Introduction

The rapid expansion of online discussion platforms and social media has generated an abundance of user-generated content, offering valuable insights into public perceptions of emerging technologies. Within this ecosystem, Artificial Intelligence (AI) emerges as a focal point of intense, often informal discourse, characterized by candid exchanges, diverse viewpoints, and real-time reactions to breakthroughs in areas such as large language models, ethical implications, and societal transformations.

For this project, we have selected informal discussions on Reddit pertaining to AI as the primary dataset. This selection is informed by the following key considerations:

1. **High Consumer Engagement**: Subreddits such as *r/ArtificialIntelligence, r/singularity, and r/Futurology* feature millions of users contributing thousands of informal comments daily. The substantial volume of unstructured, conversational data provides a robust foundation for text analysis, capturing spontaneous debates on topics ranging from AI capabilities to future timelines.
2. **Diverse User Feedback**: Participants encompass a broad spectrum of voices—from *technical experts and enthusiasts to skeptics and ethicists*—who express unfiltered opinions on AI's technical merits, potential risks, and practical applications. This variety enables a nuanced examination of sentiment fluctuations and thematic diversity within casual, peer-to-peer interactions.
3. **Relevance and Timeliness**: As of October 2025, Reddit's informal threads reflect contemporary accelerations in AI development, including discussions on superintelligence projections, ethical challenges in creative industries, and regulatory debates. Analyzing this ongoing, conversational feedback sheds light on the evolving long-term attitudes toward AI's integration into society.
4. **Opportunities for Advanced Analysis**: The dataset's *mix of optimistic hype, critical skepticism, and neutral observations*—often laced with slang, humor, and anecdotes—lends itself to sophisticated techniques such as sentiment analysis, word clustering, network diagrams, and topic modeling, which uncover latent patterns in informal language.
5. **Practical Implications**: Derived insights hold value for stakeholders, including AI developers aiming to address user concerns, policymakers shaping governance frameworks, and organizations navigating adoption strategies. This renders the analysis both academically rigorous and practically applicable.

In summary, focusing on Reddit's informal AI discussions allows us to harness a dynamic, voluminous dataset through text analytics. The project not only elucidates user perceptions of AI but also exemplifies the utility of extracting actionable intelligence from unstructured, conversational online content.

Objectives

1. ***To collect a comprehensive dataset of informal Reddit discussions on AI***: Aggregate user-generated posts and comments from subreddits such as r/ArtificialIntelligence, r/MachineLearning, and r/singularity to form a large, diverse, and representative corpus. This establishes a solid basis for analysis, encompassing a spectrum of informal viewpoints from varied user demographics.
2. ***To generate word clouds for commonly used words****:* Identify and visualize the most recurrent terms within the discussions. This facilitates an immediate overview of prevalent themes, concerns, and expressions, such as AI capabilities, ethical issues, or societal effects, as articulated in casual language.
3. ***To construct a network diagram of co-occurring keywords***: Investigate interconnections among terms and concepts that appear together in threads. This discloses relational patterns in informal exchanges, for instance, associations between "superintelligence" and "existential risks" or "AGI timelines" and "ethical safeguards."
4. ***To apply word clustering to group similar terms:*** Categorize words and phrases into coherent clusters, including sentiment-based groupings (positive, negative, neutral) and thematic ones (e.g., technical innovations, ethical considerations, economic impacts). This structures the informal discourse for enhanced interpretability.
5. ***To perform sentiment analysis***: Classify the emotional valence of posts and comments to assess overall attitudes toward AI. This quantifies whether discussions lean optimistic, cautious, or ambivalent, yielding insights into public sentiment amid 2025's evolving AI landscape.
6. ***To apply topic modeling techniques****:* Employ algorithms like Latent Dirichlet Allocation (LDA) to reveal underlying themes, such as regulatory imperatives, automation anxieties, disruptions in creative sectors, or enthusiasm for innovations. This exposes subtle drivers of informal online narratives.
7. ***To provide actionable insights through visualization and interpretation****:* Leverage tools including word clouds, network graphs, sentiment visualizations, and cluster diagrams to articulate the progression of public discourse. These findings can guide AI policy formulation, ethical design practices, strategic implementations, and research directions.

METHODOLOGY

The methodology outlines the systematic process for analyzing informal Reddit discussions on Artificial Intelligence (AI), from data collection to interpretation, tailored to the unstructured conversational data:

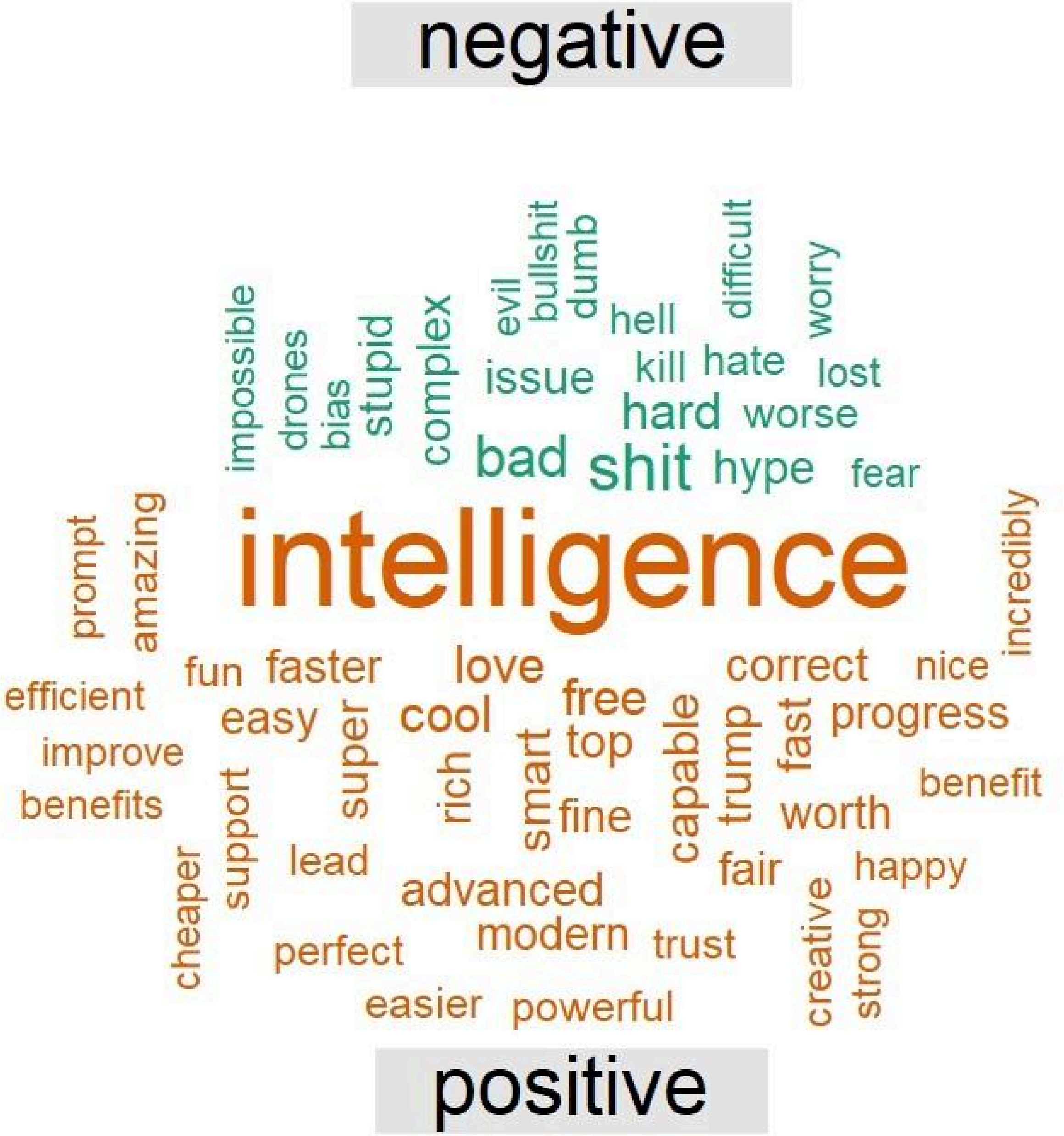
1. **Data Collection**: A custom Python scraper retrieved the top 50 posts and comments matching the query "artificial intelligence" from subreddits like r/ArtificialIntelligence and r/singularity. This generated over 20,000 rows of textual.
2. **Data Preprocessing**: Techniques such as tokenization, stopword removal, lowercasing, lemmatization, and slang/emoji replacements were applied. The informal nature (e.g., sarcasm, jargon) required extensive, iterative cleaning and code adjustments to eliminate noise, duplicates, and irrelevancies, yielding ~18,000 clean entries.
3. **Word Cloud Generation**: Word clouds visualized frequent terms, providing quick insights into key themes like AI ethics and societal impacts in casual discourse.
4. **Network Diagram Construction**: Co-occurrence networks mapped keyword relationships (e.g., "industrial" with "revolution"), highlighting debate interconnections.
5. **Word Clustering**: Clustering algorithms grouped terms into sentiment (positive/negative/neutral) and thematic clusters (e.g., innovations, ethics), structuring informal text.
6. **Sentiment Analysis**: Discussions were classified as positive, negative, or neutral to quantify attitudes, identifying hype (e.g., innovations) versus fears (e.g., job loss).
7. **Topic Modeling**: Latent Dirichlet Allocation (LDA) uncovered hidden themes like regulatory needs and automation anxieties, beyond surface-level anecdotes.
8. **Visualization and Interpretation**: Tools like word clouds, graphs, and charts depicted discourse patterns, offering actionable insights for AI policy and strategy.

DATASET

A dataset of informal Reddit discussions on Artificial Intelligence was compiled using a ***custom Python scraper***, which targeted the top 50 posts and their comments via the search query "*artificial intelligence."* The initial scrape produced over 20,000 rows, encompassing columns for ID, titles, and comment bodies. This voluminous corpus captured a broad spectrum of conversational threads from October 2025, reflecting real-time user engagement.

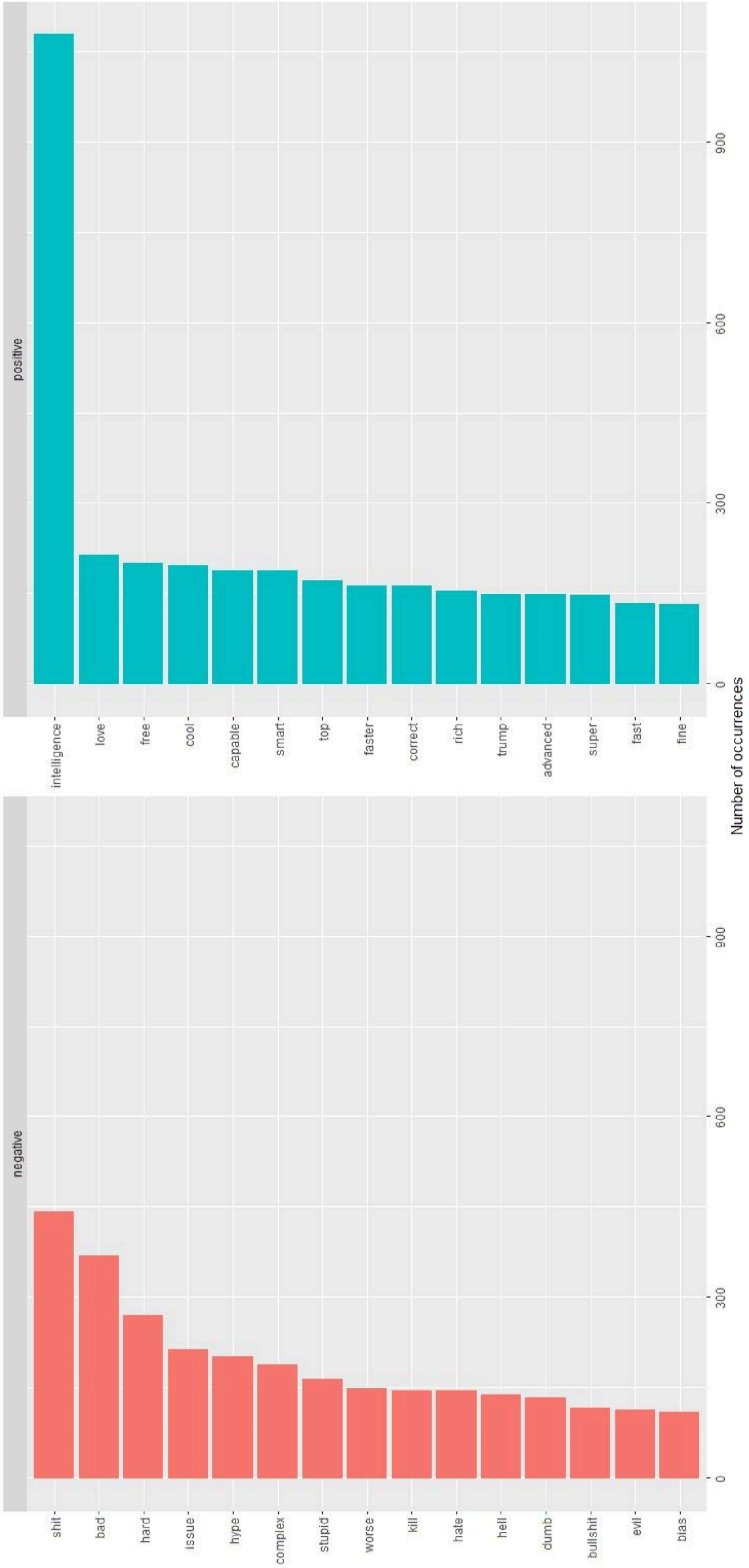
Post-collection, the dataset underwent extensive cleaning to mitigate the challenges posed by informal language. Duplicates were excised, irrelevant content (e.g., off-topic replies or promotional spam) filtered, and special characters/URLs removed. Preprocessing required a significant time investment, involving iterative stopword customization (to account for AI-specific jargon, such as "LLMs" or "ASL"), slang replacements and *code revisions for handling variable-length threads and emoji-laden text.* Optimized for robust text analysis while preserving the authentic, unfiltered essence of Reddit discourse.

Sentiment Analysis:



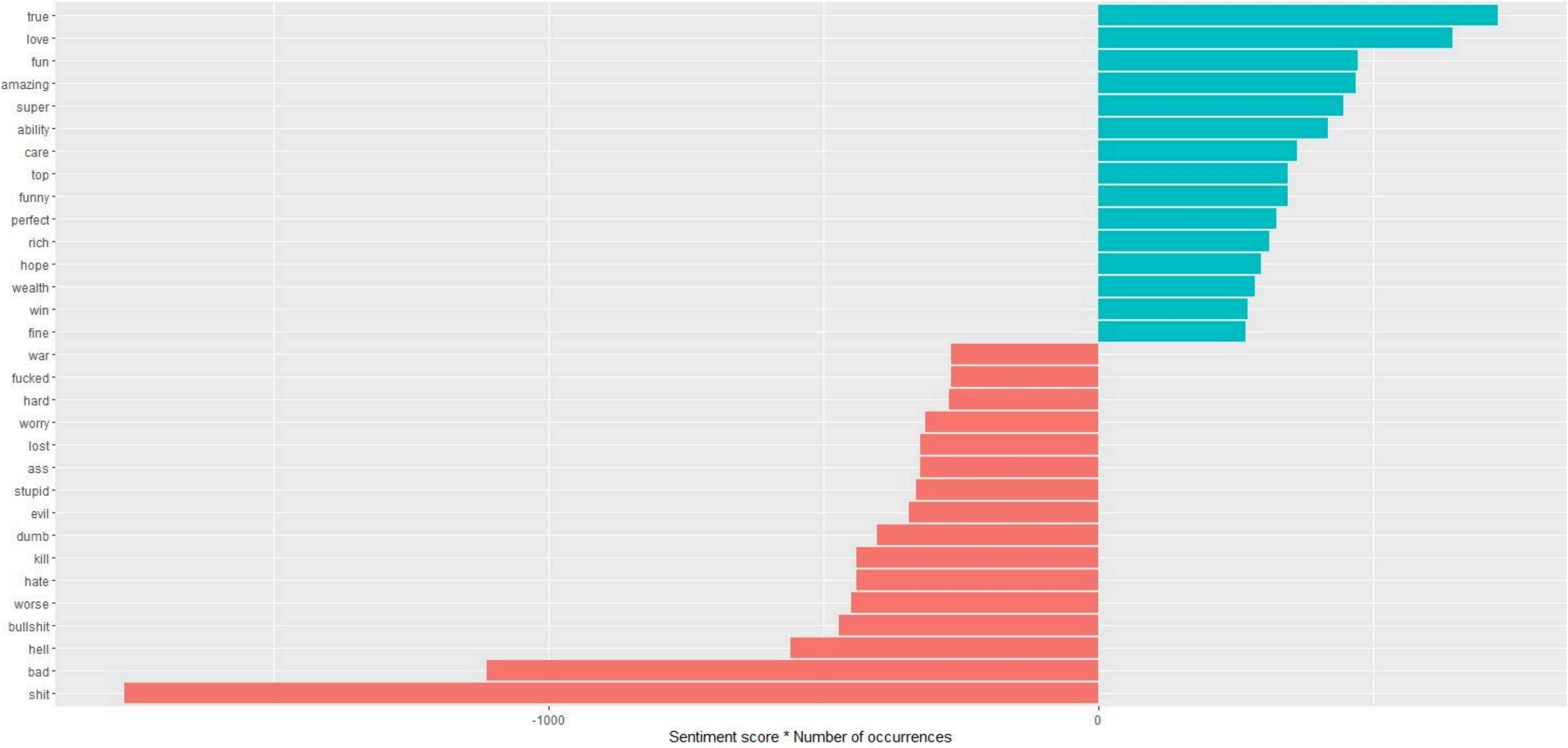
The word cloud separates the negative and positive words. The negative words are in green and the positive words in red. Influential words are bigger: *"bias", "issues", "hype", and "fear" are the most influential negative words, while "amazing", "efficient", "fun", and "love"* are the most influential positive words.

Bar Diagram



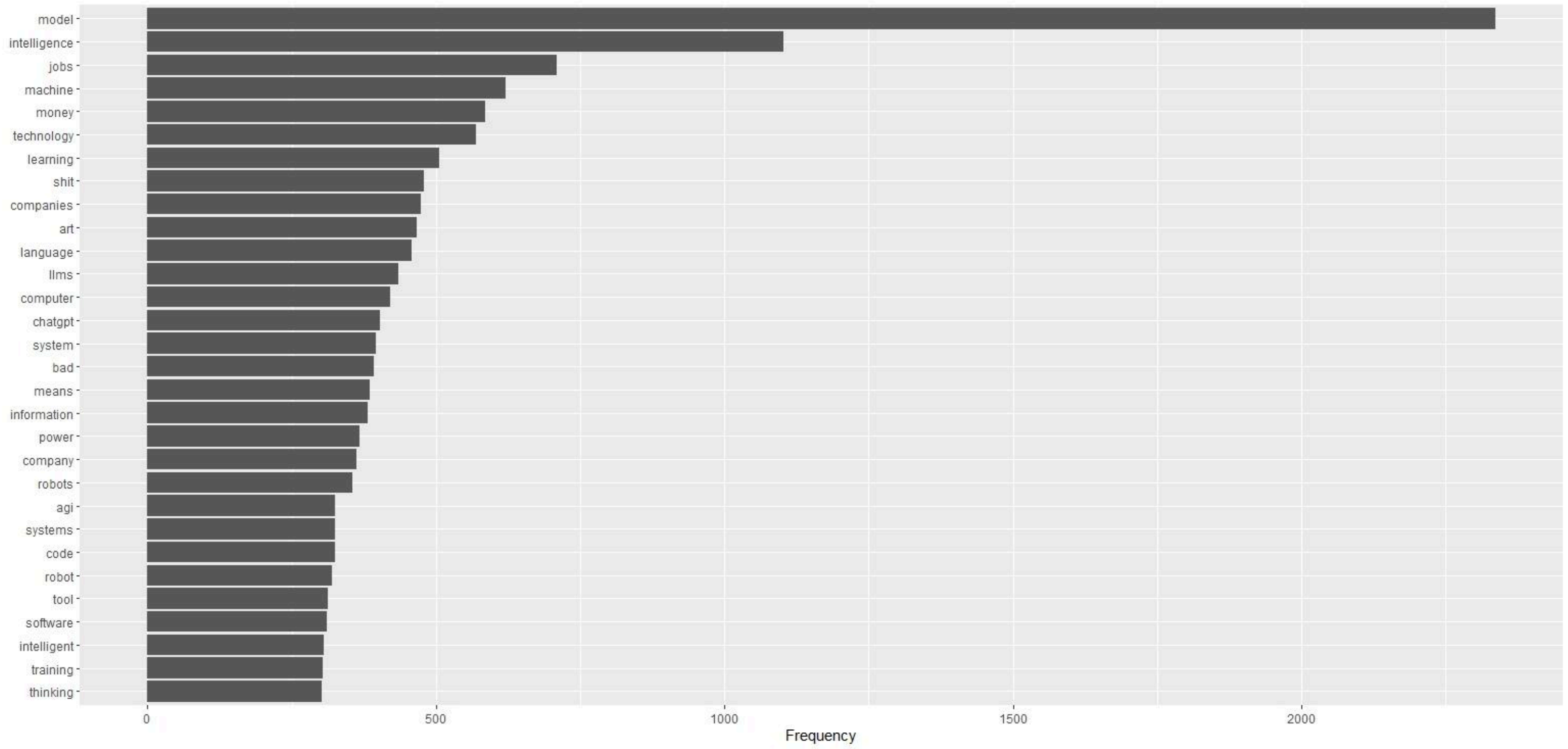
The bar diagram separates words by sentiment into negative and positive categories. The negative words are displayed in red, and the positive words are in teal. In both charts, the bars are arranged in decreasing order of their number of occurrences. The most frequent negative words are ***"shit", "bad", "hard", and "issue"****,* while the most frequent positive words are ***"intelligence", "love", "free", and "cool".***

Sentiment Score:



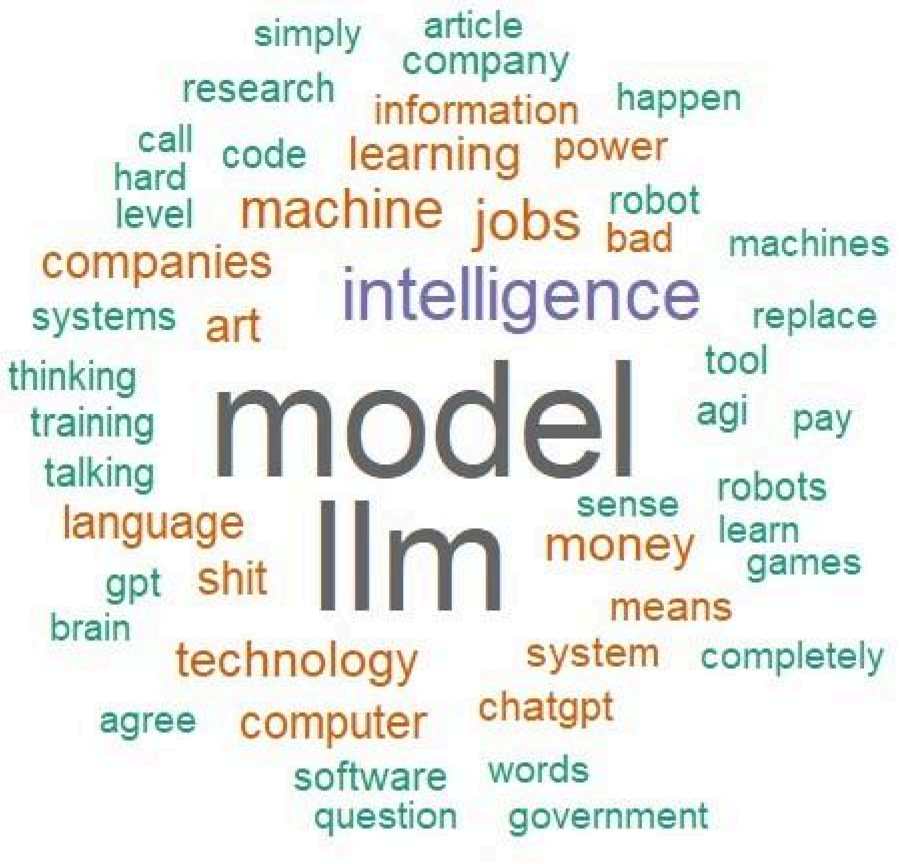
The greatest contribution chart categorizes words into positive and negative sentiment, with negative words displayed in red and positive words in teal. The length of each bar is determined by the word's sentiment score multiplied by its number of occurrences, indicating its overall impact on the sentiment. The negative words contributing the most to the sentiment are ***"shit"**, **"bad"**, and **"hell"**,* while the positive words with the greatest contribution are ***"true"**, **"love"**, and **"fun"***.

Bar Plot:



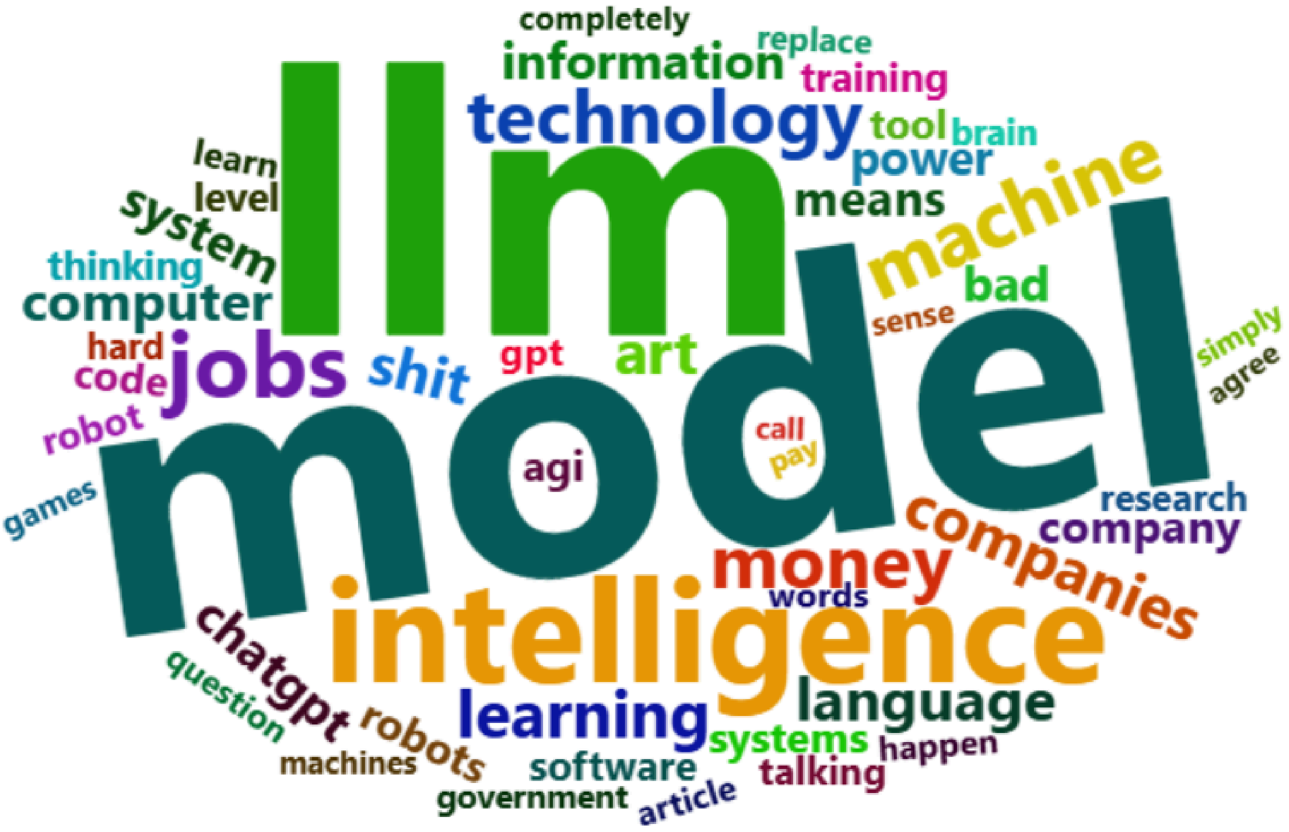
Words like ***"model," "intelligence," "agi," and "machine"*** dominate, showing that the discussion mainly revolves around artificial intelligence. Terms such as ***"software," "llms," "chatgpt," and "tool"*** indicate a focus on its applications. The presence of ***"money," "companies,", "jobs" and "power"*** reflects that many comments are likely discussing the economic and societal impact. Negative words like ***"shit" and "bad"*** suggest the conversation also includes significant concerns and criticisms.

Word Cloud:



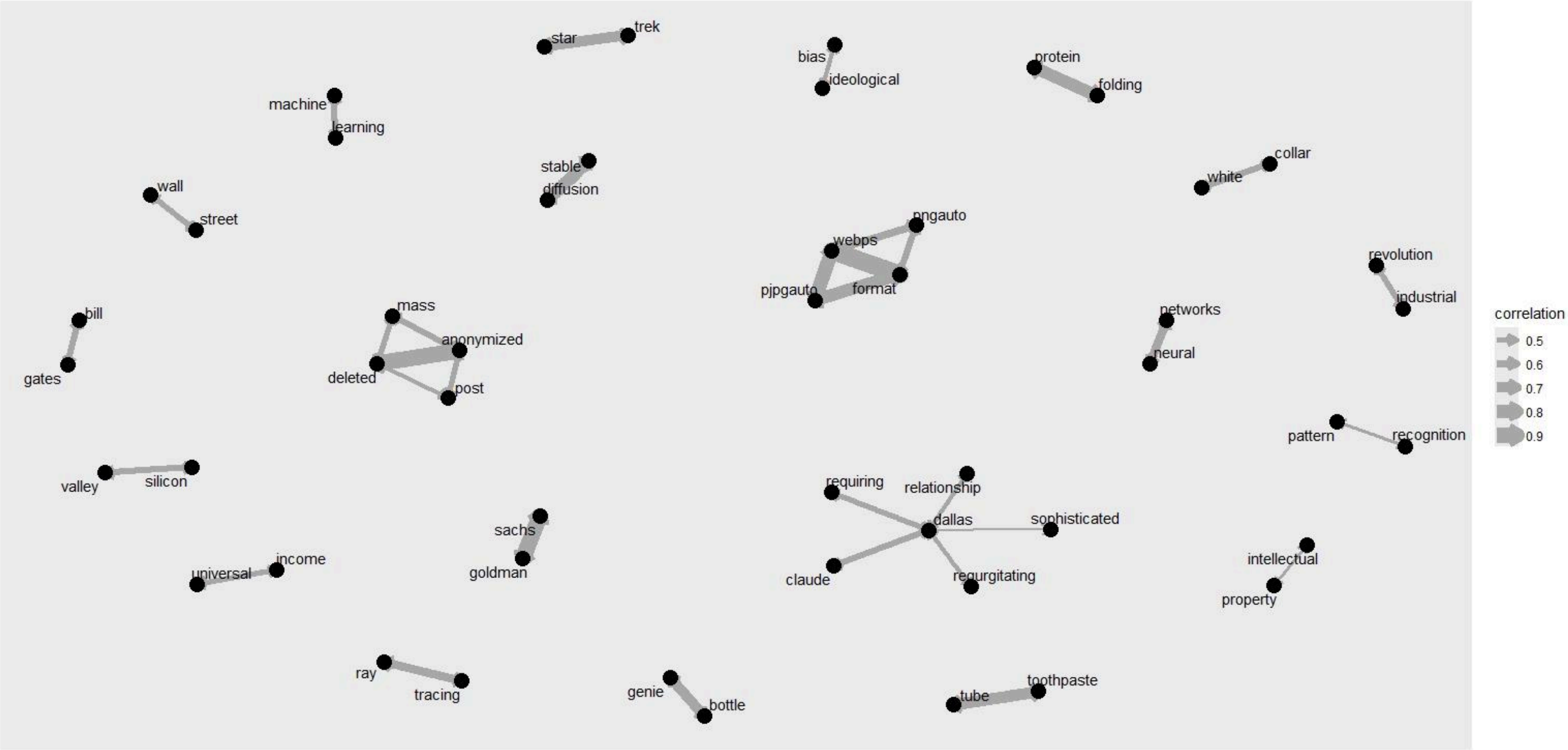
This word cloud illustrates a multifaceted conversation centered on AI ***"model"*** *and* ***"llm"*** development. One cluster of words, including ***"machine," "learning,"*** *and* ***"computer,"*** points to the technical aspects of the discussion. Another prominent theme, highlighted by ***"jobs," "money,"*** *and* ***"power,"*** focuses on the profound socio-economic consequences. The mix of analytical and emotive words like ***"thinking"*** *and* ***"shit"*** reveals a complex and often critical public discourse on the technology's impact.

HTML Word Cloud:



This word cloud reveals a deep dive into the very definition of artificial cognition. The conversation juxtaposes human concepts like ***"brain," "language,"*** *and* ***"thinking"*** with technical building blocks like ***"code," "software,"*** *and* ***"system."*** The inclusion of aspirational yet controversial terms such as ***"agi"*** *and* ***"replace"*** indicates a strong focus on the ultimate, and perhaps unsettling, potential for AI to mirror or surpass human capabilities.

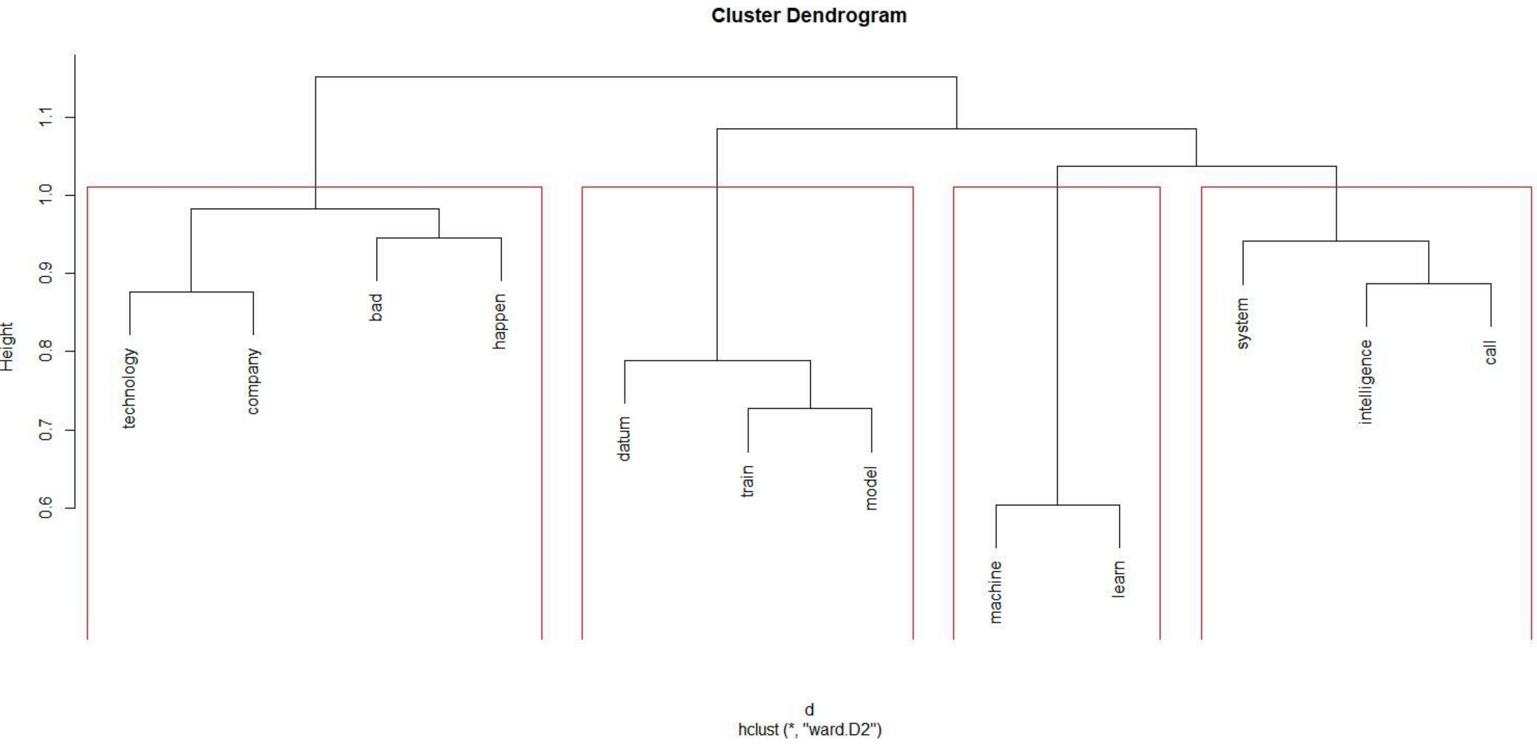
Word Network Diagram:



This word network reveals distinct thematic clusters based on the strong correlation of co-occurring terms. We observe strong associations between foundational AI concepts like ***"machine"–"learning," "neural"–"networks,"*** *and* ***"pattern"–"recognition."*** The specific AI applications linking ***"stable"–"diffusion"*** *and* ***"protein"–"folding."*** Furthermore, it captures the socio-economic discourse surrounding AI, with tight pairings like ***"universal"–"income,"*** stealing ***"intellectual"–"property,"*** and 4th ***"industrial"–"revolution."*** The link between ***"bias"*** *and* ***"ideological"*** on the ethical dimensions of the LLMs

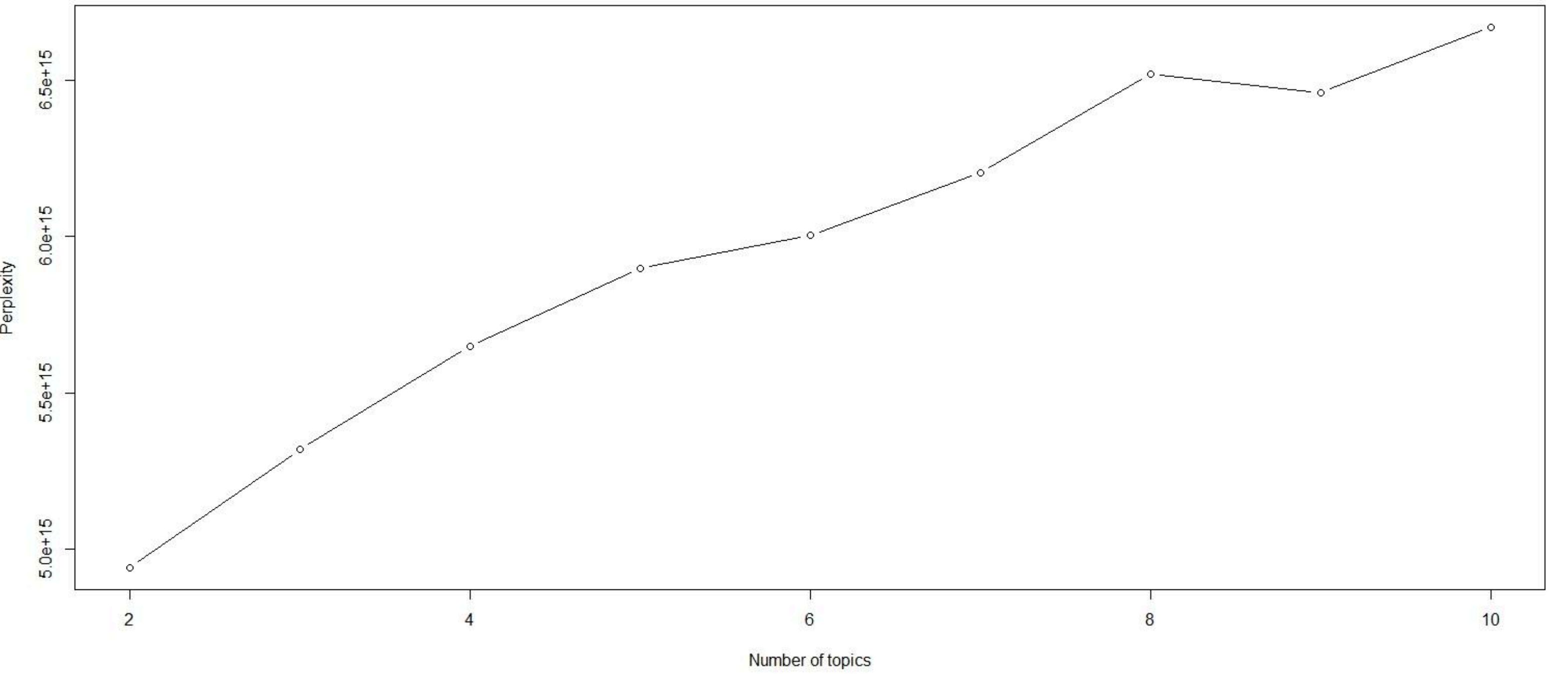
.

Cluster Dendrogram:



Four distinct thematic clusters based on contextual similarity. The cluster on the far left, containing ***"technology," "company,"*** *and* ***"bad,"*** groups concepts around the societal risks and corporate accountability in tech. The two central clusters detail the mechanics of AI: one captures the fundamental process by linking ***"machine"*** *and* ***"learn,"*** while the other outlines the development pipeline with ***"datum," "train,"*** *and* ***"model."*** The final cluster on the right, featuring ***"system," "intelligence,"*** *and* ***"call,"*** appears to represent the functional application of AI systems.

Topic Modeling and Perplexity:



Diminishing returns start from Perplexity 5.75e+15 to 6.5e+15 (Topic 4-8), and after 8 topics, the return becomes negative and rises again; therefore, ***the optimal number of topics selected for interpretation is 8***.

Topics:

Art & Coding	**Machine Learning**	**Entertainment**	**Smart Devices**
art	intelligence	games	robot
code	machine	vr	love
chatgpt	learning	shit	car
gpt	language	video	cool
artist	llms	movie	smart
training	brain	hardware	move

Geopolitics	**Research**	**Public Discourse**	**Automation & Displacement**
system	technology	government	jobs
mass	research	shit	money
kill	power	talking	replace
war	design	reddit	pay
china	agi	trump	automation
hope	field	elon	workers

Discourse around AI centers on four key areas: **core technology** (*machine learning, LLMs*), **diverse applications** (*AI art, coding, robotics, and gaming*), **societal consequences** *(job automation, workers' rights, and supply chain systems*), and **public debate** (*government policy and social media discussion*s). The topics distinctively separate the technical exploration of AI's capabilities from the critical assessment of its ***real-world impact on labor, global power dynamics, and political commentary***.

Conclusion

This comprehensive analysis reveals that the public discourse surrounding artificial intelligence is ***not a single conversation but a complex and deeply polarized tapestry of distinct, interwoven themes.*** The sentiment is sharply divided, with a juxtaposition of high optimism for AI's capabilities (e.g., *"intelligence," "love," "cool"*) and profound cynicism and concern, evidenced by the high frequency of negative terms like *"shit," "bad," and "hell."*

The key takeaways from this analysis are:

1. **The Discourse is Structurally Thematic:** The analysis consistently shows that the conversation is not random but organized into clear, predictable clusters. These range from the purely technical discussions about the mechanics of AI *("machine," "learning," "model," "datum")* to its creative applications *("art," "games," "design")* and, most significantly, its far-reaching societal consequences.
2. **Socio-Economic Impact is a Dominant Concern:** While the technology itself is the core subject, themes of ***"jobs," "money," "power,"*** *and* ***"replace"*** are dominant throughout the analysis. This indicates that the primary public focus is less on how AI works and more on how it will impact human labor, economic structures, and societal power dynamics.
3. **The Conversation is Highly Politicized and Personal:** The presence of terms like ***"government," "war," "bias,"*** and specific political figures reveals that AI is not viewed as a neutral technology. Instead, it is seen as a powerful tool with significant ethical, geopolitical, and personal ramifications, sparking passionate public debate.

In essence, the analysis paints a clear picture of a public grappling with AI on multiple fronts simultaneously: as a revolutionary technology, a powerful creative tool, a formidable economic disruptor, and a pressing ethical and political challenge.