Social Network Analysis

Professor: Qasim Pasta

Homework 2

Analysis of Animal Abuse Media Coverage in Dawn (2010-2024)

Group Members:

Syeda Hoorain Imran si08704

Muhammad Munib Sattar ms09239

Habib University

November 16, 2025

Table Of Contents:

1.0 Introduction

Public and media discourse surrounding animal welfare has become an increasingly significant indicator of social and legal change. In Pakistan, media coverage plays a crucial role in highlighting animal welfare issues, shaping public opinion, and influencing policy. This report provides a comprehensive analysis of media coverage on "animal abuse" from *Dawn*, one of the nation's most prominent English-language news outlets.

Using a dataset of article links spanning from 2010 to 2024, this study employs a computational methodology to understand the evolution of this narrative. The full text of each article was first extracted using web scraping techniques. This raw data was then systematically processed, cleaned, and lemmatized to build a structured analytical corpus, enabling a data-driven examination of the topic.

This report uses a multi-faceted approach to analyze the corpus. It begins with a thematic analysis of term frequency to identify the most prominent topics. This is followed by a detailed temporal and sentiment analysis, which correlates peaks in media coverage and emotional tone with significant real-world events, such as key legal cases and public advocacy campaigns. Finally, a co-occurrence network analysis is used to visualize the relationships between core themes, revealing the underlying structure of the media narrative. Through this analysis, the report aims to provide a clear, empirical picture of how the framing of animal abuse has evolved in Pakistan's media over the last decade.

2.0 Methodology

This study employed a computational text analysis approach, executed in the RStudio environment, to collect, clean, and analyze the media corpus. The methodology was divided into three distinct phases: (1) data collection and preparation, (2) text processing and cleaning, and (3) data analysis and visualization.

2.1 Data Collection and Preparation

The project began with the dataset provided by the assignment, which contained a list of *Dawn* article URLs related to "animal abuse" from 2010 to 2024. A preliminary manual verification of this dataset was conducted. It was observed that several older URLs were outdated and led to "Not Found" errors. A manual search was performed on the *Dawn* website to locate the new, active links for these articles, which were then updated in our dataset. Only one article link could not be found and was excluded.

With a clean list of URLs, we used a custom R script to systematically scrape the full text of each article.

2.2 Text Processing and Cleaning

Once the raw text corpus was collected, it was passed through a comprehensive cleaning pipeline. This process was essential to standardize the text for accurate analysis.

1. **Normalization:** All text was converted to **lowercase**, and common non-standard characters were normalized.
2. **Noise Removal:** Irrelevant "noise" was removed from the text, including all HTML links, numbers, and punctuation.
3. **Tokenization:** The clean text from each article was "tokenized", that is, broken down from paragraphs into a list of individual words.
4. **Stopword Removal:** Common English stopwords (e.g., "the," "is," "at," "in") were removed using the standard `tidytext` stopword library.
5. **Lemmatization:** Finally, the words were **lemmatized** using the **`textstem`** package. This process converts words to their core dictionary form (e.g., "animals" and "animal's" both become "animal"). We preferred this over stemming because this method seemed more accurate for grouping and counting terms that share the same root meaning.

2.3 Data Analysis and Visualization

The final, cleaned dataset was used to conduct four types of analysis:

1. **Frequency Analysis:** We calculated the raw frequency of all terms to identify the most dominant words in the corpus. We also calculated prevalence (the percentage of articles a word appeared in) to determine the most significant and widespread topics.
2. **Temporal & Sentiment Analysis:** Word frequencies were grouped by publication year to create time-series graphs. This allowed us to track the rise and fall of specific topics and correlate them with real-world events. In parallel, we used a sentiment lexicon to score the positive and negative emotional language in the articles, plotting the change in sentiment over time.
3. **Network Analysis:** A co-occurrence network was built through R to visualize the relationships between terms using Gephi.
    - **Nodes:** The 60 most frequent words in the corpus.
    - **Edges:** A connection was drawn between two words if they appeared in the same article together at least twice.
    - The resulting graph data was exported and visualized in the Gephi software using the Force Atlas 2 layout algorithm, which clusters nodes based on the strength of their connections.

3.0 Results and Analysis

This section presents the findings from the data analysis, structured into four parts: a thematic overview of the most common topics, a detailed temporal analysis correlating word trends with real-world events, an analysis of the corpus's emotional sentiment over time, and a network analysis to map the relationships between topics.

3.1 Thematic Overview: The Dominance of Legal and Urban Issues

The initial frequency analysis of the corpus identifies the most prominent terms, providing a "big picture" of the conversation.

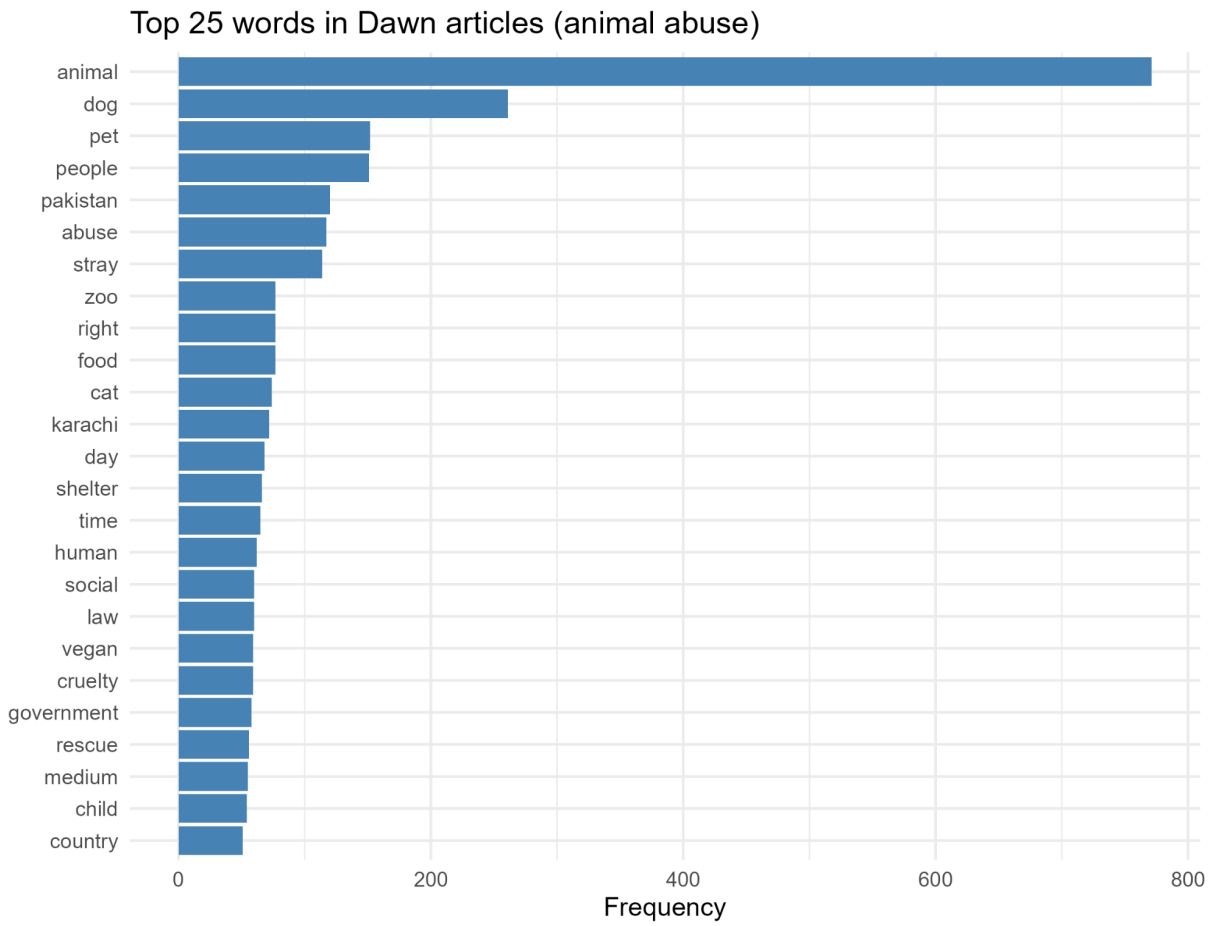
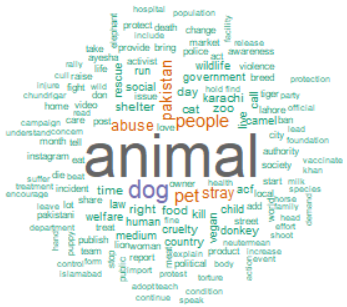


Figure 1: Top 25 Most Frequent Words in the Corpus. This chart displays the raw frequency of the most common words.



As Figure 1 clearly demonstrates, the dataset is dominated by the core terms of the topic: **`animal`** (n=1599), **`people`** (n=151), and **`abuse`** (n=117). This is expected, as it forms the very basis of the topic.

However, a more detailed look at the word frequency data reveals a more nuanced picture. Beyond these primary terms, a secondary set of high-frequency words emerges, pointing to specific, concrete narratives: **`law`**, **`dog`**, **`stray`**, **`karachi`**, **`government`**, **`rights`**.

The prominence of these terms strongly indicates that the media discussion in *Dawn* is not limited to abstract acts of cruelty. Instead, it is a concrete conversation deeply connected with the **legal and state response** to specific and recurring **urban issues**. The word cloud reinforces this by visually elevating these terms.

3.2 Temporal & Event-Driven Analysis

The temporal analysis provides context for when these conversations occurred and how their focus shifted over the last decade.

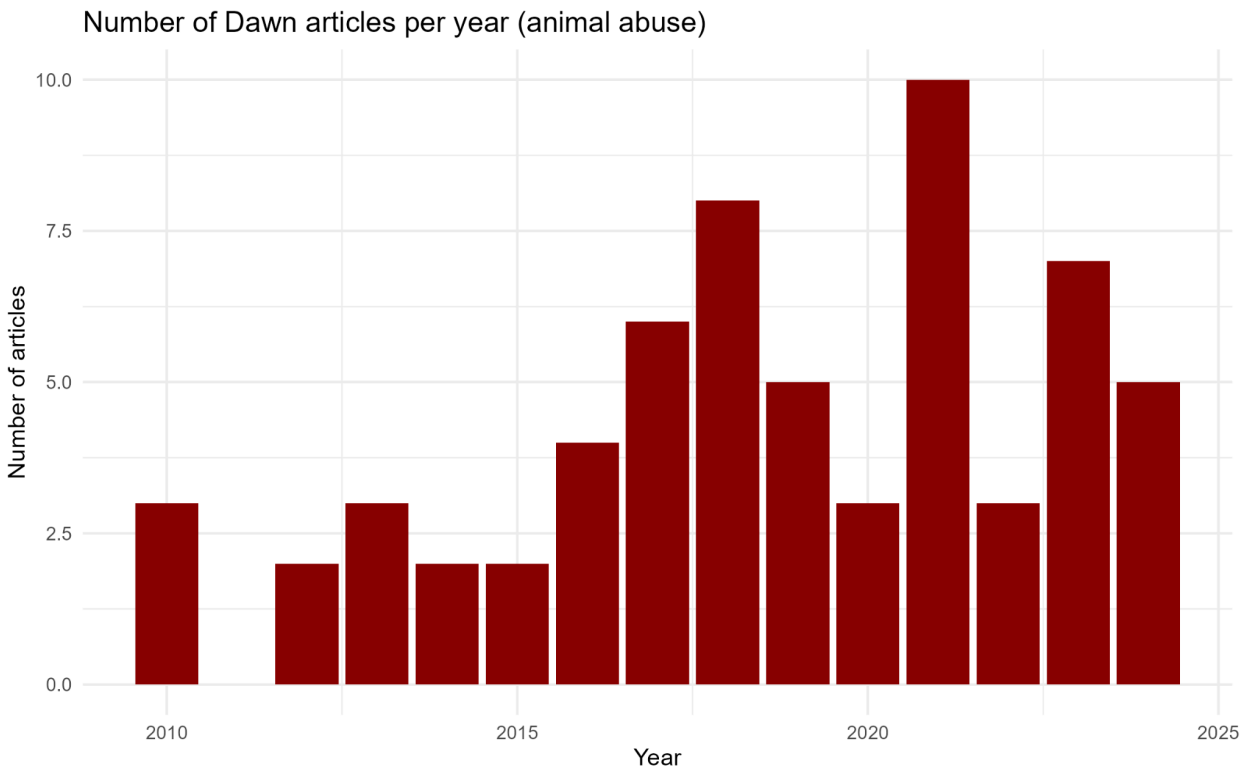


Figure 2: Number of Scraped Articles per Year. This chart shows the total volume of articles in the final, cleaned corpus by publication year, indicating peaks in media attention.

The temporal analysis reveals that the conversation was not consistent, but event-driven. As seen in Figure 2, the volume of articles was relatively sparse in the early 2010s, reflecting a period where animal welfare was a low-priority topic.

The media landscape began to change around 2017-2018, correlating with the 2018 federal amendment that increased fines for cruelty in Islamabad, as well as the widespread public outrage from the **2018 "Hero the donkey" torture incident**, which galvanized public opinion and media focus. However, the most significant finding is the **dramatic spike in articles in 2021**, which represents the absolute peak of the entire dataset. This spike is not an anomaly. It is a direct reflection of real-world events, such as the “stray dog culling controversy” in 2021.

The new faceted frequency graphs allow us to dissect this 2021 spike:

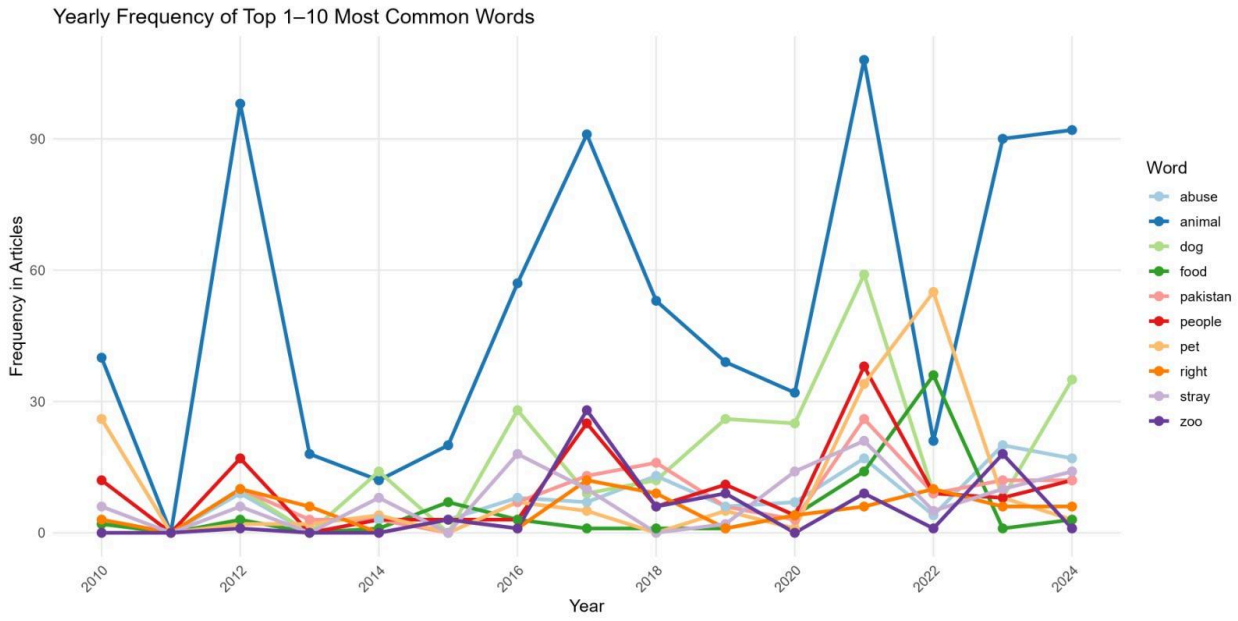


Figure 3: Yearly Frequency of "Core Theme" Words (Ranked 1-10).

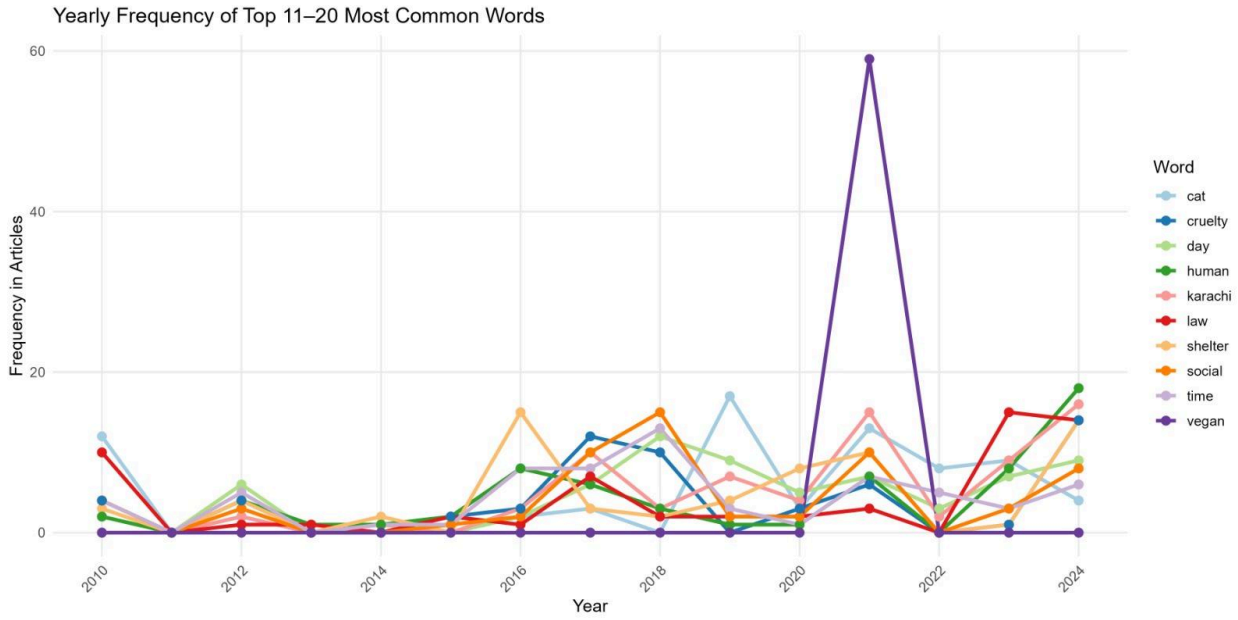


Figure 4: Yearly Frequency of "Mechanism & Society" Words (Ranked 11-20).

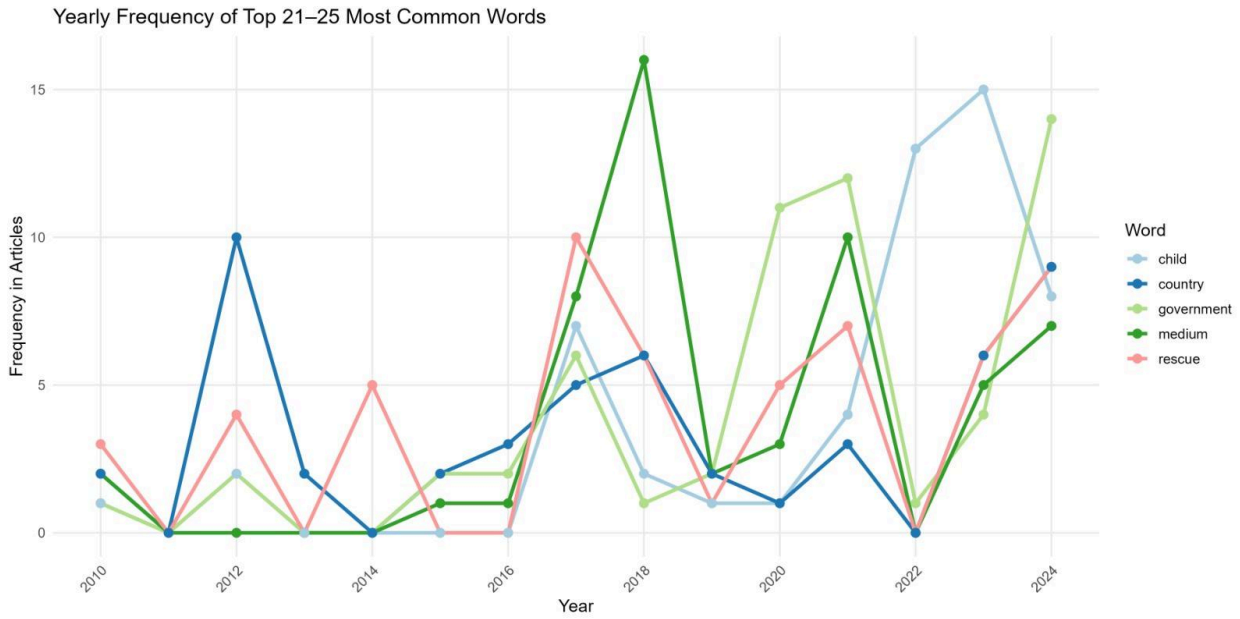


Figure 5: Yearly Frequency of "Specific Solution" Words (Ranked 21-25).

The above figures show that the 2021 spike wasn't driven by a single niche topic; it was an explosion of the entire core conversation. The terms `cruelty`, `dog`, `stray`, and `karachi` all peak or increase simultaneously. This shows a "perfect storm" of media coverage where the incident (cruelty), the subject (stray dogs in Karachi), and the proposed solution (law) all converged. This spike is directly correlated with intense public debates over mass dog culling in Sindh and the subsequent passing of a new animal welfare bill in the Sindh Assembly.

These graphs provide proof of the legal focus as well. The terms **`rescue`**, **`law`**, and **`zoo`** all show a distinct jump and sustained high frequency from 2021 onwards. This directly reflects the media's new focus on the legislative process and court cases, such as the 2021 Islamabad High Court ruling on the zoo. We also see the term like **`shelter`** begin its rise post-2021, indicating a shift from identifying the problem to discussing solutions. Furthermore, the term **`government`** appears in a relatively high frequency from 2019 onwards (with Covid being the exception), directly corresponding to the legislative efforts in Sindh (2019-2021) and Khyber Pakhtunkhwa (2024).

3.3 Sentiment Analysis: A Shift in Tone

This analysis of the corpus's emotional tone provides a new layer to the temporal findings.

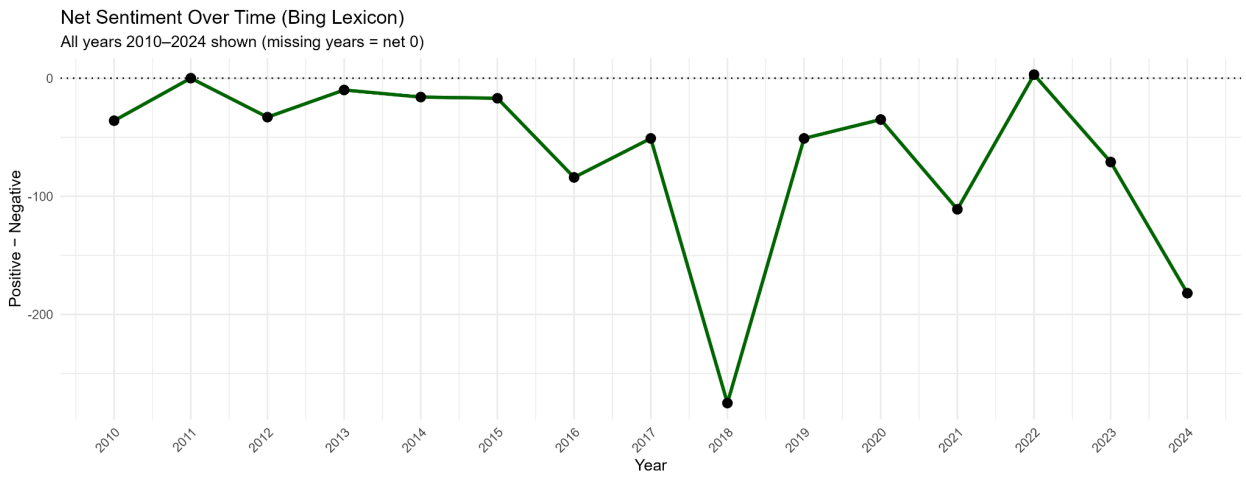


Figure 6: Change in Negative and Positive Sentiment Over Time.

As expected for a dataset on "animal abuse," Figure 6 shows that **negative sentiment consistently and significantly outweighs positive sentiment** across the entire period. However, the trend is still highly revealing.

The **sentiment line** shows a dramatic trough around **2017-2018**. This directly corresponds to the 2018 federal amendment that increased fines for cruelty in Islamabad, and confirms that the conversation was driven by public outrage over the "Hero the donkey" incident and reports of mass culling. More interestingly, the **sentiment line** begins a steady climb from 2021 onwards, Here, while the language evidently remains negative throughout, the positive-coded vocabulary after 2021 directly correlates with the rise of the positive words identified, such as **`rescue`**, and **`shelter`**. The line eventually starts dipping down again after 2022, most likely due to the latest animal cruelty incidents in the country.

This data suggests that while the media's focus is on a negative problem, the framing of the conversation has evolved. It has generally shifted from overall low reporting mostly on negative acts of cruelty to increasingly discussing positive, solution-oriented actions like legislation, animal welfare, and shelters.

3.4 Network Analysis: A Single "Core Conversation"

The co-occurrence network provides the deepest insight into the *relationships* between these themes.

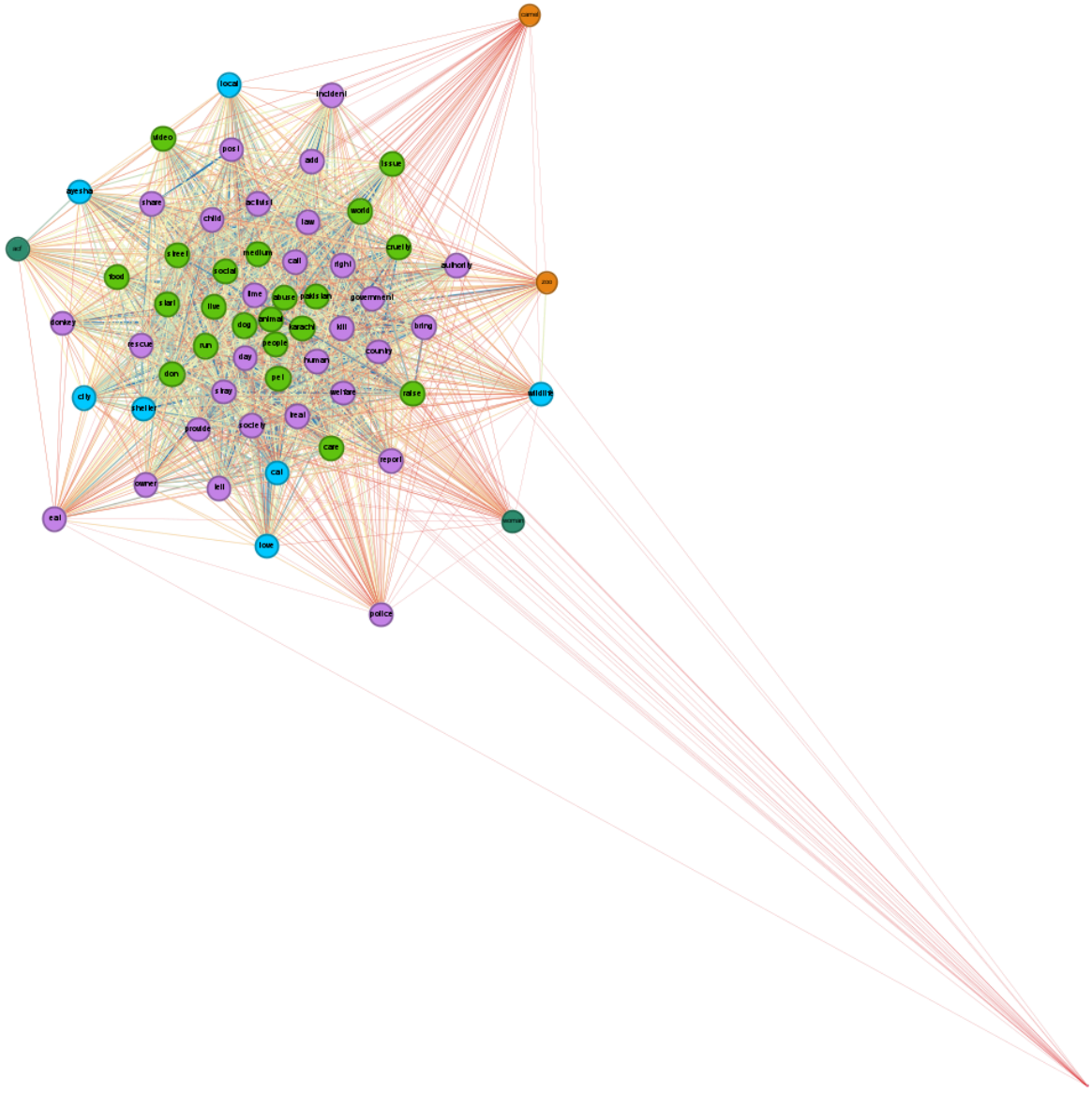


Figure 4: Co-occurrence Network (Force Atlas 2 Layout). Nodes represent the 60 most frequent words. An edge exists if two words appear in the same article at least twice. Node size is proportional to degree, as well as the color (Purple-60, Green-61, Blue-59, etc). Edge color represents weight, with blue edges being higher weight.

The network's statistical profile is its most revealing feature. With how it is constructed, it has 62 nodes, 1839 connections, and a **graph density of 0.973**, the network is hyper-dense. This means the core topics are not isolated. They are all discussed together, constantly and repeatedly. This statistical finding is **visually confirmed by the network graph in Figure 4**. The Force Atlas 2 layout algorithm, which pulls strongly connected nodes together, is unable to

find distinct, separate "neighborhoods" or "clusters." Instead, the visualization reveals a single, massive, and densely interconnected **"core conversation"**, with very few nodes at a bit of a distance.

This visual structure is useful, as it shows that the key topics are **analytically inseparable**. Looking closely at the labels in Figure 4, terms from different conceptual frames are all pulled into the same central core:

- **The Legal Frame:** `law`, `government`, `sindh`, `activist`
- **The Urban Issue Frame:** `stray`, `dog`, `karachi`, `culling`, `city`
- **The Social Frame:** `people`, `society`, `children`, `human`

The network's structure proves these are not separate conversations. The data indicates that a *Dawn* article discussing the **`culling`** of **`stray dogs`** in **`Karachi`** is also highly likely to discuss the **`law`**, the **`government`**'s response, and the impact on **`people`** and **`society`**. The media's discussion of animal cruelty in this corpus is mostly a single, indivisible narrative overall.

3.5 Analysis of Term Significance by Article Frequency

While the previous sections analyzed the raw frequency of words, this section examines their prevalence, that is the percentage of articles in which a term appears at least once. This metric is important for determining which terms are truly significant and indicative of widespread public interest, as opposed to terms that may just be repeated many times in a few articles.

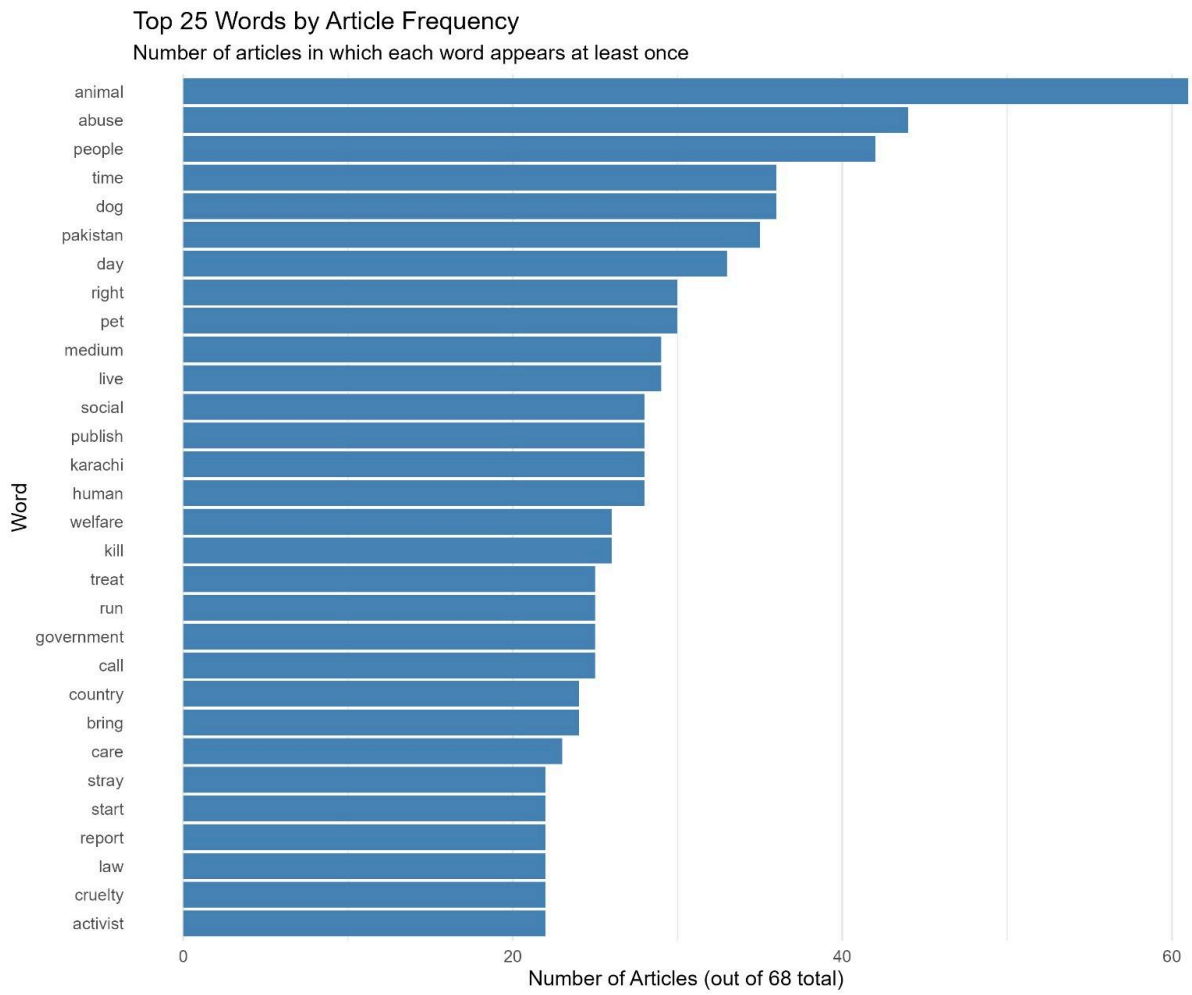


Figure 8: Prevalence of Top 25 Terms Across All Articles. This chart shows the "Share of Articles" (in percentage) that contain each of the top 25 words.

This analysis reveals clear themes of significance, which highlight the core areas of public concern:

The word **`animal`** is present in 100% of the articles. This is expected, but it confirms that every single article in the corpus is directly and explicitly about this topic. The terms like **`abuse`** (present in ~85% of articles) and **`people`** (~80%) form the emotional and social core of the conversation. The high prevalence of `abuse` shows that the discussion is not abstract but almost always framed as a concrete act of harm. The presence of `people` in 4 out of 5 articles emphasizes that this is not just an "animal" issue. It is a human and societal one too, which indicates high public concern for the relationship between people and animals.

Similarly, the terms like **`welfare,shelter,rights`** appear in a large number of articles overall, making this a dominant and highly significant theme. This confirms that public concern is not just about the problem but overwhelmingly about the solution as well. Furthermore, the

inclusion of `dog`, `stray`, and `karachi` in this list is a critical finding. It signifies that the specific urban issue of stray dogs in Karachi is not a minor sub-topic. It is a massive driver of the entire media narrative.

Finally, terms like `law`, `government`, `report`, and `activist` all appear in this list. This high frequency of appearance highlights that public interest is deeply tied to the institutional response. The public concern is not just "this is happening," but "what are the government, police, and courts doing about it," particularly in the context of Sindh.

4.0: Work distribution:

This project was completed as a two-person collaborative effort.

Member 1: Technical Implementation & Data Pipeline

- Developed the complete web scraping pipeline
- Performed IDF analysis and identified it gave no significant
- Researched real-world events (2018 Hero the donkey incident, Sindh Assembly bills, IHC zoo ruling) to contextualize the 2019 spike and sustained legislative focus
- Conducted sentiment analysis interpretation, connecting emotional tone shifts to legislative periods
- Analyzed network structure and density, interpreting why standard centrality metrics were not useful for this hyper-connected discourse
- Created and maintained the reproducible R script with comprehensive inline documentation
- Generated all visualizations: frequency bar charts, trend lines, word clouds, network graphs, and sentiment plots
- Drafted sections of the report: methodology and work distribution.

Member 2: Analysis, Contextualization & Report Writing

- Implemented all text cleaning and preprocessing functions (lemmatization, stopword removal, tokenization, date parsing)
- Constructed both word co-occurrence and article co-occurrence networks
- Calculated network metrics (density, degree centrality) and exported GraphML files for Gephi visualization
- Managed data quality control throughout, debugging scraping issues and handling missing text
- Conducted exploratory data analysis to identify key patterns in word frequencies and temporal trends
- Drafted sections of the report: introduction and results/analysis.
- Synthesized technical findings into narrative form, connecting data patterns to broader social and legal context

- Integrated all visualizations with explanatory captions and interpretive text

Collaborative Responsibilities:

Both members participated in regular planning sessions to define the analytical approach, reviewed each other's code and writing for accuracy, and jointly debugged technical issues. Key decisions, such as using lemmatization over stemming, splitting trend graphs into three panels, and setting the co-occurrence threshold at 2 articles, were made collaboratively. The final deliverable represents equal contribution and shared ownership of both the technical and analytical work.

5.0: References:

1. https://darlab-pakistan.github.io/animal-abuse-pakistan/
2. https://www.bexpress.com.pk/2018/07/donkey-tortured-by-political-partys-supporters-passes-away
3. https://voicepk.net/2025/07/beyond-legislation-has-society-failed-humane-treatment-of-animals/
4. https://www.desiblitz.com/content/donkeys-legs-mutilated-in-pakistan-sparks-outrage
5. https://tribune.com.pk/story/2322697/animal-cruelty-in-pakistan
6. https://www.dawn.com/news/1627150
7. https://gulfnews.com/world/asia/pakistan/donkey-tortured-in-pakistan-dies-social-media-users-furious-1.2255679