as avoidance of medical costs and lost productivity costs which can run into the tens of millions of dollars (Ronca, Ruff, & Murray, 2021). This research attempts to build a model that identifies when and where WNV is most likely to occur in the state. This model, along with appropriate visualizations, can assist decision makers in the public health arena in mitigation strategies for WNV in Colorado.

Observations and studies from Chicago to Texas have suggested that "enhancement of surveillance and vector control in limited geographic areas could produce an outsized impact on WNV incidence nationwide" (Petersen, 2019, p. 1457); further, Hadfield et al. (2019) state that a coordinated effort between state health departments could be used as vector control by "strategically focusing resources at a precise time and location to limit potential outbreaks" (p.10). This research addresses the "when and where" WNV might appear so that proactive mitigation strategies can be deployed.

The main variables for WNV models come under three general headings:

- Environmental (precipitation, humidity, and temperature as examples)
- Physical (elevation, water area, and land use as examples)
- Population (human, mosquito, and bird census data as examples).

Hadfield, et al. (2019) states "Dynamic extrinsic factors, such as rainfall and temperature, that influence mosquito and bird populations can be predictive of WNV intensity, yet the contributions of these extrinsic factors vary across the United States due to differences in regional ecology." (p. 7) This statement leads one to conclude that more localized models containing extrinsic factors are necessary to predict WNV occurrences. This is also the conclusion reached by Paz (2015) who states that although rainfall has a pattern of positive association with outbreaks of WNV, the literature shows mixed results, and that the response (to rainfall) "might change over large geographical regions, depending on differences in the ecology of mosquito vectors" (p. 2). Garcia-Carrasco et al. (2021) found that temperature and elevation were significant factors in the occurrence of WNV. They also suggest that mountains "could be barriers or at least filters for the spread of the disease" (p. 6). This implies that elevation could have an effect on WNV occurrences.

Other authors (Barker, 2019; Ronca et al., 2021) have worked with surveillance data (of bird and mosquito populations) in an attempt to predict outbreaks of WNV. However, as Barker (2019) states "Human and veterinary disease cases are also tracked through surveillance systems for notifiable diseases, although notifications generally arrive too late to initiate mosquito control, sometimes weeks or even months after the infection has occurred" (p. 1511).

Variables that have the potential to affect the transmission of WNV in Colorado include:

- Temperature (Ciota, 2017; Hadfield, et al., 2019; Paz, 2015; Garcia-Carrasco, et al. 2021)
- Precipitation (Ciota, 2017; Hadfield, et al., 2019; Paz, 2015)
- Humidity (Paz, 2015)
- Seasonality (Paz, 2015)
- Wind (Paz, 2015)
- Urban/rural (Ciota, 2017)
- Land use (Hadfield, et al., 2019; Paz, 2015; Garcia-Carrasco, et al. 2021)
- Presence of water (Garcia-Carrasco, et al. 2021)
- Elevation (Hadfield, et al., 2019; Garcia-Carrasco, et al. 2021)
- Mosquito abundance data (Ronca, et al., 2021)
- Bird population data (Ronca, et al., 2021)

Other variables added to this data set for their perceived usefulness for multiple types of analysis as well as possible filters for visualizations include:

- Year
- County name
- County population
- Number of WNV cases per county
- El Niño, La Niña years
- Land area of county
- Severity of WNV in county, a qualitative variable constructed from historical data on the count of WNV cases per county (primarily as a filter for geospatial analysis)

Variables which could be found are illustrated in Table 1, along with references or links to where to find the original data source.

Table 1: Variables for the Colorado WNV Models (see Appendix)

Some variables were difficult or otherwise unobtainable for this study. These include humidity, mosquito, bird, and wind data, all at a county level. Other variables that proved difficult to find include the number of irrigated acres per county (only found 2007 and 2012 data sets) and a reasonable elevation variable.

2. PROBLEM STATEMENT

This research seeks to determine what variables influence WNV presence in the counties of Colorado. By applying techniques such as:

- descriptive statistics
- correlation analysis
- hypothesis testing
- multiple linear regression analysis
- logistic regression analysis
- cluster analysis
- geospatial analysis
- graph analysis

salient variables will be selected for modeling purposes and mapping techniques will be used as an illustrative guide for deploying resources at a county level. This analysis will aid the decision maker (hospital, EMS, medical community, county/state level medical personnel) in when and where to apply mitigation measures for WNV, when and where to deploy limited personnel across the state of Colorado, where to locate sentinel devices for WNV, and when to notify public health decision makers of WNV activity.

3. METHODOLOGY AND RESULTS

Performing both multiple linear regression ($\hat{y}$= number of WNV cases in county) and logistic regression ($\hat{p}$= probability of WNV in county) with all possible variables, the best models, and their R^2 values are given in Table 2. These models were distilled using the backward elimination method, which uses all variables to construct a model, then removes the least significant variables, iteratively, until a model comprised solely of significant variables is left. (Camm, et al., 2024)

Table 2: Best models for WNV (see Appendix)

The first finding is that elevation played no significant part in the analysis, even though six different elevation variables (maximum, minimum, difference, qualitative over 7000 ft., elevation at center point of county, and high qualitative (maximum over 9000 ft.)) were tried in the models. Elevation has been shown to be significant in other models (Garcia-Carrasco, et al., 2021).

Many other variables were also not significant predictors of WNV in this county level model. They include La Nina, area of county, and an annual precipitation variable.

Evaluating the explanatory power of the models in Table 2 leads to the construction of Figure 1, a Venn diagram of the statistically significant variables from $\hat{y}$ and $\hat{p}$. Figure 1 shows the core variables of the model (the intersection) and the peripheral variables of the individual models. These variables will be used for the remainder of this work.

Figure 1: Significant variables for WNV models (see Appendix)

The variables identified in Figure 1 are defined here:

Core Variables (significant for both models)

- Urban/rural – a qualitative variable, 1 if the population of the county is greater than 150,000 people. This variable will be used to compare the two data sets for differences (urban counties versus rural counties) and as a filter for visualizations.
- Water area – a quantitative variable for the area (in mi^2) of each county that is covered by water.
- El Niño – a qualitative variable, 1 if it is an El Niño year (which occurred in 2005, 2007, 2010, 2015, 2016, and 2019). This variable will be used to compare the two data sets for differences (El Niño year versus non-El Niño years) and as a filter for visualizations.

Peripheral Variables (significant for only one model)

- Population density – number of people per square mile. With any disease of the human, population density can play a factor.
- AvgTempJulyOct and AvgRainfallJulyOct – as named, these variables measure the average temperature in Fahrenheit and the average rainfall in inches over the four-month period of July to October. This is the period in which 80-90% of the WNV cases occur as can be seen in Figure 2 from the CDC. Due to the short life cycle of the mosquito (generally a couple of weeks) looking at the environmental variables during this time period seems reasonable.

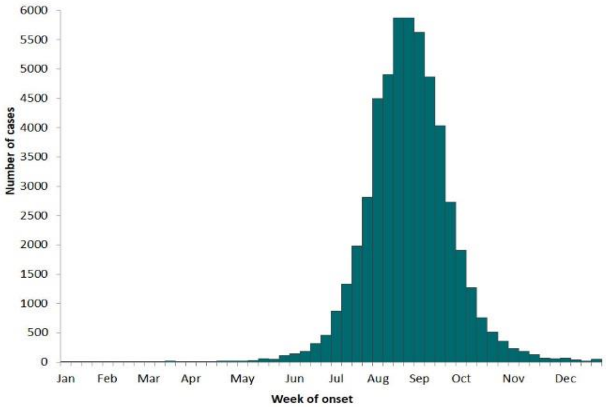
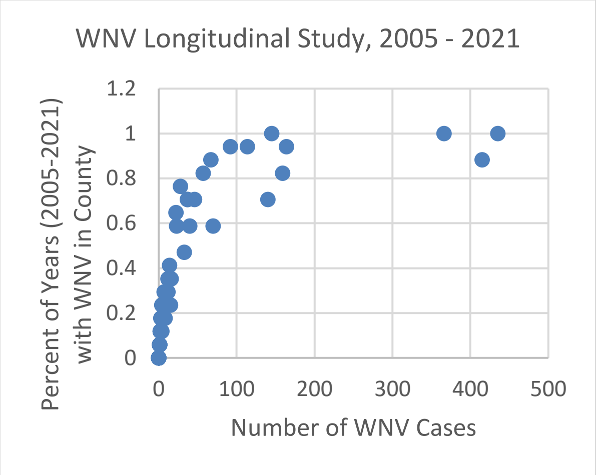


Figure 2: WNV in the USA during the year (CDC, 2021)

Time Period

- WNV arrived to New York State in 1999, then spread throughout the country. WNV made it to Colorado in 2002, and being a new virus there was a lot of testing and no natural immunity (Marzec, 2022), causing a large spike in WNV numbers for 2003 and 2004. In 2005, the pattern of WNV settled down, exhibiting peaks and valleys, but not as severe as in 2003 - 2004. This model uses the WNV case count, by Colorado county, for the time frame 2005 – 2021 assuming these years to be more representative of the WNV numbers in the state moving forward. (of course 2020 and 2021 might be skewed due to the presence of Covid in the state)

Analyzing the counties longitudinally for the percentage of years the county has had WNV cases present over the study period, Graph 1 is obtained.



Graph 1: Percent of years county has WNV, 2005-2021

Graph 1 points at a couple of interesting items, three counties have WNV every year (Denver, Boulder, Larimer) and three counties have a lot of WNV (Boulder, Larimar, Weld). Graph 1 also facilitates the development of a qualitative variable for a county's WNV status.

$$CountyLongitudinalWNVStatus = \left\{ \begin{array}{c} HighRiskCounty\ if\ Percent \geq 0.8 \\ MediumRiskCounty\ if\ 0.5 \leq Percent < 0.8 \\ LowRiskCounty\ if\ Percent < 0.5 \end{array} \right\}$$

It can be calculated that the HighRiskCounty group, when summed, represents 76.3% of the cases, leaving only 23.7% of the cases in the medium and low risk categories. This variable will be used in the visualizations as well as for separating the data set into multiple pieces for some descriptive analysis and interpretive guidance on differences between the risk groups. This analysis is presented in Table 3.

Table 3: Statistically significant differences when dividing the data set along qualitative lines (high, medium, low and El Niño dimensions) (see Appendix)

Overall, Table 3 tells us that there will be more WNV cases in El Niño years with warmer temperatures, counties with higher population, and lower elevation counties with more surface water. The data also showed more WNV cases in years with less rainfall, an unexpected result, as mosquitoes need water to reproduce.

Cluster analysis (k = 3 clusters) also indicated that counties with high WNV cases clustered around:

- counties with more water area
- counties that have denser population
- counties with a higher AvgTempJulyOct
- counties lower in elevation,

which helped to identify salient variables and well-designed filters with which to build visualizations to illustrate these primarily analytical results.

4. VISUALIZATIONS

Visualization 1 is a Tableau dashboard tracking WNV cases against four of the independent variables of the model, water area, temperature, precipitation, and population. Other variables, namely urban/rural and the CountyLongitudinalWNVStatus are used as filters for the dashboard.

Visualization 1: A Dashboard for Tracking WNV Variables (see Appendix)

Visualization 1 gives us some information not only about the status of the WNV model, but information on future directions as well. The Colorado WNV "Goldilocks Zone" for mosquitoes has been addressed in news articles (Bailey, 2022) where it is stated "Spring rain, summer drought, and heat created ideal conditions for mosquitoes to spread the West Nile Virus through Colorado last year" (p. 1). Visualization 1 yields information on the Goldilocks Zone for "High" WNV counties (this is the filter applied in Visualization 1) as:

- AvgTempJulyOct: 57 – 67 degrees Fahrenheit
- AvgRainfallJulyOct 5.5 – 7.0 inches
- Urban/Rural – 80% of the "high WNV" counties are classified as urban
- County water area in mi^2 – positive linear relationship between water area and WNV

The Goldilocks Zone might, according to the article by Bailey, also need a spring rain variable or a drought variable to identify this zone more accurately.

Visualization 1 also indicates that Denver County (formally the City and County of Denver) might be an outlier and could be removed in future analysis. It would also be prudent to address the other city/county designated area, City and County of Broomfield, for potential removal as well. These large urban areas could also be importing the virus from camping or other "out of county" activities by the county residents.

Visualization 1 also indicates that the number of WNV cases increase as the amount of surface water increases in the county (slope = 0.6 cases/mi^2 of surface water in county for counties with a "high" WNV status). This indicates that minimization of water area (particularly in urban counties) could lead to a lessening of the WNV impact in urban areas. This would also align with the West's increasing water conservation issues – reduction of water features in golf courses, urban parks, and other (arguably) unnecessary water features in the urban landscape.

While Visualization 1 focused on "High" WNV counties (counties that have had WNV present for more than 80% of the years in this study), Visualization 2 explores, geospatially, the impact of urban (greater than 150,000 population) counties in Colorado. It should be noted that over the time span of this study, two counties, Mesa, and Pueblo, changed from rural to urban (Mesa eclipsed 150,000 in 2015 and Pueblo in 2006), noting the dynamic nature of studies of this type.

Visualization 2: Geospatial analysis of the WNV load in Colorado (see Appendix)

Visualization 2 indicates that urban counties, which follow the I-25 and I-70 corridors (except in the mountains), and that WNV also follows the urban corridors particularly in the counties that have more surface water. Additionally, the population density choropleth verifies that Denver County could be a candidate for removal from this study. These maps could be utilized to effectively direct mosquito and WNV mitigation efforts.

5. FINDINGS

WNV follows the urban corridor throughout Colorado, occurring more in urban counties with large surface areas of water. WNV mosquitoes enjoy certain temperature ranges and precipitation conditions. WNV occurs during the July – October time frame and occurs more in an El Niño year. Counties can be classified as high, medium, or low depending on their WNV load, directing mitigation efforts to areas most in need. It appears that any major WNV events would be in one of Boulder, Larimar, or Weld counties, making these counties useful as sentinel counties, indicating when WNV is in the state.

WNV is in Colorado and will be with us for the foreseeable future. Major findings include the positive correlation of WNV with temperature, the positive correlation with precipitation, and the positive correlation with El Niño years. Current climate change predictions coupled with the results of this study indicate that WNV cases in select counties in Colorado can be expected to increase, indicating a need for mitigation strategies and WNV preparedness of medical personnel. These models and visualizations will assist public health decision makers in protecting the health of the citizens of Colorado.

6. BUSINESS BENEFIT

Hospitals, EMS, state, county, and local officials (health agencies) can all benefit from this analysis by using it to plan for WNV cases in the counties of Colorado.

Medical personnel can benefit by being aware of high-risk counties for WNV and being alert to the symptoms of WNV as the WNV season approaches.

State and county officials can direct limited resources to those counties most in need of mosquito mitigation strategies, surveillance activities, or water conservation efforts (in particular, surface water features). This can protect the citizens of Colorado and protect businesses from lost employee productivity.

7. SUMMARY AND FUTURE DIRECTIONS

Addressing WNV in Colorado, its causes and primary locations can protect the citizens and the labor force of Colorado. Mitigation strategies should be directed to identified high risk counties of the state. Early warning for medical personnel can be based on weather patterns observed and/or other indicator variables.

Current and future research could include the variables presented in Table 4.

Table 4: Current and Future Variables (see Appendix)

Analysis needed to enhance the model would include addition of identified variables, drill down analysis (instead of county, is there zip code level data available), and interaction effects (in particular between temperature, humidity, rainfall, and elevation – if available – in pursuit of the elusive "mosquito line" – a hypothesized but never derived "contour" line (iso-line) that would divide the state into two zones – mosquito and mosquito-free). A variable map dividing the variables into categories (environmental, physical, population, filters) which could be used to gage the impact of each of the categories would be useful.

8. REFERENCES

Bailey, M. (2022). Climate Change May Push the US Toward The 'Goldilocks Zone' For West Nile Virus. Retrieved from: https://techilive.in/climate-change-may-push-the-us-toward-the-goldilocks-zone-for-west-nile-virus/

Barker, C. (2019). Models and Surveillance Systems to Detect and Predict West Nile Virus Outbreaks. *Journal of Medical Entomology,* 56(10), p. 1508-1515.

Camm, J., Cochran, J., Ohlmann, J., and Fry, M. (2024). Business Analytics, 5th. Cengage Learning, Inc.

Centers for Disease Control (CDC). (2021). West Nile virus disease cases reported to CDC by week of illness onset, 1999-2020, graphical result. Retrieved from: https://www.cdc.gov/westnile/statsmaps/cummapsdata.html

Ciota, A. (2017). West Nile virus and its vectors. Current Opinion in Insect Science. 22, p. 28-36.

Colorado Department of Public Health and Environment (CDPHE). (2022). West Nile Virus Data. Retrieved from: https://cdphe.colorado.gov/animal-related-diseases/west-nile-virus/west-nile-virus-datahttps://cdphe.colorado.gov/animal-related-diseases/west-nile-virus/west-nile-virus-data

Garcia-Carrasco, J., Muñoz, A., Olivero, J., Segura, M., and Real, R. (2021). Predicting the spatio-temporal spread of West Nile Virus in Europe. PLoS Neglected Tropical Disease, 15(1).

Hadfield, J., Brito, A., Swetnam, D., Vogels, C., Tokarz, R., Andersen, K., Smith, R., Bedford, T., and Grubaugh, N. (2019). Twenty years of West Nile virus spread and evolution in the Americas visualized by Nextstrain. PLOS Pathogens. 15(10). p. 1-18.

Kramer, L., Li, J., and Shi, P. (2007). West Nile Virus. The Lancet Neurology. 6(2), p. 171-181.

Marzec, N. (2022). Personal email communication. MPHZoonotic Disease Unit Manager Communicable Disease Branch, CDPHE.

Paz, S. (2015). Climate change impacts on West Nile Virus transmission in a global context. *Philosophical Transactions B* 370:20130561

Petersen, L. (2019). Epidemiology of West Nile Virus in the United States: Implications for Arbovirology and Public Health. Journal of Medical Entomology. 56(6). p. 1456-1462.

Ronca, S., Ruff, J. and Murray, K. (2021). A 20-year historical review of West Nile virus since its initial emergence in North America: Has West Nile virus become a neglected tropical disease? PLoS Neglected Tropical Diseases. 15(5). p. 1-14.

Tableau. (2022). Software package. Retrieved from: https://www.tableau.com

Wildlife Futures Team. (2021). West Nile Virus. Retrieved from: https://www.vet.upenn.edu/research/centers-laboratories/research-initiatives/wildlife-futures-program/resources/fact-sheets/fact-sheet-detail/west-nile-virus#:~:text=West%20Nile%20virus%20infects%20over,bite%20of%20an%20infected%20mosquito

Zaiontz, C. (2022). Real Statistics software add-in for Excel. Available from: https://www.real-statistics.com/

Editor's Note:

This paper was selected for inclusion in the journal as 2023 ISCAP Conference Distinguished Information Systems Applied Research Paper. The acceptance rate is typically 7% for this category of paper based on blind reviews from six or more peers including three or more former best papers authors who did not submit a paper in 2023.

Appendix: Tables, Figures, and Visualizations

Variable	URL
Calendar year	https://data.colorado.gov/Demographics/Total-Population-by-County-by-Year/9dd2-kw29
County name in Colorado	
Population (2020 Census) of county	
Number of WNV cases in county for given year	https://cdphe.colorado.gov/animal-related-diseases/west-nile-virus/west-nile-virus-data
Relative measure of WNV cases against county population	
Rainfall in county	https://www.ncdc.noaa.gov/
Urban: 1 if population of county is greater than 150,000; 0 otherwise	https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural/2010-urban-rural.html
Minimum county elevation, ft	https://en.wikipedia.org/wiki/List_of_counties_in_Colorado
Maximum county elevation, ft	
Difference in elevation, Maximum - Minimum, ft	
Elevation at the geographical center of the county, ft	https://www.usgs.gov/products/data/all-data
Water area of county, square miles	https://en.wikipedia.org/wiki/List_of_counties_in_Colorado#County_data
Percent coverage of county by water	
Number of people in county per square mile	derived from population and area variables
1 if El Nino year, 0 otherwise	https://ggweather.com/enso/oni.htm
1 if La Nina year, 0 otherwise	
Area of county, square miles	https://en.wikipedia.org/wiki/List_of_counties_in_Colorado#County_data
1 if mean elevation of county is over 7000 ft, 0 otherwise	http://www.cohp.org/records/mean_elevation/mean_elevations.html
Average temperature per county for the four-month period (July – Oct.), Fahrenheit	https://www.ncdc.noaa.gov/cag/county/time-series
Rainfall average per county for the four-month period (July – Oct.), inches	https://www.ncdc.noaa.gov/cag/county/time-series
Qualitative rating of WNV historical case count per county (High, Medium, Low)	derived from WNV case data from 2005 to 2021

Table 1: Variables for the Colorado WNV Models

Model	$\mathbf{R^2}$
$\hat{y} = -2.86 + 7.61(urban/rural) + 0.29(water\ area\ mi^2) + 1.09(El\ Niño) + 0.22(AvgRainfallJulyOct)$	25.7%
$\hat{p} = 1/(1 + e\text{^}(-\hat{y}))$, where $\hat{y} = -9.56 + 2.14(urban/rural) + 0.03(water\ area\ mi^2) + 0.4(El\ Niño) + 0.001(population\ density) + 0.13(AvgTempJulyOct)$	***Pseudo** $\mathbf{R^2}$ $R_L^2 = 30.3\%$ $R_{CS}^2 = 31.8\%$ $R_N^2 = 44.3\%$

Table 2: Best models for WNV

* See the Real Statistics help page at: https://www.real-statistics.com/logistic-regression/significance-testing-logistic-regression-model/ (Zaiontz, 2022)

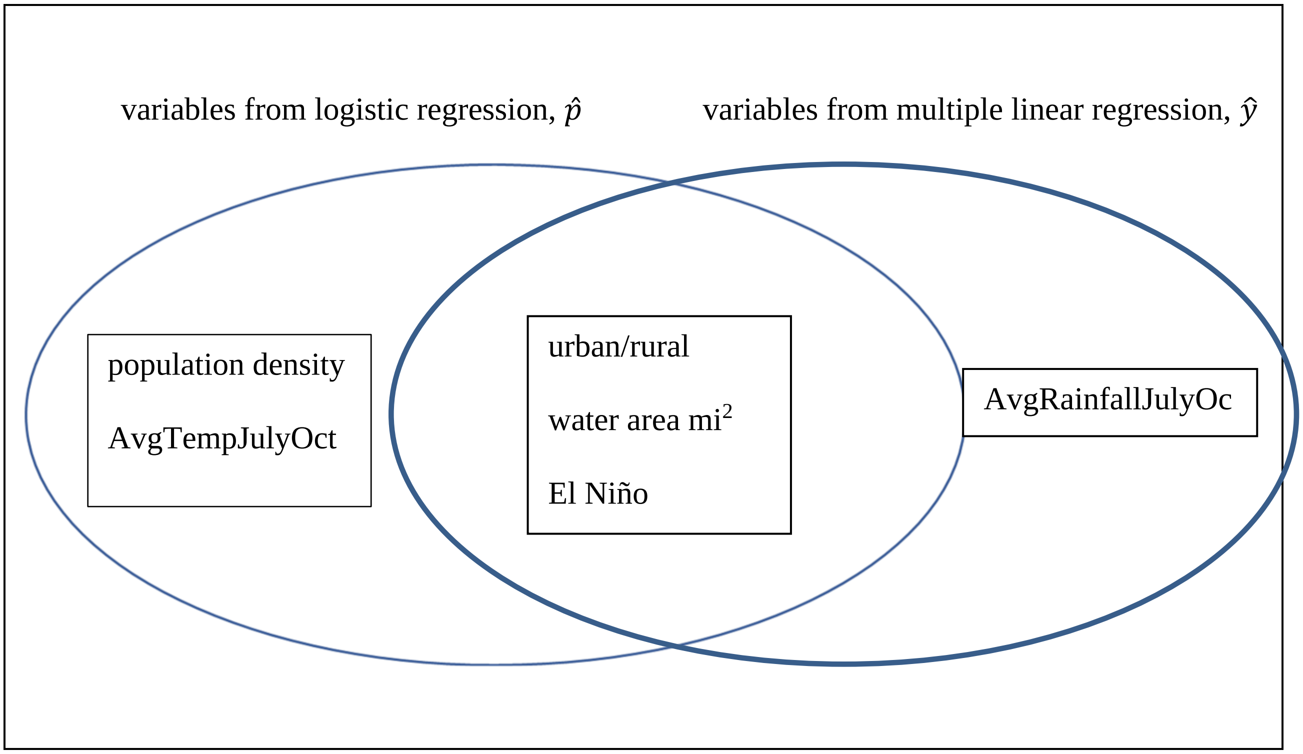
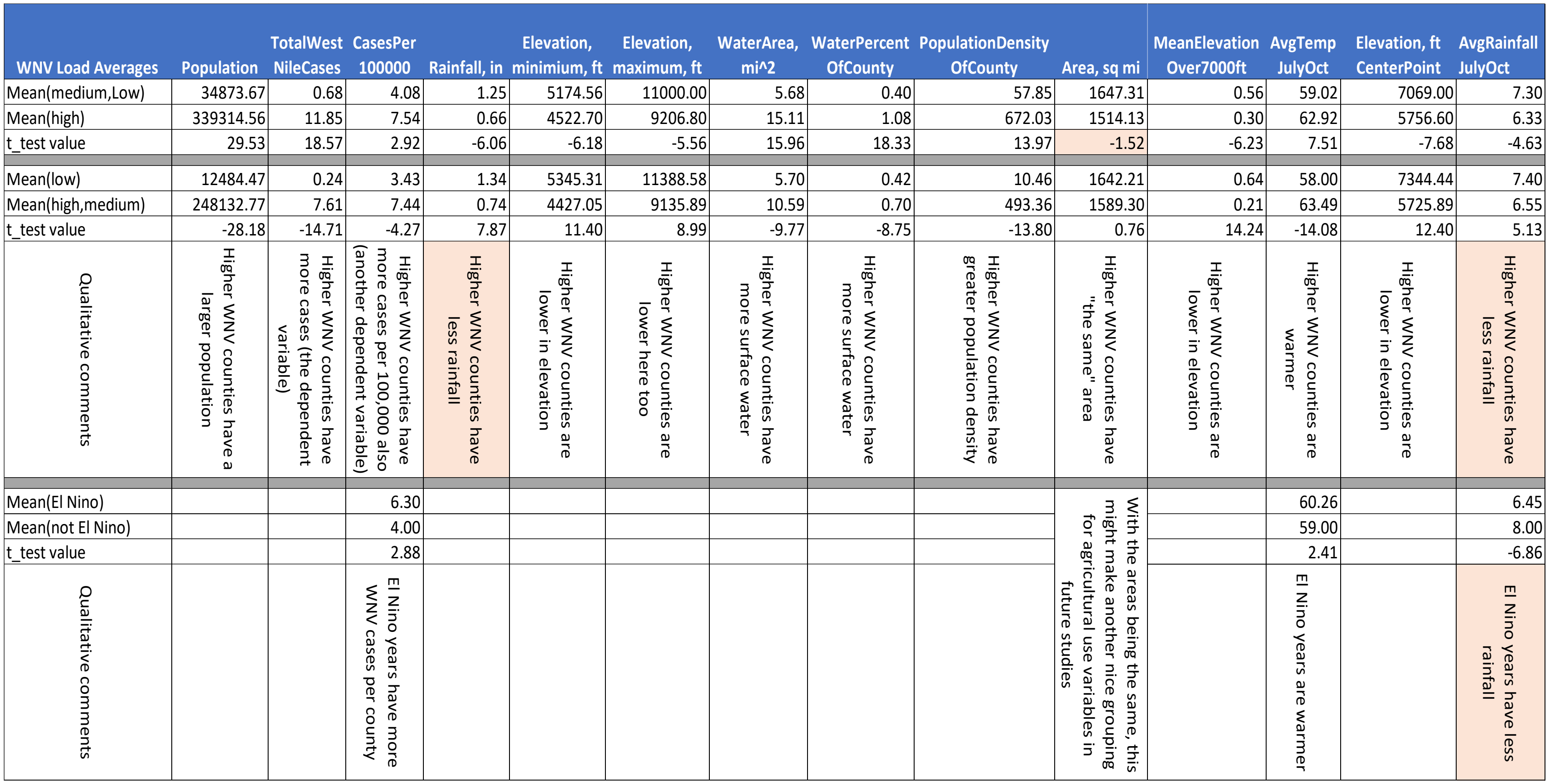
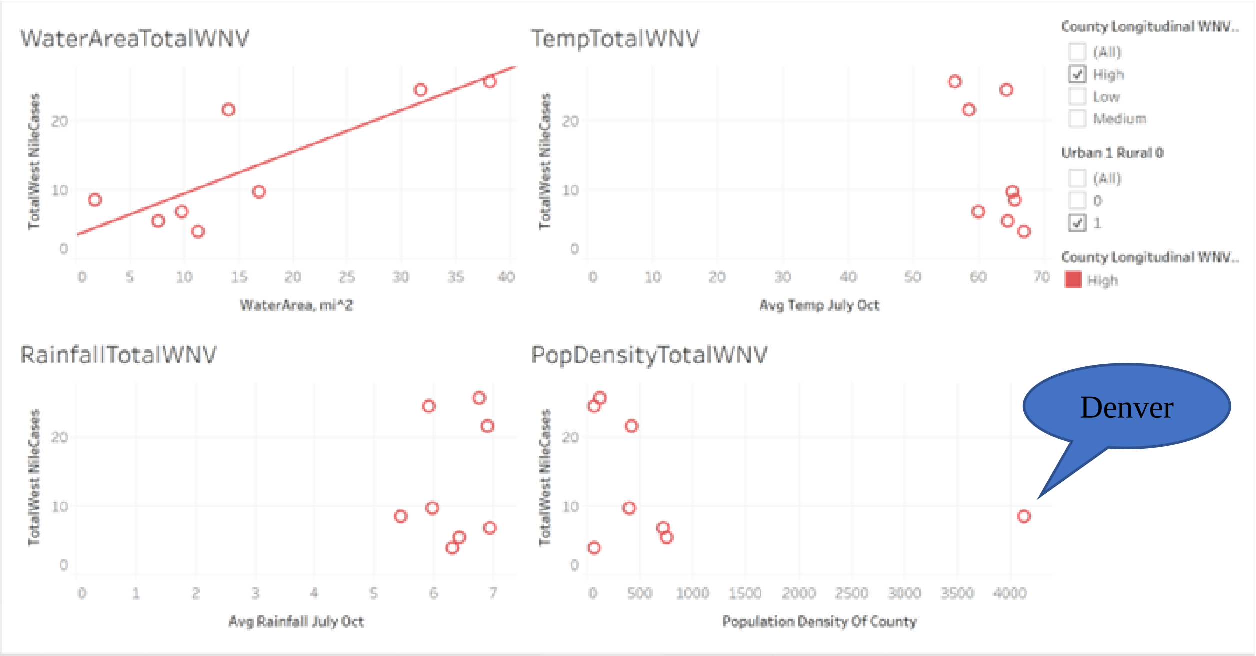

variables from logistic regression, $\hat{p}$
variables from multiple linear regression, $\hat{y}$
population density
AvgTempJulyOct
urban/rural
water area mi^2
El Niño
AvgRainfallJulyOc

Figure 1: Significant variables for WNV models

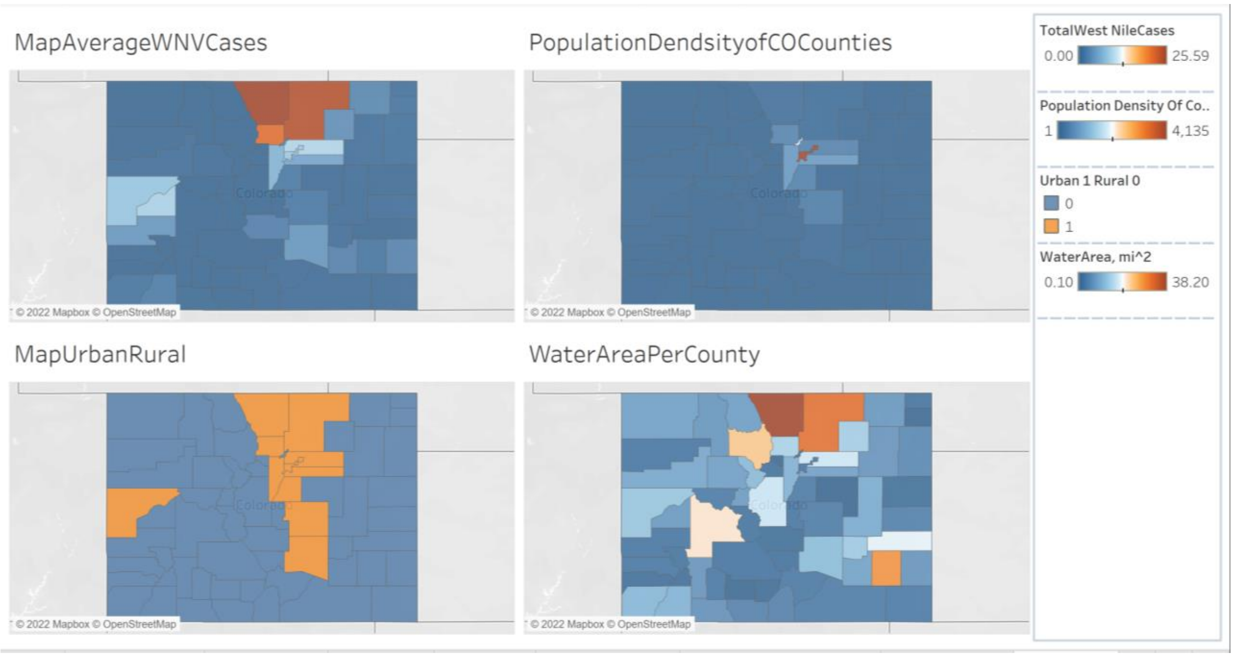


WNV Load Averages	Population	TotalWest NileCases	CasesPer 100000	Rainfall, in	Elevation, minimium, ft	Elevation, maximum, ft	WaterArea, mi^2	WaterPercent OfCounty	PopulationDensity OfCounty	Area, sq mi	MeanElevation Over7000ft	AvgTemp JulyOct	Elevation, ft CenterPoint	AvgRainfall JulyOct
Mean(medium,Low)	34873.67	0.68	4.08	1.25	5174.56	11000.00	5.68	0.40	57.85	1647.31	0.56	59.02	7069.00	7.30
Mean(high)	339314.56	11.85	7.54	0.66	4522.70	9206.80	15.11	1.08	672.03	1514.13	0.30	62.92	5756.60	6.33
t_test value	29.53	18.57	2.92	-6.06	-6.18	-5.56	15.96	18.33	13.97	-1.52	-6.23	7.51	-7.68	-4.63
Mean(low)	12484.47	0.24	3.43	1.34	5345.31	11388.58	5.70	0.42	10.46	1642.21	0.64	58.00	7344.44	7.40
Mean(high,medium)	248132.77	7.61	7.44	0.74	4427.05	9135.89	10.59	0.70	493.36	1589.30	0.21	63.49	5725.89	6.55
t_test value	-28.18	-14.71	-4.27	7.87	11.40	8.99	-9.77	-8.75	-13.80	0.76	14.24	-14.08	12.40	5.13
Qualitative comments	Higher WNV counties have a larger population	Higher WNV counties have more cases (the dependent variable)	Higher WNV counties have more cases per 100,000 also (another dependent variable)	Higher WNV counties have less rainfall	Higher WNV counties are lower in elevation	Higher WNV counties are lower here too	Higher WNV counties have more surface water	Higher WNV counties have more surface water	Higher WNV counties have greater population density	Higher WNV counties have "the same" area	Higher WNV counties are lower in elevation	Higher WNV counties are warmer	Higher WNV counties are lower in elevation	Higher WNV counties have less rainfall
Mean(El Nino)			6.30							With the areas being the same, this might make another nice grouping for agricultural use variables in future studies		60.26		6.45
Mean(not El Nino)			4.00									59.00		8.00
t_test value			2.88									2.41		-6.86
Qualitative comments			El Nino years have more WNV cases per county									El Nino years are warmer		El Nino years have less rainfall

Table 3: Statistically significant differences when dividing the data set along qualitative lines (high, medium, low and El Niño dimensions)



Visualization 1: A Dashboard for Tracking WNV Variables


MapAverageWNVCases
Colorado
© 2022 Mapbox © OpenStreetMap
PopulationDendsityofCOCounties
Colorado
© 2022 Mapbox © OpenStreetMap
MapUrbanRural
Colorado
© 2022 Mapbox © OpenStreetMap
WaterAreaPerCounty
Colorado
© 2022 Mapbox © OpenStreetMap
TotalWest NileCases
0.00
25.59
Population Density Of Co..
1
4,135
Urban 1 Rural 0
0
1
WaterArea, mi^2
0.10
38.20

Visualization 2: Geospatial analysis of the WNV load in Colorado

Variable	Comments
spring rainfall data	some authors have observed, on a country-wide basis, that these variables are indicators of WNV (a drought variable could also be a good filter variable for visualizations)
a drought variable	
WNV patient latitude and longitude	for more accurate accounting of WNV case locations
average humidity per county	difficult to obtain or compute
average elevation per county	
mosquito data (location, population density)	cuts across county boundaries
bird data (location, population density, migration patterns)	cuts across county boundaries
garden zone (planting guidelines, might be based on elevation)	cuts across county boundaries
monthly data rather than annual data (for drill down)	would be twelve times the data mining
agricultural variables (water usage)	these have shown some promise, percent of county irrigated (2007 and 2012) and number of cattle in county (2017-2021) were explored, and number of cattle was found to be significant in analytic models
interaction effects	these have shown some promise, temperature*rainfall was found to not be significant, but temperature*elevationCenterPoint was significant

Table 4: Current and Potential Future Variables

Integrating Virtual and Augmented Reality with Brain-Computer Interfaces for ADHD and ASD Management: A Preliminary Review

Maximus Streeter
mstre5@brockport.edu
SUNY Brockport
Brockport, NY 14420 USA

Zhigang Li
zli8@kennesaw.edu

Joy Li
joy.li@kennesaw.edu

Chi Zhang
chizhang@kennesaw.edu

Xin Tian
xtian2@kennesaw.edu

Selena He
she4@kennesaw.edu

Kennesaw State University
Marietta, GA 30060 USA

Abstract

To review current and past applications of brain-computer interfaces in combination with augmented or virtual reality technologies in the intervention of Attention-Deficit/Hyperactivity Disorder and Autism Spectrum Disorder, we reviewed the related literature during the past 13 years. The literature review was categorized based on the focus of the studies. While more investigation is needed to thoroughly investigate the results and impact of these experiments, it has been shown that the experiments tested in the literature are mostly successful in intervention and diagnosis. We discuss relevant observations that may help future studies and inspire collaboration among researchers and partitioners in the field.

Keywords: ADHD, ASD, brain-computer interface, virtual reality, augmented reality

Recommended Citation: Streeter, M., Li, Z., Li, J., Zhang, C., Tian, X., He, S., (2024). Integrating Virtual and Augmented Reality with Brain-Computer Interfaces for ADHD and ASD Management: A Preliminary Review. *Journal of Information Systems Applied Research.* 17(1), pp. 33-41, https://doi.org/10.62273/VDBZ9371

Integrating Virtual and Augmented Reality with Brain-Computer Interfaces for ADHD and ASD Management: A Preliminary Review

Maximus Streeter, Zhigang Li, Joy Li, Chi Zhang, Xin Tian and Selena He

1. INTRODUCTION

Attention-Deficit/Hyperactivity Disorder (ADHD) is a neurological condition with an array of symptoms, including challenges in sustaining attention, difficulties in behavior regulation, and hyperactivity (Reddy & Lingaraju, 2020). In 2022, the number of children aged 3 to 17 in the United States diagnosed with ADHD is estimated at approximately 6 million, constituting 9.8% of this age group (CDC, 2022). Additionally, Autism Spectrum Disorder (ASD) encompasses a range of neurobehavioral symptoms, such as struggles in social interaction and communication, and repetitive behavioral patterns (Arpaia, Bravaccio, et al., 2020). The reported ASD among children in the United States is about 2.8% of the same age group as of 2023 (CDC, 2023). While ADHD and ASD are two different neurological disorders, both have witnessed an increase in intervention approaches through brain-computer interface (BCI) and virtual or augmented reality (VR or AR), and the combination methods over the last decade. Through the utilization of BCI, the brainwave activity of individuals can be monitored for data collection or for neurofeedback processes. Neurofeedback is a technique for individuals to gain awareness of their own brainwave activity through external stimulants and to normalize irregular brainwave patterns (Arpaia, Duraccio, et al., 2020). However, the use of BCI alone sometimes falls short in terms of engagement, where the intervention process is extensive. However, incorporating VR/AR systems into intervention methodologies has shown that participants tend to become more engaged and involved with the intervention process (Reddy & Lingaraju, 2020). This is especially true for individuals with ASD, and it is a major reason for its popularity among the ASD community (Simões et al., 2012). The interest comes from it as an alternative to medications, as it causes fewer side effects, including anxiety, suppressed appetite, irritability, and insomnia (Reddy & Lingaraju, 2020).

In this paper, we aim to review the studies focusing on ADHD and ASD intervention through the BCI with VR or AR technologies, and discuss the challenges and accomplishments in this specialized area. We hope to provide valuable insights to researchers and practitioners with similar interests in this evolving field.

2. METHOD

To find the relevant studies, we searched databases including Web of Science, IEEE Xplore, Google Scholar, ACM, Scopus, and PubMed. We searched the databases using the query "brain computer interface" AND ("virtual reality" OR "augmented reality" OR "extended reality") AND (autism OR ASD OR ADHD). We narrowed our search to only include studies published between 2010 and 2023. This was done because results that are most relevant started to emerge around 2010. Most results before 2010 discussed virtual reality rehabilitation methods outside of a headset context, which is not part of the focus of this study. The last search for literature was done on June 26, 2023. The search results from all the databases included an initial screening process where the title and the abstract of the study were reviewed against inclusion criteria. The inclusion criteria follow the guidelines below:

- The study must have an ADHD or ASD rehabilitation focus.
- Rehabilitation methods must involve the use of brain-computer interfaces.
- Rehabilitation methods must involve virtual, augmented, or extended reality.
- The study must be available in full text or have a substantive abstract.
- The study must be a peer-reviewed academic journal or conference paper.
- The study must be in English.

If a study's title and details met two of the first three guidelines, its abstract was examined to see if the last guideline could be met. It was excluded if the studies did not meet the last three guidelines or if the VR environment used in the study was not immersive, i.e., a 3D environment displayed on a computer monitor. The initial screening process resulted in 39 studies. After removing duplicates, 19 results were examined with a more in-depth screening process. One result was excluded because the headset was not involved in their virtual reality experiment. Another one was excluded because it is a thesis instead of a peer-reviewed publication; three others were excluded because they were proposal-oriented and did not include relevant details. After the screening processes, 12 studies were included in the review. The complete screening process is presented in Figure 1.

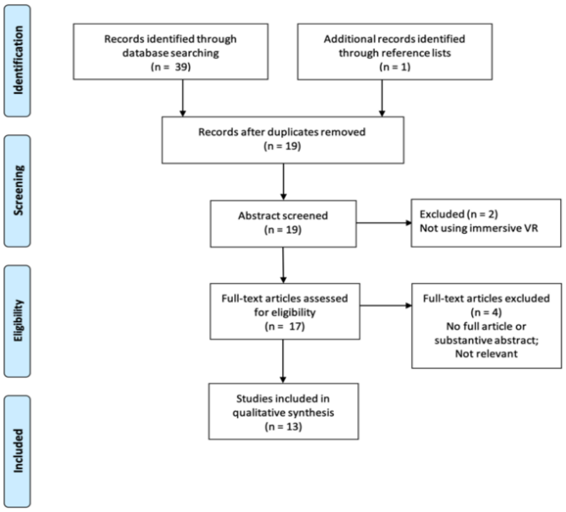

Identification
Screening
Eligibility
Included
Records identified through database searching (n = 39)
Additional records identified through reference lists (n = 1)
Records after duplicates removed (n = 19)
Abstract screened (n = 19)
Excluded (n = 2) Not using immersive VR
Full-text articles assessed for eligibility (n = 17)
Full-text articles excluded (n = 4) No full article or substantive abstract; Not relevant
Studies included in qualitative synthesis (n = 13)

Figure 1. PRISMA (Page et al., 2021) flowchart of the search and screening process

3. REVIEW RESULTS

In this section, we describe the results of the search process and summarize the findings within the literature.

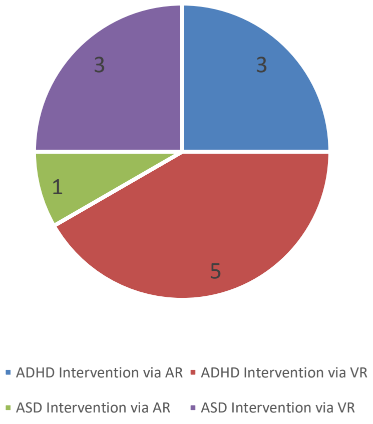


Figure 2. Number of papers per category

Overview of Search Results

Upon analyzing the search results, distinct patterns emerged within the topics covered in each paper, prompting the creation of well-defined categories to facilitate organization. The four categories are:

1. ADHD intervention through AR
2. ADHD intervention through VR
3. ASD intervention through AR
4. ASD intervention through VR

Figure 2 illustrates the distribution of papers in each category and the prevalence of specific research focus. It shows that ADHD intervention using VR emerged as the most prominent subject. In contrast, ASD intervention using AR constituted the least explored area, with only one study published in the year 2020.

A key aspect we kept track of while documenting details on the studies was the country of origin for each study. If it was explicitly mentioned in the paper as the country where that study took place, that country was designated for that particular paper. Otherwise, the country would be the authors' location at the time of publication, or the country of the university that the authors are affiliated with. The number of papers from each country is shown in Figure 3. The country that contributes the most papers is Portugal, with three papers, while China and Italy tie for the second most contributions with two papers each.

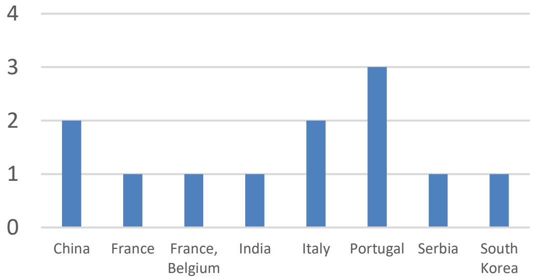


Figure 3. Number of papers per country

ADHD Intervention Through AR Methods

The articles collected that focus on ADHD intervention through AR used one of two approaches for intervention. The first approach involves transforming virtual objects through a process of regulating brainwave patterns. By allowing patients to transform a simulated world with their minds, this form of neurofeedback keeps the targeted children engaged with the intervention process. Rajshekar Reddy and G. M. (2020) discuss this form of intervention in their paper, where they developed two AR telekinetic games. Both of their games work by collecting the brainwave information of an individual via electroencephalogram (EEG), sending the brainwave data over a network connection where they are classified and the features are extracted, then the results of the classified data are sent to the AR application where they responded with positive or negative feedback. By reaching a desired psychological state, the first AR game will respond with a virtual balloon that begins to inflate. Similarly, the second game requires reaching a desired psychological state to bend a spoon. While no study has been performed yet on children with ADHD in their paper, Reddy and Lingaraju expected that their games would work well for ADHD as gamified exercises, as the ADHD neurofeedback protocol they adopted to design the game has shown to help increase children's concentration, control, and memory.

A second approach to ADHD intervention through AR was explored recently by Arpaia et al. (2021; 2020). Both articles detail their experiment on two different groups of children with ADHD. They described the method of intervention by controlling a robot to move through BCI and blink detection applications. Arpaia et al. (2021) attempted to train patients to control their focus and drive the robot according to the indication provided, which lessened the symptoms associated with ADHD, such as attention deficit, hyperactivity, and impulsivity. The system works through various applications, beginning with a pair of AR glasses that emit a flickering stimulus superimposed onto the surrounding environment. The flickering stimulus is set to two frequencies on either side of the glasses. Arpaia et al. explained that steady-state visual evoked potentials (SSVEPs) are natural frequencies emitted by the brain when focusing on a visual stimulus flickering within a certain frequency range. By focusing on one of these frequencies in the AR glasses, the patient's brainwaves can respond by oscillating at similar frequencies. These frequencies are picked up by electrodes placed on the scalp of the patient and translated into one of two commands, either move left or move right. Blink detection was also used in the process to tell the robot to cycle between three different states: idle, change direction, and move forward. All together, these two systems allow a person to control a robot through thinking and blinking. The purpose of the experiments was to observe the accuracy of the signal classifier, assess the comfort level of the patients when interfacing with the headset, and confirm the effectiveness of the system when used in a therapy setting. Results claim that the classifier worked as intended and most children responded very well to the system.

ADHD Intervention Through VR Methods

Instead of focusing on intervention, many applications of VR with a focus on ADHD in the collected literature trended toward diagnosis. However, most researchers expressed their interest in exploring ADHD intervention using the processes like neurofeedback in future work. Regardless, diagnosis in every example collected follows a similar process, which begins by placing the patient into a simulated environment via VR headset. Within this environment, several tests can be conducted while data readings are taking place physically through EEG, eye-tracking, or head position-tracking devices. The simulated environment varies from study to study, while all tend to include some sort of distracting feature whether it is traffic driving by, a bug flying around the room, or some other distraction. Some studies opted for a classroom environment, as explored by Lee et al. (2017) and Tan et al. (2019), while others chose to let the patient choose their virtual environment as a stress-reducing precaution, as explored by Devigne et al. (2020; 2020). These studies all include a similar test known as a continuous performance test (CPT). A CPT, as described by Lee et al. (2017), "is a common method for detecting sustained and selective attention in the field of neuropsychology." Sustained attention is the ability to keep focus on a stimulus for a continued period while selective attention is the ability to focus on a specific stimulus while blocking others out. Both abilities are negatively affected by ADHD and this is why they are important abilities to be tested in the diagnosis. A CPT can take on many different forms. In Lee et al. (2017) and Tan et al. (2019), the CPT involves flashing the patient a letter at a time and when the letter X appears directly after the letter A, the patient is instructed to confirm through some method that the event just took place. Lee et al. (2017) and Tan et al. (2019) also incorporate another test into their diagnosis process known as the Wisconsin card sorting test (WCST). The WCST is a test of cognitive function where the test takers are asked to sort a pile of 128 cards where each card has a picture of an object with variations in color, shape, and graphic quality. The results of the WCST are the number of correct and incorrect placements and supports the diagnosis of ADHD. The results of these studies described had a high accuracy of correct diagnosis when all factors were combined, including the results of the virtual test and readings from the physical behavioral and mental monitoring.

A study by Oh et al. (2022) is the only study we found that focuses on intervention of ADHD through VR instead of diagnosis. Their study takes place over multiple sessions with multiple patients, both with and without ADHD. During each session, the patient is immersed in a virtual simulation where they were riding a roller coaster. While on the roller coaster, balloons taking the shape of different animals as well as other distractable objects will move across the environment. An announcer will instruct the patient on which balloon animal to pop with a virtual gun while avoiding all other animals and objects. The purpose of this task is to heighten the patient's sense of presence. After each session the patient is given a questionnaire where they describe their sense of presence by answering multiple questions on a scale from 1 to 10. The results of this study show that patients with ADHD had a statistical significant increase in overall sense of presence after the sessions were complete, while those without ADHD experienced no significant change in sense of presence.

A summary of the above-mentioned studies, including title, source, data source, description of participants, whether AR or VR methods are used, and findings of each study that focuses on ADHD can be found in Appendix A.

ASD Intervention Through AR Methods

Only one article was identified to focus on ASD intervention through AR methods (Arpaia, Bravaccio, et al., 2020). This study follows a similar, if not identical, approach to intervention as seen in (Arpaia, Duraccio, et al., 2020) and (Arpaia et al., 2021). However, the difference in this study is that the three participants in the case study all have diagnosable ASD. This intervention was applied by letting patients control a robot through the use of blinking and brainwave activity alone. In this case, the intervention is believed to work on both ASD and ADHD children, as both tend to respond well to computerized activity. The results of this study are provided by the response of the 3 case study participants, who all gave positive feedback on the intervention.

ASD Intervention Through VR Methods

All studies collected that cover ASD rehabilitation through VR methods were done in Portugal. These studies all focus on training the joint attention skills of individuals with ASD. Joint attention is a social interaction between two people where nonverbal actions such as gaze and gestures are used to indicate a reference point on where the individuals should focus their shared attention (Charman, 2003). The collected literature describes that having deficits in joint attention skills is a common characteristic of individuals with ASD and that by training these skills, especially at a young age, other negative characteristics of ASD can be improved as well (C. Amaral et al., 2018; C. P. Amaral et al., 2017; Simões et al., 2012). The joint attention skill training described in these studies is done through neurofeedback. The neurofeedback process works by taking advantage of the P300 signal, a well-known neural signal created by detecting a rare item in a stimulus series (C. Amaral et al., 2018). For example, in (Simões et al., 2012), a VR scenario is described where the patient is positioned in front of a group of people. Each person in the simulated group is animated to make a movement at different instants. When one of these characters makes a movement and the patient notices using their joint attention skills, a P300 signal can be elicited. All other studies describe using similar VR scenarios to elicit a P300 signal. The first two studies in chronological order (C. P. Amaral et al., 2017; Simões et al., 2012) experiment with refining the machine learning methods used to classify these P300 signals. By increasing the classification accuracy of these signals, the neurofeedback process becomes more effective. The last article (C. Amaral et al., 2018) describes a multi-session study where individuals with ASD are subjected to VR scenarios and their behavioral characteristics are monitored. At the end of the sessions, while the rate of P300 signals stayed stagnant, participants showed a decrease in ASD related symptoms, depression, and mood disturbance.

A summary including title, source, data source, description of participants, AR or VR methods, and findings of each article that focuses on ASD can be found in Appendix B.

4. DISCUSSION

Only ADHD in the observed experiments has been reported on diagnosis, but no study was found directly focused on the diagnosis of ASD. Overall, the methods tested in the collected literature for the diagnosis of ADHD were reported successful. All the methods used a combination of head tracking, eye tracking, and EEG devices for data collection. Each of these studies revealed an interest in using their devices for neurofeedback-based therapies in the future.

The intervention of ADHD and ASD in the collected literature shows successful results. This high level of success shows the potential of complementing or replacing the traditional intervention methods or medications that causes unwanted side effects (Reddy & Lingaraju, 2020).

Current research on the intervention of ADHD and ASD through AV/VR and BCI technologies has shown promising results. However, this area is still under-researched. Given the effectiveness of the therapies described above and the accessibility of current AR/VR technologies, it makes sense for practitioners and researchers alike to explore and experiment further in the area of healthcare. An obvious furtherment of this area of research is to study the long-term effects of these therapies. Most of the studies examined in this paper are pilot studies and many describe the need for a long-term observational study to examine the results of these therapies.

An added benefit in continuing this area of research is that more data will be provided to train machine learning models. The need for more data is expressed when researchers described using deeper learning models for their intervention (Delvigne, Ris, et al., 2020). However, this is not possible due to the lack of data. As more data is collected, more accurate models will be available, and therapies will become more effective.

5. CONCLUSION

In this paper, we reviewed past studies conducted on ASD and ADHD intervention using AV/VR and BCI technologies. We systematically collected papers for review and broke the literature into four categories (ADHD intervention using AR or VR, ASD intervention using AR or VR). Each study was reviewed and analyzed. We also discussed the challenges the researchers faced during the experiments and the significant achievements made by the researchers. Overall, VR and AR technologies have shown great potential in helping children with ADHD and ASD. This paper serves as a starting point for continuing further research in the technologies in helping the therapy for ASD and ADHD.

6. ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 2244450.

7. REFERENCES

Amaral, C., Mouga, S., Simões, M., Pereira, H. C., Bernardino, I., Quental, H., Playle, R., McNamara, R., Oliveira, G., & Castelo-Branco, M. (2018). A Feasibility Clinical Trial to Improve Social Attention in Autistic Spectrum Disorder (ASD) Using a Brain Computer Interface. Frontiers in Neuroscience, 12. https://www.frontiersin.org/articles/10.3389/fnins.2018.00477

Amaral, C. P., Simões, M. A., Mouga, S., Andrade, J., & Castelo-Branco, M. (2017). A novel Brain Computer Interface for classification of social joint attention in autism and comparison of 3 experimental setups: A feasibility study. Journal of Neuroscience Methods, 290, 105–115. https://doi.org/10.1016/j.jneumeth.2017.07.029

Arpaia, P., Bravaccio, C., Corrado, G., Duraccio, L., Moccaldi, N., & Rossi, S. (2020). Robotic Autism Rehabilitation by Wearable Brain-Computer Interface and Augmented Reality. 2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA), 1–6. https://doi.org/10.1109/MeMeA49120.2020.9137144

Arpaia, P., Duraccio, L., Moccaldi, N., & Rossi, S. (2020). Wearable Brain–Computer Interface Instrumentation for Robot-Based Rehabilitation by Augmented Reality. IEEE Transactions on Instrumentation and Measurement, 69(9), 6362–6371. https://doi.org/10.1109/TIM.2020.2970846

Arpaia, S. Criscuolo, E. De Benedetto, N. Donato, & L. Duraccio. (2021). A Wearable AR-based BCI for Robot Control in ADHD Treatment: Preliminary Evaluation of Adherence to Therapy. 2021 15th

International Conference on Advanced Technologies, Systems and Services in Telecommunications (TELSIKS), 321–324. https://doi.org/10.1109/TELSIKS52058.2021.9606352

CDC. (2022, August 9). Data and Statistics About ADHD. Center for Disease Control and Prevention. https://www.cdc.gov/ncbddd/adhd/data.html#print

CDC. (2023, April 4). Data & Statistics on Autism Spectrum Disorder. Center for Disease Control and Prevention. https://www.cdc.gov/ncbddd/autism/data.html

Charman, T. (2003). Why Is Joint Attention a Pivotal Skill in Autism? Philosophical Transactions: Biological Sciences, 358(1430), 315–324. JSTOR.

Delvigne, V., Ris, L., Dutoit, T., Wannous, H., & Vandeborre, J.-P. (2020). VERA: Virtual Environments Recording Attention. 2020 IEEE 8th International Conference on Serious Games and Applications for Health (SeGAH), 1–7. https://doi.org/10.1109/SeGAH49190.2020.9201699

Delvigne, V., Wannous, H., Vandeborre, J.-P., Ris, L., & Dutoit, T. (2020). Attention Estimation in Virtual Reality with EEG based Image Regression. 2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR), 10–16. https://doi.org/10.1109/AIVR50618.2020.00012

Lee, W., Kim, S., Kim, B., Lee, C., Chung, Y. A., Kim, L., & Yoo, S.-S. (2017). Non-invasive transmission of sensorimotor information in humans using an EEG/focused ultrasound brain-to-brain interface. PloS One, 12(6). https://doi.org/10.1371/journal.pone.0178476

Oh, S. H., Park, J. W., & Cho, S.-J. (2022). Effectiveness of the VR Cognitive Training for Symptom Relief in Patients with ADHD. Journal of Web Engineering, 21(03), 767–788. https://doi.org/10.13052/jwe1540-9589.21310

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. Systematic Reviews, 10(1), 89. https://doi.org/10.1186/s13643-021-01626-4

Reddy, G. S. R., & Lingaraju, G. M. (2020). A Brain-Computer Interface and Augmented Reality Neurofeedback to Treat ADHD: A Virtual Telekinesis Approach. 2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), 123–128. https://doi.org/10.1109/ISMAR-Adjunct51615.2020.00045

Simões, M., Carvalho, P., & Castelo-Branco, M. (2012). Virtual Reality and Brain-Computer Interface for joint-attention training in Autism. Virtual Reality.

Tan, D. Zhu, H. Gao, T. -W. Lin, H. -K. Wu, S. -C. Yeh, & T. -Y. Hsu. (2019). Virtual Classroom: An ADHD Assessment and Diagnosis System Based on Virtual Reality. 2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS), 203–208. https://doi.org/10.1109/ICPHYS.2019.8780300

APPENDIX A
Article Summary of ADHD-Focused Studies

Title / Source	Data Source	Participants	AR/VR	Findings
A Brain-Computer Interface and Augmented Reality Neurofeedback to Treat ADHD: A Virtual Telekinesis Approach (Reddy & Lingaraju, 2020)	Not Available	Not Available	AR	The authors developed a virtual telekinesis game in an AR environment for the treatment of persons diagnosed with ADHD. Further studies are still needed to evaluate the effectiveness of the game.
A Wearable AR-based BCI for Robot Control in ADHD Treatment: Preliminary Evaluation of Adherence to Therapy (Arpaia et al., 2021)	Commands used, errors, and accuracy of tasks completed from adult voluneteers. Initial reluctance and task completed from children participants.	10 healthy adults for preliminary testing and 18 children 5 – 10 years old (yo) with ADHD.	AR	Children aged 8 to 10 years were able to complete the tasks successfully. A few younger kids experienced difficulties concerning the usability of the devices and exhibited reduced levels of focus when the operation was being explained.
ADHD Assessment and Testing System Design based on Virtual Reality (Lee et al., 2017)	Test results along with head position, eye tracking, EEG, and smart watch data.	120 aged between 8 and 14. 60 children diagnosed with ADHD, 60 children without ADHD.	VR	The authors designed and implemented a VR based system for the diagnoistic and test of ADHD using data from a combination of sources.
Attention Estimation in Virtual Reality with EEG based Image Regression (Delvigne, Wannous, et al., 2020)	Model accuracy, EEG, eye-tracking, and head position data.	5 healthy individuals.	VR	By using the collected data, their CNN was able to outperform previous studies on the estimation of attentional state of individuals.
Effectiveness of the VR Cognitive Training for Symptom Relief in Patients with ADHD (Oh et al., 2022)	Sense of presence questionnaire and EEG data.	8 experimental subjects with ADHD and 8 control subjects without ADHD. 8 – 13 yo.	VR	The sense of presence underwent a notable transformation following VR cognitive training among the participants with ADHD. However, the control group did not experience any comparable alterations in their sense of presence after the VR cognitive training.
VERA: Virtual Environments Recording Attention (Delvigne, Ris, et al., 2020)	EEG, eye-tracking, and head position.	Not Available	VR	The study proposed a noval framework for evaluating attention in a VR environment. The framework demonstrated promising outcomes in terms of categorizing attention states.
Virtual Classroom: An ADHD Assessment and Diagnosis System Based on Virtual Reality (Tan et al., 2019)	EEG, eye-tracking and head position data along with audio test, CPT, and WCST performance data.	100 male children (6 – 12 yo). 50 children with ADHD in the experimental group and 50 without ADHD in the control group.	VR	By analysis of the performance and sensor data, researchers could conclude whether the subject had ADHD.
Wearable Brain–Computer Interface Instrumentation for Robot-Based Rehabilitation by Augmented Reality (Arpaia, Duraccio, et al., 2020)	SSVEP detection and eye-tracking data along with time to complete task, number of commands used, errors, and accuracy results.	10 healthy adult subjects for preliminary study, 4 untrained children (6 – 8 yo) with ADHD for case study.	AR	The SSVEP/Eye blink detection algorithm achieved an average accuracy of higher than 83%. The study on ADHD patients got positive feedback on device acceptance and attentional performance.

APPENDIX B
Article Summary of ASD Focused Studies

Title / Source	Data Source	Participants	AR/VR	Findings
A Feasibility Clinical Trial to Improve Social Attention in Autistic Spectrum Disorder (ASD) Using a Brain Computer Interface (C. Amaral et al., 2018)	Questionnaires, eye-tracking, EEG data	N =15 (mean age = 22 years and 2 months, 16-38 yo) with high-functioning ASD	VR	Positive effects in all subscales of the Autism Treatment Evaluation Checklist (ATEC) and in ATEC total scores, improvement in Adapted Behavior Composite, and in all subareas from the Vineland Adaptive Behavior Scale.
A novel Brain Computer Interface for classification of social joint attention in autism and comparison of 3 experimental setups: A feasibility study (C. P. Amaral et al., 2017)	Setup time, reported comfort	N=17 (13 healthy, 4 with ASD)	VR	g.Nautilus proved to be the best-performing system among 3 systems (g.Mobilab+, g.Nautilus, V-Amp with actiCAP Xpress dry-electrodes) in accurately detecting P300, preparation time, speed, and reported comfort.
Robotic Autism Rehabilitation by Wearable Brain-Computer Interface and Augmented Reality (Arpaia, Bravaccio, et al., 2020)	SSVEP detection, eye-tracking, errors, and accuracy results	N=10 healthy adults for preliminary study, 3 children on ASD (8-10 yo with different CGI scores) for case study.	AR	SSVEP/Eye blink detection algorithm shows accuracy higher than 83%, with a corresponding time response 1.5s for adults. Positive device acceptance and attentional performance for 3 children on ASD.
Virtual Reality and Brain-Computer Interface for joint-attention training in Autism (Simões et al., 2012)	EEG data	4 subjects with no developmental disorders, mean age =22	VR	90% accuracy by a classifier to identify the target object with P300 wave in the EEG, on high-level social animations performed by virtual characters.

Investigating the Relationship between Developer Job Satisfaction and Life Satisfaction: A Global Analysis

Alan Peslak
arp14@psu.edu
Department of Information Sciences and Technology
Penn State University
Dunmore, PA 18512 USA

Wendy Ceccucci
wendy.ceccucci@qu.edu

Kiku Jones
kiku.jones@qu.edu

Business Analytics and Information Systems Department
Quinnipiac University
Hamden, CT 06518 USA

Lori N. K. Leonard
lori-leonard@utulsa.edu
Department of Accounting and Business Information Systems
The University of Tulsa
Tulsa, OK 74104

Abstract

This paper explores the association between developer job satisfaction and life satisfaction on a global scale. With the increasing prominence of the software development industry and its impact on individuals' professional and personal lives, understanding the connection between job satisfaction and overall life satisfaction becomes crucial. We present a comprehensive analysis based on data collected from diverse regions and cultures, aiming to determine the magnitude of the correlation between these two constructs.

Through a systematic review of existing literature, surveys, and empirical studies, we compile a robust dataset encompassing responses from developers across different countries and career stages. To assess job satisfaction, we used a worldwide survey from StackOverflow. Life satisfaction is evaluated from the World Values Survey, the Human Development Index, and the World Happiness Report. We grouped the world into logical geographic regions to discover key insights.

Our findings reveal a modest correlation between developer job satisfaction and life satisfaction worldwide. Despite the significant impact of job satisfaction on one's professional life, the influence on overall life satisfaction appears to be more nuanced. While several studies have reported positive associations between the two constructs, our analysis suggests that the correlation is relatively weak, indicating that job satisfaction alone cannot fully predict an individual's level of life satisfaction.

Keywords: developer job satisfaction, life satisfaction, global analysis, correlation, well-being, holistic approach

Recommended Citation: Peslak, A., Ceccucci, W., Jones, K., Learnard, L., (2024). Investigating the Relationship between Developer Job Satisfaction and Life Satisfaction: A Global Analysis. *Journal of Information Systems Applied Research*. 17(1), pp 42-52, https://doi.org/10.62273/UFMR5574

Investigating the Relationship between Developer Job Satisfaction and Life Satisfaction: A Global Analysis

Alan Peslak, Wendy Ceccucci, Kiku Jones and Lori N. K. Leonard

1. INTRODUCTION

The importance of determining variables that affect job satisfaction is an important element in determining how job satisfaction can be improved. Improving job satisfaction has many benefits for both individual workers, organizations that employ these workers and the economy of a nation overall (Faragher, Cass, & Cooper, 2005; Augner, 2015; Arnold, Coffeng, Boot, Van Der Beek, Van Tulder, Nieboer, Van Dongen, 2016). Life satisfaction and job satisfaction are interrelated constructs that play a pivotal role in individuals' mental health and overall quality of life (Diener, Oishi, & Tay, 2018). Life satisfaction, the cognitive-judgmental aspect of subjective well-being (Diener, 1984), is significantly influenced by various factors, including employment and job satisfaction (Clark, Oswald, & Warr, 1996). The latter, job satisfaction, is a multidimensional psychological response to one's job, which may include cognitive (evaluative), affective (or emotional), and behavioral components (Spector, 1997).

Understanding the relationship between these two aspects and the factors that influence them can provide valuable insights into the shaping of policies and strategies aimed at improving the overall quality of life for individuals worldwide. Although previous research has explored these aspects separately or within specific cultures or nations (Judge, Thoresen, Bono, & Patton, 2001; Helliwell & Huang, 2014), a comprehensive global analysis is still lacking. This is particularly relevant in the face of globalization and the current era of digital transformation, where work conditions and life satisfaction factors are rapidly changing (Brynjolfsson & McAfee, 2014).

This research paper aims to fill this gap by providing a comprehensive examination of worldwide life satisfaction and job satisfaction. We synthesize data from various international databases, such as the World Happiness Report (Helliwell, Layard, Sachs, De Neve, Aknin, & Wang, 2022), the Human Development Index (United Nations, 2023), and the World Values Survey (Inglehart, Haerpfer, Moreno, Welzel, Kizilova, Diez-Medrano, Lagos, Norris, Ponarin, & Puranen, 2020) to provide an overarching understanding of these constructs across different cultures, economies, and labor markets. By integrating these diverse data sources, we aim to generate an inclusive, nuanced, and up-to-date understanding of life and job satisfaction on a global scale.

2. LITERATURE REVIEW

Job Satisfaction

Organizational behaviorists and organizational psychologists have long studied the subject of employees' job satisfaction. The literature includes several facets of what variables make up job satisfaction. According to Lumley, Coetzee, Tladinyane & Ferreira (2011), job satisfaction can be defined as "an individual's total feeling about their job and the attitudes they have towards various aspects or facets of their job, as well as an attitude and perception that could consequently influence the degree of fit between the individual and the organization" (pg. 101). Employee satisfaction is "determined by subjective perceptions related to the treatment received by the organization, for instance, policies of rewards, hiring and firing policies, performance and retribution." (Crespi-Vallbona & Mascarilla-Miro, 2018, pg. 36). Sempane, Rieger & Roodt (2002), assert that job satisfaction is made up of many variables such as "structure, size, pay, working conditions and leadership", all representatives of organizational climate (pg. 23). Some of these variables may also include the "importance of job position, teamwork atmosphere, leadership, recognition and compensation, physical labor conditions and personal labor conditions as key aspects of employees' well-being." (Crespi-Vallbona, et al., pg. 37). In a study done by LeRouge, Wiley, & Maertz (2013), the authors included job security, the work itself, one's supervisor, compensation, work/life balance, and advancement/opportunities as important facets of job satisfaction.

According to an article written by David Engle (2020) and published by CompTIA, 72% of global IT professionals are satisfied with their job. One of main reasons for employee's satisfaction was professional development.

Globally, 81% of IT staff listed "build new skills" as their top reason for development.

A study completed across four countries, Austria, Germany, Slovenia, and Spain, found that differences in employee job satisfaction between IT and other sectors were not statistically significant (Cic, Bobek, & Zizek, 2018)

Life Satisfaction
Life satisfaction measures how people evaluate their life as a whole rather than their current feelings. Daniel Kahneman, Nobel Prize winner, described happiness as being happy in your life. and the experience in real time. He described life satisfaction as being retrospective. It is the happiness about your life. It is the happiness that exists when one talks about the past and the big picture. Life satisfaction is the way in which people show their emotions, feelings and how they feel about their directions and options for the future (Kahneman, 2010).

Several research papers have extensively reviewed the literature on job and life satisfaction (Tait, Padgett, & Baldwin, 1989; Rain, Lane, & Steiner, 1991; Unanue, Gomez, Cortez, Oyandedel & Mediburo-Sequel, 2017; Riche, Near, & Hunt, 1980). The summary below uses the results of some of this extensive literature review.

According to the literature, there are three main tracks of thought regarding the correlational relationship between job satisfaction and life satisfaction: segmentation, compensation and spillover.

The segmentation theory suggests that there is no relationship between job and life satisfaction Theoretical positions such as partial inclusion have been proposed to explain the link between both concepts from this perspective.

The compensation theory holds that people compensate for their job dissatisfaction by finding more satisfaction in other areas of their life, and vice versa (Iris & Barrett, 1972). This implies that there is a negative relationship between the two constructs.

The majority of previous research has supported the spillover theory (Rain et al., 1991). The spillover theory argues that there is a positive relationship between job satisfaction and life satisfaction. The relationship is based on the generalization of belief and attitudes, conditioning, and cognitive dissonance. Previous research has demonstrated that some type of reciprocal relationship exists between job and life satisfaction (Tait, 1989 ; Rain et al., 1991).

To measure life satisfaction our study uses the results of a question from the World Values Survey (WVS). WVS is an international research project developed by Professor Inglehart and his researchers from the University of Michigan. The survey is conducted globally every 5 years. The purpose of the research is to analyze people's values, beliefs, and norms over time.

Happiness
Happiness measures a person's well-being and contentment. Several studies have explored happiness. Weaver (1978) examines global happiness as it relates to job satisfaction. He finds very few correlations between job satisfaction and global happiness. However, he indicates that employees that are happy because of their related job satisfaction are likely to have more satisfaction in other parts of their life.

Martinez-Marti & Ruch (2017) indicate that happiness can be evaluated as a pleasant life, an engaging life, and a meaningful life. When exploring happiness's relationship with job satisfaction in Switzerland, they find an engaging life to be positively related to job satisfaction, where as a pleasant life and a meaningful life are not found to individually influence job satisfaction. However, the interaction of all three measures of happiness are a good predictor of job satisfaction.

Tsou & Liu (2001) examine happiness and job satisfaction in Taiwan. Their findings indicate that various domains can influence both factors. In particular, an individual's income affects happiness and job satisfaction. Married individuals are also found to have greater happiness and job satisfaction.

Taken from the Gallup World Poll (GWP) between 2005 and 2022, happiness is measured with the following question: "Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?" (World Happiness Report, 2023).

Human Development Index
The Human Development Index (HDI) is a measure of a country's average achievement in three dimensions: income, health, and education

(World Population Review, 2023). Many indicators are aggregated to determine a country's HDI. These include, but are not limited to, "life expectancy, literacy rate, rural populations' access to electricity, GDP per capita, exports and imports, homicide rate, multidimensional poverty index, income inequality, internet availability, and many more" (World Population Review, 2023).

Ngoo & Tey (2019) study life satisfaction in relation to HDI. They find that HDI attributes to variation in life satisfaction across countries. In particular, they indicate a positive correlation between HDI and life satisfaction. However, HDI alone can result in different life satisfaction across countries. Countries with similar HDI are found to have variations in life satisfaction due to variations in the dimension of well-being.

Yin, Lepinteur, Clark, & D'Ambrosio (2023) use the HDI data from 150 countries between 2005 and 2018 to examine the relationship between HDI and well-being. Looking at the three dimensions of HDI separately, they find income to be the strongest predictor of well-being. Examining all three HDI dimensions, they find that they only matter equally in "Western and rich countries".

3. METHODOLOGY

Our investigation into the remuneration and contentment of software engineers utilized data gleaned from the 2020 Stack Overflow Survey, which recorded responses from more than 65,000 individuals. The yearly Developer Survey conducted by Stack Overflow is globally recognized as the most comprehensive and extensive examination of individuals who engage in coding. Annually, the array of questions posed in their survey span various topics, from preferred technologies of developers to their favored job conditions. As stated on StackFlow's website:

> "Stack Overflow's annual Developer Survey has been the biggest survey of coders worldwide for nearly a decade. For the year 2020, we aimed to make our survey more reflective of the global diversity of programmers rather than merely being the largest. Nevertheless, the survey remains substantial, with close to 65,000 participants." (StackFlow, 2020)

The use of Stack Overflow as a data source is well recognized and has been cited in various peer-reviewed publications including those by Barua, Thomas, & Hassan (2014), Asaduzzaman Mashiyat, Roy, & Schneider (2013), and Treude & Robillard (2016). The Stack Overflow dataset provides a wealth of data, with numerous demographic, descriptive, and opinion-based questions about the current state of programming. The data was processed and scrutinized using IBM SPSS 26.

The initial step in our analysis involved refining the dataset. We eliminated data from participants who identified themselves as hobbyists as opposed to full-time developers. Out of the 65,000 participants, 47,200 identified themselves as professional developers.

4. RESULTS

We used responses from three different surveys to evaluate life satisfaction. The use of distinct datasets is well established in the literature. For example, Augner (2015) compared job satisfaction to numerous distinct data sets including healthy life years, family satisfaction, fixed term contracts, inflation rates, employment rate of women, physical activity >=1/week, etc. The reports used for this study included:

1) World Values' Survey (Haerpfer, Inglehart, Moreno, Welzel, Kizilova, Diez-Medrano, Lagos, Norris, Ponarin, Puranen, 2022).
2) the World Happiness Index (Helliwell, Layard, Sachs, De Neve, Aknin, Wang, 2022).
3) Human Development Index (United Nations, 2023).

We evaluated the results to determine which if any correlated with developer job satisfaction.

Job Satisfaction

The job satisfaction levels for the more than 60 countries represented in the Stackflow database are shown in Figure 1. Ranging from 2.7 to 4.0 the darker the blue the higher the satisfaction level. Andorra, Netherlands, United States, Canada, and Australia were the countries with the highest job satisfaction levels. Whereas Zimbabwe, Tunisia, China, Libya, and Morocco had the lowest job satisfaction levels.

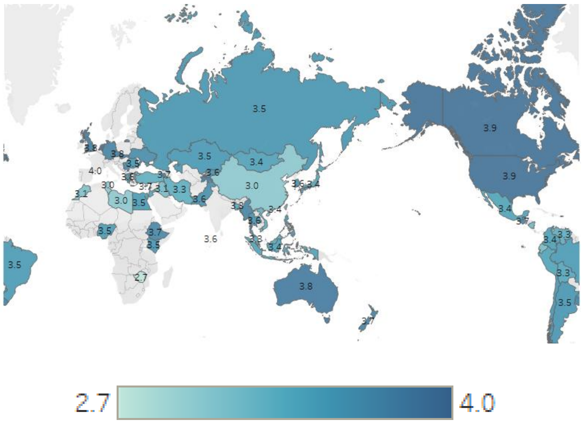


Figure 1. Job Satisfaction by Country

The next step was to divide the countries reported into world regions. We have chosen to categorize countries into the following world areas as shown in Figures 2 and 3. Primarily these subdivisions map to the World Bank Organization's groups (World Bank, 2023). We separated Russia out due to its land mass in both Asia and Europe.

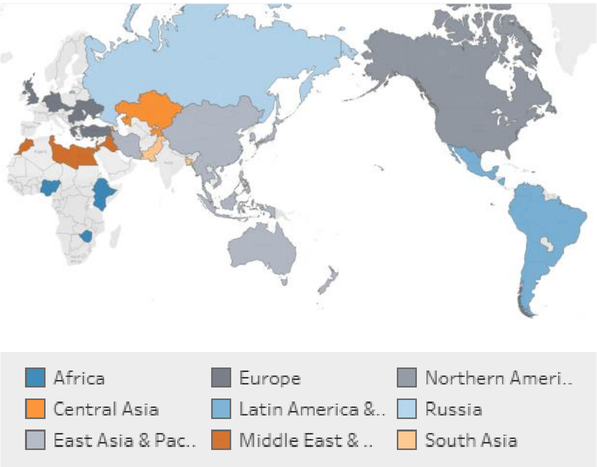


Figure 2. Countries by Geographical Areas

Figure 3 shows the overall developer job satisfaction by World Region. The highest job satisfaction was in Northern America which includes USA and Canada. The lowest developer satisfaction was in Africa, but all were above a neutral position. The differences were significant at p < .005.

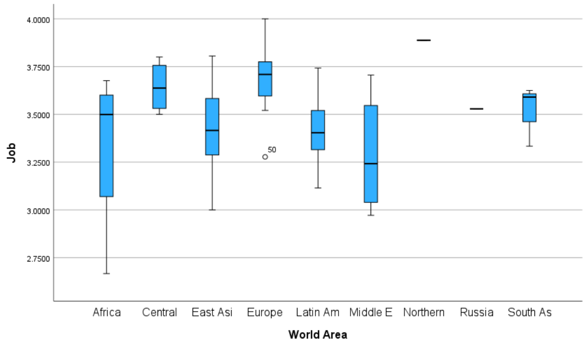


Figure 3. Job Satisfaction by Geographic Area

The economic situation in the country was then examined to see if this significantly correlated with job satisfaction. The correlation was .514 between GDP per capita and developer compensation. This suggests that 25% (R squared) of the variance may be due to the poor economy and wages within the World Region.

World Values' Life Satisfaction

We then examined the geographic areas (GA) for overall life satisfaction based on the responses from the World Values Survey. The results taken from the survey were responses to the question, "All things considered, how satisfied are you with your life as a whole these days?" The Likert scale for the question was from 1 to 10, where 1 indicated completely dissatisfied and 10 completely satisfied.

The life satisfaction levels for these same countries were used in the analysis and are shown in Figure 4. The values ranged from 4.5 to 8.4. Andorra, Netherlands, United States, Canada, and Australia were the countries with the highest job satisfaction levels. Whereas Zimbabwe, Tunisia, China, Libya and Morocco had the lowest job satisfaction levels.

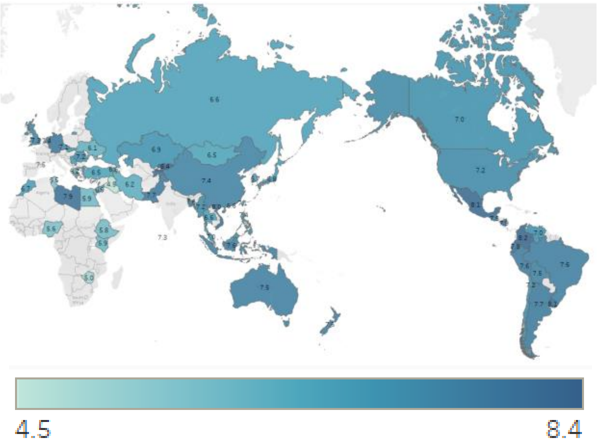


Figure 4. World Values Life Satisfaction

Results by Country

Geographical Area	Mean	N
Africa	5.56	4
Central Asia	7.47	4
East Asia & Pacific	7.07	16
Europe	6.95	11
Latin America & The Caribbean	7.69	12
Middle East & North Africa	6.33	8
Northern America	7.13	2
Russia	6.55	1
South Asia	7.52	3
Total	7.01	61

Table 1. World Values' Life Satisfaction by Geographic Area

Table 1 shows the results of dividing the World Values' life satisfaction levels into their respective geographical areas. The World Regions with the highest life satisfaction differed from those GAs with the highest job satisfaction. The highest life satisfaction was found in Latin America and the Caribbean, followed by South Asia and Central Asia. Africa, however, still had the lowest levels for life satisfaction. These results indicate there is a difference between job satisfaction and life satisfaction for IT developers.

The next analysis was to compare each World Region's World Values life satisfaction response with job satisfaction. Since the scales for each were different, we normalized the data (Figure 5).

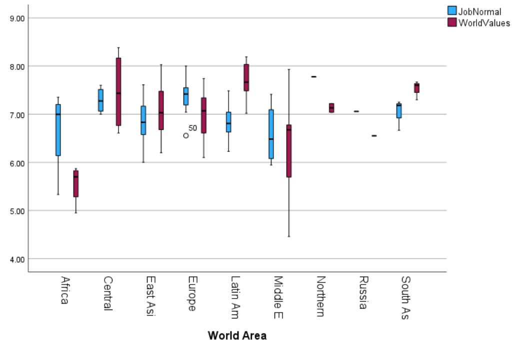


Figure 5. Job Satisfaction and World Values' Life Satisfaction

Africa as a region did not fare well in this analysis. Both job satisfaction and life satisfaction were below the average World means. Overall life satisfaction however was much lower than job dissatisfaction.

Central Asia on the other hand, had above average job satisfaction as well as life satisfaction. Their variances were very similar. East Asia and Pacific job satisfaction was below average, but life satisfaction was above average.

Europe oddly had strong above average job satisfaction but was below the world in life satisfaction. Latin Americans and the Caribbean had the largest positive variance from world life satisfaction, but their job satisfaction was below average.

The Middle East and North Africa suffered like Africa with low job satisfaction and low life satisfaction. Northern America showed the highest job satisfaction but was only marginally above average in life satisfaction.

Russia job satisfaction was about on average but was well below average life satisfaction. Finally, South Asia was on average in job satisfaction but well above average in life satisfaction.

When we examine the correlation between job satisfaction and World Value's life satisfaction overall, we find a significance level of only p < .053 and a small effect of about 12% (R squared).

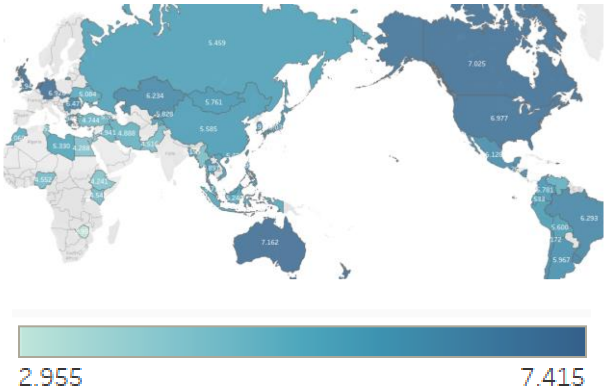


Figure 6. Happiness by Country

Happiness Index

The next measure we looked at was the Happiness Index (Figure 6). The values ranged from 2.9 to 7.45. The countries that were the happiest were the same as those with the highest world values' life satisfaction. Except, there was no data available for Andorra's happiness level. The least happy countries were Lebanon, Zimbabwe, Jordon, and Ethiopia.

The happiness results were then grouped by World Region and normalized to the same scale as job satisfaction. The results are shown in Figure 7 and Table 2.

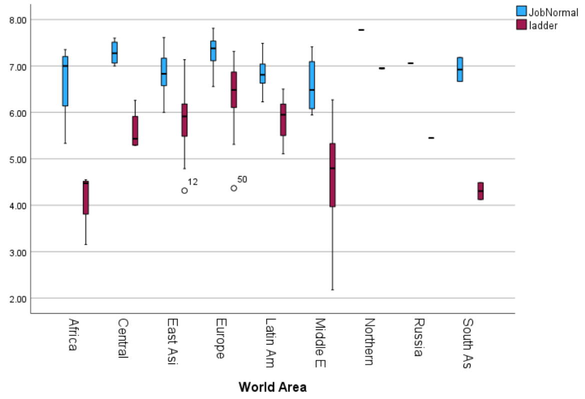


Figure 7. Happiness and Job Satisfaction

Geographical Area	Mean	N
Africa	4.16	4
Central Asia	5.60	4
East Asia & Pacific	5.87	16
Europe	6.29	10
Latin America & The Caribbean	5.86	12
Middle East & North Africa	4.58	8
Northern America	6.95	2
Russia	5.44	1
South Asia	4.30	2
Total	5.61	59

Table 2. Happiness Index by Geographic Area

Northern America had the highest happiness index values. Differences between areas were significant at p < .001.

Human Development Index

The last measure we examined was the Human Development Index (HDI) (Figure 8). The values ranged from 0 to 1. The countries with the highest index values were Hong Kong, Australia, Germany, Netherlands, and Singapore. The countries with the lowest Human Development Index were Taiwan, Ethiopia, Nigeria, Pakistan and Kenya,

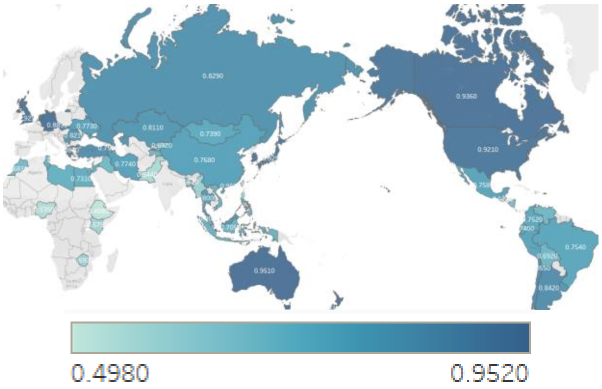


Figure 8. Human Development Index by Country

The HDI results were then grouped by World Region and normalized to the same scale as job satisfaction. The results are shown in Figure 9 and Table 3.

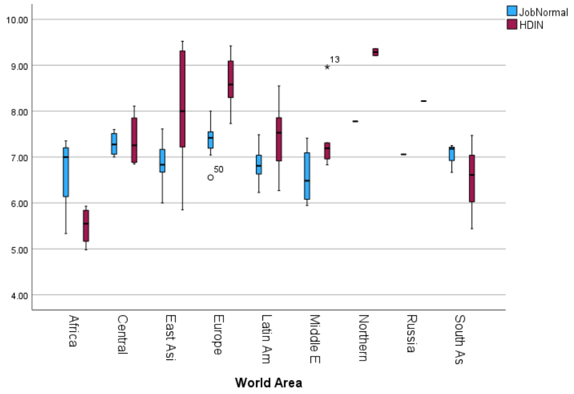


Figure 9. Human Development Index and Job Satisfaction

Geographical Area	Mean	N
Africa	0.550	4
Central Asia	0.737	4
East Asia & Pacific	0.814	15
Europe	0.866	11
Latin America & The Caribbean	0.746	12
Middle East & North Africa	0.734	8
Northern America	0.929	2
Russia	0.822	1
South Asia	0.651	3
Total	0.772	60

Table 3 HDI by Geographic Area

Developer job satisfaction and the Human Development index showed even more different results. Though Africa is consistent with job satisfaction above HDI, South Asia now shows higher job satisfaction and Europe shows much higher HDI than job satisfaction.

Overall
The results of the preceding charts are quantified in Appendix 1. We see the full differences in variances in means between job satisfaction and happiness and between job satisfaction and HDI. A World Values table was not included due to its nonsignificant correlation with developer job satisfaction.

5. CONCLUSIONS

The study highlights the importance of considering a broader range of factors beyond professional aspects when assessing life satisfaction. Individual differences, cultural influences, and personal values play significant roles in shaping overall well-being. Moreover, the complex interplay between work and personal life, as well as the varying priorities and aspirations of individuals, contributes to the intricate nature of the relationship between job satisfaction and life satisfaction.

Understanding these nuanced associations can have implications for organizations, policymakers, and individuals seeking to enhance well-being in the workplace and beyond. By recognizing that job satisfaction is just one facet contributing to overall life satisfaction, organizations can focus on holistic approaches to employee well-being, considering factors beyond traditional work-related metrics. Policymakers can use this information to design strategies that promote well-being across various domains, fostering healthier work environments and work-life integration. Lastly, individuals can gain insights into the multifaceted nature of life satisfaction, leading to more balanced and informed decisions regarding their career and personal aspirations.

6. ACKNOWLEDGEMENTS

The article includes contents edited by ChatGPT, a language model developed by OpenAI in San Francisco, CA, USA. Additionally, ChatGPT provided assistance in sentence editing.

7. REFERENCES

Asaduzzaman, M., Mashiyat, A. S., Roy, C. K., & Schneider, K. A. (2013, May). Answering questions about unanswered questions of stack overflow. In *2013 10th Working Conference on Mining Software Repositories (MSR)* (pp. 97-100). IEEE.

Augner, Christoph. (2015). Job satisfaction in the European Union. *Journal of occupational and environmental medicine, 57*(3),241-245.

Arnold, A. E., Coffeng, J. K., Boot, C. R., Van Der Beek, A. J., Van Tulder, M. W., Nieboer, D., & Van Dongen, J. M. (2016). The Relationship Between Job Satisfaction and Productivity-Related Costs. *Journal of occupational and environmental medicine*, *58*(9), 874-879.

Barua, A., Thomas, S. W., & Hassan, A. E. (2014). What are developers talking about? an analysis of topics and trends in stack overflow. *Empirical Software Engineering*, *19*, 619-654.

Brynjolfsson, E., & McAfee, A. (2014). The second machine age: Work, progress, and prosperity in a time of brilliant technologies. WW Norton & Company.

Cic, Z., Bobek S., & Zizek, S. IT Employees' Job Satisfaction (2018) Comparative Analysis between Industries. *Journal of Management & Research*. 8(1), 3-11.

Clark, A. E., Oswald, A. J., & Warr, P. B. (1996). Is job satisfaction U-shaped in age? *Journal of occupational and organizational psychology*, 69(1), 57-81.

Crespi-Vallbona, M., & Mascarilla-Miro, O. (2018, 2nd Quarter). Job Satisfaction: The Case of Information Technology (IT) Professionals in Spain. *Universia Business Review*, 58, 36-51.

Diener, E. (1984). Subjective well-being. *Psychological bulletin*, 95(3), 542.

Diener, E., Oishi, S., & Tay, L. (2018). Advances in subjective well-being research. *Nature Human Behaviour*, 2(4), 253-260.

Engle, D. (2020) IT Job Satisfaction: It's About More than Money, published by CompTIA, retrieved Aug 8, 2023 from: https://www.comptia.org/blog/it-job-satisfaction-it-s-about-more-than-money

Faragher, E. Brian, Monica Cass, and Cary L. Cooper (2005). The relationship between job satisfaction and health: a meta-analysis, *Occupational and environmental medicine* 6(2)2 105-112.

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano J., M. Lagos, P. Norris, E. Ponarin & B. Puranen (eds.). (2022). World Values Survey: Round Seven - Country-Pooled Datafile Version 5.0. Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat, retrieved from https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp.

Helliwell, J. F., & Huang, H. (2014). New measures of the costs of unemployment: Evidence from the subjective well-being of 3.3 million Americans. *Economic Inquiry*, 52(4), 1485-1502.

Helliwell, J. F., Layard, R., Sachs, J. D., De Neve, J.-E., Aknin, L. B., & Wang, S. (Eds.). (2022). World Happiness Report 2022. New York: Sustainable Development Solutions Network.

Inglehart, R., Haerpfer, C., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., ... & Puranen, B. (eds.) (2020). World Values Survey: Round Seven–Country-Pooled Datafile. JD Systems Institute & WVSA Secretariat.

Iris, B., & Barrett, G. V. (1972). Some relations between job and life satisfaction and job importance*. Journal* of *Applied Psychology,* 56, 301–304.

Judge, T. A., Thoresen, C. J., Bono, J. E., & Patton, G. K. (2001). The job satisfaction–job performance relationship: a qualitative and quantitative review. *Psychological Bulletin*, 127(3), 376-407.

Kahneman, D. (2010) *The Riddle of Experience vs. Memory*. March. Available at https://www.ted.com/talks/daniel_kahneman_the_riddle_of_experience_vs_memory (Accessed June, 2023)

LeRouge, C. M., Wiley, J. W., & Maertz, J. C. (2013). A Comparison of Job Satisfaction between IT and Non-IT Women Incumbents in Clerical, Professional, and Managerial Positions. *The DATA BASE for Advances in Information Systems*, 44(2), 39-54.

Lumley, E., Coetzee, M., Tladinyane, R., & Ferreira, N. (2011). Exploring the job satisfaction and organizational commitment of employees in the information technology environment. *Southern African Business Review,* 15(1).

Martínez-Martí, M. L., & Ruch, W. (2017).The Relationship Between Orientations to Happiness and Job Satisfaction One Year Later in a Representative Sample of Employees in Switzerland. *Journal of Happiness Studies,* 18, 1–15.

Ngoo, Y. T., & Tey, N. P. (2019). Human development index as a predictor of life satisfaction. *Journal of Population and Social Studies*, 27(1), 70-86.

Rain, J. S., Lane, I. M., & Steiner, D. D. (1991). A current look at the job satisfaction/life satisfaction relationship: review and future considerations. *Human Relations*, 44, 287–307.

Rice, R. W., Near, J. P., & Hunt, R. G. (1980). The job-satisfaction/life-satisfaction relationship: a review of empirical research. *Basic Applied Social Psychology*, 1, 37–64.

Sempane, M., Rieger, H., & Roodt, G. (2002). Job Satisfaction in Relation to Organisational Culture. SA Journal of Industrial Psychology, 28(2), 23-30.

Spector, P. E. (1997). *Job satisfaction: Application, assessment, causes, and consequences* (Vol. 3). Sage. Stack Overflow retrieved May 18, 2022, from https://insights.stackoverflow.com/survey/2020#overview

Tait, M., Padgett, M. Y., & Baldwin, T. T. (1989). Job and life satisfaction: a reevaluation of the strength of the relationship and gender effects as a function of the date of the study. *Journal of Applied Psychology,* 74, 502–507.

Tsou, M. W., & Liu, J. T. (2001). Happiness and Domain Satisfaction in Taiwan. *Journal of Happiness Studies*, 2, 269–288.

Treude, C., & Robillard, M. P. (2016, May). Augmenting API documentation with insights from stack overflow. In *Proceedings of the 38th International Conference on Software Engineering* (pp. 392-403).

Unanue, W., Gomez, M., Cortez, D., Oyanedel, J., & Mendiburo-Seguel, M. (2017). Revisiting the Link between Job Satisfaction and Life Satisfaction: The Role of Basic Psychological Needs. *Frontiers in Psychology,* 8, 680.

United Nations Development Program (2023). https://hdr.undp.org/data-center/documentation-and-downloads.

Weaver, C. N. (1978). Job satisfaction as a component of happiness among males and females. *Personnel Psychology*, 31, 831-840.

World Bank (2023). Where we Work, retrieved from https://www.worldbank.org/en/where-we-work.

World Happiness Report (2023). https://worldhappiness.report.

World Population Review (2023). https://worldpopulationreview.com/country-rankings/hdi-by-country.

Yin, R., Lepinteur, A., Clark, A. E., & D'Ambrosio, C. (2023). Life Satisfaction and the Human Development Index Across the World. *Journal of Cross-Cultural Psychology*, 54(2), 269–282.

Appendix 1

World Area	Job Satisfaction from the Mean	Happiness Index from the Mean	Human Development Index from the Mean
Africa	-4.5%	-25.75%	-28.66%
Central Asia	4.25%	-0.04%	-4.49%
East Asia & Pacific	-1.44%	4.68%	-1.11%
Europe	5.73%	12.17%	12.29%
Latin America & The Caribbean	-2.00%	4.52%	-3.32%
Middle East & North Africa	-5.62%	-18.30%	-4.86%
Northern America	11.00%	23.99%	20.37%
Russia	1.00%	-2.80%	-100.00%
South Asia	1.00%	-23.19%	-15.65%

Resilience During Times of Disruption:
The Role of Data Analytics in a Healthcare System

Elizabeth Weiss
pohanae@gmail.com

Thilini Ariyachandra
ariyachandrat@xavier.edu

Business Analytics & Information Systems Department
Xavier University
Cincinnati, Oh 45207, USA

Abstract

The COVID-19 pandemic has triggered an unprecedented transformation in society, disrupting daily life, businesses, and healthcare systems. This paper explores how organizations, particularly a healthcare institution, can demonstrate resilience amidst these challenges. The swift adaptation to changing conditions, implementation of new technologies, and strategic shifts in response to market changes underscores the importance of resilience and dynamic capabilities. In healthcare, the rapid adoption of telemedicine, reorganization of hospital operations, and implementation of new protocols highlight the need to quickly reconfigure when faced with dynamic change. Amidst such uncertainty, effective information processing, supported by data analytics, emerges as a critical survival tool. This paper applies the frameworks of dynamic capability and organizational resilience to examine how data analytics can aid organizations during major disruptions. By leveraging analytics, organizations can gain insights to inform strategies and maintain operations. To do so, it presents evidence from how a hospital system used analytics to enhance resilience and adaptability during the pandemic, providing insights into managing major disruptions.

Keywords: Data Analytics, Dynamic Capability, Resilience, Healthcare Analytics, Disruption, Pandemic

Recommended Citation: Weiss, E., Ariyachandra, T., (2024). Resilience During Times of Disruption: The Role of Data Analytics in a Healthcare System. *Journal of Information Systems Applied Research.* 17(1), pp 53-63, https://doi.org/10.62273/KYUF7967

Resilience During Times of Disruption: The Role of Data Analytics in a Healthcare System

Elizabeth Weiss and Thilini Ariyachandra

1. INTRODUCTION

The COVID-19 pandemic has precipitated a profound transformation across all strata of society, creating a pace of change that is unparalleled in contemporary history. The rapid spread of the virus and the subsequent lockdowns disrupted daily life, forced businesses to adapt or close, and put immense pressure on healthcare systems worldwide (Nicola et al., 2020). Businesses across sectors had to quickly adapt to new ways of working, quickly adopt new technologies and processes to support remote work and daily operations, while also addressing challenges related to communication, collaboration, and employee engagement (Bartik et al., 2020). They had to rethink strategies and operations to survive in a rapidly changing environment (Kramer & Kramer, 2020). Surviving disruption required a high degree of adaptability and flexibility, as well as the ability to leverage digital technologies effectively.

In addition to operational changes, many businesses had to rethink their strategies in response to changes in the market. For example, retailers had to shift to online sales as physical stores were closed, while manufacturers had to adjust their production lines to meet changes in demand or to produce essential supplies (Kaplinsky & Utecht, 2020). These strategic shifts required a strong sensing capability to identify changes in the market, as well as a seizing capability to respond to these changes effectively.

Healthcare organizations faced particularly severe challenges. Hospitals had to manage an influx of patients, protect their staff, and deal with shortages of essential supplies. Hospitals had to rapidly adapt their operations and care delivery models to respond to the crisis (Ranney et al., 2020). This required significant changes such as the use of telemedicine and more broadly telehealth, the reorganization of hospital wards, and the implementation of new protocols for infection control (Greenhalgh, Wherton, Shaw, & Morrison, 2020). These changes required a high degree of reconfiguring capability, as well as strong leadership and coordination.

In periods of such profound uncertainty, the ability to process information effectively becomes a critical survival tool for organizations. As posited by Galbraith (1974), organizations can mitigate uncertainty by augmenting their capacity to process information, which entails the collection and analysis of data to inform decision-making and action. Data analytics infrastructure and tools enable an organization to increase its information processing capacity in the face of uncertainty (Behl, Gaur, Pereira, Yadav, and Laker, 2022; Zhu, Song, Hazen, Lee, and Cegielski, 2018).

In the face of real time dynamic change and disruption such as a pandemic, analytics can be specifically useful. The frameworks of dynamic capability and organizational resilience provide theoretical lenses through which the role of enhanced information processing, supported by analytics, can be examined in the context of major disruptions, such as a pandemic. These frameworks offer a perspective on how to compete and survive when faced with rapid change. By gathering and analyzing data, organizations can gain insights that inform their strategies and actions, enabling them to adapt to changing circumstances and maintain their operations (Chen et al., 2012).

This paper uses the principles of dynamic capability and organizational resilience to explore how a hospital demonstrated resilience during the pandemic and continues to strive to attain agility and competitive advantage using analytics when faced with change. The paper hopes to provide valuable insights into how organizations can leverage analytics to enhance their resilience and adaptability in the face of major disruptions. The rest of the paper is organized as follows. First background information on data analytics, dynamic capability and organizational resilience is presented. Next the hospital system, its use of analytics infrastructure is anonymized in the description of the actual hospital that was investigated. This is followed by a discussion of how analytics enables dynamic capability and organizational resilience at the hospital. Finally, recommendations and conclusions are discussed.

2. DATA ANALYTICS

Data analytics can play a crucial role in enhancing organizational resilience and combating dynamic change in several ways. Data analytics is the process of examining, cleaning, transforming, and modeling data to discover useful information, draw conclusions, and support decision-making (Watson, 2014). It employs various techniques and methodologies drawn from fields such as mathematics, statistics, computer science, and information science, including signal processing, probability models, machine learning, statistical learning, data mining, database systems, and visualization, among others (Kelleher & Tierney, 2018). There are several types of data analytics, including descriptive, predictive, and prescriptive analytics. While descriptive analytics involves analyzing historical data to understand what has happened in the past, predictive analytics focuses on forecasting future outcomes based on historical data and analytics techniques. It uses statistical models and forecasting techniques, as well as machine learning techniques to predict the future (Shmueli & Koppius, 2011). Prescriptive analytics aims to provide advice based on the results of descriptive and predictive analytics using techniques such as optimization, simulation, decision tree, and complex event processing, among others (McAfee & Brynjolfsson, 2012).

The demand for data analytics across various industries today is unprecedented. The technology, finance, healthcare, and marketing sectors are leading the charge, using data analytics to drive decision-making, enhance customer experiences, and streamline operations. According to a report by Market Research Future, the data analytics market is projected to grow from 7.03 billion dollars in 2023 to 303.4 billion dollars in 2030 at a compound annual growth rate of 27.6 percent (Market Research Future, 2023). This surge reflects the growing reliance on data to understand market trends, predict consumer behaviors, and inform strategic decision-making. The rapid growth in big data and advancements in machine learning algorithms have accelerated the use of predictive analytics, a segment that is experiencing the sharpest increase in demand. According to Statista, the predictive analytics market is set to reach 21.5 billion dollars by 2025, up from 4.56 billion dollars in 2017 (Statista, 20200. Data analytics has had a tremendous impact in the healthcare sector. It is predicted that the healthcare analytics market will reach 50.5 billion dollars by 2024, up from 14 billion dollars in 2019 (MarketsandMarkets, 2020).

The unprecedented challenges the healthcare industry faced worldwide due to the COVID -19 pandemic led to a surge in the use of data analytics in that field. Analytics played a critical role in managing the crisis and strengthening healthcare systems. The rapid spread of COVID-19 necessitated quick and informed decision-making. Healthcare organizations used data analytics to predict virus spread, allocate resources, and identify vulnerable populations (McKee, Stuckler, Zeegers Paget, & Dorner, 2016). For example, the University of Virginia Health System developed a predictive model to anticipate COVID-19 case numbers and manage its resources accordingly (Rosenbaum, 2020). Public health surveillance was also enhanced through analytics by helping track and monitor the spread of the virus in real-time. Google and Apple, for example, developed a contact tracing app that used anonymized data to alert people if they had been in contact with a confirmed case (Ferretti et al., 2020). Finally, hospital systems relied on real-time data analytics to manage the rapid expansion of telemedicine that occurred during the pandemic. Healthcare providers used analytics to help monitor patients remotely, identify those at risk, and decide the appropriate level of care. A team at Johns Hopkins used data from wearable devices to predict early symptoms of COVID-19, allowing for quicker interventions (Radin et al., 2020). Data analytics emerged as a vital tool in the healthcare industry's response to the massive disruption caused by the pandemic. In the literature review, two lenses are described that provide more input on how analytics can be utilized to handle disruptive change.

3. LITERATURE REVIEW

Dynamic Capability View (DCV)

The dynamic capabilities theory, introduced by Teece, Pisano, and Shuen (1997), revolves around the concept that a firm's competitive advantage hinges on its ability to integrate, build, and reconfigure internal and external competences to address rapidly changing environments. The framework is rooted in the resource-based view (RBV) of the firm (Barney, Wright, and Ketchen 2001), but adds dynamism to explain how firms can achieve sustainable competitive advantage in volatile markets. While RBV suggests that unique, valuable, rare, and non-substitutable resources provide firms with competitive advantage, dynamic capabilities view (DCV) expands this by highlighting the

importance of the firm's ability to constantly adapt, integrate, and reconfigure its resources and capabilities according to the ever-changing business landscape.
Teece et al (1997) introduces three types of dynamic capabilities: sensing, seizing, and transforming. Each can play a significant role in helping a firm navigate the rapidly changing business environment. Sensing refers to the capability of a firm to identify and discover opportunities and threats in its environment. It involves the ability to continuously scan, learn, and interpret signals from the environment including changes in technology, competition, markets, and customer needs (Teece, 2007). This process requires a systematic approach to market research, trend analysis, and the detection of emerging patterns. Effective sensing also implies a good understanding of customers, their needs, and their behavior (Teece, 2016).

Once an opportunity is identified, seizing involves designing and refining business models to capture and exploit these opportunities. It involves aligning or reconfiguring the firm’s resources and capabilities with the opportunities sensed. This includes strategic decision-making processes, designing business models, coordinating resources, and defining organizational structures (Teece, 2007). Effective seizing often requires strong leadership and a supportive organizational culture that fosters innovation and risk-taking (Teece, 2012). An organization commits resources to the identified opportunities and builds new capabilities to gain a competitive advantage.

Transforming capabilities relate to a firm's ability to revamp its resource base to meet the requirements of a changing business environment. This could involve reconfiguring the firm's structure, resources, routines, or even its culture. Transformation often requires firms to unlearn obsolete practices and learn new ones, which is particularly challenging given the resistance to change that is often encountered within organizations (Teece, 2018). All three capabilities are deeply interconnected and need to be balanced for a firm to successfully navigate a dynamic business environment and achieve sustainable competitive advantage.

Dynamic capability view is a theoretical framework that has been widely applied in the field of information systems (Bozic, and Dimovski, 2019; Mikalef, Boura, Lekakos, and Krogstie 2019; Matarazzo, Penco, Profumo, and Quaglia, 2021). In the context of IS, dynamic capabilities can be seen as the firm's ability to leverage its IT resources and competencies to respond to changing environments. This includes the ability to develop new IS, adapt existing systems, and integrate systems to meet changing business needs (Wang & Ahmed, 2007). The DCV has been used to explain how firms can gain a competitive advantage through the strategic use of IS. For example, firms with strong dynamic capabilities can leverage IS to create new products or services, improve business processes, and enhance decision-making capabilities (Pavlou & El Sawy, 2006).

Analytics and DCV

Enabling dynamic capabilities within organizations by using data analytics, allows organizations to not only respond to changes in their environment but also shape those changes to their advantage. Sensing capabilities might involve the use of data analytics to detect market trends and customer preferences. For example, Netflix uses data analytics extensively to understand viewing patterns and preferences, which helps them to identify trends and opportunities for new content creation (Purkayastha, & Tangirala, 2013). Organizations that are effective at seizing opportunities use analytics to capture value from the opportunities they have sensed. For instance, Amazon uses its recommendation systems (powered by machine learning algorithms) to seize opportunities by providing personalized recommendations based on customer's browsing and purchasing history (Walker, 2015). Transformation or reconfiguration involves changing an organization's operations and strategies in response to sensed opportunities or threats. For example, IBM, facing declines in its hardware business, leveraged information systems to reconfigure its capabilities around services and software. They invested in analytics and cognitive computing, leading to the development of Watson, an AI system, redefining IBM's value proposition in the market (Mithas, Tafti, Bardhan, & Goh, 2012).

Organizational Resilience

Organizational resilience is the ability of an organization to anticipate, prepare for, respond and adapt to incremental change and sudden disruptions in order to survive and prosper (Duchek 2020). It embodies more than just the ability to bounce back from adversity; it includes an organization's capacity to anticipate and respond to changes, as well as its ability to transform itself when necessary (Lee, Vargo, & Seville, 2013). It is a proactive approach to identifying potential risks and vulnerabilities, as well as implementing strategies and measures to mitigate their impact. From a process-oriented view, organizational resilience can involve an anticipation stage, coping stage and adaptation stage.

The anticipation stage involves identifying potential threats, estimating their potential impact, and taking proactive measures to prevent or minimize potential harm (Weick & Sutcliffe, 2007). At this initial stage, the organization must be aware of its environment and keenly observe for changes that might indicate potential threats or opportunities (Weick & Sutcliffe, 2007). Once potential disruptions have been identified, preparation is required. This involves designing strategies and action plans to mitigate the identified risks. Preparation may involve creating contingency plans, strengthening existing structures, and investing in technologies or practices that enhance resilience (Linnenluecke, 2017). Moreover, resources must be made available or reserved for potential use during a crisis. These resources could be financial, such as emergency funds; physical, such as additional inventory or backup equipment; or human, such as extra personnel or expert teams. An organization's preparedness and resource availability are vital in determining its ability to respond to and recover from disruptions (Hosseini et al., 2016). Organizations that excel in this stage tend to have a culture of preparedness and a strong emphasis on continuous learning.

Coping is the organization's immediate response to a disruption. It involves managing the crisis to minimize harm and stabilize the situation as quickly as possible (Boin & McConnell, 2007). This stage involves accepting the situation, developing and implementing solutions, and leveraging social resources (Williams et al., 2017). Denial or delay in acceptance can exacerbate the impacts of the crisis. Hence, acknowledging the situation is critical for organizations to quickly and effectively mobilize their resources and enact their response plans (Horney et al., 2010). In addition, the utilization of social resources is a key aspect of this stage. These resources include relationships with stakeholders, collaborations with other organizations, and the support of the wider community. This often requires decisive leadership, effective communication, and rapid decision-making. Resilient organizations have systems and procedures in place that allow them to respond effectively in a crisis, such as emergency response teams and crisis management plans (Weick, Sutcliffe, & Obstfeld, 2005).

The adaptation stage involves learning from the disruption and making changes to avoid similar situations in the future or becoming better prepared for them (Lengnick-Hall et al., 2011). After managing a disruption, organizations must reflect on the crisis and its handling to extract valuable lessons. This process of reflection facilitates learning, enabling organizations to identify what worked well, what did not, and how they could improve their response in the future. This learning is critical to evolving the organization's practices, systems, and strategies to enhance its resilience (Linnenluecke, 2017). This could involve making changes to processes, systems, or structures, or it could involve a broader cultural or strategic shift (Hosseini et al., 2016). Adaptation also involves recognizing and redefining roles, power structures, and responsibilities within the organization. Leadership plays a crucial role in driving and managing change, but resilience also requires the engagement of employees at all levels. It involves creating a culture of resilience, where each member understands their role in managing disruptions and is empowered to act when necessary (Bhamra et al., 2011). Resilient organizations are not only able to bounce back from a disruption, but they are also able to learn and grow from it, emerging stronger than before (Sutcliffe & Vogus, 2003). The process of organizational resilience is an ongoing and cyclical process of anticipation, coping, and adaptation. Each stage is crucial and interdependent, forming the backbone of a resilient organization that can survive and thrive amidst disruptions.

Analytics and Organizational Resilience

Data analytics can provide businesses with meaningful insights that can be used to improve decision-making, optimize operations, improve customer service, and increase profitability, among other things. Therefore, analytics can enhance an organization's ability to anticipate and prepare for potential disruptions. By analyzing historical data and using predictive analytics, organizations can forecast potential risks and disruptions, and develop contingency plans accordingly (Araz, Choi, Olson, and Salman, 2020). For example, in supply chain management, predictive analytics can be used to anticipate potential disruptions and develop risk mitigation strategies (Ivanov, Dolgui, Sokolov, Ivanova, 2016).

Analytics can also enhance an organization's ability to respond to or cope with disruptions. Real-time data analytics can provide organizations with timely and accurate information during a crisis, enabling them to make informed decisions and respond effectively (Ransbotham, Kiron, & Prentice, 2015). For instance, during the COVID-19 pandemic, many organizations used real-time data analytics to monitor the impact of the pandemic on their operations and adjust their strategies accordingly (Verma & Gustafsson, 2020). Finally, data analytics can enhance an organization's ability to adapt to changes and disruptions. By analyzing data on the impact of disruptions, organizations can identify areas for improvement, learn from their experiences, and adapt their strategies and operations accordingly (Barton, Castillo, Petrie, & Wardell, 2019). This can enhance the organization's resilience and its ability to recover from disruptions.

The dynamic capability view and organizational resilience are closely related concepts that both focus on an organization's ability to adapt to changes and disruptions in the business environment. While organizational resilience primarily focuses on surviving disruptions and returning to a baseline, DCV focuses more on strategically adapting and growing amidst changing environments. The sensing, seizing, and transforming capabilities of DCV appear to align closely with the anticipation, coping, and adaptation stages of organizational resilience. However, DCV is more focused on strategic management and is a more proactive and forward-looking approach which emphasizes creating and shaping opportunities and not simply responding to the external environment. Organizational resilience can also be seen as an outcome or manifestation of dynamic capabilities. By effectively sensing, seizing, and transforming, organizations not only can gain a competitive advantage, but they can also become more resilient to disruptions (Vogus & Sutcliffe, 2007; Weick & Sutcliffe, 2007). By developing their dynamic capabilities, organizations can enhance their resilience, allowing them to effectively anticipate, respond, and adapt to disruptions and changes in their environment. Understanding and viewing the world through both lenses together can help an organization better navigate complex, volatile environments to either bounce back from adverse events or the strategically adapt to new opportunities.

4. ABC CHILDREN’S HOSPITAL SYSTEM

ABC children’s hospital, located in the Midwest, provides pediatric services to area children and is an institution that has gained national recognition for high quality care provided. ABC hospital has a strong commitment to technological innovation which is ingrained in its mission to adopt and implement new technology in pursuit of helping children in the region. The recent pandemic outbreak of the COVID-19 virus forced the hospital to re-evaluate the technology choices and methods employed for treating patients.

Prior to the pandemic, ABC was beginning to expand its IT infrastructure to a complex network of software and systems, professionals, and strategies aimed at harnessing and analyzing data for better decision-making and improved patient outcomes. As part of this strategy, it offered telehealth services in a limited capacity. As the pandemic began to spread, ABC significantly ramped up their ability to effectively deliver medical services via telehealth applications to limit patient and medical provider risk due to the COVID-19 outbreak. ABC was able to quickly connect its telehealth services to its rapidly evolving data and analytics infrastructure in order to give hospital personnel more information to make decision under rapidly changing conditions. The data and analytics architecture at ABC during the pandemic is presented in figure one. The overall architecture can be described in terms of several key components and ongoing processes.

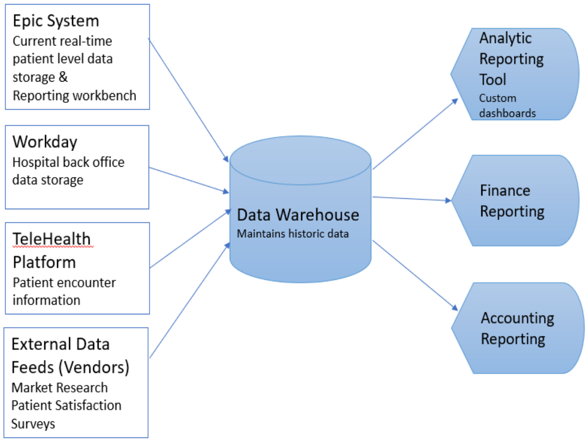

Epic System
Current real-time patient level data storage & Reporting workbench
Workday
Hospital back office data storage
TeleHealth Platform
Patient encounter information
External Data Feeds (Vendors)
Market Research
Patient Satisfaction Surveys
Data Warehouse
Maintains historic data
Analytic Reporting Tool
Custom dashboards
Finance Reporting
Accounting Reporting

Figure One: IT and Analytics Architecture

Core Systems: At the center of the architecture is the EPIC electronic medical records system, which is integral to nearly all operations and decision making. Individuals working with EPIC and information systems in clinical areas develop and maintain data required for all stakeholders. In addition, Workday is a back-office system used for hospital administrative support functions like accounting, HR, supply chain.

Together with EPIC, the Workday helps run and stores data on day to day operations.

Data Warehousing and Storage: Internal data stores are managed to facilitate long term reporting allowing for the analysis of several years of historical data. While internal data stores are currently in place, there is discussion of implementing cloud storage in the future.
Reporting and Analytics: An analytics platform enables real time reporting using data sourced from EPIC. However, these reports are typically limited to recent history to maintain speed for active patient care. For long-term reporting, different divisions such as Finance and Accounting use specific applications. All these applications are housed within a business intelligence department which was later rebranded as a centralized analytics division called Decision Analytics Division that is also responsible for developing and communication information to the executive team of the hospital.

Vendor Partnerships and Integration: ABC hospital collaborates with various analytics and market research companies, to obtain market forecasts, strategy analysis and patient satisfaction surveys. The challenge lies in centralizing the data from these diverse sources, and the hospital is actively working towards streamlining this process by investigating the possibility of integrating a centralized data analytics tool such as Power BI to disseminate information. Unfortunately, this process has challenges due to a lack of expertise and the need for significant manual data import.

Telehealth Implementation: The hospital has adopted telehealth as a strategic initiative by partnering with Zoom to integrate with the EPIC system. The hospital's strategy aims to record telehealth visits directly into patients' medical records, eliminating manual documentation. Strategy is being developed to integrate telehealth into its existing analytics strategy, which will be crucial in the post-COVID-19 era.

The adoption of telehealth in hospitals has surged in recent years, particularly in response to global events like the COVID-19 pandemic. It has been used to deliver a variety of healthcare services remotely, from routine checkups to specialist consultations, improving accessibility, efficiency, and potentially even the quality of care (Hollander & Carr, 2020). Patients were not only able to connect with their doctors for COVID-19 related concerns but also for managing chronic conditions, mental health services, and preventive health consultations.

Data analytics plays a critical role in telehealth by providing opportunities to improve patient outcomes, increase efficiency, and transform the way care is delivered (Wade, Eliott, & Hiller, 2014). On a basic level, data analytics in telehealth involves the collection, analysis, and interpretation of health-related data. With large volumes of data being collected through telehealth platforms, there is a need for robust data management systems to store, manage, and protect this data in a manner that ensures the confidentiality, integrity, and availability of patient data, in accordance with legal and ethical requirements (Hilty, Crawford, Teshima, Chan, Sunderji, Yellowlees, Kramer, O'Neill, Fore, Luo, Li, 2015). This data can include patient-reported outcomes, clinical measurements, and other data collected through telehealth technologies (Knapp, Harst, Hager, Schmitt, & Scheibe, 2021). These data can be analyzed to monitor patient progress, identify trends and patterns, and inform clinical decision-making. For example, predictive analytics can be used to identify patients who are at risk of hospital readmission or to predict disease progression, enabling early interventions that can improve patient outcomes (Shi et al., 2020). Similarly, telehealth data can be used to evaluate the effectiveness of different treatments, informing evidence-based practice (Wade, Eliott, & Hiller, 2014).

5. DISCUSSION

As disruption and change began to overwhelm society and the hospital system, ABC children's hospital began working with its existing IT and analytics infrastructure as well as ramping up new capabilities to face the onslaught of patients and care services required. It used its current analytics framework to analyze data and identify trends that might impact its operations. The ability to use EPIC and its analytics platform allowed the hospital to gain insights into patient behavior, market trends, and more. The sensing capability also extended to monitoring feedback from patient satisfaction surveys, market research data, and other sources to make informed decisions about hospital operations and strategies.

Once opportunities or threats were sensed to its existing operations, the hospital began shifting to implementing new technologies and practices based on the insights gained from their analytics. The rapid deployment of telehealth in response to COVID-19 demonstrates the hospital's ability to sense a need for a shift in healthcare delivery methods and how it seized this opportunity. Further, the incorporation of telehealth data to with existing EPIC data to gain comprehensive patient-level data for decision-making shows how the hospital is leveraging analytics infrastructure to seize opportunities for operational efficiency and financial management.

Transforming involves the adaptation or reconfiguration of the organization's operations and processes to suit the changes it has sensed and seized. ABC hospital demonstrates this capability by rebranding its business intelligence department to Decision Analytics Division, indicating a shift in its function to provide more operational and executive decision-making support. The potential shift to cloud storage for data and ongoing efforts to consolidate data reporting into centralized locations suggests ongoing efforts to transform their information management strategies.

At the time of investigating ABC Children’s hospital, it shows signs of being in the coping stage of its organization resilience journey. The hospital has experienced significant growth, expanding its clinical staff and service providing locations including the expansion of IT infrastructure indicating an adaptation to the increasing demands of data management and analytics. Having sensed future needs, they took a proactive approach to accommodate growth. The addition of new employees and the establishment of the Decision Analytics Division demonstrate the hospital's efforts to enhance its analytics capabilities. This indicates that the hospital is actively responding to the need for data-driven decision-making. The centralize data delivery and governance developed through the creation of the Decision Analytics Division helps provide organization-wide confidence in a single source of truthful answers. This step has enhanced the efficacy of data management, providing a reliable foundation for strategic decisions to be grounded in data analytics, moving away from gut-based decisions that were more prevalent in earlier practices. Consequently, the hospital landscape is seeing an evolving focus on analytics and the development of data-driven teams in different functional areas.

The hospital relies on EPIC as core system that manages its electronic medical records system. Most decision making at all levels revolve around this core system. This indicates a level of stability and integration in the hospital's analytics ecosystem. The hospital has formed partnerships with analytics companies to enhance its analytics capabilities. This suggests that the hospital is seeking external expertise and resources to adapt to the evolving analytics landscape in times of rapid change. Finally, having sensed the need for a new means of service delivery in light of the COVID-19 pandemic, the hospital recognized the importance of telehealth services. The rapid implementation of telehealth and the need for a robust analytics strategy for monitoring and improving telehealth options indicate the hospital's ability to respond to sudden and disruptive changes, which is characteristic of the coping stage.

However, it is worth mentioning that while they appear to be in the coping stage, there are elements of anticipation and adaptation present as well. They are anticipating future needs for cloud storage, more robust telehealth infrastructure, and improved data analytics. They have also shown elements of adaptation by taking steps to integrate, consolidate and streamline disparate systems and vendors with their own reporting solutions to an integrated data analytics and visualization solution. Challenges like the lack of automation and expertise with an analytics and visualization tool like Power BI suggest that the adaptation process is still early in its progress. Overall, these efforts showcase that ABC hospital is actively working on resilience by integrating technology, investing in analytics, adapting to changing healthcare practices (like telehealth), and maintaining a focus on patient experience and market trends. Further examination of analytics practices through the lenses of DCV and organizational resilience suggests the following recommendations that ABC could adopt.

(1) Implementation of a data analytics solution such as Microsoft Power BI to support hospital wide decision making. Conduct a thorough assessment of the organization's data infrastructure, security protocols, and gateway access to identify and address capacity limitations and roadblocks. Hire individuals with experience in data management and the analytics tool such as Power BI to effectively manage and optimize the data behind the tool, ensuring efficient data integration, transformation, and visualization.

(2) Expand the centralized data governance framework developed through the analytics division that clearly defines roles,

responsibilities, and sources for each type of information to deal with conflicting data in disparate sources. Implementation of data quality control measures, such as data validation and standardization, to ensure consistency and accuracy across reports and data sources.
(3) Development of a comprehensive training program and adoption plan for an analytics application such as Power BI that caters to different user groups, including leaders with varying computer skills. Promote data literacy and data-driven decision-making across the organization to continue to foster a culture of data-driven collaboration. Offer continuous learning opportunities, such as webinars or online resources, to keep staff updated with the latest features and best practices in Power BI.

(4) Strengthen deployment of surveys and feedback mechanisms to collect patient satisfaction data regarding telehealth services. Analyze provider feedback to identify pain points, challenges, and areas for improvement in telehealth implementation. Identify metrics to evaluate the impact of telehealth on provider efficiency, patient outcomes, and overall satisfaction.

(5) Collaborate with stakeholders to define key performance indicators (KPIs) and metrics relevant to telehealth success, such as appointment volume, cancellation/no-show rates, percentage of telehealth visits, patient satisfaction scores, and financial data.

(6) Develop a centralized dashboard that integrates data from various sources to provide a comprehensive view of telehealth performance. Utilize data visualization techniques to present the metrics in an intuitive and actionable format, allowing stakeholders to monitor trends, identify areas for improvement, and make informed decisions.

(7) Leverage analytics to conduct cost analyses, such as cost of illness analysis, cost-effectiveness analysis, or cost-benefit analysis, to evaluate the economic impacts of teleconsultations and telemedicine applications. Develop models that consider utilization levels, cost structures, and potential cost savings associated with telehealth implementation. Utilize data visualization and storytelling techniques to effectively communicate the economic benefits of telehealth to stakeholders, supporting strategic decision-making and resource allocation.

By implementing these solutions and leveraging analytics, ABC hospital can become more resilient and continue to hone its dynamic capabilities to meet changes and attain competitive advantage.

6. CONCLUSION

Faced with a major disruption such as a global pandemic, many organizations had to scramble and rapidly change to survive and conduct its day to day operations. Increasing information processing capacity through investment in information technology such as data analytics helped companies to combat change. Dynamics capability view and organizational resilience offers lenses to consider how analytics can improve and effectively deal with rapid change and disruption. At ABC Children's Hospital, implementation of telehealth integrated with decision support structures and analytics enabled that them to successfully navigate and survive the worst of the pandemic. Analytics enabled ABC to gain dynamic capability and cope with the disruption resulting from the pandemic. Dealing with disruption led to an estimated 8.8 percent loss of working hours worldwide which is equivalent to 255 million full-time jobs (International Labour Organization, 2021). Analytics can be one of the many tools that organizations can use to survive massive change as well as black swan events such as the Covid-19 pandemic.

7. REFERENCES

Araz, O. M., Choi, T. M., Olson, D. L., & Salman, F. S. (2020). Role of analytics for operational risk management in the era of big data. *Decision Sciences, 51*(6), 1320-1346. https://doi.org/10.1111/deci.12451

Barney, J., Wright, M., & Ketchen Jr, D. J. (2001). The resource-based view of the firm: Ten years after 1991. *Journal of management, 27*(6), 625-641. https://doi.org/10.1177/014920630102700601

Bartik, A. W., Bertrand, M., Cullen, Z. B., Glaeser, E. L., Luca, M., & Stanton, C. T. (2020). How are small businesses adjusting to COVID-19? Early evidence from a survey. *National Bureau of Economic Research*. https://doi.org/10.3386/w26989

Barton, C., Castillo, A., Petrie, J., & Wardell, D. (2019). Making data analytics work for you—instead of the other way around. *McKinsey Quarterly*.

https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/making-data-analytics-work-for-you-instead-of-the-other-way-around

Behl, A., Gaur, J., Pereira, V., Yadav, R., & Laker, B. (2022). Role of big data analytics capabilities to improve sustainable competitive advantage of MSME service firms during COVID-19–A multi-theoretical approach. *Journal of Business Research, 148*, 378-389. https://doi.org/10.1016/j.jbusres.2022.05.009

Bhamra, R., Dani, S., & Burnard, K. (2011). Resilience: the concept, a literature review and future directions. International Journal of Production Research, 49(18), 5375-5393. https://doi.org/10.1080/00207543.2011.563826

Boin, A., & McConnell, A. (2007). Preparing for critical infrastructure breakdowns: the limits of crisis management and the need for resilience. Journal of Contingencies and Crisis Management, 15(1), 50-59. https://doi.org/10.1111/j.1468-5973.2007.00504.x

Bozic, B., & Dimovski, V. (2019). Business intelligence and analytics use, innovation ambidexterity, and firm performance: A dynamic capabilities perspective. The Journal of Strategic Information Systems, 28(4), 101578. https://doi.org/10.1016/j.jsis.2019.101578

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. MIS quarterly, 1165-1188. https://doi.org/10.2307/41703503

Purkayastha, D., & Tangirala, V. K. (2013). Netflix: Leveraging Big Data to Predict Entertainment Hits (Case No. 913-006-1). IBS Center for Management Research.

Knapp, A., Harst, L., Hager, S., Schmitt, J., & Scheibe, M. (2021). Use of Patient-Reported Outcome Measures and Patient-Reported Experience Measures Within Evaluation Studies of Telemedicine Applications: Systematic Review. Journal of Medical Internet Research, 23(11), e30042. https://doi.org/10.2196/30042

Duchek, S. (2020). Organizational resilience: A capability-based conceptualization. *Business Research*, 13(1), 215-246. https://doi.org/10.1007/s40685-019-0085-7

Ferretti, L., Wymant, C., Kendall, M., Zhao, L., Nurtay, A., Abeler-Dörner, L., … & Fraser, C. (2020). Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. Science, 368(6491). https://doi.org/10.1126/science.abb6936

Galbraith, J. R. (1974). Organization design: An information processing view. Interfaces, 4(3), 28-36. https://doi.org/10.1287/inte.4.3.28

Greenhalgh, T., Wherton, J., Shaw, S., & Morrison, C. (2020). Video consultations for covid-19. Bmj, 368. https://doi.org/10.1136/bmj.m998

Hilty, D. M., Crawford, A., Teshima, J., Chan, S., Sunderji, N., Yellowlees, P. M., Kramer, G., O'Neill, P., Fore, C., Luo, J., Li, S. (2015). A framework for telepsychiatric training and e-health: Competency-based education, evaluation and implications. International Review of Psychiatry, 27(6), 569-592. https://doi.org/10.3109/09540261.2015.1091292

Hollander, J. E., & Carr, B. G. (2020). Virtually perfect? Telemedicine for Covid-19. New England Journal of Medicine, 382(18), 1679-1681. https://doi.org/10.1056/nejmp2003539

Horney, N., Pasmore, B., & O'Shea, T. (2010). Leadership agility: A business imperative for a VUCA world. People & Strategy, 33(4), 32. https://www.researchgate.net/profile/Brian-Pasmore/publication/228626764_Leadership_Agility_A_Business_Imperative_for_a_VUCA_World/links/5a0d8f5caca272b0b5d4a4d9/Leadership-Agility-A-Business-Imperative-for-a-VUCA-World.pdf

Hosseini, S., Barker, K., & Ramirez-Marquez, J. E. (2016). A review of definitions and measures of system resilience. Reliability Engineering & System Safety, 145, 47-61. https://doi.org/10.1016/j.ress.2015.08.006

International Labour Organization. (2021). ILO Monitor: COVID-19 and the world of work. Seventh edition Updated estimates and analysis. International Labour Organization. https://www.ilo.org/wcmsp5/groups/public/—dgreports/—dcomm/documents/briefingnote/wcms_767028.pdf

Ivanov, D., Dolgui, A., Sokolov, B., Ivanova, M. (2016). A dynamic model and an algorithm for short-term supply chain scheduling in the smart factory industry 4.0. International Journal of Production Research, 54(2), 386-402.

https://doi.org/10.1080/00207543.2014.999958

Kelleher, J. D., & Tierney, B. (2018). Data Science. MIT Press Essential Knowledge series. The MIT Press. ISBN: 9780262535434

Kramer, A., & Kramer, M. (2020). The potential impact of the Covid-19 pandemic on occupational status, work from home, and occupational mobility. Journal of Vocational Behavior, 119, 103442. https://doi.org/10.1016/j.jvb.2020.103442

Lee, A. V., Vargo, J., & Seville, E. (2013). Developing a tool to measure and compare organizations' resilience. Natural Hazards Review, 14(1), 29-41. https://doi.org/10.1061/(ASCE)NH.1527-6996.0000075

Lengnick-Hall, C. A., Beck, T. E., & Lengnick-Hall, M. L. (2011). Developing a capacity for organizational resilience through strategic human resource management. Human Resource Management Review, 21(3), 243–255. https://doi.org/10.1016/j.hrmr.2010.07.001

Linnenluecke, M. K. (2017). Resilience in business and management research: A review of influential publications and a research agenda. International Journal of Management Reviews, 19(1), 4-30. https://doi.org/10.1111/ijmr.12076

Walker, R. (2015). From Big Data to Big Profits: Success with Data and Analytics. Oxford University Press. ISBN: 9780199378326

Matarazzo, M., Penco, L., Profumo, G., & Quaglia, R. (2021). Digital transformation and customer value creation in Made in Italy SMEs: A dynamic capabilities perspective. Journal of Business Research, 123, 642-656. https://doi.org/10.1016/j.jbusres.2020.10.033

MarketsandMarkets. (2020). Healthcare Analytics Market by Type (Predictive, Prescriptive), Component (Hardware, Software, and Services), Delivery Mode (Cloud), Application (Clinical, RCM, Claims, Fraud, Risk, PHM), End user (Payer, Provider) - Global Forecast to 2024. MarketsandMarkets. https://www.marketsandmarkets.com/Market-Reports/healthcare-data-analytics-market-905.html

McAfee, A., & Brynjolfsson, E. (2012). Big data: the management revolution. Harvard Business Review, 90(10), 60-68. https://hbr.org/2012/10/big-data-the-management-revolution

McKee, M., Stuckler, D., Zeegers Paget, D., & Dorner, T. (2016). The Vienna Declaration on Public Health. European journal of public health, 26(6), 897–898. https://doi.org/10.1093/eurpub/ckw194

Mikalef, P., Boura, M., Lekakos, G., & Krogstie, J. (2019). Big data analytics capabilities and innovation: The mediating role of dynamic capabilities and moderating effect of the environment. British Journal of Management, 30(2), 272-298. https://doi.org/10.1111/1467-8551.12343

Mithas, S., Tafti, A., Bardhan, I., & Goh, J. M. (2012). Information technology and firm profitability: mechanisms and empirical evidence. MIS Quarterly, 36(1), 205-224. https://doi.org/10.2307/41410414

Pavlou, P. A., & El Sawy, O. A. (2006). From IT leveraging competence to competitive advantage in turbulent environments: The case of new product development. Information Systems Research, 17(3), 198-227. https://doi.org/10.1287/isre.1060.0094

Radin, J. M., Wineinger, N. E., Topol, E. J., & Steinhubl, S. R. (2020). Harnessing wearable device data to improve state-level real-time surveillance of influenza-like illness in the USA: a population-based study. The Lancet Digital Health, 2(2), e85-e93. https://doi.org/10.1016/S2589-7500(19)30222-5

Wade, V., Eliott, J., & Hiller, J. (2014). Clinician acceptance is the key factor for sustainable telehealth services. Qualitative Health Research, 24(5), 682-694. https://doi.org/10.1177/1049732314528809

Watson, H. J. (2014). Tutorial: Big data analytics: Concepts, technologies, and applications. Communications of the Association for Information Systems, 34(1), 65. https://doi.org/10.17705/1CAIS.03462

Zhu, S., Song, J., Hazen, B. T., Lee, K., & Cegielski, C. (2018). How supply chain analytics enables operational supply chain transparency: An organizational information processing theory perspective. International Journal of Physical Distribution & Logistics Management, 48(1), 47-68. https://doi.org/10.1108/IJPDLM-11-2017-0341

Using Textual Analytics to Process Information Overload of Cyber Security Subreddits

Stephanie Omakwu
so05640@georgiasouthern.edu

Hayden Wimmer
hwimmer@georgiasouthern.edu

Department of Information Technology
Georgia Southern University
Statesboro, GA 30460, USA

Carl M Rebman, Jr.
carlr@sandiego.edu
Knauss School of Business
Department of Supply Chain, Operations, and Information Systems
University of San Diego
San Diego, CA 92110, USA

Abstract

Increases in digitalization have made it possible to track and measure every click, every payment, every message, and almost everyone's daily thoughts. Companies are extremely interested in the robustness of this data, specifically regarding understanding the sentiment of consumers. Yet the amount of information being produced and processed is quite staggering causing information overload. As such, companies tend to fall into analysis paralysis which can result in missing important insights that could help their business. The goal of this study is to analyze and categorize the top posts on multiple hacking subreddits to determine the most discussed topics and to examine the sentiment of these posts expressed by the users. We began by scraping data, specifically the title, ID, score, comments, and URL for each top post from multiple hacking subreddit communities. We then used the Natural Language Toolkit (NLTK) to perform the data preprocessing techniques for an effective analytic process and bias-free results. The results of the testing allowed us to filter through the posts and determine whether sentiment was positive, negative, or neutral. In the case of the hacking subreddits, many of the posts were of a neutral opinion. This study aims to provide a contribution by utilizing Natural Language Processing methods Topic Modeling such as Term Frequency Inverse Document Frequency, Latent Semantic Analysis (LSA) algorithm, and Sentiment Analysis to gather and synthesize cybersecurity data.

Keywords: information overload, text analytics, sentiment analysis, LSA, term frequency, NLP

Recommended Citation: Omakwu, S., Wimmer, H., Rebman Jr., C., (2024). Using Textual Analytics to Process Information Overload of Cyber Security Subreddits. *Journal of Information Systems Applied Research.* 17(1), pp 64-74, https://doi.org/10.62273/AJJR5232

Using Textual Analytics to Process Information Overload of Cyber Security Subreddits

Stephanie Omakwu, Hayden Wimmer and Carl M. Rebman, Jr.

1. INTRODUCTION

More and more data is generated as a result of the rising digitalization of our daily lives. This data comes from a highly diverse range of sources. Approximately 90% of the data generated by organizations is categorized as unstructured data. The size of the data produced daily is astonishing and more than any human and many computers could handle. Most of the data contains a significant amount of text that is written in natural language that includes slang and sarcasm. This poses a number of difficult issues for extraction and synthesis during analysis (Carnot et al., 2020).

Unstructured data comes in various forms which does not usually fit neatly into traditional data models. It is also a challenge to store and manage unstructured data because it also does not easily fit into relational databases or spreadsheets like Excel. These challenges have historically hindered analysis and search efforts, thus rendering it less useful for organizations. However, with the rise of modern business intelligence tools and platforms the landscape has changed. These innovations now enable companies and organizations the ability to manage and conduct analysis of unstructured data (Jain et al., 2020).

This research uses computer-based techniques like Natural Language Processing (NLP) to study cybersecurity data. Specifically, this study uses Topic Modeling, which assists in finding out what subjects are talked about the most, and Sentiment Analysis, which helps in understanding how people feel about certain discussions. It is important for businesses to keep up with the latest cybersecurity practices, especially in our interconnected world. This way, companies can build a strong defense against changing cyber risks and ensure the safety of their data. The goal of the study is to analyze and categorize the top posts on multiple hacking subreddits to determine the most discussed topics and to examine the sentiment of the top posts expressed by users in hacking subreddits.

2. BACKGROUND

Over the past twenty years computational linguistics, also known as the rule-based modeling of human language, has developed into an intriguing field of study as well as a useful technology that is increasingly being implemented into consumer products such as Apple's Siri and Skype Translator. These advancements were made possible by four major factors: (i) a significant increase in computing power; (ii) the availability of extremely large amounts of linguistic data; (iii) the development of highly effective machine learning (ML) methods; and (iv) a much deeper understanding of the structure of human language and how it is used in social contexts (Hirschberg & Manning, 2015).

When computational linguistics is combined with statistical, machine learning, and deep learning models it forms the basis of 'natural language processing' or NLP. NLP is considered to be part of artificial intelligence (AI) technologies that is concerned with providing computers the capacity to comprehend written and spoken words. The goal of the technology is to provide computers the ability to comprehend human language in the form of text or speech data and to "understand" its full meaning, including the speaker's or writer's intention and sentiment (IBM, 2023). The very first step in the NLP project model creation is to execute text preprocessing which is transforming text into a clean and consistent format. Preprocessing procedures include the following:

Removal of Punctuation
This involves eliminating all of the text's punctuation using the Python's string library which comes pre-loaded with a number of punctuation marks, including '!"#$%&'()*+,-./:;?@[]_'|'. Removal of punctuation will aid in treating each text equally (Rahimi & Homayounpour, 2023).

Removal of Stopwords
Stop words are the words that are often excluded before processing natural language. These are the most prevalent words in any language (such as articles, prepositions,

pronouns, conjunctions, etc.), yet they do not significantly add to the text's content. The words "the," "a," "an," "so," and "what" are a few stop words in English. Any human language has an abundance of stop words(Giri & Banerjee, 2023). By removing these words, we can make our text more focused on the key information by eliminating the low-level information. A list of terms that are regarded as stop words in the English language is included in the NLTK library. Removing Stop words decreases the dataset size and, as a result, the training time because there are fewer tokens to be trained.

Tokenization

Tokenization is an easy procedure that turns raw data into a meaningful data string, it involves splitting sentences into smaller units referred to as tokens. For example, "Computers understand language with NLP." when broken into tokens will be "Computers", "understand", "language", "with", "NLP", ".". Tokenization is most known for its applications in cybersecurity and the development of non-fungible token (NFT) a specific kind of digital asset that uses blockchain technology to signify ownership and authenticity of a particular object or piece of content, because NFTs are non-fungible, each token is unique and cannot be traded for another token of the same kind. Tokenization also plays a significant role in the NLP process as it is a technique used to break down phrases and paragraphs into simpler language-assignable elements (Choo & Kim, 2023).

Stemming

By simply combining the suffix with the basic root of the word, stemming reduces a word to its stem while maintaining the word's semantic meaning. Stemming disregards grammatical conventions of the language (Pramana et al., 2022). For example, "producing, production" is stemmed into its root word "produc."

Lemmatization

Lemmatization, in contrast to stemming, must always result in a real word form. Lemmatization reduces word variants by removing inflectional endings and returning the term to its base or dictionary form (Pramana et al., 2022). For example, the word "paraphrasing" is lemmatized to its based form "paraphrase".

3. LITERATURE REVIEW

This section discusses relevant literature that was used as the basis of our methodology. Specifically, these studies address our use and selection of classifier models, sentiment analysis techniques such as Latent Semantic Analysis, Industry 4.0, fake news, social media, algorithms, and natural language processing methods.

Kumar and Subba (2020) emphasized how sentiment analysis frameworks have high energy and processing requirements, which restricts their ability to be deployed in real time. They found the frameworks performed poorly when applied to corpora of textual data that contain emoticons and other unusual texts. Kumar and Subba (2020) proposed a sentimental analysis framework based on Support Vector Machine (SVM). Real-time sentiment analysis of the text documents was performed using the trained SVM classifier model and the deployment phase where Real-time sentiment analysis of the text documents is performed using the trained SVM classifier model. The proposed sentimental analysis framework outperformed previous similar frameworks proposed in the literature, according to experimental results on the Amazon electronics item review dataset and the IMDB movie review data corpus.

Subba and Gupta (2021) presented a brand-new host intrusion detection system (HIDS) framework based on truncated singular value decomposition (SVD) and tfidfvectorizer for the real-time detection of unusual system operations. The system calls trace files that are entered into the proposed HIDS framework, which converts them into n-gram feature vector representational models and calculates the tfidf values via tfidfvectorizer. Then, depending on their tfidf values, the modified n-gram feature vectors are dimension reduced using truncated SVD. According to experimental findings using the benchmark datasets ADFA-LD and ADFA-WD, the proposed HIDS framework efficiently and effectively detects abnormal system operations with little processing cost.

Wagire et al. (2020) identified that Industry 4.0 experienced a rapid increase in research articles while being a relatively new and developing subject of study. The goal of their research was to identify "patterns of research" in the area of Industry 4.0. The enormous corpus of 503 academic paper abstracts published in various journals and conference proceedings is reviewed and knowledge is extracted using Latent Semantic Analysis (LSA). The technique used retrieves several latent elements that define the newly developing study pattern. High-loaded papers are subjected to cross-loading analysis to determine the semantic relationship between research fields and themes. The findings of the

LSA reveal 100 study topics and 13 major research fields. The survey identifies "new business model" and "smart factory" as the two most important research areas. A taxonomy is created that includes five industrial theme categories. 4.0 field.

The most significant and well-known unsupervised techniques in automatic voice recognition are Maximal Marginal Relevance (MMR) and Latent Semantic Analysis (LSA). The goal of Ramezani et al. (2023) was to determine how well the two unsupervised algorithms described above performed when transcribing summaries of Persian broadcast news. The 58 news documents in the corpus used in this study were derived from the manual transcription of Persian broadcast news over a period of more than 15 hours. Over the course of 45 days, it was gathered by four native transcribers from the news on three radio channels and four TV channels. Over 115,000 words and 7,000 sentences make up the corpus. The findings indicate that MMR performs better than LSA in query-based broadcast news summarization while LSA outperforms MMR in generic summarization.

Mayopu et al. (2023) showed that fake news is constantly created to mislead readers, spreads quickly, and has a significant negative impact on human civilization through social media. This project intended to create an efficient method that combines latent semantic analysis (LSA) and natural language processing (NLP) utilizing singular value decomposition (SVD) techniques to assist social scientists in analyzing fake news and identifying its precise components. In order to construct a summary by recognizing a latent or hidden semantic structure, latent semantic analysis (LSA) was used to extract relevant sentences from an input material. The LSA approach models gathered documents as a term-document matrix (TDM) and employs singular value decomposition (SVD) to derive concepts from collected documents. The efficacy of Mayopu et al. (2023) techniques was illustrated using a genuine scenario from the 2016 United States presidential election campaign which displayed five concepts and were taken from the LSA and are indicative of political fake news during an election.

Numerous laws are made by regulatory authorities and must be adhered to, as a result, finding regulatory non-compliance involves complex compliance requirements and time-consuming procedures. Information Technology (IT) offers a wide range of governance, management, and security frameworks to let firms conduct their processes at a much more advanced level. Huyut et al. (2022) presented a method based on Latent Semantic Analysis (LSA) to produce a particular relatedness correlation map to have an objective and statistical relationship map. A relatedness map between a banking regulation and a best practice was made by (Huyut et al., 2022), under regard to the 1202 actions under Control Objectives for Information Technologies (Cobit 2019).

Huyut et al. (2022) examined 224 statements of this regulation and they used multi-criteria decision-making (MCDM) analysis techniques to support their LSA results. Fuzzy Analytics Hierarchy Process (FAHP) was used to prioritize their criteria, and Weighted Aggregated Sum Product Assessment Method (WASPAS) was used to compare the results of similarity tests between pairs of regulations and Cobit activities.

Chiny et al. (2022) performed an exploratory study on information collected from Flixable, a search engine that displays Netflix material. The TF-IDF and Cosine similarity algorithms, which are popular models in Natural Language Processing (NLP), were also employed to construct a recommendation system used to highlight crucial information about the material offered on this site through evaluation of a dataset of 7,787 unique records. A crucial step of the analysis was the Word Cloud library, a component of an NLP software stack, used to calculate the word clouds' densities. In order to examine the similarities between the titles and descriptions of the Netflix programs, Chiny et al. (2022) applied the TF-IDF and Cosine Similarity algorithms to them. Countries including the United States, India, and the United Kingdom stood out to be the nations with the greatest availability of Netflix material based on investigation emphasized data on the distribution of programs broadcast by genre (69% movies and 31% TV shows). Their study also highlighted the word-by-word analysis of the phrases used in the televised programs.

Prasanth et al. (2022) outlined the methodology used by team CENTamil to identify offensive comments in Tamil. The goal of this investigation was to determine whether a particular comment contains offensive language. To build feature vectors, the author utilized TF-IDF with char-wb analyzers and the Random Kitchen Sink (RKS) algorithm. The YouTube comments in the datasets were first preprocessed, and the preprocessed texts were then transformed into vectors. To make the data clean, noise was removed at the preprocessing stage. For classification of the YouTube comments, a Support Vector Machine (SVM) classifier with a polynomial kernel was employed. Using this technique, Prasanth et al. (2022) was able to rank first with a f1-score of 0.32 for the Tamil dataset and seventh with a f1-score of 0.25 for the Tamil-English dataset, respectively.

Liang and Niu (2022) hypothesized that text classification, a method used in sentiment analysis, intelligent recommendation systems, and intelligent question-and-answer systems, could automatically categorize and label text in accordance with predefined rules. The bidirectional LSTM input structure and the TF-IDF method were changed in this paper. The author specifically updated the TF-IDF calculation and used a sliding window to parse the words. This study trained the neural network, validated it, and tested it using the Sohu news dataset. In an 8:1:1 ratio, the dataset was split into training sets, validation sets, and testing sets. The text features were extracted using a combination of bidirectional LSTM and Text-CNN for classification prediction, with the word2vec vector serving as the word embedding layer. While Text-CNN can extract local crucial features at the sentence level, Bidirectional LSTM treats texts as sequences to understand information. Good precision was attained as a result.

4. METHOD

It is important for businesses to keep up with the latest cybersecurity practices, especially in an interconnected world. In doing so, companies can build a strong defense against changing cyber risks and ensure the safety of their data.

Using Python's PRAW (python Reddit API wrapper) module, which enables Reddit API using Python scripts, we first began by scraping data, specifically the top posts title, ID, score, comments, and URL for each post, from multiple hacking subreddit communities into a.csv file. We used the Natural Language Toolkit (NLTK) to perform the following data preprocessing techniques for an effective analytic process and bias-free results: punctuation removal, tokenization, stop word removal, stemming, and lemmatization.

On completing the data preprocessing, we further analyzed the data by performing the Term Frequency Inverse Document Frequency (TFIDF) Vectorization Using Tfidfvectorizer, Latent Semantic Analysis (LSA) and Sentiment Analysis. The hardware machine that is enabling the implementation of this experiment is an 11th Gen Intel(R) Core (TM) i7-1165G7 @ 2.80GHz 2.80 GHz processor, 16.0 GB (15.7 GB usable) RAM, Windows 10 Home 64-bit operating system, x64-based processor and Python as the language used for analysis.

The TFIDF results for all subreddits indicate that "Hack", "Hacking", "Ethical", and "Comptia" based on their TFIDF scores were the most significant words. With the LSA results identifying the 3 most discussed topics in the subreddits as certifications, techniques on how to do ethical hacking, and training tools. Lastly the base Polarity indicated that there were more Neutral than positive and a little fewer negative sentiment from Top posts in the subreddits.

System Architecture

Figure 1 displays the system architecture on which the project was built on. Visual studio code was utilized as an enabling environment for all phases of this project.

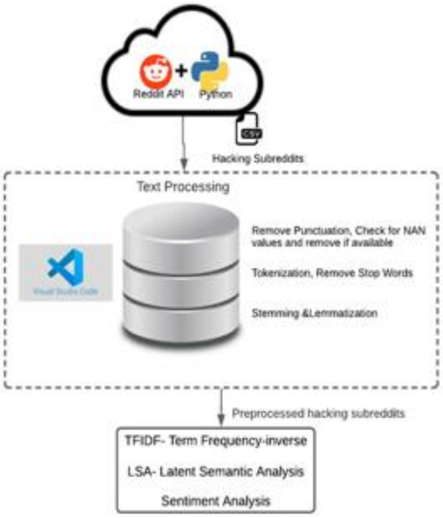

Reddit API
+
Python
CSV
Hacking Subreddits
Text Processing
Visual Studio Code
Remove Punctuation, Check for NAN values and remove if available
Tokenization, Remove Stop Words
Stemming &Lemmatization
Preprocessed hacking subreddits
TFIDF- Term Frequency-inverse
LSA- Latent Semantic Analysis
Sentiment Analysis

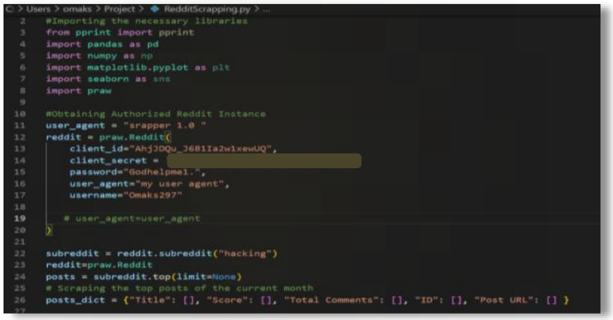
Figure 4: System Architecture

Data Selection and Preprocessing

Data for this study was pulled from Reddit which is a popular platform where experts and students can share knowledge and experiences, there are numerous groups, thousands of communities within it called subreddits where users can converse about articles on predetermined subjects whether you enjoy technology, sports, breaking news, TV fan theories, or a never-ending stream of the prettiest animals online.

Millions of individuals post, vote, and comment every day in communities/subreddits catered to their interests across the globe, with reddit having a daily active users of over 57 Million users, over a 100k active subreddits/communities and over 13 billion posts and comments (Dive Into Anything, 2023). All Top posts from their creation date were extracted from hacking subreddits dedicated to hacking content into .csv files, the subreddits were selected based on popularity.

Attributes such as title, score, id, total comments, and post url were extracted from r/ethicalhacking (389k) with 1001 top posts, r/hackintutorials (228k) with 997 top posts and r/hacking (2.6miilon) with 979 top posts. We started off by scraping data specifically the top posts title, ID, score, comments, and URL for each post from multiple hacking subreddit communities into a .csv file using python's PRAW (python Reddit API wrapper) module which allows Reddit API through python scripts(Boe., 2023).

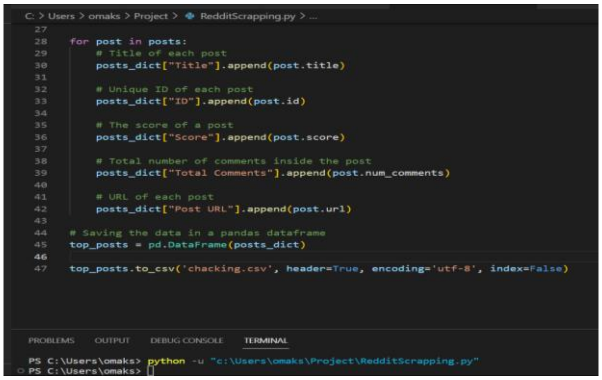


```python
#Importing the necessary libraries
from pprint import pprint
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import praw

#Obtaining Authorized Reddit Instance
user_agent = "srapper 1.0 "
reddit = praw.Reddit(
    client_id="AhjJDQu_J6B1Ia2w1xewUQ",
    client_secret = [redacted]
    password="Godhelpme1.",
    user_agent="my user agent",
    username="Omaks297"

   # user_agent=user_agent
)

subreddit = reddit.subreddit("hacking")
reddit=praw.Reddit
posts = subreddit.top(limit=None)
# Scraping the top posts of the current month
posts_dict = {"Title": [], "Score": [], "Total Comments": [], "ID": [], "Post URL": [] }
```

Figure 5*:* Creating PRAW Instance

To enable data extraction from reddit the PRAW library was installed, then we created a reddit developer account under which we created a Reddit app which generated the client_Id, secrete and user agent values which we used to connect to Reddit using Python. To give us access to perform all possible actions on reddit a PRAW authorized instance was created.

To improve the consistency, accuracy, and reliability of our dataset, for an efficient analysis process, and non-bias result Natural Language toolkit (NLTK) a leading platform for building Python programs to work with human language data was utilized to perform data preprocessing techniques such as removal of punctuations, removal of stop words, tokenization, stemming, and lemmatization. NLTK provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum("Natural Language Toolkit," 2023).



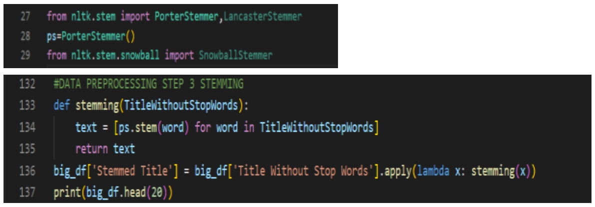
```python

for post in posts:
    # Title of each post
    posts_dict["Title"].append(post.title)

    # Unique ID of each post
    posts_dict["ID"].append(post.id)

    # The score of a post
    posts_dict["Score"].append(post.score)

    # Total number of comments inside the post
    posts_dict["Total Comments"].append(post.num_comments)

    # URL of each post
    posts_dict["Post URL"].append(post.url)

# Saving the data in a pandas dataframe
top_posts = pd.DataFrame(posts_dict)

top_posts.to_csv('chacking.csv', header=True, encoding='utf-8', index=False)
```

```
PS C:\Users\omaks> python -u "c:\Users\omaks\Project\RedditScrapping.py"
PS C:\Users\omaks> 
```

Figure 6: Scraping the Top posts from a subreddit

For data preprocessing and cleaning we began by Removing Punctuation using a customized created method in collaboration with a for loop.

Data Preprocessing–Removing Punctuation
For easier interpretation the dataset was split from sentences into smaller units called token which can be more easily assigned meaning. The the word_tokenize function under the NLTK.tokenize package was utilized for the tokenization procedure.

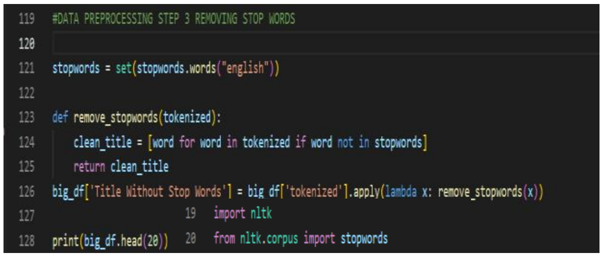


```python
from nltk.stem import PorterStemmer,LancasterStemmer
ps=PorterStemmer()
from nltk.stem.snowball import SnowballStemmer
```

```python
#DATA PREPROCESSING STEP 3 STEMMING
def stemming(TitleWithoutStopWords):
    text = [ps.stem(word) for word in TitleWithoutStopWords]
    return text
big_df['Stemmed Title'] = big_df['Title Without Stop Words'].apply(lambda x: stemming(x))
print(big_df.head(20))
```

Figure 7: Removal of Punctuation

Data Preprocessing - Removing Stop words
To create room for the ability to focus more on the most important information in the dataset we removed Stopwords using a customized created method with a for loop.



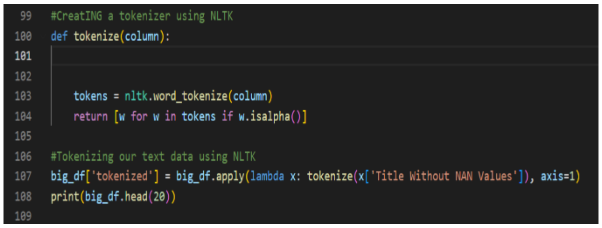
```python
#DATA PREPROCESSING STEP 3 REMOVING STOP WORDS


stopwords = set(stopwords.words("english"))

def remove_stopwords(tokenized):
    clean_title = [word for word in tokenized if word not in stopwords]
    return clean_title
big_df['Title Without Stop Words'] = big_df['tokenized'].apply(lambda x: remove_stopwords(x))

print(big_df.head(20))
```

```python
import nltk
from nltk.corpus import stopwords
```

Figure 5: Removal of Stop Words

Data Preprocessing - Tokenization

For easier interpretation I split the sentences into tokens using the nltk.word_tokeinze function.

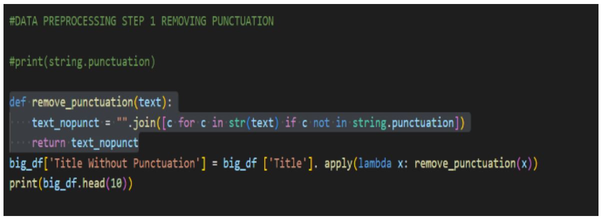


```
#CreatING a tokenizer using NLTK
def tokenize(column):


    tokens = nltk.word_tokenize(column)
    return [w for w in tokens if w.isalpha()]

#Tokenizing our text data using NLTK
big_df['tokenized'] = big_df.apply(lambda x: tokenize(x['Title Without NAN Values']), axis=1)
print(big_df.head(20))
```

Figure 6: Tokenization

Data Preprocessing - Stemming

Another data preprocessing technique that was utilized is Stemming which is a natural language processing technique used to extract the base word form of the words by removing affixes from them, it is just like cutting down branches of a tree to its stems. The PorterStemmer and SnowballStemm algorithm were compared within the NLTK package, but they approximately generated the same output.

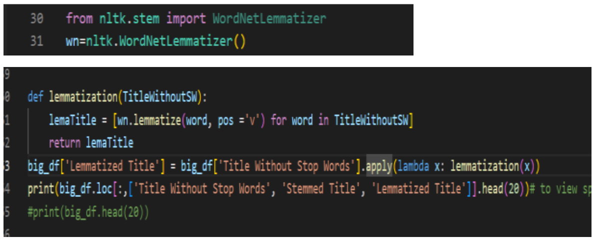


```
#DATA PREPROCESSING STEP 1 REMOVING PUNCTUATION


#print(string.punctuation)


def remove_punctuation(text):
    text_nopunct = "".join([c for c in str(text) if c not in string.punctuation])
    return text_nopunct
big_df['Title Without Punctuation'] = big_df ['Title']. apply(lambda x: remove_punctuation(x))
print(big_df.head(10))
```

Figure 7: Stemming

Data Preprocessing – Lemmatization

The last data preprocessing technique used is Lemmatization. This method was used to normalize the data, it is a similar technique to stemming but with the focus on finding a valid word. The output after lemmatization is called a lemma which is the root word. To implement the lemmatization technique, the WordNetLemmatizer class provided by the NLTK package was utilized.



```
from nltk.stem import WordNetLemmatizer
wn=nltk.WordNetLemmatizer()
```

```

def lemmatization(TitleWithoutSW):
    lemaTitle = [wn.lemmatize(word, pos ='v') for word in TitleWithoutSW]
    return lemaTitle
big_df['Lemmatized Title'] = big_df['Title Without Stop Words'].apply(lambda x: lemmatization(x))
print(big_df.loc[:,['Title Without Stop Words', 'Stemmed Title', 'Lemmatized Title']].head(20))# to view sp
#print(big_df.head(20))
```

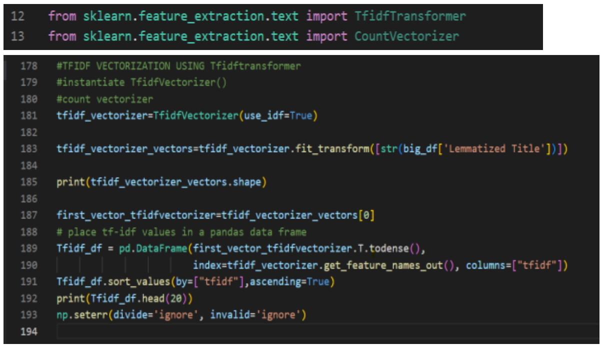
Figure 8: Lemmatization

Term Frequency Inverse Document Frequency (TFIDF)

After cleaning the dataset, the first text analysis we proceeded to perform the Term Frequency Inverse Document Frequency (TFIDF) Vectorization by importing scikit-learn python library and using sklearn.feature_extraction.text.TfidfVectorizer to convert the collection of raw documents to a matrix of TF-IDF features. Machine learning algorithms often use numerical data, so when dealing with textual data or any (NLP) task, the data first needs to be converted to a vector of numerical data by a process known as vectorization(Kumar & Subba, 2020).

TF-IDF vectorization involves calculating the TF-IDF score for every word in your corpus relative to that document and then putting that information into a vector. TF-IDF can be broken down into two parts TF (term frequency) and IDF (inverse document frequency). Term frequency works by looking at the frequency of a particular term you are concerned with relative to the document. Inverse document frequency looks at how common (or uncommon) a word is amongst the corpus. IDF also helps with filtering stopwords like "of", "as", "the", etc. since they appear frequently in an English corpus. Thus, by taking inverse document frequency, the weighting of frequent terms while making infrequent terms have a higher impact can be minimized(Irawaty et al., 2020).

To summarize the key intuition motivating TF-IDF is the importance of a term is inversely related to its frequency across documents.TF gives us information on how often a term appears in a document and IDF gives us information about the relative rarity of a term in the collection of documents. TF-IDF can be a very handy metric for determining how important a term is in a document or dataset.



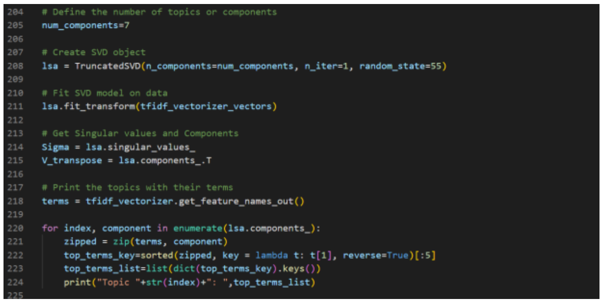
```
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.feature_extraction.text import CountVectorizer
```

```
#TFIDF VECTORIZATION USING Tfidftransformer
#instantiate TfidfVectorizer()
#count vectorizer
tfidf_vectorizer=TfidfVectorizer(use_idf=True)

tfidf_vectorizer_vectors=tfidf_vectorizer.fit_transform([str(big_df['Lemmatized Title'])])

print(tfidf_vectorizer_vectors.shape)

first_vector_tfidfvectorizer=tfidf_vectorizer_vectors[0]
# place tf-idf values in a pandas data frame
Tfidf_df = pd.DataFrame(first_vector_tfidfvectorizer.T.todense(),
                        index=tfidf_vectorizer.get_feature_names_out(), columns=["tfidf"])
Tfidf_df.sort_values(by=["tfidf"],ascending=True)
print(Tfidf_df.head(20))
np.seterr(divide='ignore', invalid='ignore')

```

Figure 9: TFIDF

Latent Semantic Analysis

Latent Semantic Analysis (LSA), also known as Latent Semantic Indexing (LSI), is a fully automated unsupervised statistical-algebraic summarization strategy that employs an extractive method to analyze texts and uncover hidden semantic relationships between the text's words and sentences. It generates a collection of concepts by conducting an analysis of the relationship between a group of terms and the documents that contain them. A machine learning approach called Singular Value Decomposition, or SVD, is used by LSA to attempt to extract the dimensions.



```
# Define the number of topics or components
num_components=7

# Create SVD object
lsa = TruncatedSVD(n_components=num_components, n_iter=1, random_state=55)

# Fit SVD model on data
lsa.fit_transform(tfidf_vectorizer_vectors)

# Get Singular values and Components
Sigma = lsa.singular_values_
V_transpose = lsa.components_.T

# Print the topics with their terms
terms = tfidf_vectorizer.get_feature_names_out()

for index, component in enumerate(lsa.components_):
    zipped = zip(terms, component)
    top_terms_key=sorted(zipped, key = lambda t: t[1], reverse=True)[:5]
    top_terms_list=list(dict(top_terms_key).keys())
    print("Topic "+str(index)+": ",top_terms_list)
```

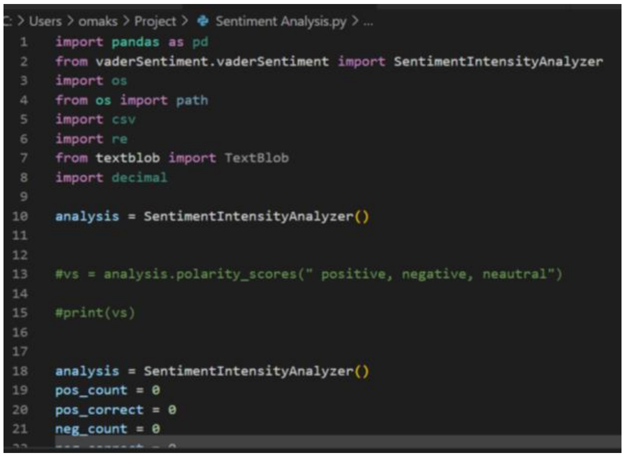
Figure 10: LSA

Sentiment Analysis

To understand the tone and how members of the individual subreddits feel about the discussions on the subreddits we performed a Sentiment analysis, also known as opinion mining, is a technique used in natural language processing (NLP) to determine the emotional undertone of a document. This is a common method used by companies to identify and group opinions about a given good, service, or concept. Data mining, machine learning, artificial intelligence, and computational linguistics are all used in sentiment analysis to sort through text for sentiment and subjective information, such as whether it's expressing positive, negative, or neutral sentiments.

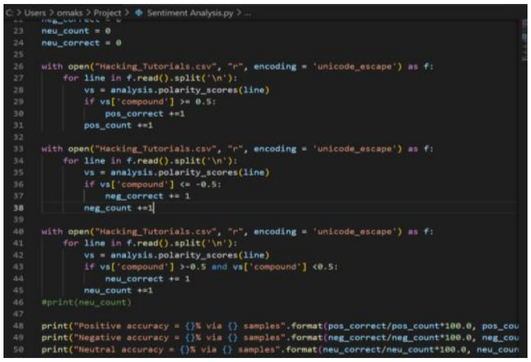
VADER (Valence Aware Dictionary for Sentiment Reasoning) an NLTK module that provides sentiment scores based on the words used was utilized for the implementation of sentiment analysis. It is a rule-based sentiment analyzer in which the terms are generally labeled as per their semantic orientation as either positive, neutral, or negative (Elbagir & Yang, 2019).

The sentiment was ascertained using the polarity scores approach, the preprocessed top posts were categorized into positive, neutral, and negative using the VADER Sentiment Analyzer. A good metric for assessing the sentiment in a particular top post is the compound value. The threshold values in the suggested approach are used to classify the top post as either good, negative, or neutral.



```
import pandas as pd
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
import os
from os import path
import csv
import re
from textblob import TextBlob
import decimal

analysis = SentimentIntensityAnalyzer()


#vs = analysis.polarity_scores(" positive, negative, neautral")

#print(vs)


analysis = SentimentIntensityAnalyzer()
pos_count = 0
pos_correct = 0
neg_count = 0
```

Figure 11: Importing the necessary Sentiment Analysis Library



```
neu_count = 0
neu_correct = 0

with open("Hacking_Tutorials.csv", "r", encoding = 'unicode_escape') as f:
    for line in f.read().split('\n'):
        vs = analysis.polarity_scores(line)
        if vs['compound'] >= 0.5:
            pos_correct +=1
        pos_count +=1

with open("Hacking_Tutorials.csv", "r", encoding = 'unicode_escape') as f:
    for line in f.read().split('\n'):
        vs = analysis.polarity_scores(line)
        if vs['compound'] <= -0.5:
            neg_correct += 1
        neg_count +=1

with open("Hacking_Tutorials.csv", "r", encoding = 'unicode_escape') as f:
    for line in f.read().split('\n'):
        vs = analysis.polarity_scores(line)
        if vs['compound'] >-0.5 and vs['compound'] <0.5:
            neu_correct += 1
        neu_count +=1
#print(neu_count)

print("Positive accuracy = {}% via {} samples".format(pos_correct/pos_count*100.0, pos_cou
print("Negative accuracy = {}% via {} samples".format(neg_correct/neg_count*100.0, neg_cou
print("Neutral accuracy = {}% via {} samples".format(neu_correct/neu_count*100.0, neu_coun
```

Figure 12: Sentiment Analysis Using VADER

5. EXPERIMENTAL RESULTS

The more important or relevant a word is, the higher its TF-IDF score. Table 1 shows the results of the most significant words for the Hacking Tutorial Subreddit. "Comptia" which is a cybersecurity certification had the highest score of 0.251976. Table 2 presents the TFIDF results for Hacking subreddit which indicate that "Hack" was the most significant with the TFIDF value of 0.353553. Likewise, Table 3 shows the results for the Ethical Hacking Subreddit which indicates that "hacking", and "ethical" were the most significant words.

Word	TFIDF Score
Banned	0.125988
book	0.125988
Bounty	0.125988
Box	0.125988
Brilliant	0.125988
Bug	0.125988
Community	0.125988
Comptia	0.251976
D-type	0.125988
Escape	0.125988

Table 1: TFIDF Results for Hacking Tutorial Subreddit – Top 10 Words

Word	TFIDF Score
Dutch	0.117851
Ethical	0.117851
E-ticket	0.117851
Every	0.117851
Exploit	0.117851
Folina Flaw	0.117851
Free	0.117851
Government	0.117851
Hack	0.353553
Hacker	0.117851

Table 2: TFIDF Results for Hacking Subreddits – Top 10 Words

Word	TFIDF Score
Ethical	0.235702
Hacking	0.471405
Hangout	0.117851
Humble	0.117851
Interested	0.117851
Joining	0.117851
Length	0.117851
Life	0.117851
Message	0.117851
Mine	0.117851

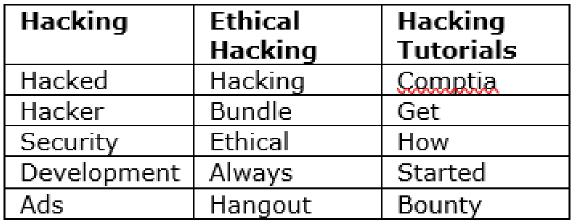
Table 3*:* TFIDF Results for Ethical Hacking Subreddit – Top 10 Words

Search engines are a frequent example of how TF-IDF is used in the field of information retrieval. A search engine can utilize TF-IDF to help rank search results based on relevance, with results that are more relevant to the user having higher TF-IDF scores. This is because TF-IDF can inform you about the relevant importance of a term based upon a document, this can be used to choose keywords (or even tags) for a document or to summarize articles more effectively.

Latent Semantic Analysis

To decompose every term and document as a vector the document-term matrix was used specifically the sklearn's Truncated SVD, because our data was from 3 different subreddits and to avoid overlapping topics we decided to have 3 topics for our text data though the number of topics can be specified using the n_components parameter. Our LSA result indicates the 3 most frequently discussed topics in the subreddits are certifications, techniques on how to do ethical hacking, and training tools. LSA in comparison to the vector space model provides better outcomes, being limited to document term matrix decomposition makes LSA faster than other available techniques.

Table 4: Comparison of LSA Topics Discussed in the Subreddits



Hacking	Ethical Hacking	Hacking Tutorials
Hacked	Hacking	Comptia
Hacker	Bundle	Get
Security	Ethical	How
Development	Always	Started
Ads	Hangout	Bounty

Semantic Analysis

This section discusses the findings of a sentiment analysis on the different hacking subreddits conducted using the NLTK and VADER sentiment analysis tools. As determined by the VADER Sentiment Analyzer, Figures 13, 14 & 15 display the sentiment percentage of the top posts in the individual subreddits as positive, negative, or neutral.

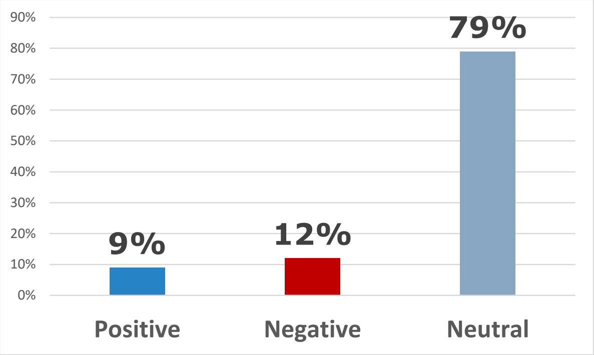


Figure 13 Hacking Subreddit Top Posts

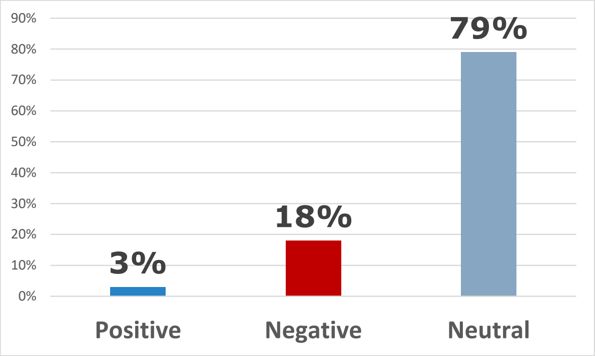


Figure 14 Hacking Tutorial Subreddit Top Posts

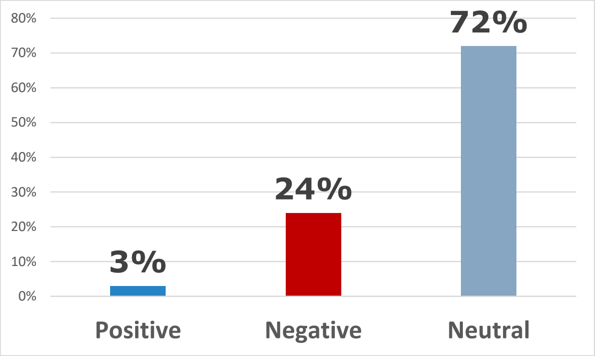


Figure 15 Ethical Hacking Subreddit Top Posts

According to the Base Polarity which is the standard scale, there were more Neutral than positive and a little fewer negative sentiment from Top posts in the subreddits data was collected from. 79% of Hacking subreddit top posts expressed neutral, 12% had negative and 9% expressed positive opinion, while in the ethical hacking subreddit 73% of the top post were neutral, 24% were positive and 3% had negative sentiments. Finally, the hacking tutorial results also indicate 79% of the top posts displayed neutral opinion, 18% were positive and 3% were negative. Based off these results we were able to confidently deduce and conclude that the hacking subreddit top posts analyzed are fact or knowledge based and less subjective as not a significant percentage of emotions and polarity is indicated in the results.

6. CONCLUSION AND FUTURE RESEARCH

To fully understand the most discussed topics in subreddits that pertain to hacking, we investigated top posts on hacking subreddits based on data collected from multiple hacking subreddits. Then we preprocessed the data by eliminating punctuation, stop words, tokenization additionally, using the stemming and lemmatization technique, all derived words in the documents are returned to their stem or base word and lemma form.

We conducted a Term Frequency and Inverse Document Frequency of the corpus which indicated "hack", "hacking" and "Comptia" were more relevant terms in our data according to their TFIDF scores. We conducted a latent sematic analysis which was utilized to analyze the association of words in the same context. We also performed sentiment analysis on the data from all 3 subreddits using the VADER algorithm, with the results indicating that there were more neutral sentiments than positive and a little fewer negative sentiment on the Top posts.

To sum up, the results of the frequency analysis highlight important topics of conversation in the context of cybersecurity. Specifically, the subjects that received the most extensive discussion involve Certifications, Techniques related to ethical hacking, and the usage of training tools. These findings illuminate significant areas of focus and exploration among individuals, showcasing a notable emphasis on improving expertise, understanding, and qualifications in the field of cybersecurity. Notwithstanding these findings, it would be helpful if future research were to examine other cybersecurity subbreddits such as r/malware, r/netsec, and r/cybersecurity to see if the same topics and more importantly same sentiment analysis appear.

7. REFERENCES

Boe., B. (2023). https://praw.readthedocs.io/en/stable/getting_started/installation.html

Carnot, M. L., Bernardino, J., Laranjeiro, N., & Gonçalo Oliveira, H. (2020). Applying text analytics for studying research trends in dependability. *Entropy*, *22*(11), 1303. https://doi.org/10.3390/e22111303

Chiny, M., Chihab, M., Bencharef, O., & Chihab, Y. (2022). Netflix Recommendation System based on TF-IDF and Cosine Similarity Algorithms. *no. Bml*, 15-20. https://doi.org/10.1080/08839514.2023.2175112

Choo, S., & Kim, W. (2023). A study on the evaluation of tokenizer performance in natural language processing. *Applied*

Artificial Intelligence, 37(1), 2175112. https://doi.org/10.1080/08839514.2023.2175112

Dive Into Anything. (2023). https://www.redditinc.com/

Elbagir, S., & Yang, J. (2019). Twitter sentiment analysis using natural language toolkit and VADER sentiment. Proceedings of the international multiconference of engineers and computer scientists.

Giri, S., & Banerjee, S. (2023). Performance analysis of annotation detection techniques for cyber-bullying messages using word-embedded deep neural networks. *Social Network Analysis and Mining, 13*(1), 1-12. https://doi.org/10.1007/s13278-022-01023-2

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science, 349*(6245), 261-266. https://doi.org/10.32604/cmc.2022.024190

Huyut, M. M., Kocaoğlu, B., & Meram, U. (2022). Regulation Relatedness Map Creation Method with Latent Semantic Analysis. *Computers, Materials and Continua*.

IBM. (2023). *What is natural language processing (NLP)?* Retrieved 04/24/2023 from https://www.ibm.com/topics/natural-language-processing#:~:text=Natural%20language%20processing%20(NLP)%20refers,same%20way%20human%20beings%20can.

Irawaty, I., Andreswari, R., & Pramesti, D. (2020). Vectorizer comparison for sentiment analysis on social media youtube: A case study. 2020 3rd International Conference on Computer and Informatics Engineering (IC2IE), DOI: 10.1109/IC2IE50715.2020.9274650

Kumar, V., & Subba, B. (2020). A TfidfVectorizer and SVM based sentiment analysis framework for text data corpus. 2020 National Conference on Communications (NCC), DOI: 10.1109/NCC48643.2020.9056085

Liang, M., & Niu, T. (2022). Research on Text Classification Techniques Based on Improved TF-IDF Algorithm and LSTM Inputs. *Procedia Computer Science, 208*, 460-470. https://doi.org/10.1016/j.procs.2022.10.064

Mayopu, R. G., Wang, Y.-Y., & Chen, L.-S. (2023). Analyzing Online Fake News Using Latent Semantic Analysis: Case of USA Election Campaign. *Big Data and Cognitive Computing, 7*(2), 81. https://doi.org/10.3390/bdcc7020081

Natural Language Toolkit. (2023). https://www.nltk.org/

Pramana, R., Subroto, J. J., & Gunawan, A. A. S. (2022). Systematic Literature Review of Stemming and Lemmatization Performance for Sentence Similarity. 2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA), DOI: 10.1109/ICITDA55840.2022.9971451

Prasanth, S., Raj, R. A., Adhithan, P., Premjith, B., & Kp, S. (2022). CEN-Tamil@ DravidianLangTech-ACL2022: Abusive Comment detection in Tamil using TF-IDF and Random Kitchen Sink Algorithm. Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, https://doi.org/10.18653/v1/2022.dravidianlangtech-1.11

Rahimi, Z., & Homayounpour, M. M. (2023). The impact of preprocessing on word embedding quality: A comparative study. *Language Resources and Evaluation, 57*(1), 257-291. https://doi.org/10.1007/s10579-022-09620-5

Ramezani, M., Shahryari, M.-S., Feizi-Derakhshi, A.-R., & Feizi-Derakhshi, M.-R. (2023). Unsupervised Broadcast News Summarization; a comparative study on Maximal Marginal Relevance (MMR) and Latent Semantic Analysis (LSA). *arXiv preprint arXiv:2301.02284*. https://doi.org/10.1109/CSICC58665.2023.10105403

Subba, B., & Gupta, P. (2021). A tfidfvectorizer and singular value decomposition based host intrusion detection system framework for detecting anomalous system processes. *Computers & Security, 100*, 102084. https://doi.org/10.1016/j.cose.2020.102084

Wagire, A. A., Rathore, A., & Jain, R. (2020). Analysis and synthesis of Industry 4.0 research landscape: Using latent semantic analysis approach. *Journal of Manufacturing Technology Management, 31*(1), 31-51. https://doi.org/10.1108/JMTM-10-2018-0349