Literature Review of Qualitative Data with Natural Language Processing

Bukuroshe Elira Epoka
University, Autostrada Tiranë-Rinas, km. 12, 1000, Albania
eliraepoka230@hotmail.com

Correspondence should be addressed to Bukuroshe Elira Epoka : eliraepoka230@hotmail.com

Article Info
Journal of Robotics Spectrum (https://anapub.co.ke/journals/jrs/jrs.html)
Doi: https://doi.org/10.53759/9852/JRS202301006
Received 16 January 2023; Revised from 28 February 2023; Accepted 03 March 2023.
Available online 12 March 2023.

Abstract – Qualitative research techniques are frequently employed by scholars in the field of social sciences when investigating communities and their communication media. The proliferation of computer-mediated communications has resulted in a substantial volume of textual content. However, the process of coding this vast amount of information necessitates significant time and effort. This article examines the potential for automating specific elements of content analysis through the utilization of natural language processing (NLP) systems, which analyze text in human languages, with a focus on extracting theoretical evidence. In this study, we present a case analysis utilizing NLP to examine the effectiveness of NLP rules in qualitative analysis. Our findings indicate that the NLP rules demonstrated strong performance across multiple codes. The utilization of a NLP system in its current developmental stage has the potential to significantly minimize the text volume, which has to be evaluated using the human coder. This reduction could potentially result in a substantial increase in coding speed, potentially by a factor of ten or more. The research is considered groundbreaking as it pioneers the application of advanced NLP approach to evaluate qualitative data, making it one of the earliest studies in this domain.

Keywords – Natural Language Processing, Cognitive Psychology, Artificial Intelligence, Qualitative Research Techniques, Free/Libre Open-Source Software.

I. INTRODUCTION

Natural Language Processing (NLP) refers to a computational methodology that is employed to analyze textual data, as described by De, Desarkar, and Ekbal [1]. In order to attain language processing capabilities that resemble those of humans in a wide range of activities and applications, NLP encompasses a collection of computationally-based techniques that are theoretically grounded. These techniques are employed to evaluate and depict texts, which occur in a natural manner, at a single or multiple linguistic analysis levels. The field of NLP encompasses a wide range of study challenges and methodologies. NLP tasks encompass a range of applications, including digitalized text summary, question answering machine search and translation.

This research paper investigates the potential of employing NLP methods to partially automate the analysis of qualitative information and data through coding, which involves identifying text segments that offer evidence for concepts of theoretical significance. The scope of our discussion is limited to the natural language processing techniques employed, acknowledging that a significant portion of these techniques have been independently developed by various researchers in diverse contexts. Several scholars, such as Lauriola, Lavelli, and Aiolli [2], offer comprehensive introductions to the field of NLP for readers who are interested in gaining a deeper understanding.

The procedures involved in the application of natural language processing (NLP) to unprocessed text for the purpose of extracting features are depicted in **Fig. 1** of the referenced publication [3]. Initially, we conducted preprocessing on both the titles and text of the reviews. One of the preprocessing techniques employed is known as tokenization, which entails the segmentation of the text into discrete units of words. Distinct tokens were generated for every punctuation mark. The subsequent procedure involved eliminating any recognized contractions associated with those terms. Tokens exhibiting a repetition of the same character more than three times were subjected to normalization. The treatment of consecutive punctuation marks remained consistent. NLP methods and techniques can be applied to various levels of analysis. Language can be analyzed and deconstructed at various levels, ranging from its fundamental phonological aspects to its more complex pragmatic elements.

The intricacy and challenges associated with language processing escalate as individuals' progress through the hierarchical levels of linguistic analysis, which align with progressively larger units of examination. The potential for subtle

interpretation expands as the level of analysis becomes more detailed, progressing from morphemes (such as prefixes or suffixes) to words, sentences, paragraphs, and ultimately entire documents. As an individual ascends from lower to higher levels, the theoretical frameworks employed to elucidate the data delve further into the domain of artificial intelligence, and cognitive psychology, thereby rendering the identification of substantial regularities, upon which to construct principles for text processing, increasingly challenging and elusive. Furthermore, the development of advanced linguistic comprehension is contingent upon the foundation of comprehension established at lower levels by novices.

Natural language processing (NLP) systems employ various techniques to derive semantic information from linguistic usage patterns. The two primary methodologies employed in this study are statistical analysis and the use of symbolic representations. The symbolic approach examines language processes within texts to extract meaning by incorporating discourse, semantic and syntactic information using human-designed lexicons and rules. An instance where this approach can be employed is in the handling of educational benchmarks, as demonstrated in the previous discourse. Corpus-based statistical approaches employ mathematical techniques to develop language models by analyzing real-world instances. In the context of statistical machine translation, it is common practice to analyze multiple multilingual papers in order to ascertain the prevailing translations of specific words or phrases across different contexts. With a sufficiently large training dataset, this approach has become widely accepted as the prevailing method for analyzing textual corpora.

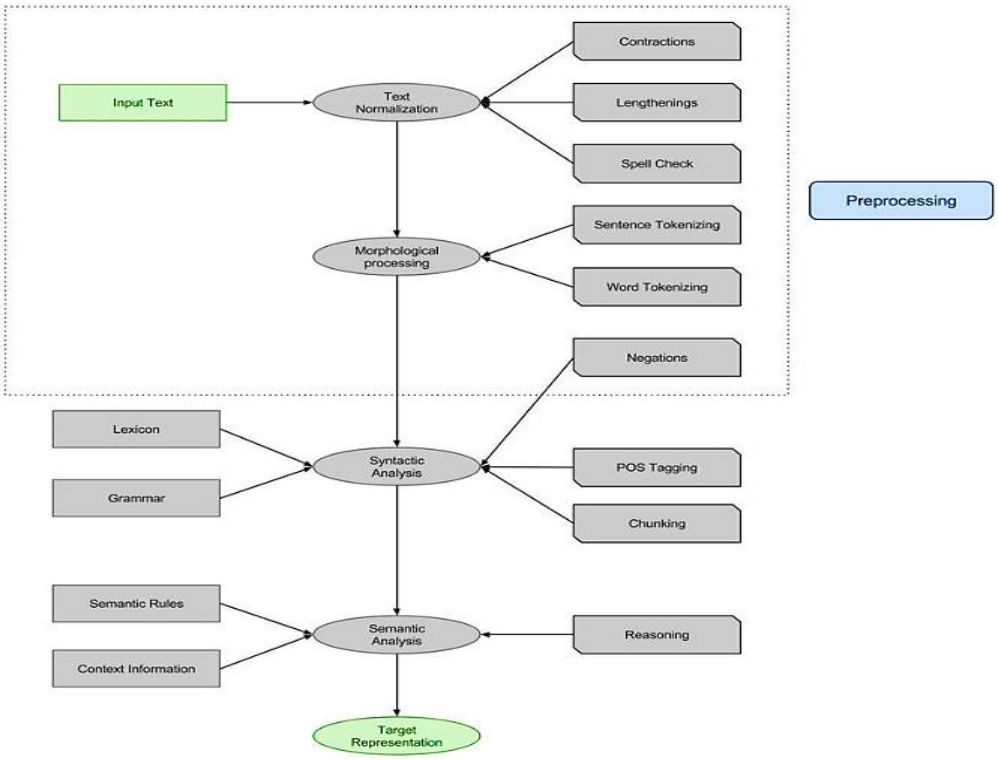

Fig 1. The NLP Procedures Used During the Text-Preprocessing Phase

This study aims to explore the potential of employing natural language processing techniques, specifically sublanguage theory, to enhance a specific task in positivist qualitative study, i.e. content analysis coding. Qualitative research employs content analysis as a means to substantiate hypotheses pertaining to subjects of interest through the examination of textual materials as the primary source of data. The act of annotating or tagging a text with codes to represent the ideas presented is the ultimate outcome of the coding process. In lieu of attempting to decode and comprehend the entirety of the text, we propose a methodology wherein codes are employed based on the attributes of individual passages.

In the context of researching group decision-making, an approach could involve coding transcripts of interactions to identify instances of theoretical underpinnings, which are of major interest, for instance, the detection of a problem or the proposal of a solution. The utilization of coding techniques enables researchers to systematically examine and evaluate the interrelationships among ideas as they are articulated within textual materials. The encoded text could potentially be utilized to empirically examine hypotheses pertaining to the effects of teamwork and member engagement in the decision-making process. Although the utilization of coded data is crucial to the overall research process, this aspect will not be addressed in this particular discussion.

In this discussion, we will present the knowledge we have acquired regarding the practical application of the symbolic method. In contrast to statistical approaches, symbolic approaches provide the benefit of not needing extensive datasets for training, making them particularly well-suited for our intended objectives. However, further research and development are required to establish a comprehensive set of rules for symbolic approaches. Furthermore, the limited applicability of porting rules across different domains restricts their utility to a singular area of study. These limitations will be reexamined in the

upcoming discourse. The article has been organized as follows: Section II presents case study of natural language processing (NLP) applied to qualitative data evaluations. Section III presents a conceptual analysis of social presence and face work in relation to the subject of the article. Section IV review the manual data analysis process, considering FLOSS (free/libre open source software). Section V focusses on the results of the article, and Section VI presents the discussions and limitations of the research. Lastly, Section VII concludes the paper and recommends directions for future research.

II. CASE STUDY

This section describes a case analysis of natural language processing (NLP) applied to qualitative data evaluations is presented. Radtke, Janssen, and Collofello [4] are engaged in a collaborative research endeavor focused on investigating the work practices employed by teams of developers involved in the creation of FLOSS (free/libre open source software). The teams exhibit geographical and temporal dispersion, with limited face-to-face interactions, relying predominantly on electronic communication channels for coordination. Extensive repositories of these communications are accessible for scholarly examination. In the subsequent sub-sections, an account is provided of the various stages encompassed in the study, commencing with the conceptual progress and subsequent formulation of the coding system. This is followed by an examination of the manual coding process, culminating in an exploration of the utilization of Natural Language Processing (NLP) for the aforementioned coding endeavor. In accordance with our emphasis on research methodologies, we hereby provide a concise account of the study, offering sufficient information for readers to comprehend the employed methodology and the involvement of Natural Language Processing (NLP), while refraining from delving into specific discussions regarding the study's findings.

III. CONCEPT ANALYSIS

The primary objective of this study was to examine the impact of team maintenance practices on productivity within Free/Libre and Open Source Software (FLOSS) communities. The concept of "group maintenance behavior" pertains to the deliberate actions undertaken by individuals within a group to enhance interpersonal connections, maintain trust, and foster cooperative efforts. In our study, we employed two theoretical frameworks to analyze and classify beneficial behaviors within organizations. Specifically, we drew upon the concepts of social presence and face work in computer-mediated communication. Each topic is addressed sequentially.

Social Presence

The concept of social presence refers to an individual's capacity to effectively communicate their personal attributes within a given community, thereby presenting themselves as genuine and authentic individuals to other participants. Sude and Dvir-Gvirsman [5] conducted a study which revealed that the presence of others played a crucial role in determining the level of happiness experienced by individuals in computer-mediated communication (CMC) supported groups. Individuals within the CMC (Computer-Mediated Communication) realm employ a diverse array of strategies to enhance their approachability. These tactics encompass the utilization of emoticons, humor, vocatives (explicit references to other individuals), phatics (verbal exchanges aimed at expressing emotions instead of conveying data), compliments, inclusive pronouns, expression of agreements and appreciation, including expressive or non-standard punctuations, and visible capitalization.

The employment of CMC tools and electronic systems for OGL (online group learning) has been found to have a positive correlation with social presence. This correlation arises from the ability of these tools and platforms to replicate, in an online setting, the same interpersonal communications, group dynamics and group learning that are observed in face-to-face interactions. The learning outcomes of groups engaged in online game-based learning (OGL) are influenced by the level of social presence exhibited by its members. Caspi and Blau [6] have demonstrated that an individual's social presence plays a significant role in shaping group learning and group dynamics through interpersonal interactions. Furthermore, it has been suggested that active participation in social activities can enhance one's social presence. Furthermore, Ribosa and Duran [7] posited that the identification of strategies to cultivate an individual's "social presence" within the educational setting is crucial for fostering a more immersive and supportive academic encounter, wherein students exhibit heightened motivation and achieve greater levels of accomplishment. In the study conducted by Gogus [8], it was found that social presence plays a crucial role in defining the levels of interaction and learning effectiveness in a digital ecosystem. The authors provide a definition of social availability as a fundamental element and emphasizes its significance as one of the key constructs in this context.

Notwithstanding its significant, the accurate definition of social availability has remained elusive due to the proliferation of divergent methodologies for its assessment. The concept of social presence is frequently subject to misinterpretation, resulting in the emergence of multiple interpretations derived from disparate theoretical frameworks. The absence of a cohesive field of research dedicated to the examination of social availability in digitalized group learning can be attributed to the absence of consensus regarding its definition, methods of measurement, and its significance. This elucidates the reason behind the divergent outcomes observed in research pertaining to social presence. In the study conducted by de Oliveira and Esteve-González [9], a comparison was made between text-based computer-mediated communication (CMC) and web-based video conferencing. Contrary to the prediction made in [10], Johnson and Hong found no discernible difference in the experiences of social presence between the two modes of communication. In contrast to the findings of [11], Tudor did not ascertain a positive correlation between web-oriented video conferencing and enhanced student learning or enhanced study performance experiences.

Face Work

The concept of face, as elucidated by Leban, Thomsen, von Wallpach, and Voyer [12], refers to the favorable social worth that individuals' attribute to the public persona they project. The concept of an individual's "face" is formed by the combination of the need for acceptance and esteem (positive face) and the need for autonomy in action (negative face). According to Ryan and Ryan [13], behaviors that involve maintaining autonomy, independence, and privacy can be seen as indicative of a negative face, whereas behaviors, which prioritize a sense, approval and respect of society can be seen as indicative of a positive face. Irrespective of the level of public decorum an individual maintains, the possibility of experiencing a loss of social standing persists when engaging in actions that challenge or undermine one's self-image or reputation, commonly referred to as face-threatening activities (FTAs). Preserving one's personal reputation as well as the reputations of others is a crucial aspect of all types of social interactions.

Politeness is a strategy employed by individuals to mitigate the potential negative consequences that may arise during interpersonal interactions. The verbal behaviors involved in computer-mediated communication (CMC) politeness encompass positive strategies aimed at eliciting positive face and negative strategies aimed at eliciting negative face. These behaviors are enacted by both participants involved in the communication process. Effective politeness strategies encompass various elements, such as the utilization of vocatives, inclusive pronouns, compassion and agreement. Negative politeness is a concept in sociolinguistics that encompasses various linguistic strategies employed to mitigate potential threats to face or social harmony. These strategies include the application of hedgers, formal language, indirect questions, honorifics, subjunctives, passive voice, and justifications for free trade agreements (FTAs).

A coding system was initially developed through an inductive approach, drawing upon relevant ideas and their discourse in the existing literature. This coding system aimed to assess group maintainability behaviors in the view of FLOSS data. In order to facilitate the work of programmers, this coding system established a comprehensive set of terms, accompanied by illustrative examples and detailed explanations of their respective applications.

IV. MANUAL DATA ANALYSIS

For the purpose of this research, two Free/Libre and Open Source Software (FLOSS) projects, namely Gaim and Fire, were selected due to their shared objectives of developing an Instant Messaging client. In order to establish a comprehensive evaluation of the relative effectiveness of the two initiatives, they were selected on the basis of their common goals, activities, and intended recipients. In their study, Di Iorio and Vantini [14] employed a range of multivariate indicators to arrive at the conclusion that the Gaim project exhibited greater overall success. The sustained presence of Gaim, now known as Pidgin, serves as evidence of its ongoing triumph, while the progress of Fire ceased in the early months of 2007.

The study utilized a collection of posts sourced from the mailing groups of the two project developers. The mailing lists were chosen based on their role as primary channels of communication among developers, making them the focal point for group administration activities. The selection of messages preceding 60 choices was conducted in a random manner, with each of the two projects contributing an equal number of decisions (resulting in 120 decisions). Specifically, 20 decisions were sampled from the first, middle, and final stages of each project's lifecycle. To exemplify the problem, we aim to address through Natural Language Processing (NLP), we have solely implemented a portion of the dataset and assessed the corresponding messages. The dataset contained a total of 84,870 words, with an average of 106.5 words per message.

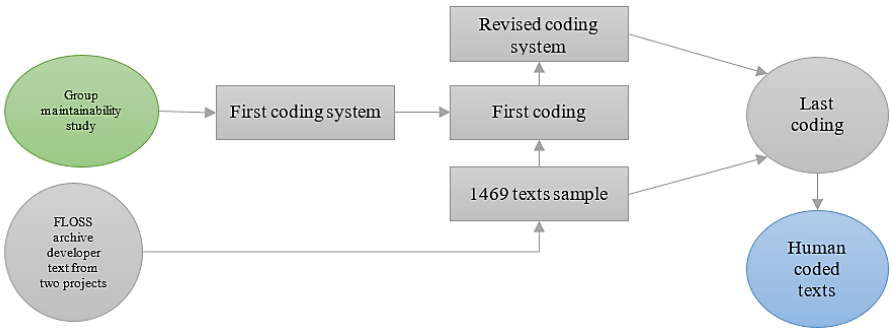

Fig 2. Traditional Process of Data Coding

Fig 2 illustrates the comprehensive methodology employed in this study. The manifestation of the indications in the data was acquired through an iterative process of coding, examination, debate, and correction, employing an inductive learning approach. Holsti's coefficient of reliability was employed in this study to evaluate the level of agreement between raters. The formula used for this assessment is $2m/ (n_1 + n_2)$, whereas m signifies the proportion of coded items on which both raters agree, n_1 denotes the judgement numbers executed by the first coder, and n_2 signifies the decisions executed by the second coder. Zhou, Tu, and Xiao [15] describe this approach of defining inter-rater reliability as "the simplest and most common approach." Although alternative measures, such as kappa, consider the potential occurrence of random agreement, the specific task of identifying short pieces of text within a larger corpus reduces the likelihood of random agreements to a minimum. Therefore, this straightforward measure was selected.

The coders underwent learning until they reached a 0.80 inter-rater reliability or higher, which is considered the benchmark level of consensus for manually conducted qualitative data evaluation. The procedure of learning necessitated two iterations of double coding on 400 texts. In qualitative data analysis, it is common for certain indicators, such as "Humor," to be excluded from a study due to the lack of consensus among coders regarding its interpretation. The programmers encountered difficulties in comprehending the code due to their limited understanding of programming humor. After establishing confidence in the remaining codes, the coders proceeded to independently code the remaining messages.

Rule-Writing Effort

All of the aforementioned features have been facilitated by the utilization of the Natural Language Processing (NLP) tool, necessitating minimal adjustments. The final phase of the study entails the implementation of specifically tailored Information Extraction (IE) rules to the textual data. This process aims to extract relevant excerpts of text that correspond to theoretically significant concepts. The text-processing tool possesses an inherent capacity to implement rules on text. However, the formulation of these rules necessitates a unique development process for each specific task. In this section, we provide a comprehensive account of the developmental process that culminated in the formulation of these guidelines.

The NLP analyst established information extraction instructions for the codes related to group maintainability as presented in [17]. Due to temporal limitation, our team managed to construct natural language processing (NLP) rules solely for 12 out of the 15 manual codes. The three remaining manual codes, namely "vocatives," "disclaimers," and "stating rationale for free trade agreements," were not included in our NLP rule development. The initial investigation into these three codes shows that the digitalization of coding may encounter comparable difficulties as with the remaining twelve codes. However, the inclusion of vocatives presents distinct challenges, which we will examine in the subsequent analysis. A set of rules was initially devised to encode the frequently occurring and easily identifiable occurrences within the encoded text. Subsequently, these rules underwent iterative enhancements to enhance both their scope and precision.

The rule of capitalization heavily depends on regular expressions for the identification of uppercase instances. Additional criteria, such as the concept of Apology, focused specifically on a compilation of words or phrases, such as "sorry" or "apologies." The provided visual representation, **Fig 2**, illustrates a fundamental exemplar rule utilized in the process of identifying Agreement. It is important to note that the identification of Agreement necessitates the incorporation of supplementary rules that encompass the entirety of Natural Language Processing (NLP) components, such as a segment of speech, semantic class, actual word, and syntax. Although this rule serves as a rudimentary demonstration of the potential efficacy of NLP rules in syntactic management, the computational demands associated with handling more complex linguistic structures surpass the capabilities of conventional CAQDAS tools.

V. RESULTS

The rule set that was created was evaluated using the GS data set that was specifically reserved for this purpose. The test set messages were subjected to manual examination in order to determine the accuracy of the coding for group maintenance activities. This examination aimed to identify any instances that were either correctly or incorrectly coded by the automated process, as well as any missed instances. Additionally, the examination sought to determine the underlying causes of any coding errors.

The performance of the automated system was evaluated based on two widely used measures in Information Extraction (IE), namely Precision and Recall. Recall determines the coverage of the system in successfully detecting and retrieving GS codes, expressed as a percentage. In the case of the Hedges code, human coders were able to identify a collective count of 156 text segments within the reserved test messages that were deemed to be representative of hedges. Conversely, the system's identification yielded a lower count of 116. This results in a Recall rate of 74% (116/156). Precision determines the size of accurately coded data, which has been automatically extracted, relative to the accurately coded data from the gold standard. In the case of the code Hedge, the structure identified a collective count of 155 text portions in a stored test texts as a hedge case. However, it was determined that only 116 of these identifications were accurate, aligning with the judgments made by human coders. Consequently, the Precision value can be calculated as 116 divided by 155, resulting in a percentage of 75%. Achieving high precision in the outputs leads to a decrease in the target data coverage by recall, and vice versa. Consequently, it is rare to achieve excellent performance on both criteria simultaneously.

In order to achieve complete automation of coding, it is imperative to attain a high level of proficiency in both metrics. In our pursuit of developing an assisting tool, our primary objective was to enhance the Recall of the automated system. During the formulation of the rules, we set a target of achieving 80% Recall. The rationale behind selecting this particular strategy was based on the assumption that it would be more straightforward for a human reviewer to identify and remove inaccurately coded data, which is a result of low Precision. In contrast, it would be more challenging for the reviewer to manually search through the unprocessed text data to determine the proof, which did not undergo coding at all, which is a consequence of lower Recall.

Crowston, Allen, and Heckman [17] depicts the results of the system integrating 12 chosen group maintainability codes. The system performance on the test data and learning data is compared in the columns designated for training and testing, respectively. Overall, the training performance demonstrates improvement as a result of refining the rules with the utilization of this information. Upon closer examination of the codes, it becomes evident that Emoticon, Inclusive Pronouns, and Formality exhibit the highest level of Recall. This observation indicates the consistent utilization of these textual

constructions. It is susceptible to idiosyncratic protocols such as slang and expressions of admiration. The reduced Precision of the findings reflects our preference for Recall against Precision. Nonetheless, the Precision is remarkably high for specific codes like Salutations or Emoticon, and all of them are at acceptable levels except for Capitalization and Punctuation. The following section delves into the factors that contribute to the unpredictably lower Precision of the codes.

The review article presents an analysis of the system's performance on the test data pertaining to a specific construct, namely Hedges. This analysis involves a comparison between the judgments made by the gold standard (GS) and the system. According to the data presented in [18], there were a total of 156 occurrences of Hedges in the GS test data, which consisted of reserved test messages. The system accurately identified and classified 116 instances, while it failed to detect 40 instances. The authors present the results of the system's identification of Hedges in 155 text segments. Among these segments, 116 were found to be consistent with the Gold Standard (GS), while 39 did not align with the GS [19]. For the sake of clarity, the number of corpus text segments that were not classified manually or automatically as hedges is not included in the last column. The decision to use thematic units considered as a unit of coding present limitations, which prevents us from providing a specific number for the cell [20].

Nonetheless, the test dataset integrated approximately 80 messages containing a total of 8037 words, indicating that the true number of units likely exceeded the thousands [21]. Hence, given its present performance level, the system exhibits the capacity to significantly enhance coding efficiency by reducing the volume of text requiring human coder examination by a factor of ten or more (specifically, hundreds of units of raw data to about 150 units notifiable within the system). Codes that are less commonly used, and therefore result in a more significant narrowing, would have a more pronounced impact on performance. Conversely, codes that are less precise would have a lesser effect [22].

VI. DISCUSSION AND LIMITATIONS

The demonstrated efficacy of the established rules in effectively processing diverse codes indicates significant potential for utilizing Natural Language Processing (NLP) techniques in coding qualitative data. By conducting a meticulous analysis of the outcomes of our endeavors, we successfully identified numerous variables that exerted an influence on performance. The aforementioned concerns are subsequently discussed, followed by an evaluation of the benefits and costs associated with this particular method.

Inadequate Pre-Processing

The allocation of time towards the organization and cleansing of data is a substantial and labor-intensive aspect of natural language processing. Prior to commencing this task, various techniques were employed to prepare the messages. These techniques included forwarded messages, signature blocks, and segregating headers. Human analysts typically refrained from incorporating these techniques when manually coding the messages. Detecting and removing messages transferred from external sources, such as lines of code, error logs, source file comparisons (diffs), and similar elements, posed a significant challenge in terms of accuracy during the processing phase. Regrettably, the inclusion of such material had an adverse effect on the precision of punctuation, capitalization, and emoticon recognition. This was primarily due to the frequent occurrence of sequences of capitalized phrases, characters, and punctuation marks that bear resemblance to emoticons. In addition, it is worth noting that email communications frequently suffer from illegibility, thereby posing challenges in adhering to formal standards.

Unit of Coding

The researchers employed the theme unit, a commonly utilized coding unit in qualitative data analysis, for the purpose of manually coding the data. In the realm of programming, theme units can encompass various linguistic elements such as words, phrases, sentences, or even paragraphs. These units are deemed as theme units by programmers when they are perceived to substantiate a specific concept or notion. The precise delineation of text boundaries for capture using NLP rules is challenging due to the diverse array of potential contexts. In accordance with standard procedure for comparing human coders, we considered any agreement between manually coded and machine coded text as a match for the aforementioned findings. Nevertheless, it would be more desirable to employ a more defined analysis unit for the purpose of coding, like a phrase or conceivably a whole message. This approach would enhance future comparison between machine and human coding, thereby simplifying the process.

Adequate Training Examples

The available data for specific codes was insufficient, rendering it unreliable for training purposes. The presence of a substantial quantity of accurately encoded textual samples is crucial for the successful implementation of natural language processing (NLP), particularly in the context of statistical methodologies. The observed variations in performance underscore the existing disparity in our dataset. It is worth noting that, contrary to expectations, the Formal Verbiage code exhibited a strong performance. However, for codes with fewer than 100 training cases, the disparity between training and testing results is more pronounced.

Manual Coding Error

The efficacy of automatic coding was assessed by conducting a comparative analysis between its generated output and that of a human coder, as well as against the reliable ground truth data. Although efforts were made to maintain consensus among human coders during the manual coding process, it was discovered that the generated GS data contained coding errors due to various factors. Initially, a period of delay was observed until a state of reliability was established among various programmers. The coding quality of the GS data utilized in our analysis exhibits variation due to its collection occurring both before and after the stabilization phase. Furthermore, it is important to note that achieving complete accuracy in programming within a reasonable timeframe is a challenging task for human programmers. This is primarily due to the labor-intensive and mentally exhausting nature of coding, which increases the likelihood of errors occurring, both through unintentional omissions and deliberate actions. One limitation faced by human programmers was their inability to effectively evaluate certain code elements, such as those containing slang or humor. This was primarily due to their non-native speaker status and lack of familiarity with the specific jargon and cultural references employed by developers.

However, errors made in human coding are moved to the digital processing phase, as instructions are designed according to possible unreliable set of data. In addition, it is important to note that drawing comparisons between the outcomes of NLP coding and inaccurate information can potentially lead to misleading conclusions. An attempt was made to quantify the influence of human coding errors on our research outcomes by reassessing the accuracy of false positive results in natural language processing (NLP) through the utilization of the codebook instead of the gold standard (GS) data. Based on the identification of multiple instances where the NLP rules detected codes that were missed by human coders, our NLP analyst, who possesses expertise in the field, has reached the conclusion that implementing this modification would have resulted in an increase in Precision for all codes.

The codes in the GS data that have the fewest instances and are highly sensitive to even minor errors would exhibit the most substantial enhancements. While automated systems demonstrate proficiency in identifying regular forms, human coders encounter significant challenges in this regard. This is evident from the observed disparity of 30% between the Precision obtained and the analyst's estimation, after accounting for errors made during human coding, particularly when handling Inclusive Pronouns. The utilization of human coders to verify the NLP output has the potential to produce more accurate coding outcomes. This observation suggests that in specific situations, the automated approach may be more reliable than human coders in identifying instances.

Language and Meaning

The vast diversity of language poses significant challenges when it comes to automated content analysis. The degree of specificity fluctuates based on the specific code under examination. The codes known as Formal Verbiage, Apology, and Agreement exhibit consistent patterns in their representation, requiring only a minimal number of rules to attain a high level of performance. In contrast, Hedges and Vocatives were found to present greater difficulty due to various syntactic and semantic factors, resulting in their exclusion from formal examination.

Context

Maintaining awareness of the contextual framework is crucial when conducting a content analysis. The utilized processing engine, TextTagger, currently does not consider text that is located beyond a sentence boundary, unless it pertains to co-reference. Hence, it is imperative to solely consider the explicit significance of an individual statement, without incorporating any additional factors. The complete realization of a message's discourse structure and context is hindered by the presence of this technological barrier.

Syntactic Variety and Synonymy

Various synonymous expressions, syntactic structures, and embellishments like adjectival and adverbial clauses enable natural language to exhibit a nearly infinite range of possibilities for organizing and conveying meaning. Effective automation often requires a larger amount of training data than what was previously available in order to capture the full complexity of language. However, it is important to note that the sublanguage used in software engineering does not encompass the entirety of linguistic diversity.

Multiple Aspects of Meaning

Identifiers such as "probably" and "possibly" were instrumental in enabling law enforcement to narrow down their focus on Hedges. Certain words, such as 'seem,' 'would,' and 'of course,' posed more challenges due to their ambiguous nature, as they could be indicative or non-indicative depending on the context. Automated systems may encounter difficulties in comprehending subtle shifts in meaning that arise from contextual factors within and beyond a phrase. The issue of differentiating between singular and plural forms of the second-person pronoun, such as in the sentence "When you open up the file, you will see two items," poses a significant difficulty in identifying the Vocative case. To date, no viable solution has been investigated. The task of coding vocatives with a high recall rate and sufficient accuracy poses a challenge in the absence of contextual cues or a corresponding response to the message.

Implicit Meaning

The domain of natural language processing (NLP) has only begun to explore the extraction of implicit meaning from textual data. Given the difficulties encountered by our analysts in deciphering the Humor code, it is evident that even humans face challenges in this domain. Consequently, the study of humor represents a currently active but intellectually rigorous area of research.

Cost/Benefit

The cost-benefit analysis plays a pivotal role in ascertaining the suitability of employing NLP-enabled content analysis for a specific research endeavor [16]. While the automation of certain aspects of the coding process through natural language processing (NLP) shows promise, further research and validation of a set of rules are required. The methodology employed in our study involved the utilization of a meticulously crafted qualitative coding codebook, which is considered indispensable for conducting qualitative research. This codebook served as a foundation for the creation of natural language processing coding. However, the generation and testing of additional rules for the NLP rule set required the expertise of a proficient NLP analyst, along with the necessary time to assess its performance. In order to effectively manage extensive volumes of text data, certainly discourses that span significant time periods, it is essential to allocate sufficient development time to address various factors. These factors include adapting to transformations in the format of data, incorporating novel discoveries, and accommodating the changes of both analytical concepts and data content underlying a large-scale analysis system.

A significant amount of time, spanning an entire year, was dedicated to the process of data preparation, rule creation, development, and testing by a software engineer and a language analyst for the specific instance being discussed. In contrast, despite allocating fifty percent of their time to the project, two human coders were only capable of coding two projects within the same temporal span. Nevertheless, a portion of the allocated time was dedicated to the enhancement of the codebook, which functioned as the fundamental reference for both manual and NLP coding. Furthermore, it was also utilized for coding supplementary messages that were not included in the analysis presented in this study. After receiving instruction and establishing a consistent codebook, a single coder invested more than 100 hours in manually coding a total of 700 messages across all 15 codes.

In the case of small-scale data sets, typically consisting of around a thousand unique messages, which can be efficiently processed by content coders within a relatively brief period, the utilization of a natural language processing (NLP)-supported approach may not be justified due to the additional human effort it would entail. It is important to note that the application of the NLP method is limited to abstract concepts that can be effectively articulated in a standardized manner through written language. The current implementation of Natural Language Processing (NLP) is unlikely to be beneficial for code that heavily depends on human judgment and contextual understanding. Nevertheless, by implementing appropriate algorithms and allocating significant resources towards its development, extensive research endeavors can potentially benefit from the ability to efficiently process and analyze vast volumes of data, thereby minimizing the time and effort required from human coders.

By allocating resources towards the development of rule-writing techniques, there is a potential to significantly decrease the amount of time and effort needed for coding additional text, potentially resulting in a ten-fold reduction. This, in turn, would enable the analysis of numerous groups consisting of hundreds of thousands of messages. The implementation of such a high degree of automation is deemed essential due to the immense scale of the material involved, which would require an extensive amount of time and effort if processed manually, potentially spanning hundreds of programmer years.

VII. CONCLUSION AND FUTURE RESEARCH

This article aimed to assess the viability of employing natural language processing (NLP) techniques for conducting content analysis on communications artifacts consequential from digital groups. Such analysis falls under the purview of qualitative data analysis, but this paper reviews a literature texts and presents its findings in the results section. Future work should encompass three facets. The initial step involves the development of a framework for the NLP text processor, which will enable users to efficiently examine the codes implemented by the system. Additionally, the system will be utilized to facilitate research on the manner in which Free/Libre and Open Source Software (FLOSS) teams manage group maintenance. Preliminary theories have been formulated based on observed trends in human coding. However, due to the limited quantity of manually coded data available, it is currently only feasible to conduct a comparative case study involving the two groups. The objective of employing natural language processing is to analyze extensive datasets of teams, thereby providing a more robust basis for our findings.

One significant constraint of the current study is the requirement for the involvement of a proficient NLP analyst in the development and fine-tuning of the rule sets. In order to overcome this obstacle, our objective is to explore the feasibility of employing machine learning (ML) methodologies for the purpose of formulating rules. A significant limitation of machine learning methods is the requirement for a larger quantity of ground truth data to be utilized as input. In contrast to the limited number of instances typically available for most codes, a substantial quantity of cases, typically in the hundreds, is required. The specific number is contingent upon various factors, including the learning method employed, the number of codes and labels integrated, and the intricacy of the phenomena under investigation.

Data Availability
No data was used to support this study.

Conflicts of Interests
The author(s) declare(s) that they have no conflicts of interest.

Funding
No funding agency is associated with this research.

Ethics Approval and Consent to Participate
The research has consent for Ethical Approval and Consent to participate.

Competing Interests
There are no competing interests.

References

[1]. A. De, M. S. Desarkar, and A. Ekbal, "Towards improvement of grounded cross-lingual natural language inference with VisioTextual Attention," Natural Language Processing Journal, no. 100023, p. 100023, 2023.
[2]. I. Lauriola, A. Lavelli, and F. Aiolli, "An introduction to Deep Learning in Natural Language Processing: Models, techniques, and tools," Neurocomputing, vol. 470, pp. 443–456, 2022.
[3]. J. A. Baktash, M. Dawodi, M. Zarif, and N. Hassanzada, "Tuning traditional language processing approaches for Pashto text classification," Int. J. Nat. Lang. Comput., vol. 12, no. 2, pp. 53–66, 2023.
[4]. N. P. Radtke, M. A. Janssen, and J. S. Collofello, "What makes Free/Libre Open Source Software (FLOSS) projects successful? An agent-based model of FLOSS projects," Int. J. Open Source Softw. Process., vol. 1, no. 2, pp. 1–13, 2009.
[5]. D. J. Sude and S. Dvir-Gvirsman, "Different platforms, different uses: testing the effect of platforms and individual differences on perception of incivility and self-reported uncivil behavior," J. Comput. Mediat. Commun., vol. 28, no. 2, 2023.
[6]. A. Caspi and I. Blau, "Social presence in online discussion groups: testing three conceptions and their relations to perceived learning," Soc. Psychol. Educ., vol. 11, no. 3, pp. 323–346, 2008.
[7]. J. Ribosa and D. Duran, "Students' feelings of social presence when creating learning-by-teaching educational videos for a potential audience," Int. J. Educ. Res., vol. 117, no. 102128, p. 102128, 2023.
[8]. A. Gogus, "Adaptation of activity theory framework for effective online learning experiences: Bringing cognitive presence with teaching and social presences in online courses," Online Learn., vol. 27, no. 2, 2023.
[9]. J. M. de Oliveira and V. Esteve-González, "Navigating choppy discourses: A conceptual framework for understanding synchronous text-based computer-mediated communication," Text Talk - Interdiscip. J. Lang. Discourse Commun. Stud., vol. 40, no. 2, pp. 171–193, 2020.
[10]. E. K. Johnson and S. C. Hong, "Instagramming social presence: A test of social presence theory and heuristic cues on Instagram sponsored posts," Int. J. Bus. Commun., vol. 60, no. 2, pp. 543–559, 2023.
[11]. C. Tudor, "The impact of the COVID-19 pandemic on the global web and video conferencing SaaS market," Electronics (Basel), vol. 11, no. 16, p. 2633, 2022.
[12]. M. Leban, T. U. Thomsen, S. von Wallpach, and B. G. Voyer, "Constructing personas: How high-net-worth social media influencers reconcile ethicality and living a luxury lifestyle," J. Bus. Ethics, vol. 169, no. 2, pp. 225–239, 2021.
[13]. W. S. Ryan and R. M. Ryan, "Toward a social psychology of authenticity: Exploring within-person variation in autonomy, congruence, and genuineness using self-determination theory," Rev. Gen. Psychol., vol. 23, no. 1, pp. 99–112, 2019.
[14]. J. Di Iorio and S. Vantini, "How to get away with statistics: Gamification of multivariate statistics," J. Stat. Data Sci. Educ., vol. 29, no. 3, pp. 241–250, 2021.
[15]. J.-L. Zhou, R.-F. Tu, and H. Xiao, "Large-scale group decision-making to facilitate inter-rater reliability of human-factors analysis for the railway system," Reliab. Eng. Syst. Saf., vol. 228, no. 108806, p. 108806, 2022.
[16]. A. Rogachev, E. Melikhova, and G. Atamanov, "Building artificial neural networks for NLP analysis and classification of target content," in Proceedings of the conference on current problems of our time: the relationship of man and society (CPT 2020), 2021.
[17]. K. Crowston, E. E. Allen, and R. Heckman, "Using natural language processing technology for qualitative data analysis," Int. J. Soc. Res. Methodol., vol. 15, no. 6, pp. 523–543, 2012.
[18]. Md. Z. Hussain, M. Ashraf, D. K. Singh, A. Haldorai, D. K. Mishra, and T. N. Shanavas, "Intelligent data post and read data system like to feed for IoT sensors," International Journal of System Assurance Engineering and Management, Jun. 2022, doi: 10.1007/s13198-022-01683-5.
[19]. D. Ramkumar, M. Ashraf, K. Sathesh Kumar, Md. Z. Hussain, A. Haldorai, and D. K. Mishra, "Defining multiple geometrical areas with modeling of elementary geometrical volumes in robot-environment interaction," International Journal of System Assurance Engineering and Management, Jun. 2022, doi: 10.1007/s13198-022-01708-z.
[20]. A. Haldorai and K. K, "An Analysis of Software Defined Networks and Possibilities of Network Attacks," Journal of Machine and Computing, pp. 42–52, Jan. 2022, doi: 10.53759/7669/jmc202202006.
[21]. S. R and A. H, "Adaptive fuzzy logic inspired path longevity factor-based forecasting model reliable routing in MANETs," Sensors International, vol. 3, p. 100201, 2022, doi: 10.1016/j.sintl.2022.100201.
[22]. G. D. Vignesh, A. Ramu, J. T. Raja, P. Ponmurugan, A. Haldorai, and G. Senthilkumar, "Sensor data fusion techniques in the construction of generalized VORONOI graph for on-line motion planning in robot navigation," International Journal of System Assurance Engineering and Management, Sep. 2022, doi: 10.1007/s13198-022-01773-4.