# How We Talk about AI: Exploring Textual Differences Between Medium Articles and Traditional News Sources Using Scattertext

**Ibrahim Bukhari** **Charlotte Puopolo**
UPV/EHU University of the Basque Country
Donostia, Basque Country, Spain

## Abstract

This paper utilizes Scattertext to compare Medium articles with traditional online news sources, specifically focusing on discussions surrounding artificial intelligence, deep learning, and data science. Through this analysis, it uncovers distinct thematic orientations, with Medium articles often emphasizing technical concepts and educational resources, while Guardian articles tend to frame AI within socio-political and ethical contexts. The findings highlight contrasting tones and focuses between the platforms, with Medium's discourse being optimistic and technically oriented, and The Guardian's coverage leaning towards ethical considerations and societal implications of AI. Suggestions for future research include exploring additional traditional news sources and tracking the evolution of media narratives around AI in response to societal shifts and technological advancements.

## 1 Introduction

In an increasingly digitized world, new methodologies have emerged to study human language. Scattertext is one of these methodologies for visualizing and analyzing textual patterns. It allows researchers to identify salient linguistic features of their text, such as the term frequencies, relevance, and lexical associations within a corpus. Scattertext makes the information easily digestible by generating interactive and colorful scatter plots. In this paper, we use Scattertext to examine the distinctive characteristics of Medium articles in comparison to traditional online news sources in the domain of artificial intelligence (AI).

Medium, a popular online publishing platform, was originally created in 2012 by the co-founder of Twitter and Blogger. Medium is an example of a "social journalism" platform as it publishes a mix of content from both paid and unpaid contributors (Hermida, 2012). The paid, professional journalists ideally provide credibility, clout, and bring in readers to the platform, while the unpaid contributors cheaply produce the bulk of the content. Since 2021, unpaid contributors have made up an increasingly large share of the authors [1]. In practice, Medium houses articles written by both experts and beginners spanning a diverse range of topics and genres, including AI, ML and Data Science. Unlike traditional news sources, Medium authors often write about their niche for their own specific audience. Moreover, the platform does not moderate its content; the posts must only be legal in US Courts (eg. copyright, not inciting voilence) and no third-party marketing.[2] As such, Medium articles present a unique opportunity to investigate how language is employed in a potentially less formal and more personalized context than traditional news sources.

For a traditional news source, our study compares articles from The Guardian, a well-established British daily newspaper known for its comprehensive coverage of global affairs. We chose to focus this analysis on articles that discuss AI, ML, and Data Science because they encompass a rapidly advancing field of technology that we are familiar with. The field has far-reaching social, political and economic implications, so it is a popular topic in the media.

Against this backdrop, our study seeks to elucidate the textual disparities between Medium articles and traditional news sources. By leveraging the analytical capabilities of Scattertext, we aim to uncover the linguistic features that distinguish Medium articles from their counterparts in established news outlets.

## 2 Related work

Scattertext is a versatile and intuitive tool for visualizing how language differs between corpora. This work draws heavily on Kessler's study of 2012 US Political Conventions, in which he compared speeches by Democrats and Republicans. He employs multiple Scattertext graphics including term association, corpus characteristics, and using correlation to explain classifiers. Many of the discrepancies speak to the larger narratives that the Democrats and Republicans were each reinforcing. In the term association plot, for example, Kessler shows that the Top Democratic terms are about the auto indus-

[1] Robertson. 2021. Medium offers buyouts to editorial employees. *The New York Times.*

[2] Medium. 2023. Medium Rules. *Medium Help Center.*

try and pell grants, whereas the Top Republican term is ”unemployment.” This reflects how the Democratic party often speaks to working class interests, and the Republican party often represents itself as the more fiscally-concerned party. In the Terms by Corpus Characteristicness plot, Kessler illustrates the terms that are common across the *entire* corpus of Democratic and Republican speeches. This effectively filters out some terms that may occur many times in some speeches but are not representative of the corpus overall. Consistent with the previous analysis, the Top Democratic terms are ”middle” and ”class” while the Top Republican terms are ”government” and ”business”. One final example of Kessler's work is how he trains a Support Vector Machine (SVM) classifier to give a prediction score for each document in the corpus, and then use Pearson's $r$ to visualize the distance of a term from the SVM decision line. This is an interesting analysis from a machine learning perspective because it shows the terms that most influence a classifier's decision-making process.

Muñoz and Iglesias used Scattertext in their study detecting psychological stress to identify the most salient words in texts labeled as expressing ”stress” or ”non-stress”. They also used Scattertext with Empath (Fast et al., 2016) to visualize the frequency and stress-feature of topics. From this Empath topic analysis, Muñoz and Iglesias identified the top stress topics as cursing and strong negative emotions, while the top non-stress topics centered on leisure activities and food.

Khan et al. likewise employed Scattertext term association plots to visualize how hate speech differs from offensive speech. In their term association Scattertext, the Top Hate terms are racist and homophobic pejoratives, whereas the Top Offensive terms are not. These works illustrate the utility of Scattertext in distinguishing and highlighting differences between even very similar corpora.

# 3 Data Analysis

Two data sources were utilized to amass a collection of articles pertinent to our study. The primary data source selected was the Medium Article Dataset [3], which comprises articles curated to encompass topics such as Artificial Intelligence (AI), Machine Learning (ML), and Data Science. This dataset encompasses a total of 337 entries, each consisting features such as “authors, claps reading time, link, title, and text”.

The secondary data source was The Guardian Article Database [4], which inherently contains articles published by The Guardian between the years 2015 and 2023, amassing a total of 183,247 entries. Given this project's focus on discerning trends within AI, ML, and Data Science, we decided to include articles only from the most recent years, specifically 2021 to 2023, resulting in a refined collection of 58,423 entries. The features associated with this database include “url, title, description, content, time and tags”.

Given that the primary dataset was already refined to include articles relevant to AI, and the secondary dataset encompassed a broader spectrum of topics, a filtration process was applied to the latter in order to isolate articles related to AI, ML, and Data Science. To ensure a balance between the primary and secondary datasets, the 337 most relevant articles were selected from the secondary dataset.

Consequently, both the primary and secondary datasets were merged to contain 337 articles each and a total of 674 articles. All articles are focused on the domains of AI, ML, and Data Science.

# 4 Data Pipeline

Prior to employing Scattertext for the visualization of scatter plots, it was important to process the articles in a data pipeline. This preliminary step guaranteed that the textual content of the articles were standardized and optimized. Such uniformity in the data format was crucial for the subsequent extraction and visualization of discernible patterns.

## 4.1 Basic Cleaning

The initial phase of the data processing pipeline involves a cleaning of the article text. This step included the elimination of newline characters (‘\n’) from the text, followed by the conversion of all textual content to lowercase. The final action in this phase was the removal of numerical digits. By removing numbers, we ensure that the dataset comprises exclusively of textual data, thereby mitigating variability and enhancing the overall data quality for analysis.

## 4.2 Filtering the Articles

This phase addresses two significant challenges previously identified within The Guardian Article Dataset. The first issue relates to the broad spectrum of topics covered by The Guardian articles. Our objective was to refine this scope, isolating articles that exclusively discuss topics related to Artificial Intelligence (AI), Machine Learning (ML), and Data Science. The second challenge involves dataset imbalance—The Guardian dataset significantly outnumbers the Medium article entries, which could potentially skew the analysis, rendering the visualizations and correlations disproportionately influenced by The Guardian articles.

To mitigate the inclusion of extraneous topics, we compiled a keyword list which encompasses terms frequently associated with AI, ML, and Data Science. Subsequently, we evaluated each article for the presence of these keywords, with a count maintained for each article's keyword. The articles were then ordered descendingly based on their keyword count. To address the dataset imbalance, we only included the top 337 articles from each with the highest keyword counts.

[3] Kaggle - Medium Article Dataset

[4] Kaggle - Guardian Article Dataset

This approach not only ensures the relevance of the selected articles to AI but also prioritizes those most closely aligned with the specified subjects, thereby enhancing the specificity and balance of our datasets.

### 4.3 Labeling the Articles

After filtering and balancing both the primary and secondary datasets, we combined them to create a single, comprehensive dataset in this step. To differentiate between Medium articles and Guardian articles, we have assigned labels to each entry: 'medium' for Medium articles and 'guardian' for Guardian articles.

### 4.4 Further Cleaning

Next, we further refined our dataset by removing punctuation and stopwords, employing the Spacy library for tokenization. This step came after the initial filtering step due to the considerable time and computational resources required for processing the extensive corpus of Guardian articles. The reduction and refinement of the corpus now rendered this task feasible.

The process began with the removal of punctuation and any special characters from the text of our articles to prevent noise in our results. Subsequently, we used Spacy to tokenize the text. Following tokenization, we identified and remove stopwords from the text, further refining our data and minimizing the potential for noise.

Having carefully executed all steps within our data processing pipeline, we are now equipped to advance to the visualization and analysis phase, utilizing Scattertext.

## 5 Results

To find patterns within the article dataset, we use Scattertext, a specialized library designed for the generation of scatter plots. These plots help analyze significant correlations within each dataset, effectively mapping them in a visual format. This allows us to not only identify the most common terms within the datasets but also to explore deeper connections, such as how specific terms may be uniquely associated with one dataset over another.

### 5.1 Term Associations

In the analysis of terms and their affiliations with the article datasets, our objective is to identify and examine single words that manifest in both Medium and Guardian articles, mapping them based on their occurrence frequency within each dataset. It is common for a term to be more dominant in one set of articles than in the other, revealing distinct thematic emphases. Conversely, certain terms may be universally prevalent across both datasets. Hence, Scattertext generates a scatter plot that visually represents the distribution of terms within the data, as seen in Figure 2. In this type of Scattertext, the most frequent words appear in three key areas: bottom right corner (remarkably common in The Guardian), top left corner (remarkably common in Medium articles), and top right (frequent in both corpora). From the term association plot in Fig. 2, the predominant terms within Medium articles include ”function,” ”Neural Network,” ”Deep Learning,” ”layer,” and ”python.” This contrasts with The Guardian's top terms, such as ”ChatGPT,” ”OpenAI,” ”Safety,” ”Risks,” ”Generative AI,” and ”Summits.” These differing term clusters offer insights into the thematic orientations of the articles and their respective publishing platforms. Medium articles tend to delve into the technological concepts of Machine Learning and Data Science, as reflected by the usage of technical terms. On the other hand, Guardian articles appear to frame AI within an ethical context, focusing on contemporary tools and issues. This is evidenced by the emphasis on terms related to safety, ethics, and ChatGPT, which has emerged as a significant AI tool in recent years.

### 5.2 Corpus Characteristics

Identifying patterns within the article datasets, it is crucial to highlight terms that demonstrate consistency across the dataset, as opposed to those that merely exhibit high frequency in the ”Term Association” analysis. The distinction here lies in the fact that a term's high frequency could potentially be skewed by its overuse in a limited number of articles, rather than its consistent presence across the dataset.

In Figure 3, we delve into this aspect by examining the distribution of terms with consistent frequencies. This analysis reveals that while the top terms in both Medium and Guardian articles may share similarities, significant variations emerge when we extend our analysis to terms with medium and low frequencies. This allows us to discern patterns that might not be apparent when focusing solely on the most frequent terms, providing a more comprehensive understanding of the thematic consistency and diversity within the articles.

### 5.3 Empath Topics and Categories

In this particular visualization, we delve into the analysis of topics and categories within the textual terms of each article, utilizing Fast et al.. This approach contrasts with Term associations, offering insights into tonal variations and abstract meanings across different articles through a scatterplot representation, as seen in Figure 4.

The results in Fig. 4 distinctly highlight the tonal shift between the topics and categories prevalent in Medium and Guardian articles. Medium articles feature Empath topics such as ”cheerfulness,” ”toy,” ”tool,” ”philosophy,” and ”politeness,” which collectively suggest a more positive and constructive view of AI, ML, and Data Science. These topics imply that Medium's discourse tends to frame these technologies in terms of their potential and utility, emphasizing their beneficial applications.

Conversely, the Empath topics associated with Guardian articles, such as ”crime,” ”law,” ”terrorism,”

”stealing,” and ”white-collar job,” paint a starkly different picture. These topics indicate a narrative that is more cautious of AI and its implications. The emphasis on terms related to legality, crime, and societal issues suggests that the Guardian’s coverage may focus on the ethical dilemmas, potential misuses, and regulatory challenges posed by advancements in AI technology.

### 5.4 Correlation using Support Vector Machine

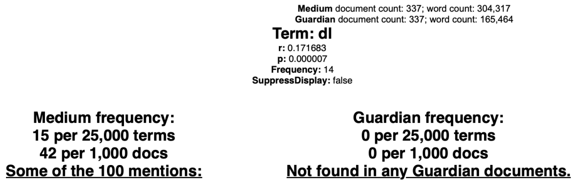
In this concluding analysis, we shift our perspective from examining term frequencies and consistency to exploring the correlation between terms and the articles they characterize, through a distinct approach. Here, the focus is on employing a classifier (Support Vector Machine) trained on our labeled datasets to identify terms that significantly influence the classifier’s decision-making process.

This methodology is instrumental in pinpointing words that capture the essence of articles published by Medium or The Guardian. It is important to note, however, that this approach may prioritize terms that may be specific to the publishers’ distinct narrative styles or thematic focuses, potentially sidelining specific AI, ML, and Data Science topics. Despite this, the analysis remains valuable, as it sheds light on the differences in how each platform approaches and frames discussions around these technological fields.

As evident in Figure 5, the terms that markedly impact the classifier’s decisions reveal the thematic divergences between the two sets of articles. For The Guardian, terms like ”government,” ”regulation,” ”safety,” ”risks,” and ”ChatGPT” suggest a narrative that leans towards the socio-political implications of AI, highlighting concerns related to governance, ethical considerations, and potential risks. In contrast, Medium’s influential terms such as ”python,” ”course,” ”unsupervised,” ”learning,” and ”data” highlight a more technical and educational perspective, focusing on the methodologies, tools, and learning resources relevant to the AI, ML, and Data Science communities.

## 6 Discussion

Medium articles often focus on the technical aspects of AI, ML and Data Science. This reflects the niche audience A curious feature of the Medium corpus in Fig. 3 is the abundance of one- or two-character terms including ”tf, h, x, y, dl” littered throughout the graph. Upon inspection, these largely come from coding and math examples within the Medium articles. As shown in Fig. 1, ”dl” is an abbreviation for Deep Learning that is relatively common in Medium articles but never appears in The Guardian dataset. The abbreviation functions as shorthand (and in-group speak) for experts in the field. Such a highly technical lexicon underscores how Medium often serves as a platform to crowd-source knowledge about new technologies in these fields.



**Medium** document count: 337; word count: 304,317
**Guardian** document count: 337; word count: 165,464

**Term: dl**
**r:** 0.171683
**p:** 0.000007
**Frequency:** 14
**SuppressDisplay:** false

**Medium frequency:**
**15 per 25,000 terms**
**42 per 1,000 docs**
**Some of the 100 mentions:**

**Guardian frequency:**
**0 per 25,000 terms**
**0 per 1,000 docs**
**Not found in any Guardian documents.**

Figure 1: Term Frequency for the term ”dl” by Scattertext

Another salient feature of Medium articles is ”claps,” a term which comes up in every Scattertext plot that refers to how readers ”like” or applaud a post. This term highlights the interactive and communal nature of readers with Medium posts. Moreover, the Top Medium topics in the Empath analysis (Fig. 4)are decidedly more optimistic than The Guardian’s. This pronounced divergence in thematic focus underscores the contrasting perspectives of the two platforms: while Medium articles tend to celebrate the technological advancements and possibilities within AI, ML, and Data Science, Guardian articles appear to approach these subjects with a more critical eye, highlighting the need for careful consideration of ethical, legal, and social implications.

Meanwhile, The Guardian’s discussion of AI, ML and Data Science places them in a political, ethical, and economic framework. The Guardian centers its articles on concrete details of the industry by highlighting ChatGPT and OpenAI, arguably the most famous AI technology and company during the early 2020s. The Guardian’s most salient terms in every analysis focus on the safety concerns and risks of the new technology. This fits into the larger narrative of reporting to the public on the potential risks of AI, but it could also reflect the doomsday shock appeal that some news outlets use to attract views online. Finally, ”says” and ”said” are consistently part of the Top Guardian terms, which point to the journalistic convention of how to report quotes by industry leaders and officials.

## 7 Conclusion

This paper investigated through Scattertext the distinctive characteristics of Medium articles in comparison to a traditional online news source, The Guardian, in their discussion of AI. We found several contrasting tones and focuses between the platforms, with Medium’s discourse being optimistic and technically oriented, and The Guardian’s coverage centering the ethical considerations and societal implications of AI.

Further research should be done to evaluate more traditional news sources, apart from The Guardian. We are curious about how The Guardian would compare to an American news source or a non-western news source like Al Jazeera in their discussion of AI. It could also be interesting to explore the evolution of how the media talks about AI after different inflection points, such as how (Power and Crosthwaite, 2022) compared Australian and New Zealand Prime Ministers’ speeches through the evolution of the COVID-19 pandemic.

## References

Ethan Fast, Binbin Chen, and Michael S. Bernstein. 2016. Empath: Understanding topic signals in large-scale text. *CoRR*, abs/1602.06979.

Alfred Hermida. 2012. Social journalism: Exploring how social media is shaping journalism. *The handbook of global online journalism*, 12:309–328.

Jason S. Kessler. 2017. Scattertext: a browser-based tool for visualizing how corpora differ. *CoRR*, abs/1703.00565.

Muhammad U. S. Khan, Assad Abbas, Attiqa Rehman, and Raheel Nawaz. 2021. Hateclassify: A service framework for hate speech identification on social media. *IEEE Internet Computing*, 25(1):40–49.

Sergio Muñoz and Carlos A. Iglesias. 2022. A text classification approach to detect psychological stress combining a lexicon-based feature framework with distributional representations. *Information Processing Management*, 59(5):103011.

Kate Power and Peter Crosthwaite. 2022. Constructing covid-19: A corpus-informed analysis of prime ministerial crisis response communication by gender. *Discourse & Society*, 33(3):411–437.

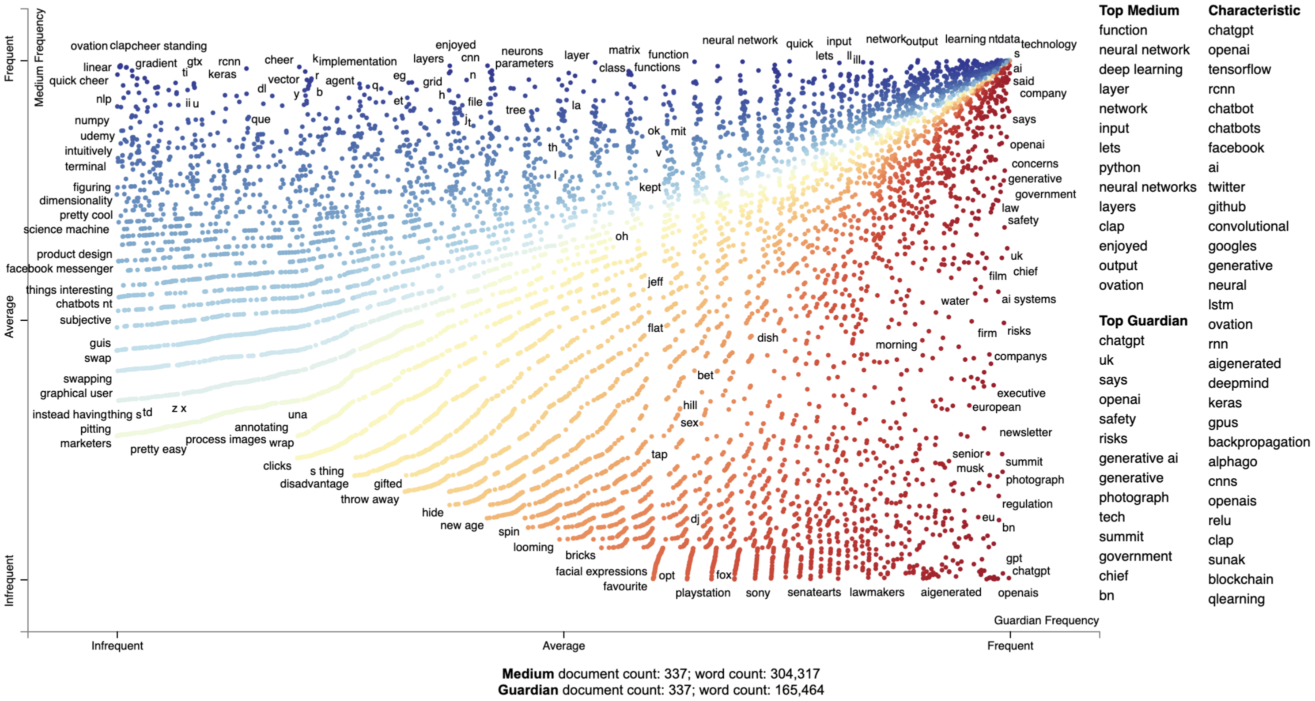


Figure 2: Scattertext Term Association

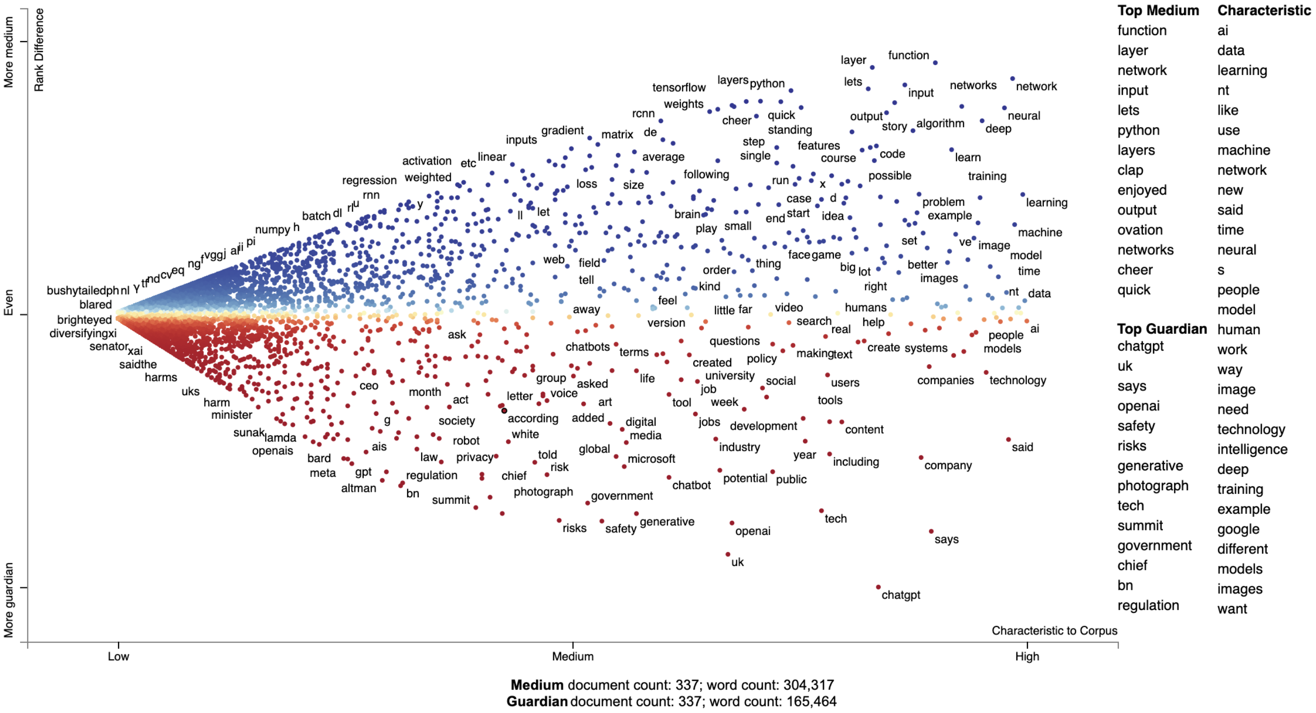


Figure 3: Scattertext Corpus Characteristics

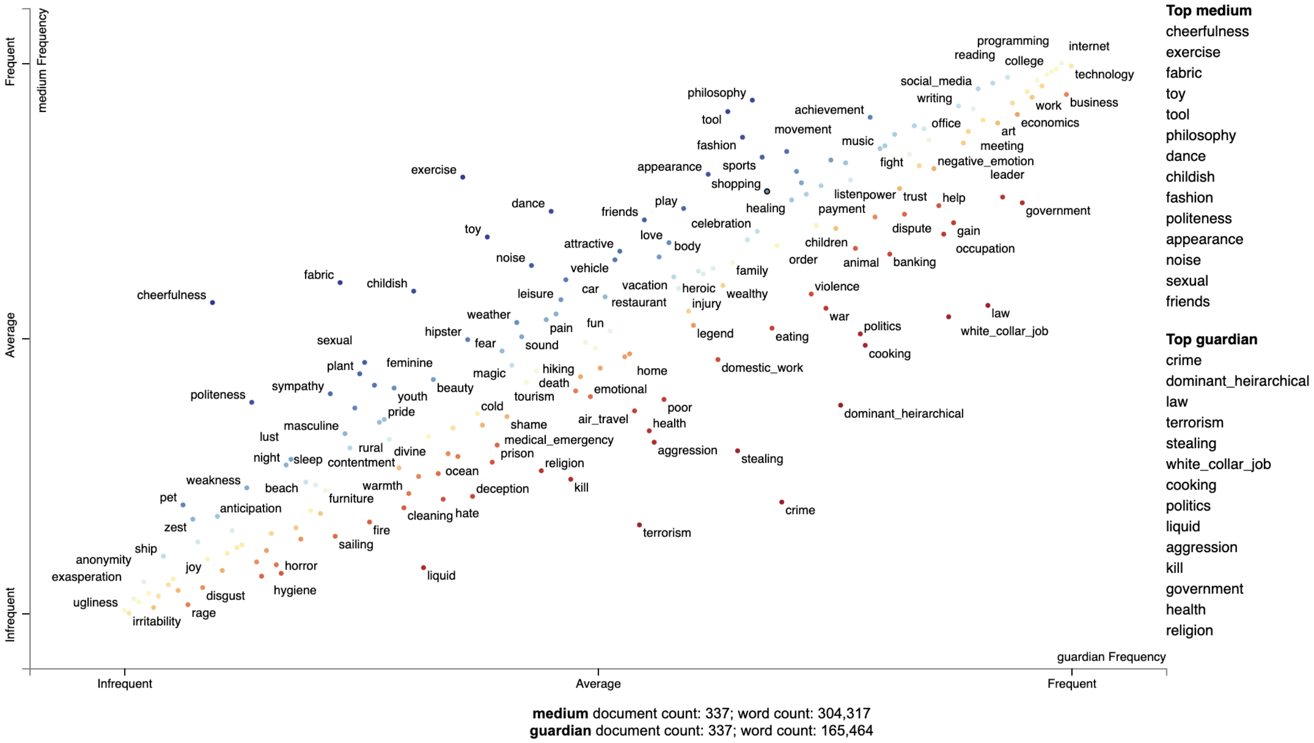


Figure 4: Scattertext with Empath Topics

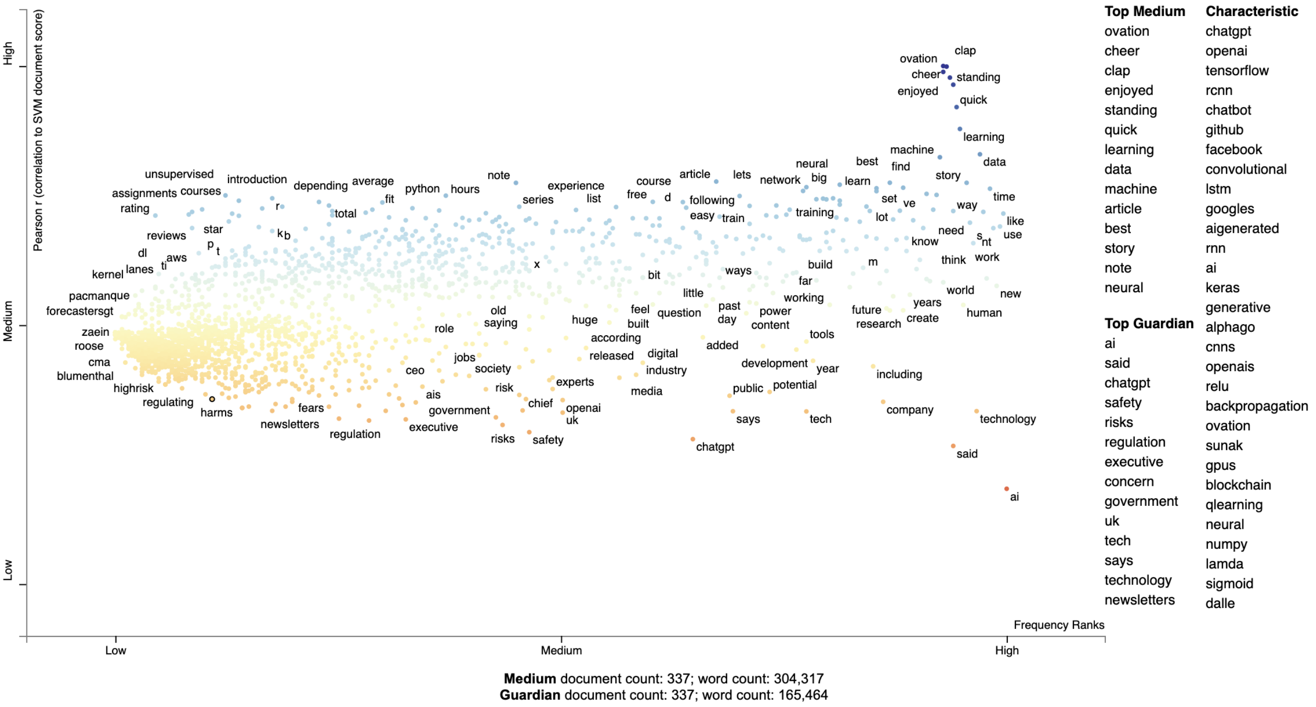


Figure 5: Scattertext Correlation using SVM