

MULTIMEDIA UNIVERSITY OF KENYA

FACULTY OF COMPUTING & INFORMATION TECHNOLOGY

PROJECT DOCUMENTATION

TrendLens

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:03/04/2023

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](#_Toc159371029)

[Abstract 3](#_Toc159371030)

[Chapter 1 introduction 6](#_Toc159371031)

[Background 6](#_Toc159371032)

[Problem statement 6](#_Toc159371033)

[Aim of study 6](#_Toc159371034)

[Research objectives 6](#_Toc159371035)

[Significance 6](#_Toc159371036)

[Scope 7](#_Toc159371037)

[Assumptions 7](#_Toc159371038)

[Limitations 7](#_Toc159371039)

[Chapter 2 literature review 8](#_Toc159371040)

[Introduction 8](#_Toc159371041)

[Related systems 8](#_Toc159371042)

[Limitations/ weaknesses 9](#_Toc159371043)

[How the proposed solution handles these weaknesses 9](#_Toc159371044)

[Chapter 3 methodology 11](#_Toc159371045)

[Introduction 11](#_Toc159371046)

[The methodology 11](#_Toc159371047)

[Data collection 11](#_Toc159371048)

[Project resources 11](#_Toc159371049)

[Project schedule 11](#_Toc159371050)

[APPENDIX 12](#_Toc159371051)

[References 12](#_Toc159371052)

Table of Figures

[Figure 1 Gantt Chart 12](#_Toc159374962)

# 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

The project will employ a supervised machine learning approach. We will use historical data to train our model, which will then be used to predict future trends. The model will be designed to analyze data from diverse sources, identify patterns, and make predictions based on these patterns. We will use a combination of classification and regression techniques to achieve this.

## Data collection

Data will be collected from various public datasets such as the Stack Exchange API, web traffic time series, GitHub trending repositories, and World Bank data catalogs. We will also utilize sentiment analysis datasets from platforms like Twitter and Reddit. To capture real-time trends, we will employ web crawling techniques on job boards, tech blogs, and Twitter threads.

## Project schedule

The project is scheduled to be completed in a structured manner over specific time frames. The task of data collection will commence on 12th December 2023 and conclude by 15th February 2024. Preprocessing of the collected data will start from 1st January 2024 and will be completed by 25th February 2024.

The development of the model will begin on 1st February 2024 and will be completed by 11th March 2024. Concurrently, the training of the model will start from 12th February 2024 and end on 11th March 2024.

Following this, the testing of the model will be carried out from 11th March 2024 to 31st March 2024. The evaluation phase will then take place from 18th March 2024 to 31st March 2024.

Finally, the report summarizing the project’s findings and outcomes will be prepared from 1st April 2024 and will be completed by 10th April 2024. This timeline ensures a systematic and efficient execution of the project.

Figure 1 Gantt Chart

# CHAPTER 4 SYSTEM ANALYSIS

## Detailed analysis

## System requirements

### Functional requirements

### Non-functional requirements

# CHAPTER 5 SYSTEM DESIGN

## Architectural design

## Database design

## User interface design

# CHAPTER 6 IMPLEMENTATION AND TESTING

## Development environment

## System components

## Test plan(test data, test cases, test results)

# CHAPTER 7 CONCLUSION

## Achievements and lessons

## Conclusions

## recommendations

# REFERENCES

* Broz, M. (n.d.). How Many Videos Are on YouTube? <https://www.youtube.com/playlist?list=PLBE338967F8DB7F2A>
* 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/>
* Internet Live Statistics. (n.d.). Google Search Statistics. Retrieved from
* <https://www.internetlivestats.com/>
* BuzzSumo. (n.d.). BuzzSumo website: <https://buzzsumo.com/>
* Hashtagify. (n.d.). Hashtagify website: <https://hashtagify.me/>
* 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/>]
* 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

# APPENDIX