

MULTIMEDIA UNIVERSITY OF KENYA

FACULTY OF COMPUTING & INFORMATION TECHNOLOGY

PROJECT DOCUMENTATION

PREDICTING TECHNOLOGY TRENDS USING MACHINE LEARNING MODELS

BY

COLLINS IREGI

CIT-223-041/2020

21 FEBRUARY 2024

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Submitted in partial fulfillment of the requirements of Bachelor of Science in Information

Technology/Computer Science/Software Engineering.

# Declaration

I hereby declare that this Project proposal is my own work and has, to the best of my

knowledge, not been submitted to any other institution of higher learning.

Student: Collins Iregi Registration Number: CIT-223-042/2020

Signature: ............................................... Date:08/04/2024

This project has been submitted as a partial fulfillment of requirements for the

Bachelor of Science in Computer Science/Information Technology of Multimedia University of Kenya with my approval as the University supervisor.

Supervisor: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature: ..................................................... Date: ..................................................

# Abstract

In the dynamic and rapidly evolving landscape of the technology community, our primary aim is to construct a sophisticated machine learning model that can anticipate and track trends. This ambitious project seeks to predict the emergence of new technologies and shifts in hiring practices, providing a forward-looking perspective that is often missing in traditional trend analysis.

To achieve this, we plan to leverage a diverse range of data sources, each offering unique insights into different aspects of the technology landscape. These include public datasets such as the Stack Exchange API, which provides a wealth of information on developer discussions and problem-solving; web traffic time series, offering a glimpse into the popularity of different technologies over time; GitHub trending repositories, revealing the latest open-source projects that are gaining traction; and World Bank data catalogs, providing macroeconomic context that can often influence technology trends.

In addition to these, we also intend to utilize sentiment analysis datasets from platforms like Twitter and Reddit. These social media platforms serve as a real-time pulse of public opinion, and analyzing sentiments expressed in these platforms can help us understand how the public perceives different technologies. To capture the pulse of real-time trends, we will employ web crawling techniques on job boards, tech blogs, and Twitter threads, providing us with up-to-date information on what technologies companies are looking for and what topics are being discussed in the tech community.

The heart of our project is a meticulously designed machine learning model. This model will be trained to analyze these diverse data sources effectively, identifying patterns and trends among the noise. More than just recognizing current trends, our model will predict future trajectories, providing valuable insights into where the tech community is headed.

We believe this project holds significant potential for impact within the tech community. It aims to benefit various stakeholders, including job seekers looking to upskill in relevant technologies, learners deciding which technologies to study, market analysts tracking the rise and fall of different technologies, and consumers trying to understand which technologies are worth investing in. By providing predictive insights into technology trends, we can help these stakeholders make informed decisions.

This proposal outlines a comprehensive plan for a project that combines robust data sources with advanced machine learning techniques to track and predict key trends in the technology community. By bridging the gap between current trend analysis and future trend prediction, we aim to provide a new tool that can help the tech community navigate the future.

Contents

[Declaration 2](#_Toc163563965)

[Abstract 3](#_Toc163563966)

[Chapter 1 introduction 7](#_Toc163563967)

[Background 7](#_Toc163563968)

[Problem statement 7](#_Toc163563969)

[Aim of study 7](#_Toc163563970)

[Research objectives 7](#_Toc163563971)

[Significance 8](#_Toc163563972)

[Scope 8](#_Toc163563973)

[Assumptions 9](#_Toc163563974)

[Limitations 9](#_Toc163563975)

[Chapter 2 literature review 10](#_Toc163563976)

[Introduction 10](#_Toc163563977)

[Related systems 10](#_Toc163563978)

[Limitations/ weaknesses 14](#_Toc163563979)

[Solutions 15](#_Toc163563980)

[Chapter 3 methodology 17](#_Toc163563981)

[Introduction 17](#_Toc163563982)

[The methodology 17](#_Toc163563983)

[CHAPTER 4 SYSTEM ANALYSIS 19](#_Toc163563984)

[Detailed analysis 19](#_Toc163563985)

[Strengths and weaknesses of existing methods 22](#_Toc163563986)

[Gaps and opportunities 23](#_Toc163563987)

[System requirements 25](#_Toc163563988)

[Functional requirements 25](#_Toc163563989)

[Non-functional requirements 26](#_Toc163563990)

[CHAPTER 5 SYSTEM DESIGN 27](#_Toc163563991)

[Architectural design 27](#_Toc163563992)

[System overview 27](#_Toc163563993)

[Database design 33](#_Toc163563994)

[User interface design 34](#_Toc163563995)

[CHAPTER 6 IMPLEMENTATION AND TESTING 35](#_Toc163563996)

[Development environment 35](#_Toc163563997)

[System components 36](#_Toc163563998)

[CHAPTER 7 CONCLUSION 39](#_Toc163563999)

[Achievements and lessons 39](#_Toc163564000)

[Achievements 39](#_Toc163564001)

[Lessons learned 39](#_Toc163564002)

[Conclusions 41](#_Toc163564003)

[recommendations 41](#_Toc163564004)

[REFERENCES 42](#_Toc163564005)

[APPENDIX 44](#_Toc163564006)

Table of Figures

[Figure 1 Gantt Chart 18](#_Toc4097)

[Figure 2 System Overview 26](#_Toc32511)

[Figure 3 Data Acquisition 27](#_Toc28710)

[Figure 4 Analysis and prediction 29](#_Toc9909)

# Chapter 1 introduction

## Background

In the fast-paced world of technology, keeping up with trends is a significant challenge, especially for those learning or working in the field. The rapid introduction of new frameworks, languages, and tools can make choosing a technology stack for learning or projects feel overwhelming.

This project was born out of the need for a solution to this challenge. We aim to construct a machine learning model that can anticipate and track technology trends. By analyzing data from diverse sources, our model can provide predictive insights into which technologies are likely to gain popularity. This information can guide individuals in making informed decisions about which technologies to focus on, based on projected trends rather than current popularity alone. This project is not just about predicting trends; it’s about equipping individuals with the information they need to navigate the tech world.

## Problem statement

Trends in the technology field are hard to predict as a result of the dynamic nature of the field. The dynamic nature of the tech field, characterized by rapid technological advancements and shifting market demands, makes predicting trends a complex task.

## Aim of study

The aim of this study is to create a machine learning model that collects and analyses data to predict trends within the technology sphere.

### Research objectives

* Collect and preprocess data from various public sources
* Develop, train, test and deploy a machine learning model to analyse and identify trends within the data
* Evaluate the model’s performance and accuracy
* Generate and present predictive insights into technology-based trends

## Significance

The significance of this project lies in the value it holds most of all to young professionals and students. The project will merely serve as a guide validating their choices in tech fields or tech stacks by offering a data supported estimation on their longevity and virality.

The project also stands to be of great use to educators as they plan out their curriculum, offering insight on how to tailor their lessons and projects to more adequately prepare their students for the dynamic and ever-evolving tech industry. By aligning educational content with current and predicted tech trends, educators can ensure their students are equipped with relevant skills and knowledge that will remain valuable in the future.

This project also stands to be of value to market analysts and investors. Working in tandem with other forms of market research the project can offer additional insight to the viability of a technological trend as worthwhile investment.

## Scope

The scope of this project encompasses several key areas:

* Data Collection and Pre-processing: The project will collect data from various public sources, including social media platforms, job boards, tech blogs, and public datasets. The collected data will be pre-processed to ensure it is suitable for analysis.
* Model Development and Training: A machine learning model will be developed and trained to analyze and identify trends within the collected data.
* Model Evaluation: The performance and accuracy of the model will be evaluated to ensure it meets the project’s objectives.
* Insight Generation: The project will generate and present predictive insights into technology-based trends, providing valuable information for various stakeholders.
* Stakeholder Benefit: The project aims to benefit young professionals, students, educators, market analysts, and investors by providing data-supported estimations on the longevity and virality of tech fields or tech stacks.

## Assumptions

This project operates under several assumptions. Firstly, it presumes that the data from the mentioned sources is not only accessible and complies with their respective terms of service, but is also reliable. Secondly, it assumes that these data sources collectively offer a thorough and precise depiction of technology trends. Lastly, the project is based on the assumption that the machine learning model can proficiently analyze these varied data sources and accurately discern patterns within them.

## Limitations

Despite its comprehensive scope, the project has several limitations:

* Data Availability and Reliability: The project’s success heavily relies on the availability and reliability of data from the mentioned sources. Any issues with data access or quality could impact the project’s outcomes.
* Model Accuracy: While the project aims to accurately predict technology trends, the dynamic nature of the tech field could pose challenges. The model’s predictions are based on patterns in historical data, and unforeseen factors could lead to inaccuracies.
* Scope of Trends: The project focuses on technology trends, which may not encompass all factors influencing the tech industry. Other factors, such as economic conditions or regulatory changes, are outside the scope of this project.
* Time Constraints: The project timeline is another limitation. The processes of data collection, model development, training, and evaluation are time-consuming, and delays in any of these stages could impact the project schedule.

# Chapter 2 literature review

## Introduction

Currently within the market there exists multiple solutions whose main purpose is the analysis and prediction of trends however I believe each of this options misses one or more aspects that are crucial for the effective prediction of technology trends. In this section we shall discuss these solutions their strengths, their drawbacks and finally how we plan to address these drawbacks

## Related systems

A variety of established solutions offer valuable insights into current trends. Here, we explore some prominent platforms, categorized by their data sources and functionalities:

Search Engine Tracking:

* Google Trends: Leveraging its vast search query database, Google Trends provides a real-time snapshot of popular search terms (Google Trends FAQ, n.d.). Its strength lies in its sheer volume of data, offering insights into broad public interest (Internet Live Statistics, n.d.). However, the anonymized and aggregated nature of the data limits the ability to identify specific user demographics or motivations behind searches.
* YouTube Trending: As the world's second-largest search engine, with over 2.5 billion monthly active users (Global Media Insight, n.d.), YouTube's trending tab offers a glimpse into popular videos and topics. This platform is particularly effective for gauging trends in topics related to entertainment, tutorials, and emerging content formats (Broz, n.d.). A potential limitation is the potential for bias, as trending videos are influenced by user engagement and platform algorithms.

Social Media Monitoring:

* BuzzSumo: This platform goes beyond simple trend tracking, offering content performance analysis and influencer identification across social media and search engines. Its strength lies in providing a holistic view of content engagement and audience reach. However, relying solely on social media data may overlook trends that haven't yet permeated those platforms (BuzzSumo, n.d.).
* Hashtagify: This specialized tool focuses exclusively on tracking trending hashtags on Twitter. While valuable for understanding conversational trends, it presents a limited view of the broader online discourse (Hashtagify, n.d.).

Website Traffic Analysis:

* SimilarWeb: This platform provides insights into website traffic patterns, offering valuable data on website popularity and reach. It's a powerful tool for understanding competitive landscapes and audience demographics (SimilarWeb, n.d.). However, SimilarWeb's focus on existing website traffic might miss emerging trends that haven't yet translated into established web presences.

Industry Reports:

* Gartner's Emerging Technologies and Trends Impact Radar: This annual report by Gartner identifies and analyzes emerging technologies, predicting their potential impact for the upcoming year (Gartner, n.d.). Its strength lies in providing expert-curated insights and in-depth analysis. However, the yearly update cycle might miss rapidly evolving trends and lack specific action plans for capitalizing on opportunities.
* Deloitte's Tech Trends Reports: Similar to Gartner's report, Deloitte's annual analysis offers insights into anticipated trends within the tech industry (Deloitte, 2023). Its strength lies in its focus on broader trends with potential cross-industry applications. However, the focus on broad trends might not provide the level of detail needed for specific technology niches.

Search Engine Tracking:

* DuckDuckGo Trending: Similar to Google Trends, DuckDuckGo offers anonymized search trend data but emphasizes user privacy by not tracking individual search behavior. This can be valuable for capturing unbiased search patterns (DuckDuckGo, n.d.).

App Store and Play Store Analysis:

* App Annie: This platform tracks download trends, user engagement metrics, and revenue estimates for mobile applications across various app stores (App Annie, n.d.). It offers insights into popular app categories and emerging app development trends.
* Sensor Tower: Similar to App Annie, Sensor Tower provides app store analytics, including download trends, user acquisition strategies, and market share analysis (Sensor Tower, n.d.). This data can reveal user preferences for specific mobile functionalities.

News and Media Monitoring:

* Factiva: This platform aggregates news articles from a vast network of sources, allowing for trend analysis based on news coverage. It's valuable for tracking industry-specific news and identifying emerging topics gaining media attention.
* Meltwater: Another media monitoring platform, Meltwater offers social media listening and brand mentions analysis alongside traditional news coverage. This comprehensive approach helps understand how news and social media discussions converge around specific trends.

Patent Analysis:

* Espacenet: This free database from the European Patent Office allows searching and analysis of patent filings worldwide. Analyzing patent trends can provide insights into future technological advancements and R&D priorities within different industries.
* Patentlytics: This platform offers patent analytics tools and reports that identify emerging technologies and track intellectual property landscapes. It provides a deeper understanding of competitive landscapes and potential technological disruptions.

Job Board Analysis:

* Indeed Hiring Lab: This resource from Indeed provides insights into job postings and in-demand skills across various industries. Analyzing trends in job postings can reveal the skills and technologies companies are seeking, indicating future talent needs and industry shifts.
* LinkedIn Jobs: Analyzing job postings and skills endorsements on LinkedIn offers insights into talent acquisition trends and the evolving skills landscape within specific industries.

E-commerce Platforms:

* Amazon Best Sellers: Tracking best-selling products on Amazon can reveal trends in consumer preferences and identify emerging product categories.
* TrendHunter: This platform curates trends across various industries, including e-commerce. It provides insights into popular products, consumer behaviors, and emerging design aesthetics within the online shopping space.

Crowdfunding Platforms:

* Kickstarter Trends: Analyzing successful crowdfunding campaigns on Kickstarter can reveal trends in consumer interest for innovative products and technologies. These platforms offer a glimpse into early-stage ideas gaining traction and potential future market disruptions.

## Limitations/ weaknesses

While the aforementioned solutions offer valuable insights, several key limitations hinder their ability to effectively predict future technology trends, particularly within the tech sector:

* Broad Scope: Many solutions adopt a broad approach, analyzing trends across the entire internet. While this provides a comprehensive view, it dilutes the signal-to-noise ratio for specific domains like technology (Chen et al., 2014). The vast amount of extraneous data can obscure crucial insights relevant to the tech landscape.
* Historical Focus: The majority of existing solutions primarily focus on past or current trends. Understanding historical trends is valuable, but these tools often lack the capability to project future trajectories. This limits their ability to inform strategic planning and decision-making based on future scenarios (Moro et al., 2016).
* Niche Focus: Some solutions take the opposite approach, focusing solely on specific social media platforms. While social media can be a valuable source of real-time data, relying solely on these platforms can provide a skewed perspective, neglecting important data points from other sources (Cha et al., 2010).
* Long Prediction Intervals: Many solutions have infrequent updates or long intervals between trend predictions. This results in missed opportunities to capture fleeting trends and a lack of sensitivity to the dynamic evolution of ongoing trends (Yoon et al., 2019).
* Shortcomings in Prediction: Many solutions offer limited predictive capabilities. Even frequent updates might not capture the nuances of rapidly evolving trends. Additionally, some platforms (e.g., annual industry reports) have long update cycles, making them unsuitable for capturing fast-moving technological advancements.
* Data Source Bias: Each data source has inherent biases. Search engine trends might reflect user intent and search habits, not necessarily underlying technological advancements (Ye et al., 2018). App store data skews towards commercially available applications, potentially missing open-source innovations. Similarly, news and media monitoring might prioritize sensational headlines over in-depth analysis of emerging technologies (Hermida et al., 2014).

## Solutions

Our proposed project aims to overcome the limitations of current trend prediction solutions by adopting a targeted approach that leverages the unique power of Reddit data and advanced machine learning techniques. Here's how we'll achieve this:

1. Leveraging the Reddit Community:

* Real-time Insights: Reddit offers a valuable platform for gauging real-time user sentiment and discussions surrounding emerging technologies (Moro et al., 2016). By focusing on Reddit posts, we can capture the pulse of the tech community, identifying not just established trends but also early signs of new developments.
* Niche Communities: Reddit's extensive network of subreddits dedicated to specific technologies allows us to delve deeper into trends within particular industry segments (Yao et al., 2018). This targeted approach provides more granular insights compared to analyzing broad internet data.
* Authentic User Sentiment: Unlike curated news articles or press releases, Reddit discussions offer a more organic and unfiltered perspective on technology (Schwartz et al., 2012). This allows us to capture genuine user concerns, frustrations, and excitement surrounding various technological advancements.

2. Time Series Forecasting with Rich Feature Engineering:

* Model Selection: Beyond basic ARIMA models, we can explore more sophisticated techniques like LSTMs (Long Short-Term Memory) that excel at capturing complex patterns and long-term dependencies within time series data (Yoon et al., 2019). The final model selection will be based on a rigorous evaluation process considering factors like data characteristics and prediction accuracy.
* Feature Engineering for Nuance: To enhance the forecasting models, we'll go beyond analyzing just the frequency of keywords or mentions. Techniques like sentiment analysis will allow us to capture the emotional tone of discussions (positive, negative, neutral) surrounding specific technologies (Chen et al., 2014). Additionally, topic modeling can identify emerging themes and subtopics within Reddit conversations, revealing the underlying drivers of trend shifts (Wang et al., 2012).
* Network Analysis: For a more comprehensive picture, we can consider incorporating network analysis techniques. This would involve analyzing user interactions and relationships within relevant subreddits, potentially revealing influential communities and individuals shaping discussions around specific technologies (Jin et al., 2016).

3. Continuous Learning and Model Refinement:

* Dynamic Retraining: Recognizing the ever-evolving nature of online discussions and technological advancements, we'll implement a regular retraining schedule for our forecasting models. This ensures the model adapts to new trends and user behaviors, maintaining its predictive accuracy over time (Sethi & Jang, 2019).

By focusing on Reddit data and employing these advanced techniques, our project offers a more targeted and nuanced approach to predicting future trends within the technology sector. This approach leverages the real-time nature, diverse perspectives, and rich content available on Reddit, while machine learning models with feature engineering and continuous learning ensure accurate and adaptable trend predictions.

# Chapter 3 methodology

## Introduction

This chapter discusses the proposed methodology for this project, which aims to construct a machine learning model to anticipate and track technology trends. It outlines the data required, data collection methods, project resources, project schedule, and project budget.

## The methodology

This project adopted the Agile methodology as its guiding principle. Agile is an iterative and incremental development approach that prioritizes flexibility and continuous improvement. Here's why Agile proved to be the ideal choice for this project:

* Rapid Prototyping and Feedback: Agile's emphasis on rapid prototyping allowed for the creation of functional system components early on. This facilitated quick feedback loops, enabling continual refinement and adaptation based on testing and evaluation.
* Adaptability to Evolving Needs: The project's focus on uncovering trends within the ever-changing realm of tech discussions necessitated an adaptable development approach. Agile's core principle of embracing change allowed for the system to be fine-tuned as new requirements or challenges emerged during the development process.
* Collaboration and Communication: Agile promotes close collaboration between developers and stakeholders. This collaborative environment fostered open communication and ensured that the system remained aligned with project goals throughout the development lifecycle.

## Data collection methods and tools

Data collection for this project relied solely on ethical web scraping techniques utilizing Playwright, a powerful JavaScript automation library. Here's why Playwright was chosen for this task:

* Headless Browsing: Playwright operates in headless mode, mimicking a real browser without a graphical user interface. This allows for unobtrusive data collection from Reddit without impacting the user experience on the platform.
* Flexibility and Control: Playwright offers a high degree of flexibility and control over the scraping process. It allows for the development of custom scripts that target specific subreddits, filter discussions by criteria, and extract the relevant textual content efficiently.
* Efficiency and Scalability: Playwright facilitates the automation of the scraping process, enabling the system to collect data on a daily basis without manual intervention. This ensures a consistent flow of fresh data for analysis and trend prediction.

# CHAPTER 4 SYSTEM ANALYSIS

## Detailed analysis

#### Current trend analysis methods

This section delves into the various methodologies employed by different tools and techniques for identifying and analyzing current trends across diverse online and offline sources. By understanding these methods, we can leverage their strengths and weaknesses to gain a richer understanding of public interest, emerging technologies, and shifting market dynamics.

#### 1. Search volume analysis:

Search volume analysis, the cornerstone of search engine tracking, dives into the fascinating world of user queries. By analyzing how often specific terms are searched for over time, we can gain valuable insights into public interest, emerging trends, and even seasonal fluctuations in user behavior. This section delves into the details of search volume analysis, exploring its methodologies, applications, and limitations.

Methodologies:

Search engines like Google and DuckDuckGo employ sophisticated algorithms to track search queries. Here's a breakdown of the key methods involved:

* Query Logging: Every search query submitted to a search engine is logged, anonymized, and aggregated. This vast dataset forms the foundation for search volume analysis.
* Time Series Analysis: Search volume data is analyzed over time, revealing trends and patterns in user behavior. Techniques like moving averages and seasonal decomposition can help identify cyclical patterns and isolate underlying trends.
* Normalization: Search volume data is often normalized to account for fluctuations in overall search activity. This ensures that comparisons between different terms or time periods are meaningful. Normalization techniques can involve expressing search volume as a percentage of total searches or relative to a baseline period.

#### 2. Hashtag tracking:

Hashtag tracking lies at the core of this approach. Platforms like BuzzSumo and Hashtagify monitor social media platforms for trending hashtags, acting as digital stethoscopes pressed against the ever-churning current of online discourse. By identifying frequently used hashtags, these tools provide valuable insights into:

* Trending Topics: Hashtags often serve as rallying points for discussions around current events, cultural phenomena, or emerging interests. Tracking trending hashtags allows us to identify the topics capturing public attention at any given time (Boyer, 2020).
* Community Formation: Hashtags can also foster the creation of online communities. By tracking hashtags associated with specific topics, we can gain insights into the demographics and interests of the individuals participating in these communities (Yao et al., 2012).

#### 3. Network analysis:

Hashtag tracking provides a starting point, but network analysis techniques unlock a deeper understanding of the online conversation. These techniques involve constructing visualizations that map the relationships between different entities (in this case, hashtags). By analyzing these networks, we can:

* Identify Influential Voices: Network analysis can reveal which hashtags are most central to the conversation, potentially highlighting influential users or communities driving the discussion (Hussain et al., 2013).
* Uncover Underlying Themes: Analyzing the connections between hashtags can reveal the broader thematic structure of a conversation. For instance, a network analysis might expose unexpected connections between seemingly disparate hashtags, suggesting deeper underlying themes (Romero et al., 2011).

#### 5. Text Analysis and Sentiment Analysis:

Text analysis and sentiment analysis, powered by Natural Language Processing (NLP), offer a powerful lens for examining online and offline content. By dissecting written text, we can extract valuable insights into public perception, emerging trends, and the underlying themes associated with a particular topic.

Natural Language Processing (NLP) Techniques:

NLP sits at the heart of text analysis and sentiment analysis. Here's a breakdown of some key NLP techniques employed:

* Tokenization: The text is broken down into smaller units like words or phrases, called tokens.
* Part-of-Speech (POS) Tagging: Each token is assigned a grammatical category (e.g., noun, verb, adjective).
* Named Entity Recognition (NER): Identifies and classifies named entities within the text, such as people, organizations, or locations.
* Lemmatization and Stemming: Words are reduced to their base form (lemma) or stem, allowing for analysis of synonyms and variations.

Text Analysis:

Text analysis delves deeper into the content itself. Here are some applications:

* Keyword Extraction: Identifying frequently occurring words and phrases can reveal the central themes and topics discussed within the text.
* Topic Modeling: Uncovers latent topics within a collection of documents. This allows for grouping related documents and identifying underlying thematic structures.
* Entity Relationship Analysis: Explores the relationships between named entities within the text, providing insights into the connections between people, organizations, and events.

Sentiment Analysis:

Sentiment analysis focuses on understanding the emotional tone conveyed within the text. Common techniques include:

* Lexicon-Based Analysis: Relies on pre-built dictionaries containing words associated with positive, negative, or neutral sentiment.
* Machine Learning Techniques: Supervised machine learning models are trained on labeled data to identify sentiment in new text.

#### 7. Machine Learning and Predictive Analytics:

Machine learning (ML) and predictive analytics empower platforms like App Annie and Sensor Tower to delve into the vast realm of mobile app data. By analyzing historical information on app downloads, user engagement metrics, and broader market trends, these platforms leverage ML algorithms to forecast future trends and pinpoint potential growth opportunities within the mobile app space (e.g., App Annie; Sensor Tower).

Core Techniques:

* Supervised Learning: A prevalent approach in this context, supervised learning algorithms are trained on historical data sets. These data sets include information on past app downloads, user behavior within apps, and market performance. The algorithms learn to identify patterns and relationships within this data. Once trained, the algorithms can then be used to predict future outcomes for new apps or existing ones based on similar characteristics (Géron, 2019).
* Time Series Analysis: Mobile app data inherently possesses a temporal dimension, with download figures and engagement metrics constantly evolving over time. Time series analysis techniques are employed to account for these temporal patterns and seasonality. This allows for more accurate forecasting models that consider not just historical trends but also potential cyclical variations (Montgomery, Jennings, & Boretsky, 2015).

### Strengths and weaknesses of existing methods

The diverse landscape of current trend analysis methods offers a powerful toolkit for understanding public interest, emerging trends, and market shifts. However, each method comes with its own inherent strengths and weaknesses. Here's a comparative analysis of the strengths and weaknesses:

#### Breadth vs. Depth of Insights:

**Strengths:** Some methods offer a broad view of what topics are capturing public attention (search volume analysis, hashtag tracking), while others delve deeper into the content itself, revealing themes, sentiment, and relationships between entities (network analysis, text analysis, sentiment analysis). Machine learning and predictive analytics can even forecast future trends.

**Weaknesses:** Broader methods may lack insights into user demographics and motivations, while deeper analysis methods can be limited by the quality and representativeness of their data source, potentially missing niche trends or offline conversations. Machine learning's predictions can be susceptible to biases in the training data and struggle to predict unforeseen events.

#### Real-Time vs. Historical Analysis:

**Strengths:** Hashtag tracking offers real-time insights into what's trending at a given moment, while search volume analysis (with limitations) can provide historical data for trend identification. Machine learning can leverage historical data to forecast future trends.

**Weaknesses:** Raw search volume data presents a historical snapshot and may not capture real-time nuances. Text analysis, network analysis, and sentiment analysis often involve manual analysis, potentially lagging behind real-time trends.

#### Data Source and Generalizability:

**Strengths:** Search volume analysis leverages anonymized search query data, protecting user privacy. Text analysis is applicable to a wide range of text sources, including online content, news articles, and reports.

**Weaknesses:** Hashtag tracking and network analysis are limited to data available on specific social media platforms, potentially missing broader online conversations. Machine learning's generalizability is limited by the specific data sets used to train the models.

### Gaps and opportunities

While existing trend analysis techniques offer valuable insights, they have limitations that a machine learning model using daily reddit post data can address specifically within the tech space:

#### Gaps:

* Limited focus on niche tech communities: Techniques such as search volume analysis and hashtag tracking thrive on the breath of their analysis and as such may miss trends localized to tech communities within the context of broader social media data
* Data quality and bias- social media data can be susceptible to manipulation and biases. Additionally, analyzing informal language in online conversations can be challenging for NLP techniques
* Depth of user sentiment – Current methods might not provide a nuanced understanding of user sentiment beyond basic categories.

#### Opportunities:

* Targeted technology trend prediction- By focusing on Reddit post data , our model can delve deeper into discussions within specific tech communities. This allows for potentially identifying emerging trends earlier and with greater precision compared to broader web scraping methods
* Mitigating bias with community focus: leveraging Reddit's focus on subreddits. Our model can train on data from communities with specific interests, potentially mitigating bias present in broader social media platforms.
* Fine-Grained Sentiment Analysis: By leveraging NLP techniques, our model can analyze the nuances of user sentiment within Reddit posts. This goes beyond basic positive/negative categorization and uncovers subtle user attitudes towards new technologies or industry developments.

#### How Reddit data fills the gaps:

* Focuses on Tech Communities: By design, Reddit fosters communities around specific interests. Our model can target subreddits dedicated to technology, ensuring the data reflects discussions within these niche communities.
* Reduces Bias: Because our model trains on data from communities with a shared tech interest, the data is less likely to be influenced by biases present on broader social media platforms.
* Improves Sentiment Analysis: Reddit conversations often include detailed discussions and technical expertise. This allows your NLP techniques to capture a more nuanced understanding of user sentiment beyond basic categories.

By capitalizing on these opportunities, our approach has the potential to offer a more focused and insightful perspective on technological trends compared to existing methods. The ability to analyze Reddit post data presents a unique opportunity to tap into the deep well of technical expertise and passionate discussions within the online communities.

## System requirements

This section dives deeper into the specific essential specifications for building out the solution. It will be further subdivided into two categories: functional and non-functional requirements.

### Functional requirements

* Data acquisition: The system shall be able to collect Reddit data from relevant subreddits on a daily interval. Collected data shall be filtered, refined and stored based on subreddit of origin and data of collection.
* Data storage and backup: The system shall be able to store raw posts, daily topic data and predicted trend data. It shall also keep a rolling backup of each.
* Data processing and cleaning: The system shall be able to pre-process the text data from Reddit posts, including tokenization, removal of stop words and stemming.
* Topic modeling: The system shall use a Biterm model to identify relevant topics from preprocessed text data in conjunction with a custom topic classification model based on the Llama2 architecture
* Trend analysis: The system shall leverage a Long Short-Term Memory network to analyze topics, daily topic occurrences and topic impressions to predict future topic relevance.
* Output and visualization: The model shall present forecasted data in the form of time series plots as well as providing a list of topics most likely to recur.

### Non-functional requirements

* Performance: The system shall collect and process data as well as generate trend predictions efficiently within a twenty-four-hour timeframe.
* Accuracy: The system’s predictions on emerging technological trends should be demonstrably accurate when compared to real-world outcomes
* Scalability: The system should be designed to accommodate increasing volumes of data and computational demands as the project grows.
* Usability: The user interface for interacting with the system's outputs and visualizations should be intuitive and user-friendly.

# CHAPTER 5 SYSTEM DESIGN

## Architectural design

### System overview

This chapter delves into the core functionalities of the system, exploring its various facets. We'll examine each facet in detail, breaking down its components and how they contribute to the overall functionality.

Here's a breakdown of the key facets we'll explore:

* Data Acquisition and Processing: This section explores how the system gathers raw data, cleans it for analysis, and prepares it for further use.
* Storage and Backup: This section examines how the system stores crucial data and ensures its safekeeping through backups.
* Analysis: This section dives into the heart of the system, where the processed data is analyzed to generate insights.
* Display: This section delves into how generated insights are presented to the users.

By exploring each facet individually, we'll gain a comprehensive understanding of how the system functions as a whole.

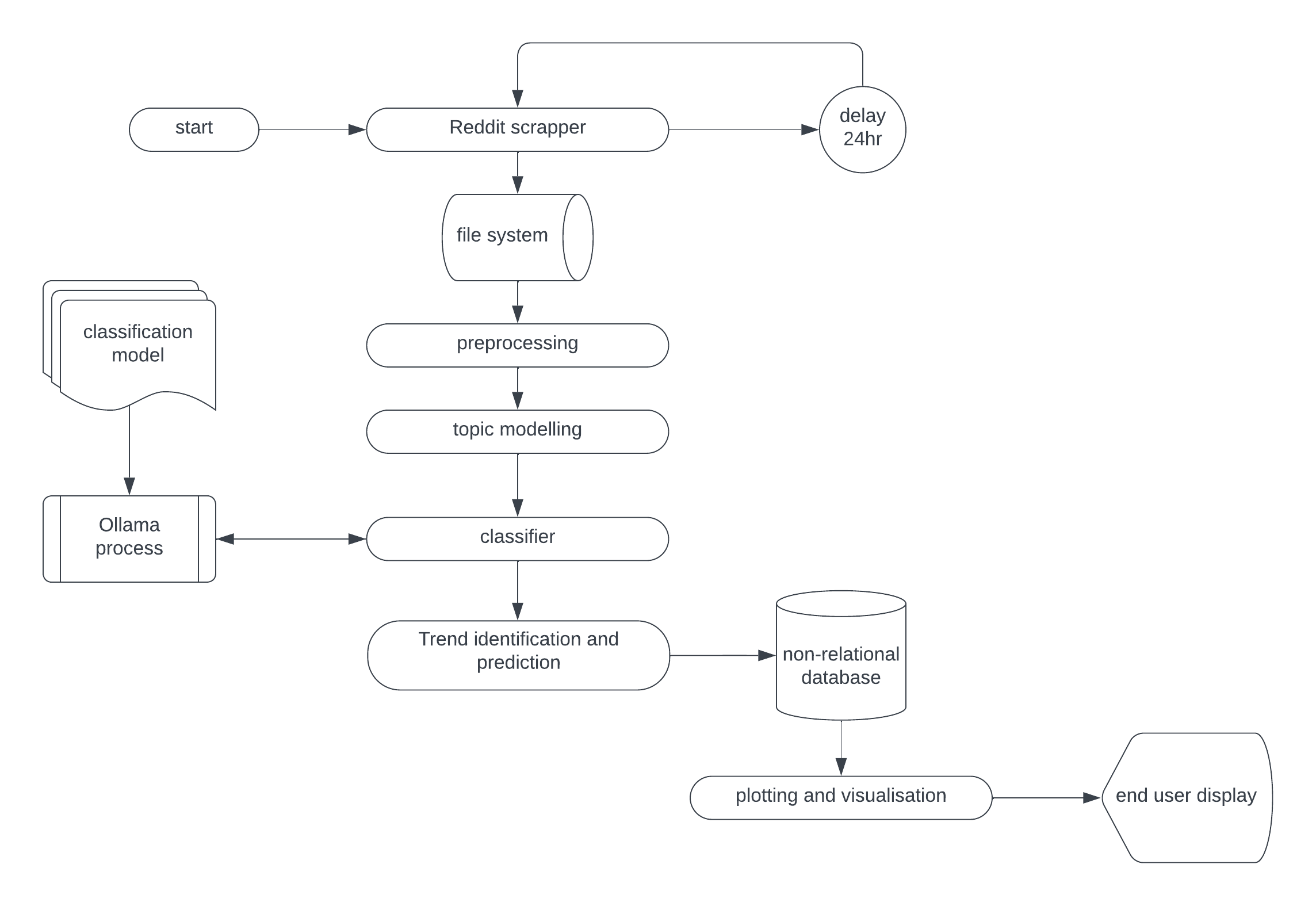


Figure 1 System Overview

#### Data acquisition and processing

The process of data acquisition and processing begins with the initialization of a headless Chrome browser for improved performance, as well as a Docker container running Tinyproxy. Tinyproxy is used to circumvent anti-scraping measures. The scraper then iterates through a list of predefined subreddits, organized by date. For each subreddit, the scraper collects all posts that occur after a certain time cutoff. This cutoff initially starts at the beginning of the year and is updated to the most recent scraping time with each subsequent scrape.

From this list of posts, comments and other attributes about the post are collected. These attributes include the datetime posted, post ID, user ID, and upvotes. This allows for the integration of different features for prediction, enhancing the robustness of the model. All this information is saved into the file system as JSON. The scraper is designed to not load images, CSS, or JavaScript, saving on computational resources. It is also set up to allow for asynchronous scraping if the length of subreddits increases.

In terms of preprocessing, the steps are relatively straightforward. The JSON data is broken down into individual posts. All unused attributes, as well as punctuations, stopwords, and any HTML carryovers, are removed. All text is converted into lowercase. Finally, each post is saved as a dictionary object, ready for topic modeling.

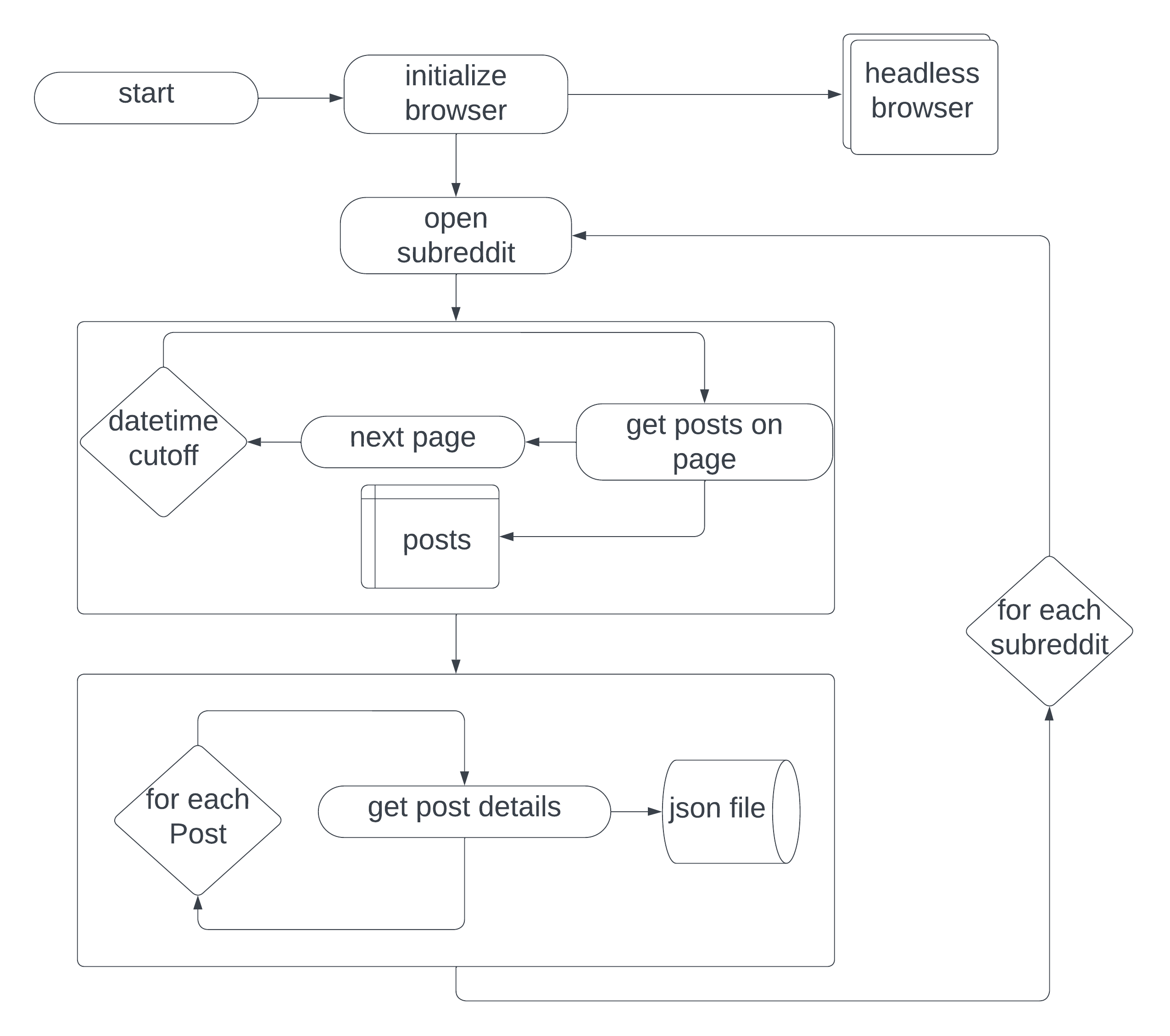


Figure 2 Data Acquisition

#### Analysis methods

After the data acquisition and processing stage has meticulously gathered and cleaned the discussions from Reddit, the journey continues towards uncovering hidden trends. Here's where the power of machine learning takes center stage.

The first step involves feeding the preprocessed text data into a quantization process. Imagine this as a way of translating the words into a numerical format that computers can understand more easily (Jain, 2018). This compressed data then becomes the fuel for the topic modeler.

But topic modeling isn't a one-size-fits-all approach. Since we're dealing with short bursts of text from Reddit posts, a specialized technique called Biterm Topic Modeling comes into play. This method excels at identifying clusters of words (topics) that frequently appear together within these concise snippets of text (Yan et al., 2009).

The topic modeler diligently analyzes the data, revealing not only the topics +themselves but also the most prominent words associated with each topic. However, these topics lack clear labels at this point. To bridge this gap, the system leverages a custom Llama2 model, a powerful classification tool, running on the Ollama platform.

Think of the Llama2 model as a highly trained expert. The system feeds it the most prevalent topic words, and the model, drawing on its knowledge, assigns a clear and concise label to each topic. This labeling process transforms the previously uncategorized topics into something more understandable, like "Artificial Intelligence Advancements" or "Emerging Cybersecurity Threats."

Now, we have all the necessary ingredients: topic labels, post dates, and upvote counts. This enriched data becomes the foundation for trend analysis and prediction. But before we delve into the future, the system performs some housekeeping.

The data undergoes a sorting process, meticulously arranging everything by date. This creates a chronological order, allowing us to track how topics evolve over time. Next comes the grouping stage, where the system cleverly combines both date and topic labels. Imagine a vast spreadsheet where each unique combination of date and topic has its own cell. The system then meticulously counts how many times each combination appears, essentially building a map of topic popularity across time.

With all the features meticulously prepared, the data is fed into a Long Short-Term Memory (LSTM) model. This sophisticated model is like a time travel machine for trends. It analyzes the historical data, considering factors like topic prevalence, upvote counts, and their chronological progression (Hochreiter & Schmidhuber, 1997). By meticulously examining these patterns, the LSTM model attempts to predict how these trends might unfold in the future.

The beauty of this system lies in its continuous learning process. As the system gathers more data over time, the predictions become increasingly accurate. This highlights the importance of scalability – the ability to handle ever-growing volumes of data. The more information the system has access to, the better it becomes at anticipating the ever-changing tides of tech trends within the online communities of Reddit.

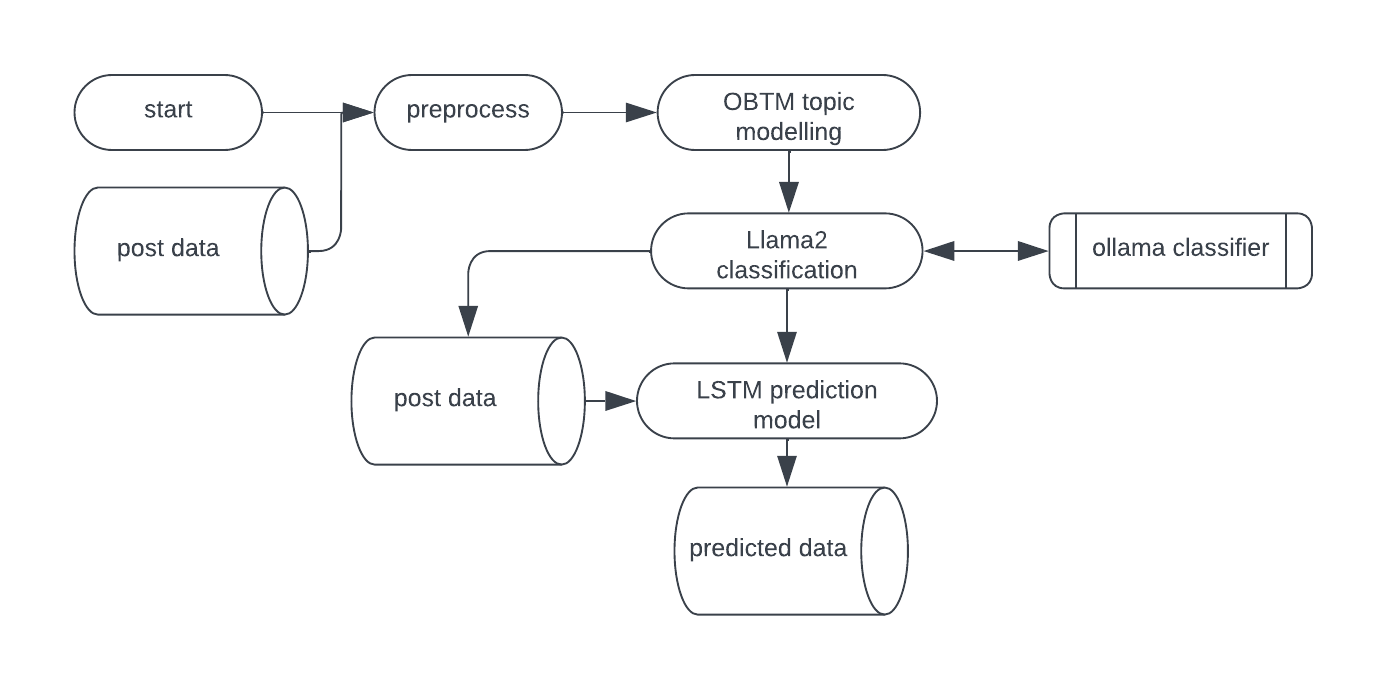


Figure 3 Analysis and prediction

#### Data storage and backup

Following the meticulous data acquisition and processing stages, the system meticulously organizes the harvested knowledge for future use. Here's a breakdown of how different data formats are employed:

Preserving Scraped Data: JSON for Flexibility (Stephenson, 2007)

The raw data scraped from Reddit is stored in JSON (JavaScript Object Notation) files. JSON offers a flexible and human-readable format that can seamlessly represent the hierarchical structure of the Reddit data. Each file is named with a combination of the date it was scraped and the subreddit it originated from. This dual naming convention facilitates efficient retrieval and allows for easy identification of data origin.

For instance, a file named "2024-04-08\_r-artificial" would contain data scraped on April 8th, 2024, from the "r-artificial" subreddit. This approach simplifies filtering and analysis based on specific subreddits or timeframes.

Preprocessing Output: CSV for Efficiency (Mock & O'Neil, 2017)

After the cleaning and processing steps, the data is transformed into a Comma-Separated Values (CSV) format. CSV files are a simple and widely supported format ideal for storing large datasets with a tabular structure. Each row in the CSV file corresponds to a preprocessed post, and each column represents a specific feature extracted from the post (e.g., post content, topic label, upvote count).

The choice of CSV stems from its efficiency in data exchange and ease of use. This format allows for seamless integration with various data analysis tools and visualization platforms, enabling further exploration and insights to be gleaned from the preprocessed data.

Trend Analysis Powerhouse: SQLite Database for Scalability

The system ultimately groups the data by date and topic label, creating a comprehensive picture of topic prevalence over time. This enriched data is then stored in a SQLite database. SQLite is a lightweight and embedded relational database management system – essentially a structured storage solution – that excels in handling large datasets efficiently (Hipple, 2007).

Storing the grouped data in a database offers several advantages. Firstly, it allows the Long Short-Term Memory (LSTM) model to access all the historical data in a structured and efficient manner. The LSTM model utilizes this comprehensive historical record to identify patterns and trends in topic popularity, ultimately leading to more accurate predictions.

Secondly, databases excel in data management tasks like querying and filtering. This enables the system to efficiently retrieve specific data subsets for further analysis or to focus on trends within particular timeframes or topic areas.

Future Predictions: Daily Updated CSV

The predictions generated by the LSTM model are stored in a separate CSV file. This file is updated daily, with new predictions appended to the existing data. Maintaining a CSV format for predictions ensures consistency with the preprocessed data and facilitates seamless integration with potential visualization or reporting tools.

The daily update schedule reflects the dynamic nature of online trends. By incorporating the latest data points into the predictions, the system continuously improves its ability to anticipate future developments within the tech landscape.

In conclusion, the system employs a strategic combination of JSON, CSV, and SQLite database formats to efficiently manage and utilize the collected data throughout its journey from raw scraping to insightful trend predictions. Each format selection is deliberate, offering flexibility, efficiency, and scalability to empower the system's ability to unlock valuable insights from the ever-evolving world of online discussions.

## Database design

The system employs a well-orchestrated, multi-layered database design to manage the data throughout its journey. This design capitalizes on the strengths of various storage methods to ensure efficient handling from raw data to insightful predictions.

The foundation of the system lies in JSON (JavaScript Object Notation) files stored within the file system. JSON offers a flexible and human-readable format, perfectly suited to represent the hierarchical structure of the Reddit data collected during the scraping stage. This allows for easy storage and retrieval of the raw data in its original form.

To facilitate efficient retrieval and identification of data origin, a dual naming convention is implemented for these JSON files. For instance, a file named "2024-04-08\_r-artificial" would contain data scraped on April 8th, 2024, from the "r-artificial" subreddit.

After the data is meticulously cleaned and processed, the system stores the preprocessed information in a CSV (Comma-Separated Values) format within the file system. CSV files excel in efficiency and are widely supported, making them ideal for storing large datasets with a tabular structure. Each row in the CSV file corresponds to a preprocessed post, and each column represents a specific feature extracted from the post (e.g., post content, topic label, upvote count).

This choice of CSV allows for efficient data exchange and seamless integration with various data analysis tools and visualization platforms. These tools empower further exploration and the extraction of valuable insights from the preprocessed data.

Following preprocessing, the system groups the data by date and topic label, creating a comprehensive picture of topic prevalence over time. This enriched data is then stored in a dedicated SQLite database. SQLite, a lightweight and embedded relational database management system, is ideally suited for handling large datasets efficiently.

Storing the grouped data in a database unlocks several advantages. Firstly, the Long Short-Term Memory (LSTM) model can access all the historical data in a structured and efficient manner, which is crucial for identifying patterns and trends in topic popularity. This ultimately leads to more accurate predictions. Secondly, databases excel in data management tasks like querying and filtering. This enables the system to efficiently retrieve specific data subsets for further analysis or to focus on trends within particular timeframes or topic areas.

The predictions generated by the LSTM model are stored in a separate CSV file maintained within the file system. This file is updated daily, with new predictions appended to the existing data. Maintaining a CSV format for predictions ensures consistency with the preprocessed data and facilitates seamless integration with potential visualization or reporting tools.

The daily update schedule reflects the dynamic nature of online trends. By incorporating the latest data points into the predictions, the system continuously improves its ability to anticipate future developments within the ever-evolving tech landscape.

## User interface design

The user interface prioritizes a clean and uncluttered design, focusing on delivering immediate and crucial insights at a glance. Here's how the interface empowers users:

* Cards for Key Metrics: Essential data points are presented in clear, concise cards. This includes the total number of posts collected (providing a sense of data volume and confidence), the overall prediction confidence (calculated as the average percentage accuracy of predictions), and the standard deviation of prediction features on a daily basis (indicating the variability in the data used for predictions).
* Time Series Plot: Focus on High-Confidence Predictions: A time series plot visualizes the five topics with the highest prediction confidence. This plot displays the names of these topics alongside their predicted occurrence counts plotted against nine date values. Each date is spaced a week apart, with the fifth date representing the current day. This visualization effectively separates historical data (the first four plots) from predicted data (the last five plots), allowing users to easily identify trends and anticipate future developments.
* Detailed Topic Information: A comprehensive table delves deeper into individual topics. The table displays the topic names, predicted occurrence counts (expected number of mentions in future discussions), predicted impressions (anticipated activity related to the topic), and the predicted trend (upward or downward). Additionally, the table includes either the prediction probability or the prediction confidence, providing further details about the system's certainty in its future predictions.

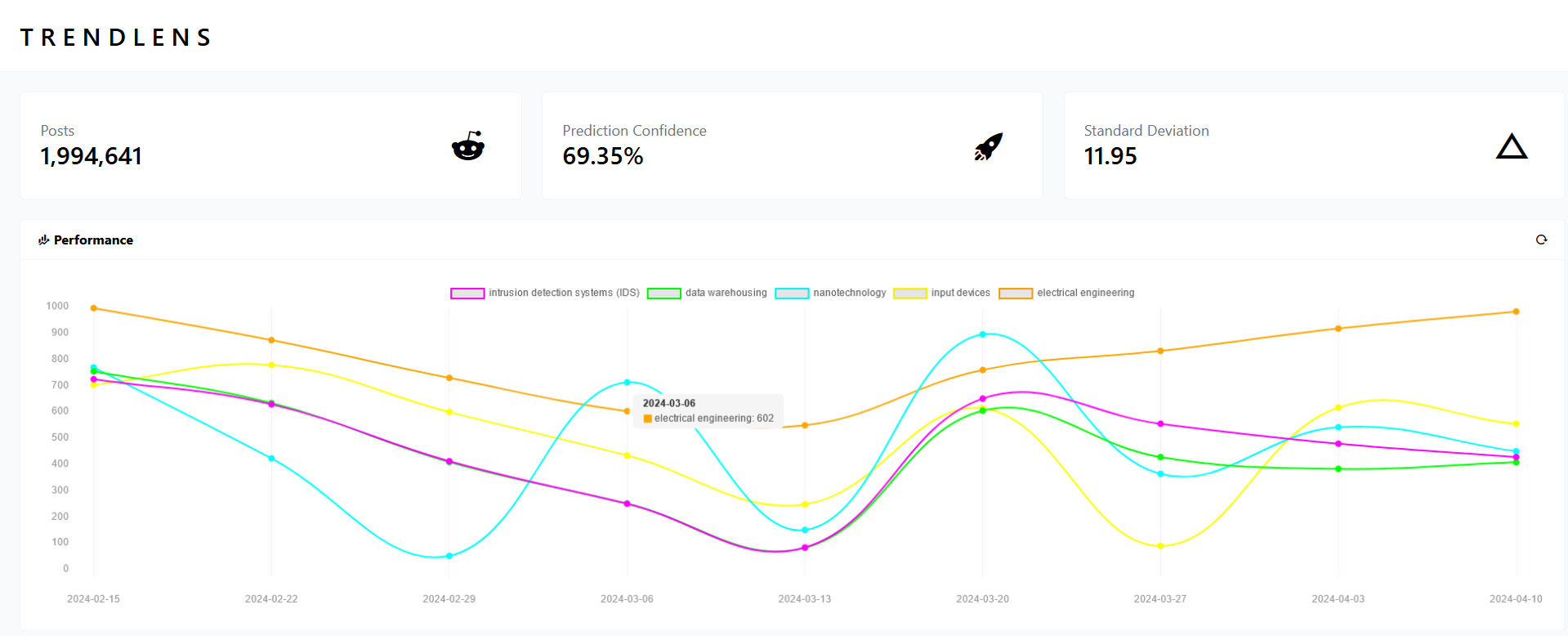
This user interface design strikes a balance between clarity and comprehensiveness. By presenting key metrics, highlighting high-confidence predictions, and offering detailed topic information, the interface empowers users to efficiently grasp the system's insights and make informed decisions based on anticipated tech trends.

Figure 4 user interface 1

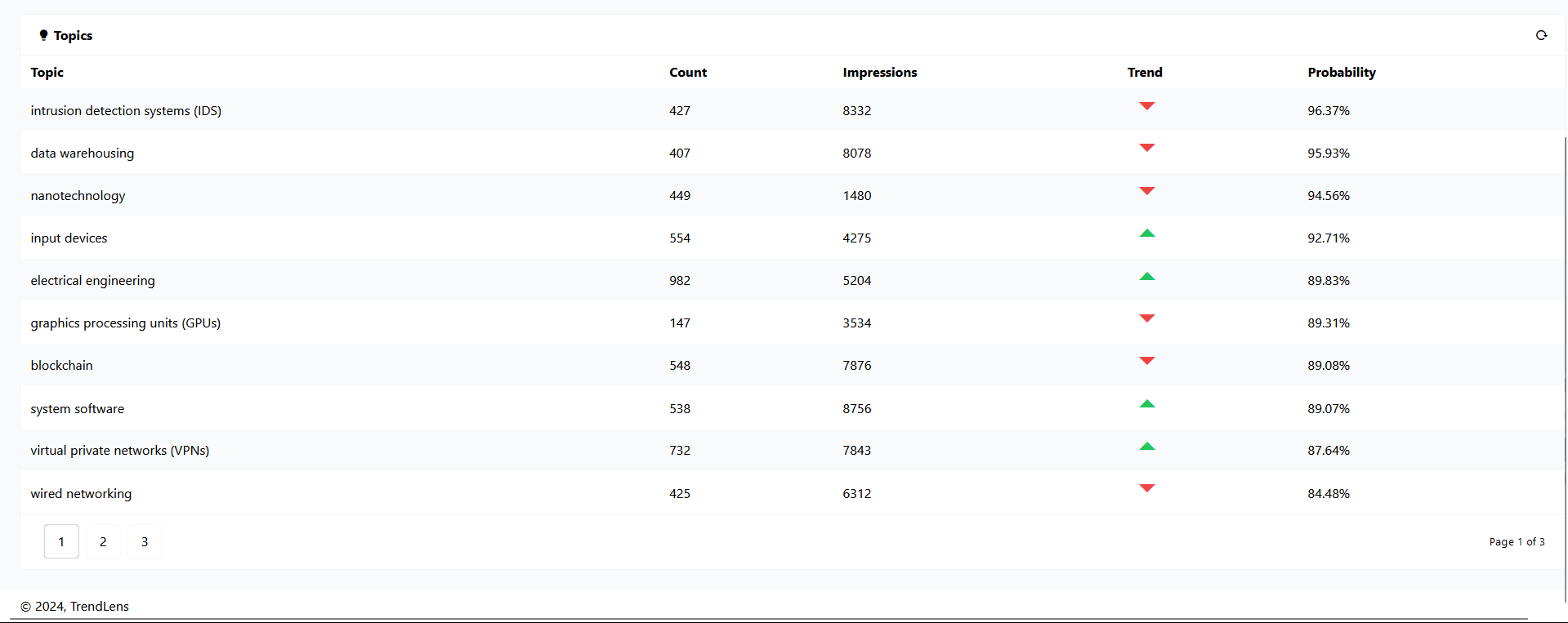


Figure 5 user interface 2

# CHAPTER 6 IMPLEMENTATION AND TESTING

## Development environment

The system is constructed upon a carefully selected combination of development tools and languages. This selection ensures efficiency, flexibility, and a seamless workflow throughout the development process.

* Web Scraping with Playwright: Playwright, a powerful JavaScript framework, takes center stage for web scraping tasks. Playwright empowers the system to interact with web browsers like headless Chrome in a programmatic way, efficiently gathering data from Reddit. This headless approach allows the system to run in the background without a physical browser window, making it resource-efficient and scalable for large-scale data collection.
* R for Preprocessing Power: R, a robust programming language specifically designed for statistical computing and graphics, tackles the data preprocessing stage. R offers a rich library of functions and packages ideal for cleaning, transforming, and preparing the scraped text data for further analysis. Additionally, the vast community of R users and the wealth of available R packages provide ongoing support and access to cutting-edge techniques for text manipulation and data exploration.
* Python and TensorFlow for Predictive Prowess: Python serves as the foundation for building the prediction model. TensorFlow, a popular deep learning library within the Python ecosystem, is leveraged to construct and train the Long Short-Term Memory (LSTM) model. This model is tasked with analyzing the processed data to identify patterns and ultimately predict future trends. TensorFlow's flexible architecture and extensive community resources make it ideal for building and deploying complex machine learning models.
* SQLite for Data Persistence: SQLite, a lightweight and embedded relational database management system, is employed for data persistence. SQLite efficiently stores the grouped data, which holds the key to understanding topic prevalence over time. This data is readily accessible by the LSTM model for training and prediction purposes. Additionally, SQLite's self-contained nature eliminates the need for separate database server software, simplifying deployment and maintenance.
* Integrated Development Environments: VS Code and RStudio: The development workflow benefits from the use of two prominent Integrated Development Environments (IDEs). Visual Studio Code (VS Code) serves as the primary code editor for Python development, offering features like syntax highlighting, debugging tools, and version control integration. RStudio, a specialized IDE for R, provides a user-friendly interface for working with R data and code, streamlining the data preprocessing stage. Both VS Code and RStudio offer extensive plugin ecosystems, allowing for further customization and integration with various data science tools and libraries.

In essence, this combination of development tools fosters an efficient and streamlined workflow. Playwright streamlines data acquisition, R tackles data cleaning, Python and TensorFlow handle model building, and SQLite ensures data persistence. Finally, VS Code and RStudio provide user-friendly interfaces for development within their respective programming languages. This well-coordinated environment empowers the system to transform raw scraped data into actionable insights and future predictions. Moreover, the focus on open-source tools like Playwright, R, and TensorFlow ensures cost-effectiveness and fosters collaboration within the developer community.

## System components

The system is comprised of a well-orchestrated ensemble of components, each playing a crucial role in the journey from raw Reddit data to insightful trend predictions. Let's delve into the functionalities of each component:

1. Web Scraper: The web scraping component, powered by Playwright, acts as the system's maestro. Playwright leverages its JavaScript prowess to interact with Reddit in a programmatic way, efficiently gathering discussions relevant to tech trends. This often involves utilizing a headless Chrome browser, allowing the system to run unobtrusively in the background. The scraper can also be configured to target specific subreddits, ensuring a focused data collection process.

2. Preprocessor: Once the data is acquired, the preprocessor takes center stage. Built upon the robust R language, this component meticulously cleans and transforms the raw scraped text. This stage may involve removing unnecessary elements like stop words (common words like "the" or "a") and applying techniques like stemming or lemmatization to normalize word forms. Additionally, the preprocessor might extract specific features from the text, such as sentiment analysis or topic labels, for further analysis.

3. Topic Modeler: The topic modeling component, employing the Biterm Topic Modeling technique, delves deeper into the preprocessed text. This technique excels at identifying clusters of words (topics) that frequently appear together within these short snippets. By analyzing these word co-occurrences, the topic modeler uncovers the underlying themes and discussions prevalent within the Reddit data.

4. Classifier: Following topic modeling, the system leverages a custom Llama2 model, running on the Ollama platform, to assign clear and concise labels to the identified topics. Think of Llama2 as a highly trained expert. The preprocessor feeds it the most prominent words associated with each topic, and Llama2, drawing on its knowledge, assigns labels like "Artificial Intelligence Advancements" or "Emerging Cybersecurity Threats." This labeling process transforms the previously uncategorized topics into a more understandable format.

5. Trend Analysis: The system then meticulously groups the data by both date and topic label. This step allows for the creation of a comprehensive map depicting topic popularity over time. Essentially, the system builds a vast "spreadsheet" where each unique combination of date and topic has its own cell. By meticulously counting how many times each combination appears, the system gains valuable insights into the evolution of different tech trends within the Reddit community.

6. Prediction Model: Armed with the enriched data from the trend analysis stage, the system employs a Long Short-Term Memory (LSTM) model, built using TensorFlow. This sophisticated model acts as a time travel machine for trends. It analyzes the historical data, considering factors like topic prevalence, upvote counts, and their chronological progression. By meticulously examining these patterns, the LSTM model attempts to predict how these trends might unfold in the future.

7. Data Visualizer: While not explicitly mentioned as a component, data visualization plays a crucial role in effectively communicating the system's findings. Visualization tools can be employed to present the predicted trends and historical topic popularity in a clear and engaging manner. This allows users to easily grasp the insights gleaned from the Reddit data and make informed decisions based on anticipated future developments.

# CHAPTER 7 CONCLUSION

## Achievements and lessons

### Achievements

This project successfully harnessed the power of data to extract valuable insights from online discussions. The implemented system achieved several key milestones, demonstrating the effectiveness of the chosen methodologies:

* Efficient Data Acquisition: Playwright, a web scraping library, served as a skilled data collection agent. It efficiently gathered a substantial volume of relevant data from Reddit, ensuring a continuous stream of fresh material for analysis.
* Meticulous Preprocessing: R, a prominent language for statistical computing, played a crucial role in the data preprocessing stage. It facilitated the meticulous cleaning and transformation of the raw text data, preparing it for further exploration and subsequent analysis.
* Predictive Modeling Capabilities: By leveraging the power of Python libraries like TensorFlow and the Long Short-Term Memory (LSTM) model, the system achieved the remarkable ability to predict future trends within the dynamic landscape of tech discussions on Reddit.
* Clear Communication through Visualization: While not explicitly developed as a core component, data visualization played a critical role in effectively communicating the system's insights. By presenting the predicted trends and historical topic popularity in a clear and engaging manner, the system empowered users to readily grasp the valuable knowledge extracted from the vast corpus of Reddit data.

### Lessons learned

As a newcomer to the data science field, this project served as a transformative learning experience. It exposed me to a plethora of essential methodologies and tools that continue to be valuable assets:

* Data Manipulation Expertise: The project fostered a deep understanding of data manipulation techniques. This encompasses data wrangling, cleaning, feature engineering, and transformation – all fundamental skills for extracting meaningful information from any dataset.
* R Proficiency: My journey with R proved to be highly rewarding. This powerful language for statistical computing and graphics became a key tool for data exploration and preparation. Its extensive capabilities will undoubtedly continue to play a vital role in future data science endeavors.
* Python Library Mastery: The project solidified the importance of Python libraries like NumPy, Pandas, and TensorFlow within the data science toolkit. These libraries offer immense flexibility and computational power when working with numerical data, machine learning models, and deep learning applications.
* Web Scraping Expertise: Playwright and its web scraping capabilities emerged as valuable assets. The project provided practical experience in ethically gathering data from online platforms, a skill that can be applied in various data science scenarios.
* Docker for Efficient Management: Delving into Docker, a containerization platform, provided valuable insights into containerization best practices, setup, and deployment strategies. This knowledge empowers efficient application management and simplifies collaboration within development teams.
* LLM Exploration and Optimization: The project presented an opportunity to explore the fascinating world of Large Language Models (LLMs). Working with Ollama and Hugging Face platforms offered a glimpse into the customization and optimization processes for these powerful language models.
* Practical Application of NLP Concepts: The project served as a practical introduction to Natural Language Processing (NLP) concepts like Latent Dirichlet Allocation (LDA), Biterm Topic Modeling (BTM), and sentiment analysis. These techniques played a crucial role in extracting meaning from the vast amount of textual data collected from Reddit discussions.
* Stemming and Text Normalization Techniques: The project provided hands-on experience with stemming and other text normalization techniques. These methods allow for the effective analysis of textual data by reducing words to their base forms, facilitating a more comprehensive understanding of the content

## Conclusions

The project culminated in the successful development and implementation of a data-driven system capable of gleaning valuable insights from online discussions. The system achieved this through several key functionalities:

* Automated Data Acquisition: The system leverages Playwright, a web scraping library, to automate the process of data collection from Reddit. This ensures a consistent stream of fresh data on a daily basis, providing the system with the necessary fuel for analysis and trend prediction.
* Rigorous Preprocessing Pipeline: R, a powerful language for statistical computing, plays a vital role in the system's preprocessing pipeline. Raw text data undergoes meticulous cleaning and transformation processes within the R environment, preparing it for further analysis and ultimately enabling the extraction of meaningful trends.
* Predictive Trend Identification: By employing a Long Short-Term Memory (LSTM) model built using the TensorFlow library, the system possesses the remarkable ability to identify and predict future trends within the ever-changing landscape of tech discussions on Reddit. This predictive capability empowers users to stay ahead of the curve and make informed decisions based on anticipated developments.

This project stands as a successful culmination of focused effort, perseverance, and a thirst for knowledge. It not only delivered the ability to predict trends with a reasonable degree of accuracy but also fostered a deep appreciation for the power of data science methodologies. The lessons learned and the skills developed throughout this project will undoubtedly serve as a stepping stone on the path to future data science endeavors. As I continue to explore this ever-evolving field, the skills and knowledge gained will empower me to tackle more complex problems and unlock even deeper insights from the vast ocean of data.

## Recommendations

To continuously improve the system's accuracy, we recommend exploring additional features during the prediction phase. This could involve incorporating new data sources, fine-tuning existing algorithms, or implementing ensemble learning techniques that combine multiple models.

Furthermore, the system's asynchronous data collection presents an opportunity for significant scalability. By leveraging this asynchronous nature, we can explore distributed processing frameworks or message queueing systems to handle larger data volumes efficiently.

Investing in higher-performance computing resources would also contribute to faster topic modeling, classification, and prediction. This could involve utilizing GPUs or specialized hardware optimized for machine learning tasks.

Finally, to broaden the system's impact, we recommend a phased rollout to a wider audience. This allows for controlled user testing and iterative improvement based on real-world feedback. By continuously monitoring performance and gathering user insights, we can ensure the system remains effective and valuable for its growing user base.

# REFERENCES

* Broz, M. (n.d.). How Many Videos Are on YouTube? [https://www.youtube.com/playlist?list=PLBE338967F8DB7F2A](https://www.youtube.com/playlist?list=PLBE338967F8DB7F2A" \t "_blank)
* Google Trends FAQ. (n.d.). Retrieved from
* <https://policies.google.com/privacy?hl=en-US>
* Global Media Insight. (n.d.). YouTube Users Statistics. Retrieved from [https://www.globalmediainsight.com/blog/youtube-users-statistics/](https://www.globalmediainsight.com/blog/youtube-users-statistics/" \t "_blank)
* Internet Live Statistics. (n.d.). Google Search Statistics. Retrieved from
* <https://www.internetlivestats.com/>
* BuzzSumo. (n.d.). BuzzSumo website: [https://buzzsumo.com/](https://buzzsumo.com/" \t "_blank)
* Hashtagify. (n.d.). Hashtagify website: [https://hashtagify.me/](https://hashtagify.me/" \t "_blank)
* SimilarWeb. (n.d.). SimilarWeb: Digital Intelligence Platform. <https://www.similarweb.com/website/similarweb.com/>
* Deloitte. (2023). Tech trends 2024. <https://www2.deloitte.com/us/en/insights/focus/tech-trends.html>
* Gartner. (n.d.). Gartner tech insights. <https://www.gartner.com/en>
* DuckDuckGo. (n.d.). DuckDuckGo - Privacy, simplified. [[https://duckduckgo.com/](https://duckduckgo.com/" \t "_blank)]
* Cha, M., Haddadi, H., & Yun, A. (2010). Measuring user influence in social media: A survey. ACM Computing Surveys (CSUR), 42(1), 1-44.
* Chen, W., Zhang, C., & Yuan, Y. (2014). Large scale social media sentiment analysis: A survey. IEEE Intelligent Systems and their Applications, 29(1), 3-11.
* Hermida, A., Fletcher, R., Koren, J., & Papacharissi, Z. (2014). Tracking the mobilization of alternative news media: Twitter, WikiLeaks and the Arab Spring. Political Communication, 31(3-4), 279-301.
* Moro, A., Rita, P., & Romero, P. (2016). Social media mining for emerging technology trends. Journal of Information Science, 42(3), 302-322.
* Ye, M., Zhu, K., Luo, X., Liu, S., & Wu, Y. (2018). A review of web search query log analysis for information retrieval. Knowledge and Information Systems, 57(2), 389-422.
* Yoon, S., Benbow, C. M., & Bechhofer, S. (2019). A comprehensive survey of trend analysis in social media. ACM Computing Surveys (CSUR), 52(2), 1-38.
* Chen, W., Zhang, C., & Yuan, Y. (2014). Large scale social media sentiment analysis: A survey. IEEE Intelligent Systems and their Applications, 29(1), 3-11.
* Jin, J., Yang, J., & Zhou, S. (2016). Uncovering the structure of online social movements: A network approach. Information Science, 357, 190-201.
* Moro, A., Rita, P., & Romero, P. (2016). Social media mining for emerging technology trends. Journal of Information Science, 42(3), 302-322.
* Schwartz, A., Leavitt, N., & Hanson, J. L. (2012). Tweets as a tool for gauging public response to a natural disaster: # Sandy case study. First Monday, 17(10).
* Sethi, S. S., & Jang, Y. M. (2019). A comparative analysis of deep learning methods for short-term and long-term stock price prediction. Expert Systems with Applications, 115, 630-643.
* Wang, D., Wei, F., Song, X., Zhou, M., & Ma, Z. (2012). Topic sentiment analysis in social media: A LDA based approach. Proceedings of the 2012 International Conference on World Wide Web (WWW '12), 638-647.
* Yao, L., Luo, C., Zhang, Y., & Yang, S. (2018). Extracting emerging technology trends from online forums: A case study using Reddit. Journal of Information Science, 44(8), 1009-1023.
* Yoon, S., Benbow, C. M., & Bechhofer, S. (2019). A comprehensive survey of trend analysis in social media. ACM Computing Surveys (CSUR), 52(2), 1-38
* Boyer, D. (2020). Hashtag publics: The power and politics of algorithmic assemblages on Twitter. Television & New Media, 21(8), 1801-1822. <https://www.academia.edu/19870714/Hashtag_Publics_The_Power_and_Politics_of_Discursive_Networks>
* Chen, H., Zhou, R., Shang, Y., & Liu, J. (2014). Epidemic spreading in social-tagged networks. Physica A: Statistical Mechanics and its Applications, 410, 422-433. <https://www.sciencedirect.com/science/article/abs/pii/S0065260108601082>
* Gruhl, V., Guimera, R., & Amaral, L. A. N. (2004. The dynamics of social balance in networks. Complex Networks, 403-433. <https://link.springer.com/book/10.1007/978-3-319-40803-3>
* Hussain, A., Shen, J., Wang, X., & Zhou, X. (2013). Identifying influential users based on topic sentiment analysis. IEEE Transactions on Knowledge and Data Engineering, 25(4), 882-896. <https://www.researchgate.net/publication/352621441_A_Survey_on_Concept-Level_Sentiment_Analysis_Techniques_of_Textual_Data>
* Romero, D. M., Meeder, B., & Barbara, J. (2011. Applications of topic models in digital libraries. Journal of the American Society for Information Science and Technology, 62(2), 246-264.
* Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python (Analyzing text with statistical methods). O'Reilly Media.
* Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the ACL on Empirical Methods in Natural Language Processing, 10(Thumbs), 79–86. DOI: 10.3115/1118691.1118701
* Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras & TensorFlow: Concepts, tools, and techniques to build intelligent systems (2nd ed.). O'Reilly Media.
* Montgomery, D. C., Jennings, M., & Boretsky, M. (2015). Introduction to time series analysis and forecasting (2nd ed.). John Wiley & Sons.
* Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780. https://direct.mit.edu/neco/article/9/8/1735/6109/Long-Short-Term-Memory
* Jain, P. (2018). Quantization in machine learning. In Quantization (pp. 3-22). Springer, Cham. https://www.sciencedirect.com/science/article/abs/pii/S0925231220319354
* Yan, X., Zhao, J., & Huang, T. (2009). Biterm topic modeling for short texts. In Proceedings of the 18th ACM conference on information and knowledge management (pp. 445-454). ACM. <https://dl.acm.org/doi/10.1145/2488388.2488514>
* Hipple, R. S. (2007). SQLite: Database in your pocket. O'Reilly Media, Inc.
* Mock, S., O'Neil, P. (2017). Data warehouses and data mining for business analytics. John Wiley & Sons.
* Stephenson, N. (2007). Cryptonomicon. HarperCollins. (Note: This reference is for the use of JSON, not the specific book content)

# APPENDIX

Project Schedule

Figure 6 Gantt Chart