Public Sentiment on Security Cameras

M. Fahad Khalid
ITEE
University of Oulu
Oulu, Finland

Usman Muhammad
Dept. of Computer Science
Aalto University
Espoo, Finland

R. Saugmann Andersen
Faculty of MBP
Tampere University
Tampere, Finland

Mourad Oussalah
CMVS / Faculty of ITEE
University of Oulu
Oulu, Finland

***Abstract*—The use of video surveillance in public spaces has increased rapidly, extending to residential areas and even non-metropolitan regions. However, there is limited research on public attitudes toward CCTV, and few studies have focused on how these attitudes have changed over the years. This study examines public sentiment toward surveillance cameras from 2000 to 2020 using data from Suomi24, Finland's largest forum. We begin by formulating a set of pertinent hypotheses and then employ the advanced Snowflake-artic-embedding model to validate these hypotheses. Simultaneously, exploratory data analysis is conducted on the collected dataset to understand trends, thematic distribution, the impact of the GDPR introduction, the effects of Snowden's revelations, and the overall sentiment polarity. The analysis revealed key shifts in sentiment, particularly during global events like the Snowden revelations and the introduction of the GDPR. Moreover, the findings highlight a complex balance between privacy, security, and surveillance, providing insights for policymakers and technology developers aiming to navigate public concerns around surveillance practices.**

***Index Terms*—Surveillance, Sentiment Analysis, Suomi24, Privacy, Security, Corpus analysis, natural language processing.**

I. Introduction

Surveillance cameras, also known as closed-circuit television (CCTV) systems, have become an integral part of modern urban landscapes and have been accompanied by intense debates about privacy, security, and the implications of pervasive monitoring on societal behaviour [1, 2, 3]. While cameras play an important role in enhancing security, these debates have attracted considerable attention in recent years. Technologies such as facial recognition and data analytics have enhanced the capabilities of surveillance systems, amplifying both their potential benefits and risks. Since surveillance cameras are also susceptible to privacy and equity concerns, a study was conducted to estimate their spatial distribution [4]. Their findings reveal that surveillance cameras are disproportionately concentrated in commercial, industrial, and minority-dominated urban zones, raising concerns about the disparate impact of surveillance on different communities [4]. This geographic concentration raises important questions about the equity of surveillance practices and the potential for systemic biases in their deployment.

In contrast, the authors in [5] demonstrated that observation cues, such as eyes and surveillance cameras, enhance the feeling of being watched, which can alter human behavior. For instance, subtle cues like images of eyes or cameras can induce a heightened sense of surveillance, leading to changes in behaviour, such as increased cooperation or self-regulation. These findings suggest that the mere presence of surveillance technologies can have profound psychological effects on individuals, influencing their sense of privacy and autonomy [5]. Duffy and Chan explored how the perception of surveillance on social media platforms can influence user behaviour, suggesting that the awareness of being watched can lead to self-censorship and altered online interactions [6]. This aligns with broader concerns about the hindering effects of surveillance on free expression and the potential for pervasive monitoring to stifle democratic participation. Considering these various concerns, it is crucial to understand how individuals perceive surveillance cameras and the factors that influence these perceptions. Moreover, the availability of large-scale dataset and corpus opens up new horizons in utilizing big-data like approach for uncovering useful insights [7]. The surge in the use of the Internet, social networks, and connected devices has further amplified this trend.

Motivated by the aforementioned discussions, our study leverages insights from Suomi24, Finland's largest online discussion forum, to analyze public opinion toward surveillance cameras. To achieve this, data was collected from the year 2000 to 2020. By examining this data, we aim to uncover how public attitudes towards surveillance have evolved, particularly in response to significant events such as the Snowden revelations and the implementation of the General Data Protection Regulation (GDPR). We also explore the potential impact of social media penetration and its changing landscape. For this purpose, we have drawn a set of hypotheses that explore various dimensions of public sentiment toward surveillance based on a preliminary exploration of Suomi24 data. Specifically, these hypotheses reflect a range of concerns and perceptions commonly discussed in the context of surveillance.

Meanwhile, we integrated a natural language processing (NLP) pipeline employing state-of-the-art text embedding and aspect sentiment analysis to evaluate each hypothesis. In addition, we conducted a temporal and thematic analysis of the Suomi24 data to comprehend the scope and scale of surveillance-related discussions. This analysis also examined the impact of Snowden's revelations on public perception of online surveillance, as well as the introduction of the General Data Protection Regulation (GDPR), among other factors. Since the hypothesis formulation thoroughly explores the scope of the study, we summarize it in the following subsection.

Hypothesis formulation

Personal encounters with surveillance technology, whether positive or negative, can strongly influence individual perceptions. Tene et al. [8] indicate that people who have had direct interactions with surveillance tend to form stronger opinions about its broader societal implications.

H_0: Personal experiences with surveillance cameras significantly shape people's views on public monitoring.

Similarly, in the context of its profound impact on public behavior, a study claimed that people are more accepting of surveillance in high-risk areas, such as shopping centers or transportation hubs, compared to residential or workplace environments where the perceived need for surveillance is lower [9]

H_1: Attitudes toward surveillance cameras vary significantly depending on their placement in different settings, such as shopping areas, residential blocks, and workplaces.

Regarding technical performance issues in surveillance systems, such as poor image quality, camera coverage, and maintenance, these are frequent points of public dissatisfaction. For instance, a study conducted in [10] shows that these technical limitations often lead to calls for improvements or more robust surveillance solutions.

H_2: There are common technical complaints about surveillance cameras, and people suggest specific enhancements to address these issues.

Public attitudes toward surveillance cameras describe the dual perception of these technologies. For instance, while they can enhance security, they also raise concerns about privacy and personal freedoms, as discussed by Foucault [11] and Koskela [12].

H_3: Surveillance cameras make people feel more secure rather than feeling watched.

Although many people agree on the benefits of surveillance for safety, there are also significant concerns about privacy and the potential overreach of state power, leading to debates about how to regulate and implement surveillance in a manner that respects civil liberties [13].

H_4: There is a noticeable relationship between the density of surveillance cameras in an area and public opinion regarding their presence.

Literature on privacy concerns often highlights that individuals perceive surveillance cameras as intrusive, particularly in public spaces where the expectation of privacy is already low [14, 15]. Studies have shown that the awareness of being watched can lead to discomfort and a sense of invasion of privacy [1].

H_5: People feel that surveillance cameras in public spaces negatively impact their privacy.

The impact of surveillance cameras on crime reduction has been a significant area of research, yielding mixed results. Some studies, such as those by Welsh and Farrington [16], suggest that visible surveillance can reduce crime rates by deterring potential offenders. However, other research indicates that the presence of cameras may merely displace crime to less monitored areas [17].

H_6: Individuals believe that surveillance cameras effectively reduce crime in their neighbourhoods and public areas.

The importance of transparency and accountability in fostering trust suggests that organizations must engage with the public and communicate openly [18]. A lack of transparency can lead to suspicion and resistance, whereas clear communication and accountability can enhance public trust [19].

H_7: Trust in the organizations that manage surveillance cameras is influenced by specific factors such as transparency and accountability.

II. Related work and methodological justification

In recent years, the NLP domain has seen significant advancements in techniques for extracting and analyzing large volumes of text data. One of the most influential developments in this domain is the use of text embeddings, such as those generated by models like BERT (Bidirectional Encoder Representations from Transformers) [20], GloVe (Global Vectors for Word Representation) [21], and Snowflake Arctic [22], which are utilized in this study. Text embeddings are particularly effective for capturing the semantic meaning of words in context, which is crucial for analyzing complex discussions, such as those around surveillance cameras, where sentiment and meaning can shift depending on the topic. Research has shown the effectiveness of text embeddings in tasks like sentiment analysis, topic modelling, and text classification [20, 21, 23], making them an ideal choice for our analysis of public opinion mining in Suomi24 discussions.

Aspect-based sentiment analysis has also emerged as a powerful tool in sentiment analysis, particularly when the goal is to understand sentiment at a more granular level than just positive or negative polarities. By dividing the discussion into specific aspects, such as privacy concerns, security benefits, and technical complaints, we can obtain a more detailed and nuanced understanding of public opinion. This approach is particularly useful in contexts where opinions may vary widely depending on the specific issue being discussed. For instance, public sentiment towards surveillance cameras might be positive when considering crime prevention but negative when considering privacy invasion. This method aligns with works like those of Pontiki et al. [24] who demonstrated that aspect-based sentiment analysis could provide valuable insights in complex domains like customer reviews.

To accurately assign text data to the relevant aspects, we employed the commonly employed metric of cosine similarity, between the vectorized representations of text embeddings and predefined aspect embeddings. Research by Mikolov et al. [23], among others, has shown that cosine similarity between words or text embeddings is a reliable method for clustering and classifying text data based on semantic similarity.

Aspect-sentiment using embedding matching is chosen for its robustness and ability to handle the complexities inherent in analyzing large datasets of forum discussions. Traditional methods of sentiment analysis, which often rely on keyword matching or basic lexicon-based approaches [25], might fail to capture the nuanced meanings and contextual dependencies in such data. By contrast, the combination of text embeddings, aspect-based sentiment analysis, and cosine similarity provides a sophisticated and scalable approach to understanding public sentiment at granular level.

In summary, the use of these methodologies allows us to dissect the multifaceted nature of public discourse on surveillance cameras. By leveraging advanced NLP techniques, we aim to provide a more accurate and comprehensive analysis, thereby contributing to the existing literature on surveillance, privacy, and public sentiment.

III. Methodology

Our main objective is to test the research hypotheses and provide a comprehensive understanding of the discussions related to surveillance on Suomi24. To achieve this, our proposed methodology includes the following steps: data collection, data preprocessing, exploratory data analysis (EDA), conducting a literature review for hypothesis generation, hypothesis selection and aspect development, and aspect-based sentiment analysis. Figure 1 provides an overall summary of our approach.

A. Data Collection

The data were collected from Suomi24, which is Finland's largest online discussion platform. In particular, the data spans a period of 20 years, from 2000 to 2020, and captures a wide range of discussions related to surveillance cameras and associated topics. Given that the primary language of the forum is Finnish, data extraction was performed using a carefully curated list of keywords in both English and Finnish to ensure comprehensive coverage of relevant discussions. The keywords were selected to align with the research hypotheses and capture a broad spectrum of discussions on surveillance, privacy, and security. Examples of the keywords used include:

- **English Keywords:** "public camera privacy", "CCTV surveillance," and "security camera."

- **Finnish Keywords:** "julkinen kamera yksityisyyttä" and "kamera valvontaa."

These keywords were used to filter the dataset, extracting specific discussions that directly pertained to the study's focus on surveillance cameras, their impact on public and private spaces, and the concerns or sentiments of people discussing these issues in the forum. Specifically, the extracted data includes discussions from various settings such as public spaces (e.g., streets, shopping areas), residential areas, and workplaces. This wide range of contexts provides a comprehensive view of how surveillance is perceived across different environments, aligning with the research hypotheses related to the placement and density of surveillance cameras.

In order to analyze the posts written in Finnish, we employed the Helsinki-NLP/opus-mt-fi-en model, a transformer-based neural machine translation model developed as part of the Finnish OPUS-MT project. The model is specifically designed for translating between Finnish and English and is known for its high accuracy in maintaining the contextual integrity of the original text. Automating the translation process ensured consistency across the dataset, which was critical for enabling the application of NLP techniques on the forum data. Figure 2 provides sample posts from the Suomi24 forum in their original Finnish, demonstrating a range of perspectives on surveillance cameras—from positive to neutral to negative sentiments. This selection of examples offers a balanced view of public opinion, illustrating the diversity of attitudes toward surveillance technologies. Below the figure, we provide the English translations of these posts to ensure that readers can understand the content and context of the discussions.

This figure helps to contextualize the qualitative data gathered from the forum, providing insights into public sentiment and the nature of discussions surrounding surveillance cameras over the two decades covered by this study.

B. Data Preprocessing

The data preprocessing phase was a crucial step in preparing the extracted discussions from Suomi24 for subsequent analysis. This phase involved key steps, including text normalization and contextual integration, ensuring that the data was in a suitable format for analysis.

1) Text Normalization: Our proposed text normalization phase involved several key steps to ensure that the data was in a suitable format for analysis:

- **Conversion of HTML Entities**: Entities such as '&' (converted to '&') and ''' (converted to " ' ") were present in the text and were appropriately converted to their corresponding characters. Other entities like '<' (representing 'less than sign') and special characters (e.g., '#', '@', '*') were removed to retain only the meaningful content necessary for analysis.
- **Elimination of Non-Alphabetic Characters**: Non-alphabetic characters, except for necessary punctuation, were removed to reduce noise in the data.
- **Lowercasing and Abbreviation Expansion**: All text was converted to lowercase to ensure consistency, particularly for keyword matching and sentiment analysis. Additionally, common abbreviations were expanded to their full forms to avoid ambiguity and ensure clarity during analysis. The specific abbreviations used for expansion were based on a predefined dictionary, which included common English contractions such as "it's" to "it is", "can't" to "cannot". This dictionary was manually curated to address common contractions and abbreviations typically found in informal discussions.
- **Manual Review**: A manual review of approximately 20% of the cleaned and normalized dataset was performed. This involved a cursory examination of the dataset to identify common anomalies and errors not caught by the automated process. While not exhaustive, this step helped address subtle issues that might have affected the analysis.

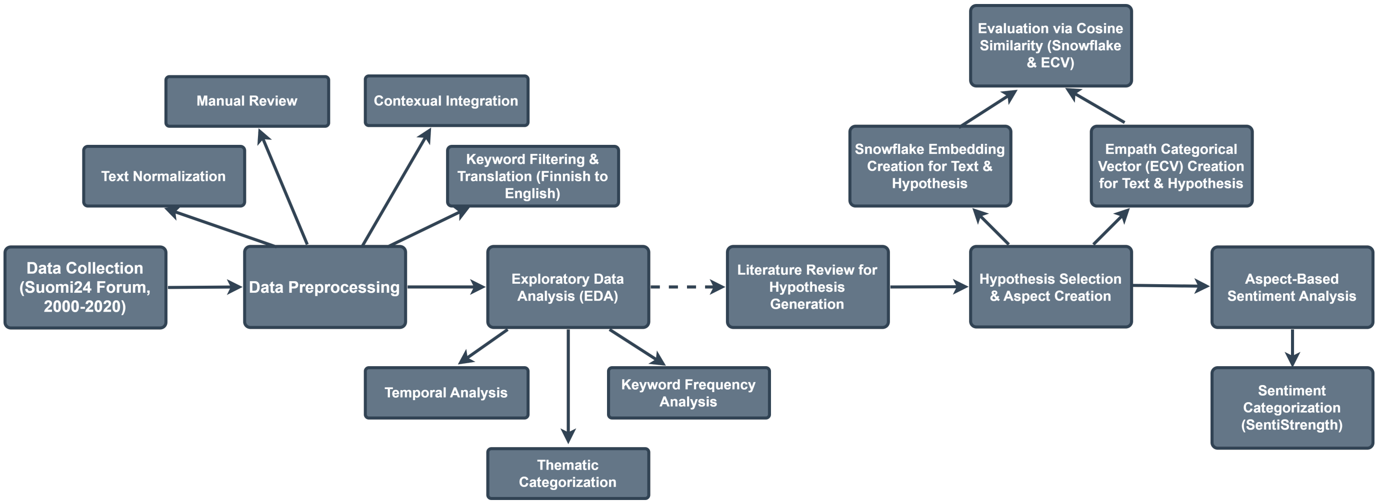

Fig. 1. Data Pipeline for Public Sentiment Analysis on Surveillance Cameras. Horizontal arrows indicate the main steps in the process, while vertical arrows show sub-steps feeding into a larger process. The dashed arrow between EDA and Literature Review highlights that, although conducted independently, EDA insights were essential for Hypothesis Selection and Aspect Creation.

The Python code used for the automated text normalization process is available in a public Github[1] repository.

2) Contextual Integration: To maintain the context of discussions, the titles and texts of posts were combined in cases where the post was identified as the thread starter (referred to as `thread_start` in the dataset). This integration preserved the complete context of the initial post, which often sets the tone and subject of the discussion thread.

For responses to these initial posts, only the comment text was used without the title, as these comments generally relied on the context established by the initial post. A combined text field (`combined_text`) was created where both the title and the text of the post were integrated. This combined field was experimented with during aspect creation and sentiment mining, but it was ultimately not used in the final analysis due to minimal differences observed between using combined and non-combined approaches.

C. Exploratory Data Analysis (EDA)

The EDA provided a foundation for understanding the trends and themes within the discussions on the Suomi24 platform regarding surveillance cameras. EDA involved various methods, including temporal analysis, keyword frequency analysis, and thematic categorization, to uncover patterns in public discourse over the years. The following methods have been used for EDA:

- **Temporal Analysis:** This involved examining the frequency of discussions over time to identify significant peaks and trends related to global and local events.
- **Keyword Frequency Analysis:** Keywords related to surveillance, privacy, and security were analyzed to track their usage over time.
- **Thematic Categorization:** Discussions were categorized into key themes based on the content of posts, such as privacy concerns, security benefits, and technical complaints.

[1]https://github.com/fahadchauhan/Public-Sentiment-on-Security-Cameras

D. Aspect Creation and Hypothesis Selection

In our study, we aimed to explore specific aspects of public discourse related to surveillance cameras on the Suomi24 platform. The creation of these aspects was informed by a combination of data-driven insights and theoretical considerations, resulting in the formulation of a broad set of hypotheses. Initially, eight hypotheses were identified based on recurring themes and topics within the dataset. However, to maintain the focus on the most relevant and impactful aspects, we employed a systematic selection process to narrow these down to the five most pertinent hypotheses.

1) Embedding Creation and Data Representation: To handle the large volume of discussion data effectively, we utilized advanced text embedding techniques. Specifically, we employed the Snowflake Arctic Embedding Model (*Snowflake/snowflake-arctic-embed-s*) to generate embeddings for both the discussion texts and the hypotheses. This model was selected due to its superior performance in capturing nuanced semantic content, which is crucial for analyzing complex discussions [22]. Embedding techniques like these are widely recognized for their ability to represent textual data in a high-dimensional vector space, making them particularly effective for tasks involving semantic similarity [26].

The Snowflake Arctic model has been benchmarked against other leading models, demonstrating its superior performance across key enterprise intelligence tasks, such as Coding (HumanEval+ and MBPP+), SQL Generation (Spider), and Instruction Following (IFEval). These benchmarks have shown that Arctic offers top-tier enterprise intelligence while maintaining a cost-effective training process, outperforming models like Llama 3 8B, Llama 3 70B, and DBRX, especially in terms of performance/cost ratio. These results justify the choice


2020-04-24 09:05:19

Maaseudulla kiertelee paljon rikollista väkeä. Ja itsekin asennutin valvontakamerat viimeisimmän asuntomurron jälkeen. Varoituskyltit kuuluivat kauppaan.
Osa naapureista osti myös kameran tai kamerat, samalla kun asentajat olivat kulmakunnalla. Ja onpa noista kameroista ollut jo paljon hyötyä. Hauskaa Vappua.

Äänestä Kommentoi

2012-03-21 07:59:04

Minulla on autossa kamera, joka tallentaa ajonaikaiset tapahtumat (gps-sijainti, nopeus ja video...). Pienikokoinen eikä herätä kenenkään mielenkintoa.

Äänestä Kommentoi

2003-08-12 20:28:35

on toisen lyöminen, potkiminen ja henkinen väkivaltakin laitonta. En ymmärrä, miksi kutsutaan poliisit paikalle (ehkä) jos aikuinen lyö/potkii toisia tai nappaa toisen omaisuutta itselleen, mutta jos lapsi tekee sitä toiselle, se on vaan sellasta mikä kuuluu asiaan...
Ja mitä ihmeen salaista kukaan voisi tehdä koulussa? vessakoppeihin ei kai kameroita laitettaisi, vaikka niissä todennäköisesti suurimmat mopotukset tapahtuukin!!

Äänestä Kommentoi

Fig. 2. Examples of comments from the Suomi24 forum in Finnish discussing surveillance cameras, with English translations for context:
- Positive example (top): "There are a lot of criminals roaming the countryside. And I myself installed surveillance cameras after the last burglary. The warning signs were part of the deal. Some of the neighbors also bought a camera or cameras, while the installers were around the corner. And those cameras have already been very useful. Happy May Day."
- Neutral example: I have a camera in my car that records driving events (GPS location, speed, and video...). Compact and won't catch anyone's fancy."
- Negative example (bottom): "Hitting, kicking, and mental violence against another is also illegal. I don't understand why the police are called to the scene (maybe) if an adult hits/kicks others or grabs someone else's property, but if a child does it to another, it's just what's appropriate... And what the hell could anyone do in school? I guess you wouldn't put cameras in the toilet cubicles, even though the biggest incidents probably take place there!!"

of the Snowflake Arctic model for our study's embedding creation.

In addition to the Snowflake Arctic embeddings, we also extracted empath categorical vectors (empath embeddings) to enhance the contextual understanding of the text data. These embeddings allowed us to capture emotional and psychological dimensions in the discussions, which are crucial for analyzing public sentiment [27].

2) Hypothesis Selection Process: Given the breadth of the initial hypotheses, a systematic selection process was necessary to ensure that our analysis remained focused and relevant. This process involved using semantic similarity and emotional context to assign data points (i.e., posts and comments) to the most relevant hypotheses.

We utilized cosine similarity measures to compare the embeddings of the discussion posts with the embeddings of each hypothesis. Cosine similarity is a widely used metric in information retrieval and text mining for measuring the similarity between two vectors in a multi-dimensional space [26]. This method was applied in two key steps:

Text Embedding Similarity: The initial step in the hypothesis selection process involves calculating the cosine similarity between the snowflake embeddings of the posts and those of each hypothesis. Let (T_s) represent the snowflake embedding of the text from a post, and $(H_s(i))$ denote the snowflake embedding of the i-th hypothesis. These embeddings capture the semantic content of the post and hypothesis, respectively. The cosine similarity between these vectors is calculated as shown in Eq. (1):

$$\cos(T_s, H_s(i)) = \frac{T_s \cdot H_s(i)}{\|T_s\| \|H_s(i)\|} \tag{1}$$

Here, $\cdot$ represents the dot product of the vectors, while $\|T_s\|$ and $\|H_s(i)\|$ denote the magnitudes of the vectors. This similarity measure helps to determine how closely the post relates to each hypothesis. Given a set of hypotheses, the post is assigned to the hypothesis that yields the highest cosine similarity score as shown in Eq. (2):

$$H_k = \text{argmax}_{i \in I} \left(\cos(T_s, H_s(i))\right) \tag{2}$$

where H_k is the most relevant hypothesis for the given post.

Empath Embedding Refinement: In cases where the similarity scores for multiple hypotheses are close, within a margin (δ) of 0.3, the empath embeddings of the post and the hypotheses are used for further refinement. Let T_e and $H_e(i)$ represent the empath embeddings of the text and the i-th hypothesis, respectively. The margin of 0.3 was selected based on a statistical analysis of the cosine similarity scores across the dataset. Specifically, this value corresponds approximately to the third quantile of the observed cosine similarity distribution, which was determined to effectively distinguish between semantically close yet distinct hypotheses. This approach is inspired by techniques in statistical text mining where quantile-based thresholds are commonly used to balance precision and recall in classification tasks [28]. These embeddings capture the emotional or psychological dimensions of the content. The cosine similarity between the empath embeddings is computed similarly as shown in Eq. (3):

$$\cos(T_e, H_e(i)) = \frac{T_e \cdot H_e(i)}{\|T_e\| \|H_e(i)\|} \tag{3}$$

To finalize the assignment, a combined similarity score is calculated by averaging the text and empath embedding similarities as shown in Eq. (4):

$$\text{CS(s,i)} = \frac{\cos(T_s, H_s(i)) + \cos(T_e, H_e(i))}{2} \tag{4}$$

The post is then assigned to the hypothesis with the highest combined score as shown in Eq. (5):

$$H_k = \text{argmax}_{i \in I} \left(\text{CS(s,i)}\right) \tag{5}$$

This detailed process ensures that each post from the Suomi24 dataset is accurately categorized under the hypothesis that most closely aligns with its content, thereby enabling a focused and relevant analysis of public sentiment towards surveillance cameras.

Pseudocode for Hypothesis Assignment Process: The pseudocode below outlines the process used to assign posts to the most relevant hypothesis:

Algorithm 1 Hypothesis Assignment Process

Inputs: T_s (text snowflake embedding), T_e (text empath embedding), δ (margin)
Output: Hypothesis index H_k
$sim_s = cosine\_similarity(T_s, H_s(i))$
$sorted\_sim = \text{sort}(sim_s, \text{descend} = \text{True})$
if $(sorted\_sim[0] - sorted\_sim[1]) \leq \delta$ **then**
  $sim_e = cosine\_similarity(T_e, H_e(i))$
  $combined\_sim = (sim_s + sim_e)/2$
  $H_k = argmax(combined\_sim)$
else
  $H_k = argmax(sorted\_sim)$
end if
Return H_k

E. Sentiment Analysis

Sentiment analysis was conducted to assess the emotional tone of discussions on the Suomi24 platform, specifically relating to surveillance cameras. The sentiment analysis aimed to categorize the sentiments expressed in the posts into three categories: positive, neutral, and negative. This analysis provides insights into how public opinion and emotions towards surveillance have evolved over time. For this study, **SentiStrength** was selected as the primary tool for sentiment analysis due to its capability to accurately capture sentiment nuances in social media and forum discussions. SentiStrength is known for its effectiveness in analyzing sentiment in short texts, such as the posts and comments on Suomi24, and has been validated in numerous studies for its accuracy in classifying sentiments as positive, negative, or neutral.

The choice of SentiStrength was further evaluated by manual verification of its outputs on a sample of the dataset, where it demonstrated superior performance in identifying the emotional tone consistent with the context of the discussions. As the dataset had been translated into English, SentiStrength was particularly suited to handling the text, ensuring that the sentiment of discussions was accurately captured.

IV. Results

In this section, we present key findings from the analysis, including trends identified during the EDA, the distribution of sentiments across different aspects, and detailed insights from sentiment analysis.

A. Results from Exploratory Data Analysis

The EDA revealed several significant trends and patterns in public discussions surrounding surveillance over the past two decades. Key findings include:

1) Early Internet Era and Initial Discussions (2001-2004): The early years of the dataset, particularly from 2001 to 2004, show relatively few discussions, as depicted in both the monthly trend analysis (Figure 3) and the heatmap (Figure 4). This low volume can be attributed to the limited internet penetration in Finland and the early development stage of online forums. However, the global context of the 9/11 terrorist attacks in 2001 sparked a significant shift in public consciousness about surveillance and security, which began to manifest in online discussions as internet adoption increased.

2) Growth of Online Discourse and Influence of Global Events (2004-2010): From 2004 onwards, there is a noticeable increase in discussions, coinciding with broader internet adoption and increasing public awareness of surveillance issues. The significant peak around 2010 likely corresponds to several factors, including Finland's technological advancements and the growing public discourse driven by platforms such as social media. The WikiLeaks disclosures in 2010, which exposed sensitive government data, may have also heightened public concerns about privacy and surveillance, contributing to the surge in discussions.

3) The Snowden Revelations and Their Aftermath (2013–2015): The most pronounced increase in discussion volume occurs between 2013 and 2015, as seen across various visualizations. This period aligns with the Snowden revelations in 2013, which had a global impact by revealing the extent of government surveillance programs [29]. The revelations led to widespread public outrage and intense debates about privacy, security, and the role of the state in monitoring citizens. This surge in interest is reflected in two key trends. Firstly, the yearly trends of Snowden discussions (Figure 5) show a significant spike during this period, indicating an increased public focus on Snowden's disclosures and the broader issues of state surveillance. Secondly, the yearly trends of camera sightings (Figure 5) reveal a parallel rise in mentions of surveillance cameras. This suggests a heightened public awareness and concern regarding physical surveillance technologies as the implications of the Snowden revelations permeated societal consciousness. Together, these trends highlight the lasting influence of

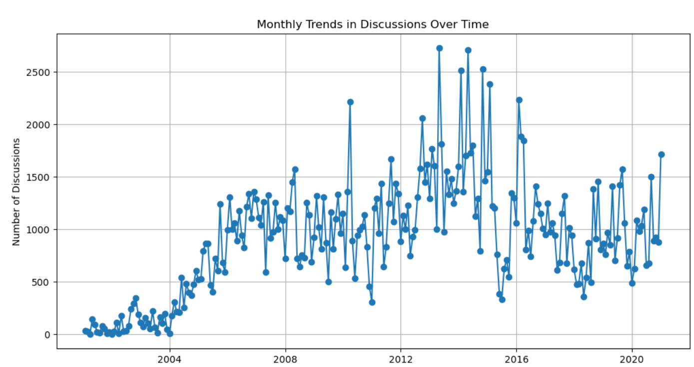

Fig. 3. Monthly Trends in Discussions Over Time

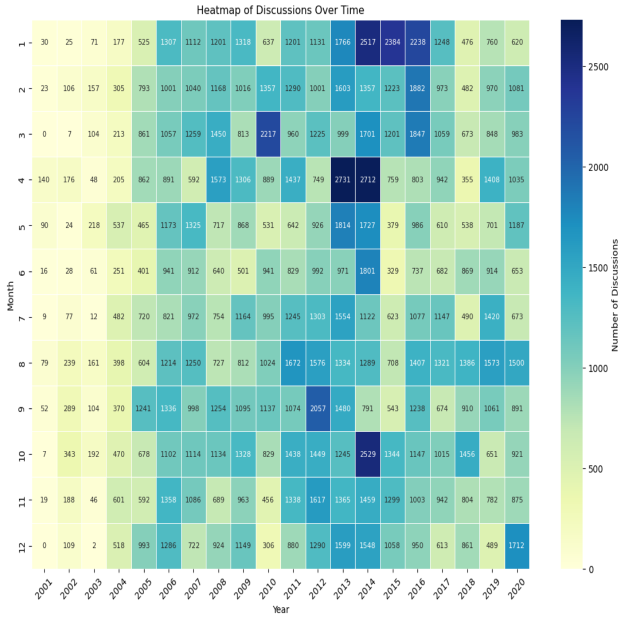

Fig. 4. Heatmap of Discussions Over Time

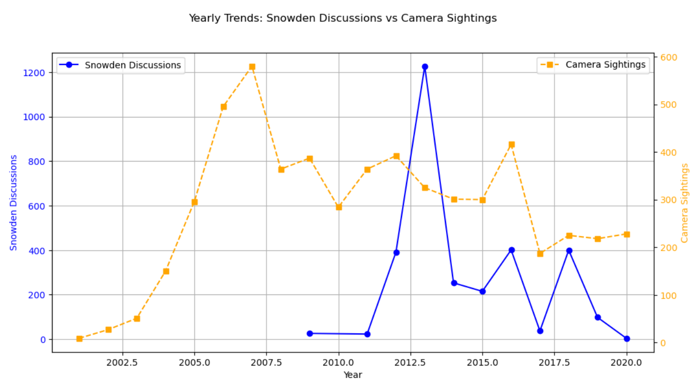

Fig. 5. Yearly Trends of Camera Sightings and Snowden Discussions

Snowden's disclosures on public discourse about surveillance, both digital and physical.

4) The Impact of GDPR and the Maturation of Public Discourse (2016-Present): Post-2016, the introduction of the GDPR by the European Union appears to have influenced a decline in discussions. The GDPR, formally adopted on April 27, 2016, and enforced from May 25, 2018, represented a significant milestone in data protection and privacy law[2]. It established comprehensive rules regarding the handling of personal data, granting individuals greater control over their data and imposing strict requirements on organizations that process this data.

The GDPR's emphasis on transparency, accountability, and individual rights likely played a role in reducing public anxiety about surveillance, as reflected in the tapering off discussion volumes in subsequent years. The regulation's impact is seen not only in the reduced frequency of discussions but also in the nature of the discourse, which became more focused on the legal and ethical implications of data privacy rather than on surveillance per se.

5) Overall Analysis of Key Findings from EDA: The exploratory analysis of the Suomi24 dataset has revealed several significant trends and patterns in public discussions surrounding surveillance over the past two decades. The data shows a noticeable increase in discussions following key global and local events, such as the 9/11 terrorist attacks and the introduction of GDPR, underscoring the public's heightened awareness and concern about surveillance practices. This claim is substantiated by observed spikes in discussion volumes immediately following these events, along with a substantial increase in the use of keywords related to privacy and surveillance in forum posts during these periods. For instance, the frequency of terms such as "privacy," "CCTV," and "surveillance" surged particularly after the Snowden revelations in 2013, reflecting a growing public discourse around the implications of surveillance technologies. Additionally, qualitative analysis of the forum content highlights recurring themes of concern regarding privacy invasion and the impact of pervasive surveillance on civil liberties, further indicating that these topics have become increasingly central to public awareness and debate.

The heatmap and monthly trends analyses highlighted the fluctuating nature of these discussions, with notable spikes corresponding to major privacy and security incidents. For example, the introduction of the USA PATRIOT Act following the 9/11 terrorist attacks led to widespread public discourse on the balance between national security and individual privacy, as the Act significantly expanded the government's surveillance capabilities. Similarly, the Snowden revelations in 2013, which exposed the extent of global surveillance programs by the NSA, triggered a surge in discussions around privacy, government overreach, and the ethics of mass data collection. These incidents catalyzed public concern and debate, reflected in the increased frequency of forum discussions during these periods. Additionally, the early surge in camera sighting discussions around 2007 aligns with the growing implementation of surveillance technologies, such as the increased use of CCTV in public spaces following heightened security measures. The subsequent stabilization in these discussions may reflect a possible normalization of these practices in the public consciousness, where surveillance technologies have become a routine aspect of everyday life, despite ongoing concerns about their implications for privacy and civil liberties.

These insights from the EDA form a critical foundation for the subsequent hypothesis testing and sentiment analysis, as they provide context for understanding the shifts in public discourse on surveillance and privacy over time.

B. Results from Aspect Creation and Hypothesis Selection

The final distribution of data across the five selected hypotheses, which had sufficient relevant data, is illustrated in Figure 6, demonstrating the relative prevalence of each aspect within the discussions. These five hypotheses are referred to

[2]https://www.edps.europa.eu/data-protection/data-protection/legislation/history-general-data-protection-regulation_en

as H_0, H_1, H_2, H_3, and H_4, as previously outlined in the hypothesis formulation subsection.

The final selection of these five hypotheses represents the core themes identified in public discussions on the Suomi24 platform. Notably, the hypothesis concerning technical complaints and enhancements (H_2) had the highest data frequency, reflecting the practical nature of many discussions. Conversely, the hypothesis focusing on personal experiences with surveillance cameras (H_0) had the lowest frequency, likely due to the subjective and private nature of such discussions.

These selected hypotheses provide a focused framework for analyzing how public sentiment towards surveillance cameras has evolved over the past two decades.

C. Sentiment Analysis Results

The sentiment analysis revealed key trends in public sentiment regarding surveillance cameras over time and across different aspects:

1) Yearly Trend of Sentiments: The sentiment analysis reveals key trends in public sentiment regarding surveillance cameras over time. Notably, there is a general increase in both positive and negative sentiments, with significant peaks corresponding to global and local events. For instance, the marked increase in negative sentiments during 2013-2015 aligns with the Snowden revelations and the subsequent public debate on surveillance practices, as shown in Figure 7.

Further analysis of the sentiment trends across hypotheses reveals more granular insights. Figure 8 illustrates the yearly trends of sentiments specifically for each hypothesis, indicating how public opinion has shifted over time. For example, the hypothesis related to the perception of surveillance cameras as enhancing security (H3) shows a steady increase in positive sentiment up until 2015, after which it stabilizes, reflecting a normalization of the technology's presence in public spaces.

Figure 9 highlights the yearly trends of negative sentiments across hypotheses, showing a marked increase in discussions concerning technical complaints (H2) and privacy concerns (H4) following the Snowden revelations. This suggests that these aspects of surveillance are particularly sensitive to public sentiment shifts in response to global surveillance incidents.

2) Distribution of Sentiment Labels per Hypothesis: The distribution of sentiment labels (positive, negative, neutral) across the five hypotheses reveals significant differences in public sentiment. As shown in Figure 10, the data indicates varying sentiment trends across different aspects of surveillance.

V. Discussion

This study provides a comprehensive analysis of public sentiment toward surveillance cameras, as captured in discussions from the Suomi24 platform. The sentiment analysis, conducted using SentiStrength, revealed distinct patterns across different aspects of surveillance, including technical complaints, placement concerns, and the security vs. privacy debate.

A. Interpretation of the Results

The analysis highlights key insights into public sentiment across the different hypotheses:

- H_0**: Personal Experiences with Surveillance Cameras** revealed a balanced sentiment distribution, highlighting the varied nature of personal encounters with surveillance.
- H_1**: Placement of Surveillance Cameras** attracted significant negative sentiment, indicating public unease with camera placement. This fluctuating sentiment suggests an ongoing concern, especially in sensitive areas like residential blocks or workplaces.
- H_2**: Technical Complaints and Enhancements** emerged as the most negatively perceived aspect. This likely reflects widespread frustration with the functionality of surveillance cameras, such as issues with video quality or maintenance. The peak in negative sentiment during 2013-2015 corresponds with the Snowden revelations, which brought attention to both governmental and technological shortcomings.
- H_3**: Security vs. Privacy Concerns** saw higher levels of positive sentiment, reflecting the public's recognition

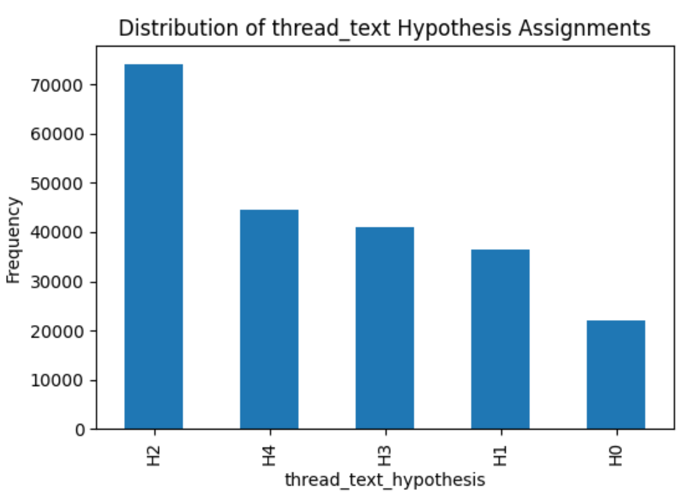

Fig. 6. Distribution of Hypothesis Assignments in the Data

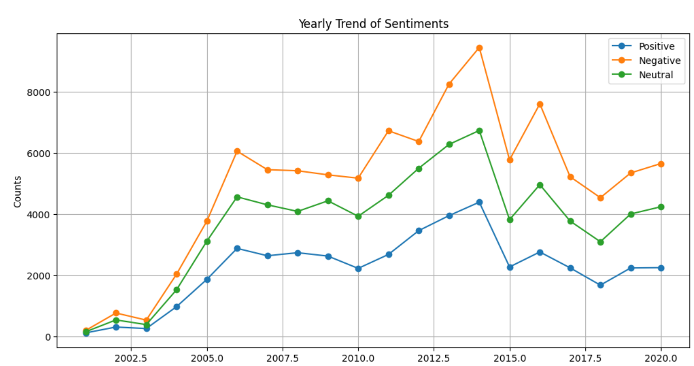

Fig. 7. Yearly Trend of Sentiments (SentiStrength)

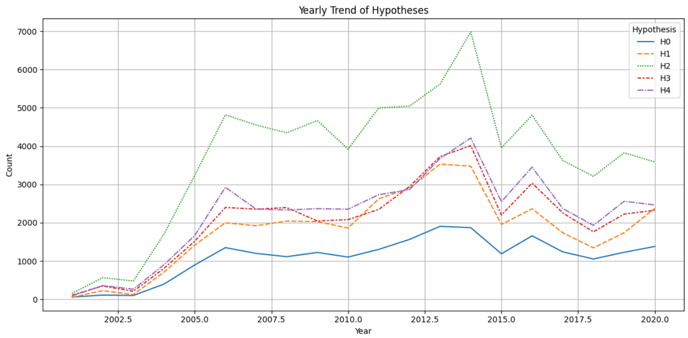

Fig. 8. Yearly Trends of Hypotheses Sentiments (Positive, Negative, Neutral)

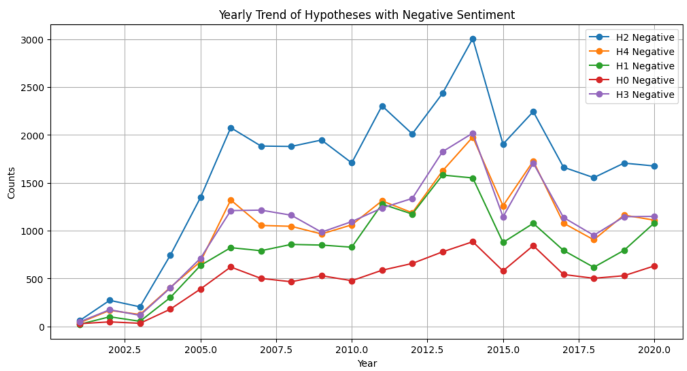

Fig. 9. Yearly Trends of Hypotheses with Negative Sentiment

of the security benefits of surveillance cameras, though tempered by privacy concerns.

- H_4: **Density of Cameras and Public Opinion** showed mixed sentiments, indicative of ongoing public debate on whether higher camera density equates to better security or constitutes over-surveillance.

B. Comparison with Related Work

Our findings align with and extend those of prior studies in the field of surveillance research. The emphasis on technical complaints about surveillance cameras mirrors results from Gill and Spriggs (2005), who also identified significant public dissatisfaction with the quality and maintenance of surveillance technology. Similarly, the concerns regarding privacy invasion, highlighted in our study, corroborate earlier work by Lyon (2001), who discussed the growing public unease with pervasive monitoring systems.

The debate between security and privacy, which emerged as a key theme in our analysis, has also been a prominent topic in related literature. Welsh and Farrington (2009) explored this tension in their review of CCTV effectiveness, noting that while surveillance can enhance security, it often does so at the expense of personal privacy. Our findings support this view, as discussions frequently centered around the trade-offs between these two priorities. Furthermore, our use of advanced embedding techniques to analyze large volumes of text data contributes to the methodological advancements in this field. Previous studies, e.g., [23] demonstrated the efficacy of text

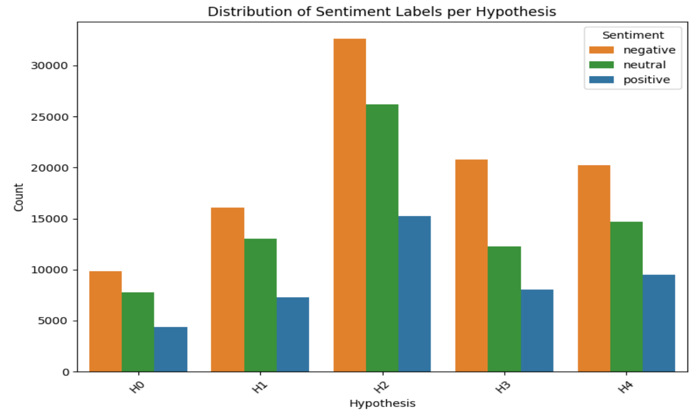

Fig. 10. Distribution of Sentiment Labels per Hypothesis

embeddings in capturing semantic content, which is critical for nuanced sentiment analysis. By employing the Snowflake Arctic Embedding Model, we were able to achieve a higher level of precision in categorizing public sentiment, further validating the utility of this approach in surveillance studies.

Finally, our results regarding the normalization of surveillance practices over time, particularly following significant global events like the Snowden revelations, resonate with trends observed in other studies. For example, Greenwald's analysis [29] of the Snowden leaks also highlighted a shift in public discourse towards a more critical stance on government surveillance. Our findings suggest that while initial reactions to such events are strong, there is a gradual acceptance or normalization of surveillance technologies, which aligns with broader societal trends discussed in the literature.

C. Implications of the Findings

The findings have important implications for both policymakers and technology developers. The negative sentiment towards technical issues (H_2) highlights the need for improved surveillance technologies and better public communication. The mixed reactions to camera placement and density (H_1 and H_4) suggest that careful consideration must be given to how and where cameras are deployed to balance security benefits with privacy concerns.

D. Limitations of the Study and Future research

Our proposed study has several limitations, which we acknowledge as follows:

- **Data Scope:** The translation from Finnish to English may have affected sentiment nuance. Although, we believe such impact is very minor after careful manual examination on a small sample.
- **Sentiment Analysis Tools:** While SentiStrength performed well, no automated tool can fully capture the complexity of human emotions, and its inherent limitations are well-documented.
- **Temporal Focus:** The study focused on data up to 2020, potentially missing trends related to newer surveillance technologies, which are ultimately worth investigating.

Future research could expand on this study by i) conducting cross-cultural comparisons to see if trends observed in Finland hold true globally; ii) performing longitudinal studies to track changes in sentiment as surveillance technology and policies evolve, and iii) integrating qualitative insights for a deeper understanding of public sentiment.

Acknowledgment

This is supported by Kone foundation UNDO project and AoF Fellowship No.330345

References

[1] D. Lyon, *Surveillance Society: Monitoring Everyday Life*. Buckingham, UK: Open University Press, 2001.

[2] L. D. Introna and D. Wood, "Picturing algorithmic surveillance: The politics of facial recognition systems," *Surveillance & Society*, vol. 2(2/3), pp. 177–198, 2004.

[3] C. Norris and G. Armstrong, *The Maximum Surveillance Society: The Rise of CCTV*. Oxford: Berg Publishers, 1999.

[4] Y. Sheng, D. Wong, J. Sanyal, and L. Wiggins, "Surveilling surveillance: Estimating the prevalence of surveillance cameras with street view data," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–23, 2021.

[5] V. M. Hesslinger, L. Goldbach, and C. C. Carbon, "The sense of being watched is modulated by arousal and duration of the perceptual episode," *Scientific Reports*, vol. 7, no. 1, pp. 1–12, 2017.

[6] B. E. Duffy and N. K. Chan, "You never really know who's looking: Imagined surveillance across social media platforms," *New Media & Society*, vol. 21, no. 1, pp. 119–138, 2018.

[7] D. Laney, "3d data management: Controlling data volume, velocity, and variety," META Group Research Note, Tech. Rep., 2001.

[8] O. Tene and J. Polonetsky, "Privacy in the age of big data: A time for big decisions," *Stanford Law Review Online*, vol. 64, pp. 63–69, 2012.

[9] A. Smith, "The impact of cctv on privacy and civil liberties," *British Journal of Criminology*, vol. 54, no. 2, pp. 280–298, 2014.

[10] M. Gill, *CCTV Handbook*. UK: Perpetuity Press, 2005.

[11] M. Foucault, *Discipline and Punish: The Birth of the Prison*. Pantheon Books, 1977.

[12] H. Koskela, "The gaze without eyes: Video-surveillance and the changing nature of urban space," *Progress in Human Geography*, vol. 24, no. 2, pp. 243–265, 2000.

[13] D. Lyon, "The search for surveillance theories," in *Theorizing Surveillance: The Panopticon and Beyond*, D. Lyon, Ed. Willan Publishing, 2006, pp. 3–20.

[14] D. J. Solove, "I've got nothing to hide and other misunderstandings of privacy," *San Diego Law Review*, vol. 44, pp. 745–772, 2007.

[15] H. Nissenbaum, "Privacy as contextual integrity," *Washington Law Review*, vol. 79, pp. 119–158, 2004.

[16] B. C. Welsh and D. P. Farrington, "Public area cctv and crime prevention: an updated systematic review and meta-analysis," *Justice Quarterly*, vol. 26, no. 4, pp. 716–745, 2009.

[17] M. Gill and A. Spriggs, "Assessing the impact of cctv," Home Office Research, Development and Statistics Directorate, Tech. Rep. Home Office Research 292, 2005.

[18] J. A. Colquitt and J. B. Rodell, "Justice, trust, and trustworthiness: A longitudinal analysis integrating three theoretical perspectives," *Academy of Management Journal*, vol. 54, no. 6, pp. 1183–1206, 2011.

[19] K. Murphy, "Procedural justice and the regulation of policing," *Criminology & Criminal Justice*, vol. 7, no. 2, pp. 223–238, 2007.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[21] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1532–1543.

[22] Snowflake AI Research Team, "Snowflake arctic: The best llm for enterprise ai, 2024," https://www.snowflake.com/en/blog/arctic-open-efficient-foundation-language-models-snowflake/.

[23] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. of International Conference on Learning Representations (ICLR)*, 2013, pp. 1–12.

[24] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "Semeval-2014 task 4: Aspect based sentiment analysis," in *Proc. of International Workshop on Semantic Evaluation*, 2014, pp. 27–35.

[25] S. Dasgupta, Y. Himeur, and S. Al-Khayatt, "Analyzing sentiments in e-reputation management using lexicon-based approach: A big data context," in *11th International Conference on Innovations in Information Technology*, 2015, pp. 218–223.

[26] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd ed. Prentice Hall, 2009.

[27] E. Fast, B. Chen, and M. S. Bernstein, "Empath: Understanding topic signals in large-scale text," *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 4647–4657, 2016.

[28] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, "Performance measures for information extraction," in *Proceedings of DARPA Broadcast News Workshop*, 1999, pp. 249–252.

[29] G. Greenwald, *No Place to Hide: Edward Snowden, the NSA, and the U.S. Surveillance State*. Metropolitan Books, 2014.