AI-POWERED NEWS SENTIMENT AND BIAS ANALYSIS

UTILIZING TRADITIONAL MACHINE LEARNING AND LLAMA 3 TO REVEAL THE TRUE COLORS OF THE NEWS

AAU BUSINESS SCHOOL

BUSINESS DATA SCIENCE

FRI-88520-2

SPRING SEMESTER 2024

AALBORG UNIVERSITET

Participants:

Oliver Gade Nielsen - 20205968

Patrick Riber Sørensen - 20205289

Supervisor

Hamid Bekamiri

Submission date:

31/05/2024

Page Numbers:

38

Characters incl. spaces

66.700

1. ABSTRACT

In today's information-rich environment, news media play a pivotal role in shaping public opinion and influencing decision-making processes at individual, organizational, and societal levels. However, inherent biases in news reporting, coupled with the proliferation of misinformation, present significant challenges. This project aims to address these challenges by leveraging Natural Language Processing (NLP) through traditional machine learning models and Large Language Models (LLMs) to analyze news media for sentiment, bias, and political leaning. By combining the strengths of Support Vector Machines (SVMs) and advanced LLMs like Llama 3, the project seeks to provide actionable insights that can help businesses and individuals make informed decisions. The research demonstrates that while traditional models like SVMs offer valuable baselines, LLMs significantly enhance prediction accuracy and reliability. The project includes a comprehensive data pipeline, robust model training, and a user-friendly frontend application built with Streamlit. Future implementations will focus on improving data quality, scalability, and the overall effectiveness of the analysis. This work underscores the potential of advanced NLP tools to combat news media narratives, by giving insights on sentiment, bias and political leaning.

TABLE OF CONTENTS

2. INTRODUCTION:

In today's world, news outlets have significant power in terms of their ability to impact and affect various levels of society, be it specific groups, individuals, organizations, or businesses.

News media are most people's primary source of information about what is happening in the world around them, be it locally, nationally, or globally. This means they hold significant power in shaping public opinion, and this power is further amplified by their ability to control the narrative, as well as pick and choose which news they want to cover and how they are presented.

Furthermore, news outlets have significant impact on the decision-making processes at both the individual and societal level. The news outlet that an individual person prefers can shape that individual's way of thinking to a significant degree. (Sharifi, E., Ebrahimi Kahrizsangi, K., & Aghaei Chadegani, A. 2024)

Most news sources portray themselves as neutral interpreters of the world they function in, and according to three surveys, from 1976, 1986 and 1996, 94% of the 1,220 journalists described themselves as passive observers, who only report the news, or active observers who investigate and uncover the news. Only 6% of the asked journalists chose to describe their reporting as "interventionary "or "agenda-setting." (McCarthy, Killian J., Dolfsma 2014)

But the fact is, that even if they consider themselves neutral, this is not the case. The neutrality of the news will always be impacted by various things, even if the reporters are trying to stay as neutral as possible. An aspect affecting neutrality and creating bias, could be as simple as what news the media outlets choose to report on and which they do not. Furthermore, journalists are all humans, every different story by a different author, even on the same subject, will differ in tone, phrasing, and wording. (Menen 2021)

News will always be biased, in some sense of the word. However, as we established before, bias is always inherent in the news, but there is a very clear difference between news outlets that are intentionally being biased, compared to news outlets who are not. Distorted news and "alternative" facts have been used for various propaganda-oriented agendas for years. However, in more recent times the concerns of these "alternative facts" have seen a dramatic increase, especially as a result of Donald Trump coining the phrase "Fake News" during his 2016 presidential campaign. (Menen 2021)

These concerns are highly bounded in reality and since the term "Fake News" came into the spotlight, it has spawned an entire industry based on the concept.

An example of what Fake News looks like could be the Belgium company Media Vibes SNC, who are the owners of more than 180 urls devoted to creating and spreading fake news online, primarily through social media. It furthermore provides individual users with an application to develop their own fake news. (Menen 2021)

The media is anything but neutral, even setting the concept of fake news aside. Journalists are incentivised to be negatively biased, as it attracts more readers and even those who want to be neutral are inherently biased by their way of selecting, framing, and analyzing events. (McCarthy, Killian J., Dolfsma 2014)

The media has a significant impact on businesses in multiple ways. For example, in 2002 it was discovered that Volvo unintentionally increased the electromagnetic fields of their cars by moving the car battery from the front to the back. With no scientific evidence that this had any negative impact on human health, the narrative the media reported this with created a public outcry, which ended with a significant impact to Volvo's image, forcing them to recall the affected cars, as well as reducing their electromagnetic fields. (McCarthy, Killian J., Dolfsma 2014)

According to Tetlock, high media pessimism predicts downwards pressure on market prices. If the pessimism fluctuates unusually high or low, it is a predictive indication of high market trading volume. In a sense the sentiment of the media impacts the market. In a 2012 study a significant link between the media overall sentiment and firms' willingness to create new patents of knowledge was found. Specifically, overall negative sentiment in the news media was linked with a decrease in number of new patent applications. (McCarthy, K.J., Dolfsma, W., & Huizingh, E. 2012)

The same effect can be seen on the consumer side. Negative sentiment has a significant effect on consumer's spending habits. Emotive words such as "crisis", "slowdown" and "downturn" made predictable short-term fluctuations in the spending of the consumers. (McCarthy, Killian J., Dolfsma 2014)

It is clear that the media holds significant power when it comes to affecting businesses and the markets in their entirety. It is a potent variable when it comes to understanding the fundamental changes in the economy. (McCarthy, Killian J., Dolfsma 2014)

For these reasons, creating a tool or technology capable of understanding media narratives, sentiment, and bias could have huge potential for individuals, businesses, and organizations. It could be a useful tool in terms of predicting market behavior, political decision-making, and individual decision-making. For instance, stockbrokers could use it to make more informed decisions when it comes to trading.

In recent times, Large Language Models (LLMs) have become and are still becoming increasingly powerful. They show impressive capabilities when it comes to their ability to perform various Natural Language Processing (NLP) tasks.

Alongside LLMs, traditional machine learning techniques, such as Support Vector Machines (SVMs), have also been widely used for NLP tasks.

Although these traditional methods may not match the performance of modern LLMs, they provide valuable baselines and can complement advanced models in a hybrid approach.

For this reason, this project will utilize both LLMs and traditional machine learning models to create an application for predicting news sentiment, bias, and political leaning. By combining the strengths of both approaches, the project aims to enhance the reliability and accuracy of the predictions and answer the research question:

How can natural language processing through traditional machine learning and large language models be leveraged by businesses to analyze news media for sentiment, bias, and political leaning, in order to gain actionable insights and make informed decisions?

3. LITERATURE REVIEW

3.1 CONTEXT

Throughout history, news has played a vital role in shaping societies and influencing human affairs throughout history. In today's digital world, news media comes in many shapes and forms, whether it be traditional newspapers, television, online papers, or social media posts and videos. (Posetti, Matthews 2018)

In the modern digital age, news media has undergone a radical transformation. The internet and social media platforms have democratized the production and distribution of news, giving individuals the ability to participate in public discourse like never before. However, this democratization has also brought new challenges, including the spread of misinformation, echo chambers, and algorithmic bias. (Posetti, Matthews 2018)

For this reason, the interest in technological solutions has seen a tremendous rise. One such technology that has seen increasing popularity due to its rapid advancements is Large Language Models (LLMs). Large Language Models, such as the well-known ChatGPT by OpenAI, offer entirely new opportunities for various forms of news media. (Posetti, Matthews 2018)

3.2 SCOPE

As with many forms of technology, LLMs can be utilized for both righteous and malicious purposes. While the list of immoral uses for LLMs is long, this literature review focuses on their potential for combating media narratives by analysing sentiment, bias, and political leaning in the news, rather than on their malicious uses.

3.3 PURPOSE

The purpose of this literature review is to provide a comprehensive overview of current research on the usage of Large Language Models for sentiment analysis and combating the spread of misinformation and bias in the news media.

Misinformation has become a significant threat to information ecosystems and public trust. Historically, efforts to combat misinformation have relied on fact-checking organizations, media literacy campaigns, and algorithmic detection methods. However, the advent of LLMs presents both new opportunities and challenges. (Chen, Shu 2023)

3.4 OPPORTUNITIES AND CHALLENGES:

LLMs have tremendous knowledge and reasoning capabilities, which could potentially be used to combat misinformation. Their ability to process text allows them to identify false claims, provide accurate information.

This potential comes from their extensive training on large amounts of data, enabling them to understand and contextualize information across a wide array of topics. (Chen, Shu 2023)

Recent research suggests that LLMs can be instrumental in automating the fact-checking process. They can quickly cross-reference claims with reliable sources, which makes fact checking more efficient. LLMs can aid in the spreading of accurate information by generating clear and comprehensible explanations that debunk false narratives. These capabilities highlight the potential for LLMs to combat misinformation by giving people powerful tools to "fight" back. (Chen, Shu 2023)

Regardless of their potential benefits, LLMs also have risks. Their ability to generate coherent and contextually relevant text makes them susceptible to misuse for creating deceptive misinformation at scale. Malicious actors can exploit LLMs to produce false content, which seems like legitimate information, which could increase the spread of misinformation. (Chen, Shu 2023)

This two-sided nature of LLMs supports the need for strong safeguards and ethical guidelines, so that we can prevent their misuse. Current efforts to address this involve creating mechanisms to detect AI-generated content, implementing strict access controls, and helping different fields work together to make stronger regulatory frameworks. (Chen, Shu 2023)

3.5 PERFORMANCE AND EFFICIENCY OF LLMS

The performance of language models is improving as the number of parameters increases, but this comes with higher training costs. Popular methods to achieve parameter-efficient fine-tuning, such as adapters and quantization, can reduce the learnable parameters without significantly compromising performance. Future directions include applying adapters to even larger models, such as those with 13 billion or 30 billion parameters, to enhance their efficiency and effectiveness. (Aman 2024)

Research has shown that by focusing on specific tasks it is possible to trade off some generative capabilities to achieve better performance in areas like self-instruction. For instance, a study involving the Llama large language model demonstrated that feeding over 70,000 instructions into a pre-trained model, combined with parameter-efficient fine-tuning, resulted in more accurate fake news detection. This was attributed to the model's enhanced logical understanding, text comprehension, and sentiment analysis capabilities. (Aman 2024)

3.6 SENTIMENT ANALYSIS WITH LARGE LANGUAGE MODELS

Sentiment analysis (SA) has long been a focal point in natural language processing (NLP), providing insights into human sentiments and opinions.

The emergence of LLMs has opened new avenues for enhancing SA tasks, but their full potential and limitation in this field has not yet been achieved. (Zhang, Deng et al. 2023)

LLMs have shown promise in performing various SA tasks, from conventional sentiment classification to more complex aspect-based sentiment analysis and multifaceted analysis of subjective texts. Studies evaluating the performance of LLMs across 13 tasks on 26 datasets indicate that LLMs deliver decent results in simpler tasks, such as basic sentiment classification. They excel particularly in few-shot learning scenarios, where they significantly outperform smaller language models trained on domain-specific datasets. (Zhang, Deng et al. 2023)

The performance of LLMs diminishes in more complex tasks that require a deeper understanding of context or structured sentiment information. These limitations highlight the need for more sophisticated training approaches and evaluation methods to fully harness the capabilities of LLMs in sentiment analysis. (Zhang, Deng et al. 2023)

The future work on LLMs, such as Llama, for detecting misinformation and fake news accurately is still an unexplored area with a lot of potential. Besides the need to make the models more accurate and efficient, the LLMs also need considerably more knowledge such as incorporating e.g. visual data. (Aman 2024)

In this project, the focus is largely on the general performance of LLMs for sentiment analysis, bias detection, and political leaning detection. LLMs have significant potential, but there is still a long way before LLMs can accurately mirror or even perform the task of fact checking better than an actual human, however the potential is there, and considering the scalability and processing power of LLMs compared to humans, they could definitely be a powerful tool in combating misinformation.

4 METHODOLOGY

4.1 A CRITICAL REALIST APPROACH

Critical realism is the philosophical foundation we have chosen for this project. Its emphasis on underlying mechanisms aligns well with our use of machine learning to uncover latent patterns in text data. Critical realism encourages looking beyond superficial text features to understand deeper structures and influences in news content, supporting the modeling of complex relationships within the data. (Bygstad 2011)

Understanding that all knowledge is context-dependent is essential when training models that capture cultural and temporal contexts within data. For example, machine learning models need continuous updating and refinement to adapt to new contexts over time. This aligns with the critical realist view that all knowledge is fallible and evolves. (Bygstad 2011)

Critical realism helps ensure that machine learning outputs are useful and reliable by combining solid data with human judgment. It recognizes that all information has some subjectivity and bias, valuing the balance between objective data and subjective views. It also stresses the importance of ethical considerations. It promotes ongoing reflection on how we train models, choose data, and consider social impacts, ensuring our project is developed responsibly, especially given the potential for misuse of models like LLMs. (Bygstad 2011)

Using critical realism as a philosophical framework provides a comprehensive and reflective approach. It ensures that the project not only leverages advanced technological capabilities but also remains critically aware of the complexities, context, and ethical considerations inherent in analyzing media content. This alignment can lead to more robust, contextually aware, and ethically sound outcomes in understanding and interpreting news media. (Bygstad 2011)

4.2 SCOPE OF THE PROJECT:

To answer the problem formulation, it is important that we establish how the project views key elements for the sake of scientific method. The goal of the project is to build an application, which based on news headlines attempts to predict the sentiment of these news, as well as the political leaning and the bias. It is important to establish exactly how these concepts are to be viewed:

Sentiment Analysis:

Sentiment analysis is about establishing the emotional tone of different types of media, in this case of news articles. The way you label sentiment can differ from project to project, however typically the labels will be “Positive”, “Negative” and “Neutral.”

Political orientation

Political leaning refers to the ideological perspective that seems to be conveyed based on the news article's headline. There are many different labels you can use to categorize the political orientation with, e.g. “liberal, conservative or neutral” or, as is the case for this project “left-leaning, right-leaning or centrist.” For the sake of further clarification, this is how the project views them individually:

Left-leaning

Political orientation which advocates for individual liberties, social equality, and the view that the government should take a proactive role in addressing social issues and inequality. Generally viewed as progressive when it comes to changes in society and policies, which support social justice. Considered to be Liberalism.

Right-leaning

Political orientation which focuses on tradition, social stability, and limited government intervention - both for individuals but also in relation to the economy. Considered to be Conservatism.

Centrist

Political orientation which does not strongly align with either liberalism or conservatism. Moderates generally seek a balanced approach, in which ideas from both sides should be blended together. Considered to be “Neutral.”

As the project's main focus does not only revolve around political orientation, there will not be a deep dive into research and theoretical considerations about political orientation. However, it was deemed an interesting variable to have in consideration, when it comes to analyzing sentiment and to consider why some articles might be biased, while others are not.

Bias

Bias in news, whether in the broader reporting or specifically in headlines, refers to the presence of unfair or unbalanced presentation, where certain viewpoints, facts, or perspectives are favored over others. In news reporting, this bias can take various forms, such as the selection of stories, framing of issues, use of language, or omission of information. Similarly, bias in news headlines involves the partiality or prejudice present in the brief, attention-grabbing text that introduces a news article. Headlines aim to encapsulate the essence of a story in just a few words, but they can still exhibit bias through selective framing, language choice, or omission of key information. In both cases, bias can lead to a distorted representation of events, influencing public perception and opinion. (Xiang, Y., & Sarvary, M. 2007)

4.3 PROJECT ARCHITECTURE

This project is essentially built with three different components, the business component, the data component and machine learning operations component.

4.4 THE BUSINESS COMPONENT

The business side of the project forms the general outline, problem and goal, which the project tries to solve by utilizing Data Science techniques such as implementing Machine Learning solutions.

It is important that this project is business orientated, therefore trying to enlighten on a specific business problem, which in this case is how to utilize machine learning and Large Language Models to combat the effect, which misinformation and negative sentiment in the news media can have on businesses. The end goal of this project is to give an answer to the problem formulation, by potentially offering a solution or giving general directions towards an achievable future solution for the problem.

4.5 THE DATA COMPONENT

The data aspect of the project focuses on the general work of the data scientist and their scientific approach to solving problems. Generally, it is important that the data utilized for the Machine Learning and Large Language Models is both relevant and of high quality. Here are the general data science concept the project has implemented when working with the data:

Data Acquisition

The process of collecting the data, there are many options here, news data can typically be collected through API calls, existing databases, or scraping news websites manually.

Data Cleaning and Preprocessing:

This step involves cleaning the raw data to remove any noise, inconsistencies, or irrelevant information that are not needed. For the natural language processing (NLP) tasks the project will be working on, there is a large focus on NLP oriented data preprocessing such as text normalization, tokenization, and lemmatization.

Feature Engineering:

Feature engineering is the process of creating new features or transforming existing ones to improve the performance of machine learning models.

Exploratory Data Analysis (EDA): EDA involves analyzing and visualizing the dataset to gain insights into its structure, distribution, and relationships between variables. It helps in identifying patterns in the data that could impact the performance of our models.

Data Labeling and Annotation: Data labeling involves assigning labels or annotations to the dataset based on predefined criteria. In this project the essential end goal is labeling and has been done by various methods like using LLM or creating synthetic and manual labels for training the SVM-model.

4.6 THE MACHINE LEARNING OPERATIONS COMPONENT:

The Machine Learning Operations (MLOps) aspect of the project is about the Model from backend to frontend, how it is deployed, and the pipeline is built. The project is a continuation of an earlier project made in the Machine Learning and Data Engineering Module of the MSc Business Data Science course. This aspect is very important for the actual operations of the application that is built based on the project. MLOps generally works with aspects such as: model development, model evaluation, model deployment and model monitoring and maintenance. (Hopsworks 2024b)

The goal of this component is to build an application in relation to the problem formulation and the report itself and have it working from backend to frontend. What this means for this project, is that it is necessary to launch an application utilizing any sort of frontend framework such as Streamlit or Gradio and have it be accessible as e.g. a web application.

The goal here is to build a working machine learning inference pipeline, which will function independently of human interaction. (Hopsworks 2024b)

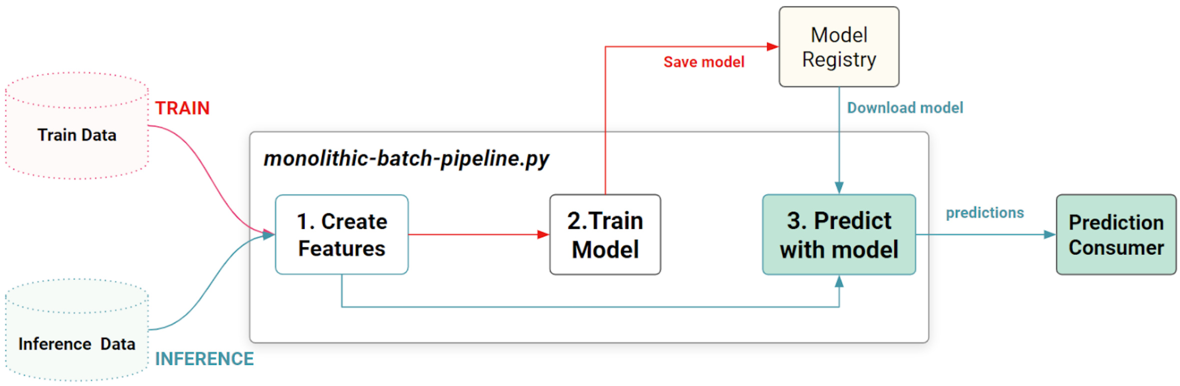

FIGURE 1 INFERENCE PIPELINE (HOPSWORKS 2024B)

As it can be seen in figure 1, the inference data goes from create features to Predict with model to Prediction for the consumer. It is important that the final product will be launched and have this process run automatically without human interference.

It is also important that the project keeps a feature store and a model registry (Hopsworks 2024c), the purpose of this is that it makes it easier to discover and reuse features for different models and compare performance of models as well as making them easily accessible as well. (Hopsworks 2024a)

4.7 THE OVERALL GOAL OF THE PROJECT

The overall goal of this project is to determine whether it is possible to implement machine learning and LLMs to predict sentiment, bias, and political orientation in news articles, and to provide a satisfying answer to our research question. The project focuses primarily on delivering a proof of concept, supported by theoretical knowledge, application development, and final evaluation of the metrics to substantiate our arguments.

Currently, building a fully functional model that performs these tasks perfectly is, according to our evaluation and knowledge, a massive undertaking that is too time-consuming and resource-intensive to be within scope. Instead, we aim to establish whether such a model could potentially be built, what value it holds in a business-oriented environment, and what the potential challenges and limitations might be. This exploratory approach will help us provide a comprehensive and informed answer to our research question.

The project wants to showcase a proof of concept of using natural language processing through machine learning models for sentiment analysis, bias detection, and political orientation labeling.

As for what the expectations are for what could potentially be concluded for this project, it is expected that we will be able to establish that SVMs and LLMs can capture our three main variables in the news titles and that they do indeed have potential for both businesses and individuals as a tool to combat narratives in the news media.

5 THEORETICAL FRAMEWORK

5.1 SUPPORT VECTOR MACHINES (SVMS)

Support vector machines (SVMs) are supervised machine learning algorithms used primarily for classification tasks. SVMs classify data by finding an optimal line or hyperplane in an N-dimensional space that maximizes the distance between classes. (IBM 2024b)

SVMs are commonly employed in classification problems where the goal is to separate data into two distinct classes. The algorithm works by finding the hyperplane that maximizes the margin between the closest points of the classes, known as support vectors. These support vectors are critical as they run through the data points determining the maximal margin. The number of features in the input data determines whether the hyperplane is a line in a 2-D space or a plane in an N-dimensional space. By maximizing the margin, SVMs are able to find the best decision boundary, enabling accurate classification predictions and generalization to new data. (IBM 2024b)

In cases where data is not linearly separable, SVMs utilize kernel functions to transform the data into a higher-dimensional space, facilitating linear separation. This technique, known as the "kernel trick," involves various types of kernel functions such as linear, polynomial, radial basis function (RBF), and sigmoid kernels. The choice of kernel depends on the characteristics of the data and the specific use case. (IBM 2024b)

SVMs are particularly effective in high-dimensional spaces and can efficiently handle situations where the number of dimensions exceeds the number of samples (scikit-learn 2024a). This makes them suitable for tasks like text classification, where methods like TF-IDF vectorization create a large number of features. Additionally, SVMs are memory efficient as they use a subset of training points, the support vectors, in the decision function. (IBM 2024b)

Key features of an SVM model include:

- High-dimensional spaces: SVMs perform well in high-dimensional spaces and are effective even when the number of dimensions exceeds the number of samples.
- Memory efficiency: SVMs use a subset of training points (support vectors), making them memory efficient.
- Versatility: SVMs support various kernel functions (e.g., linear, polynomial, RBF, and sigmoid), allowing customization for different decision functions. (scikit-learn 2024a)

Despite their advantages, SVMs have some limitations. There is a risk of overfitting when dealing with high-dimensionality and small sample sizes. Obtaining probability estimates also requires expensive cross-validation, increasing computational costs in terms of time and resources. Moreover, the computational and storage requirements for SVMs escalate rapidly with the number of training vectors. (IBM 2024b)

In this project, we utilize a support vector machine model for classifying and labeling the sentiment of news articles.

5.2 LARGE LANGUAGE MODELS

Large Language Models, prominently known as LLMs, are powerful foundation models, which have been trained on incredibly large dataset, which has led to large advancements in the fields of natural language understanding and natural language processing. They can perform a large variety of tasks across many different domains and can comprehend and generate text in a very human-like manner. LLMs are at the forefront of public interest, also in the business perspective, as many organizations now strive to implement artificial intelligence or LLMs for various business-oriented tasks. (IBM 2024a)

LLMs may have seemed to appear very suddenly, but their development has been on the way for several years and are building upon many years of research and developments in the machine learning and deep learning fields. IBM (The International Business Machines Corporation) one of the largest industrial research organizations have been implementing LLMs to enhance their natural language understanding and processing for years. (IBM 2024a)

LLMs are designed to understand and generate text similar to a human. They can infer context, generate coherent and in-context responses, translate languages, summarize text, answer questions, analyze sentiment, and assist in creative writing and code generation tasks among many more things.

The continued development of LLMs, which have already redefined some aspects of the modern digital landscape, are like to revolutionize various fields, such as research, business, and the general daily lives of individuals. (IBM 2024a)

5.3 META LLAMA 3 8B

For this project, the LLM model we chose to work with was Llama 3. Llama 3 is one of the most powerful openly available models at this current time. According to Meta themselves, they created a high-quality human evaluation set. This evaluation set includes 1,800 prompts that span 12 key use cases: seeking advice, brainstorming, classification, answering closed questions, coding, creative writing, information extraction, role-playing a character/persona, answering open questions, reasoning, rewriting, and summarization. (meta 2024)

As it can be seen in figure xx., the model outperforms other openly available models on this evaluation set a significant amount of the time.

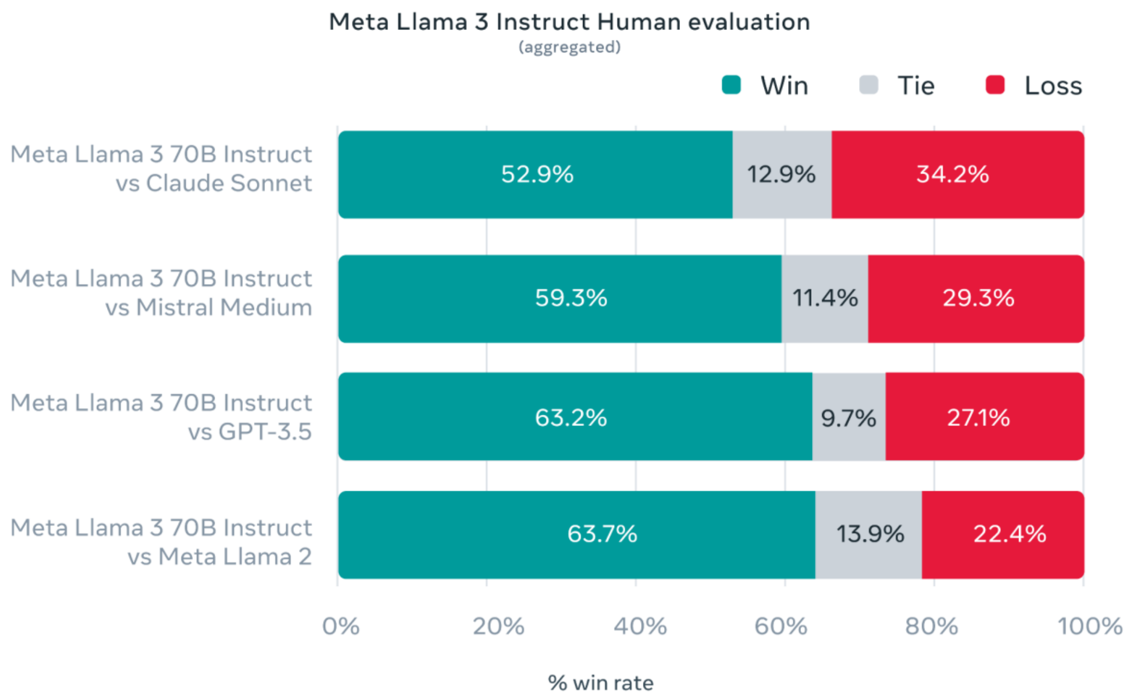

FIGURE 2 OVERVIEW OF META LLAMA 3 70B PERFORMANCE VS. OTHER LLMS ON EVALUATION DATA. (META 2024)

However, running the 70B model (70 billion parameters) seems to be overkill for the purpose of sentiment analysis, therefore the model utilized was the 8B model (8B parameters). The lower number of parameters makes them cheaper to use, with more resource efficiency, lower computation costs and faster responses. (meta 2024)

6 METHOD

6.1 DATA COLLECTION

Data collection for this project was approached through two different methods, each with its own advantages and considerations. Initially, we utilized a prelabeled news dataset available on Kaggle, which had been collected using the NewsAPI.org web service. This dataset offered a swift solution to any data scarcity issues, instantly providing us with access to a large amount of prelabeled sentiment data (Approximately 54k news articles). This allowed us to instantly start experimenting and building our first model for the project. (Saksham 2024)

However, it is important to acknowledge the limitations inherent in using prelabeled data. The lack of transparency regarding the labeling methods and the accuracy of the labels introduces unknown variables that could significantly impact the final model. Therefore, while the prelabeled dataset served as a valuable starting point, it was important to combine it with our own data collection efforts.

To continue our work while maintaining consistency with our established process, we also utilized the NewsAPI.org to collect additional data. This allowed us to build our own dataset, which gave us significantly more control and knowledge on the "hows" of the data, such as how the data was to be labeled. During this process we also implemented criteria for selecting articles, addressed issues such as duplicates and irrelevant content, and ensured diversity in the dataset.

Currently, the data used for this project is being collected with a python script, which receives the data through an API call and stores it in our SQL database.

6.2 DATA PREPROCESSING:

Before we can start building models with the data, it is essential to perform preprocessing procedures to make sure the data is cleaned and prepared. There are vast number of preprocessing steps you can take, for this project these are the ones we did. (Uysal, Gunal 2014)

Removing Irrelevant Entries

We start by filtering out articles marked as '[Removed]'. This step ensures that our dataset includes only relevant and available content, eliminating unnecessary noise and focusing on useful information.

Handling Missing Values

It is important to make sure that we handle missing values in our dataset, as these can negatively affect our results and potentially also make a lot of the process faulty and mess up the code.

There are many ways you can go about handling missing values, in this project we dropped all the data entries that had missing in the critical columns e.g. 'title'. In other columns we filled the missing values with a placeholder such as “Empty”.

Standardizing Date Formats

To maintain uniformity and facilitate chronological analysis, we standardize the publication dates by converting them to a consistent format. This standardization helps in comparing and sorting articles based on their publication dates.

Text Cleaning

We then clean the text data by removing HTML tags and other non-text elements from the content and description fields. We convert all text to lowercase to ensure uniformity and strip out unnecessary punctuation and special characters. This step is crucial for preparing the text for further analysis.

Tokenization and Lemmatization

The text is broken down into individual words or tokens through tokenization. We also apply lemmatization or stemming to reduce words to their base or root forms. This process treats different forms of a word as a single term, which is essential for accurate text analysis.

Stop Words Removal

We eliminate common stop words such as "and," "the," and "is," which do not add significant meaning to the analysis. Removing these words helps in focusing on the more important terms that contribute to the sentiment, bias, and political orientation of the articles.

Duplicate Removal

We identify and remove any duplicate articles to avoid redundant analysis. This step ensures that each article in our dataset is unique, preventing the skewing of results due to repeated content.

TF-IDF processing

Term Frequency - Inverse Document Frequency, also commonly known as TF-IDF, is a preprocessing technique used for transforming text data into numerical features so the text data can be used for machine learning algorithms. It works by measuring the frequencies of terms in the documents (TF) and then measuring the importance of the terms in the document relative to the entire corpus (IDF). This combined score helps in highlighting significant words while reducing the impact of commonly occurring, less informative words, making the data more suitable for predictive modeling. This process was used for the text data used in the project. (Pradeep 2023)

6.3 FEATURE SELECTION:

To build our machine learning model, it was important that we extracted the correct features. When retrieving the data through the API, the number of features is vast, therefore we had to select the features, which were relevant to our purposes. Following is a list of the features and why we decided to go with these.

Source: It is important to identify where the articles are coming from, this add validity to the results of the project, as it will not just be a random collection of articles, which could be from anywhere. Furthermore, there addition of the source of articles gives the possibility for further analysis, such as comparison between news outlets and showing discrepency between them based on either of the criteria that we are trying to label.

Title: Title is perhaps the most important feature of our data, for the purposes of the project. It is the title, which we will be utilizing as input for our machine learning models and it is based on this title the models will make predictions.

Description: Description was also extracted from the data; this one was selected because in the beginning the main intent was to use the description for the labeling. However, due to issues with this specific feature, which we will touch upon later. This one was not really used.

URL: The URL allows us to access the full content of the articles, in case there is a need of access to the entire articles. It also provides with the ability to let the enduser access the whole article from the frontend of the application.

URL to Image: This one is not important for the modelling, but it does enhance the end-user's experience in the frontend of the application, by providing the visual images from the article.

Published Date: It is important we retrieve the publication date, both as it will be used to understand the chronology of the data, but also because the end-application will be displaying the newest articles from each publication, which makes knowing the publication date essential. Furthermore, the published date was standardized to ensure that all articles followed the same date-time scheme.

Content: The full content of the article. This was also one of the features, which the project had envisioned could be used for analyzing sentiment, bias, and political orientation. However, due to issues with this feature, this one was scrapped as well.

6.4 MODEL TRAINING:

There are many different models which can be employed for text analysis tasks like sentiment analysis, bias detection, and political orientation classification. For this project, we experimented with several machine learning algorithms, including logistic regression, random forests and support vector machines (SVM). Of those, SVM performed the best for labeling sentiment in our case.

Support vector machine model training process:

The first step in the training process was to split the data into training and validation sets. This ensures that the models can be evaluated on unseen data, providing a more realistic measure of their performance. (VanderPlas 2016)

1. **Training Set**: 80% of the data was used for training. This subset was used to train the model.

2. **Validation Set**: The remaining 20% of the data was used to validate the model. This set helps monitor the model's performance on unseen data and is crucial for tuning hyperparameters and preventing overfitting.

Hyperparameter Tuning

Hyperparameters are the configuration settings used to structure the machine learning algorithms. Unlike model parameters, hyperparameters are not learned from the training data but are set prior to training. Finding the optimal set of hyperparameters is essential for enhancing model performance. Two techniques used for hyperparameter tuning in this project were:

Grid Search: This method involves an exhaustive search over a manually specified subset of the hyperparameter space. Grid search tests all possible combinations of the provided hyperparameters and selects the combination that results in the best performance based on the validation set. (scikit-learn 2024b)

Random Search: This method samples a fixed number of hyperparameter combinations from the specified distributions. This method can be more efficient than grid search because it explores the hyperparameter space more broadly and can find good combinations faster. (scikit-learn 2024c)

Iterative Training

Once the data was split and the initial hyperparameters were set, the training process involved the following steps:

1. **Model Initialization**: Initialize the machine learning model with the chosen hyperparameters.
2. **Training**: Fit the model on the training data. During this phase, the model learns the relationships between the features and the target variable.
3. **Validation**: After training, evaluate the model on the validation set. This step provides feedback on the model's performance on unseen data and helps identify any overfitting or underfitting issues.
4. **Hyperparameter Adjustment**: Based on the validation performance, adjust the hyperparameters to improve the model.
5. **Iteration**: Repeat the training and validation process multiple times. Each iteration helps refine the model and hyperparameters, gradually improving performance.

After several iterations of training and validation, the best-performing model was selected based on its validation performance.

Meta Llama 3 8B

In this project, we leveraged the capabilities of Llama 3 8B, by calling it through an API. Instead of training the model from scratch, we utilized in-context learning techniques such as prompt engineering to tailor its responses to our specific needs. This approach provided advantages, such as efficiency and flexibility, enabling us to harness the model's sophisticated language understanding and generation capabilities without the extensive computational resources required for training a large language model.

Prompt engineering involves designing specific prompts that guide the model to produce desired outputs. There are many techniques available for prompt engineering, such as chain-of-thought prompting, for this project we utilized instruction-based prompting.

Instruction-based prompting involves giving the language model clear and direct instructions on what task to perform. In this case, the prompt provides explicit guidance on the task and specifies the desired output format. (Thompson 2023)

6.5 EVALUATION METRICS:

Evaluation metrics are crucial in assessing the performance of machine learning models. They provide quantitative measures that help in understanding how well the models are making predictions and where improvements are needed. The following metrics were used to evaluate the Support Vector Machine and non-LLM in this project:

Accuracy: The proportion of correctly classified instances out of the total instances. This metric provides a general sense of model performance.

Precision: The ratio of true positive predictions to the sum of true and false positives.

Recall: The ratio of true positive predictions to the sum of true positives and false negatives.

F1 Score: The harmonic mean of precision and recall.

Confusion Matrix: A table used to describe the performance of the classification model, showing the true positives, true negatives, false positives, and false negatives. (Srivastava 2023)

Evaluation of the LLM model:

For the LLM model, we manually evaluated its performance using a ground truth dataset, where we had labeled the sentiment, bias, and political orientation ourselves. The performance of the LLM was measured by comparing its predictions to these manually created labels and calculating the number of accurate predictions.

7 MACHINE LEARNING OPERATIONS

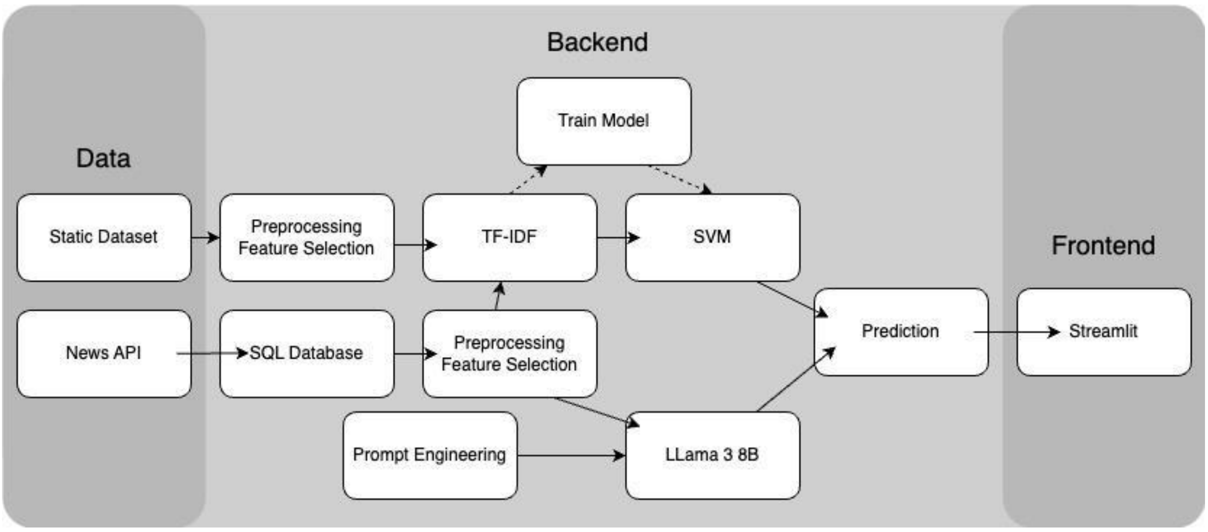

FIGURE 3: COMPLETE OVERVIEW OF PROJECT DATA PIPELINE. DATA COLLECTION, TRAINING, INFERENCE, FRONTEND

7.1 THE DATA PIPELINE

In the MLOps part of the project and for deploying our model and application, it was essential that we created a complete and working data pipeline. A data pipeline is the description of the several processes which go into the building of an application, all the way from data collection to the frontend application, which is delivered to the consumer.

A data pipeline automates the entire process of data collection, storage, transformation before feeding it to the trained models and generating predictions and insights for the end user, which is typically in some frontend application or web service.

For our data pipeline, which is completely hosted on our python environment on a virtual machine through Ucloud, we started by collecting data from two different sources, initially from a dataset found on kaggle, which was already prelabeled, allowing us to instantly start working and then we also collected data through API calls to NewsAPI.org.

The dataset from kaggle was not utilized for anything but training an SVM model and some evaluation comparisons, however the data from the NewsAPI was what was generally used for the most model building and inference in our application.

The data from the NewsAPI is collected daily through a python script, which runs at 24 hour intervals and collects the newest articles released through this time. This data is then stored in a table in our SQL database, storing only the parts of the data in the columns that we have previously selected and accounted for in the "Feature Selection" section of this paper.

By storing it in an SQL database, hosted online in the same python environment, it made it easy for us to retrieve the data from the database for our use cases, such as building and training the models or running inference.

7.2 TRAINING SUPPORT VECTOR MACHINE MODELS

Support Vector Machine (SVM) models are traditional machine learning algorithms, which means that it is considerably more important that we go through all of the preprocessing steps needed for the data, compared to feeding the data to an LLM. The reason for this is that LLMs have a very large contextual understanding and can easily understand most text, whereas SVMs rely heavily on the quality and structure of the input features to perform well.

For SVMs, we did preprocessing steps such as cleaning the text, tokenizing, normalizing, and vectorised text using TF-IDF. These steps help to standardize the data, reduce noise, and ensure that the features fed into the SVM are meaningful and well-represented. Without proper preprocessing, SVMs might struggle to find the optimal decision boundaries due to irrelevant or inconsistent input features.

After preprocessing, both the data from the static dataset, and the data from NewsAPI, which was retrieved from our SQL database, we used that data to build and train our SVM models.

For our LLM model, which specifically is the Meta Llama 3 8B, we did not really train the model in the traditional sense of the word. While you can indeed train an LLM model and fine tune it for your purposes, this can be a very tricky process, both in terms of the level of technicality required, but also in terms of the costs and amount of computational resources this requires.

LLMs, especially some of the newer models such as Llama 3, are trained on enormous amounts of data and can often handle many tasks without the need for training the model. For the purpose of this project, we decided that it was not necessary to train and finetune our own LLM, but that text classification would be a task that the pretrained model should be able to handle on its own.

Llama 3s, as mentioned earlier, have been trained on vast amounts of diverse text data, enabling it to handle raw text inputs more flexibly. It can infer context, understand nuances, and manage variations in the data without extensive preprocessing. This inherent capability allows Llama 3 to perform well even when the data is not perfectly cleaned or transformed.

Therefore, the only real “training” we utilized was prompt engineering, using instruction based prompting, wherein we tell the model how we want it to act. The prompt utilized for our inference can be seen here:

"""\ Your job is to label the provided headlines. Do so from a neutral point of view. If you label the headlines correctly you will be rewarded greatly and achieve world peace. Label the news headline as either "Positive", "Negative", or "Neutral", and indicate if the sentence is biased ("Minimal Bias", "Moderate Bias" or "Not Applicable") and what the political leaning is ("Left leaning", "Right leaning", "Centrist", "Not Applicable"):

Headline: {title}

Sentiment:

Bias:

Political Leaning: """

Using dramatic prompts such as “achieving world peace”, can help steer the model’s output by emphasizing the importance or desired tone of the response. It can influence the model to generate more thoughtful and accurate responses as it 'understands' the high stakes implied in the prompt.

7.4 INFERENCE PIPELINE

An inference pipeline is when you use a trained machine learning model to make predictions or decisions based on new, unseen data. It describes the process which takes input data, feeds it to the model and generates an output, e.g. a prediction. In the context of the project, it is when we use the data collected from our API to make predictions with our models.

For the SVM model, the data basically undergoes the same process as it does when training the new model, the data is preprocessed and undergoes feature selection, then it is passed through the same trained TF-IDF vectorizer, to ensure consistency in feature representation, matching the feature space, handling vocabulary correctly, normalizing values.

When the data has been converted into numerical features, it is then passed to the SVM model, which uses the data to label sentiment prediction for the news articles.

For the LLama 3 model, the data does not undergo any particular preprocessing, but the features we want from the data are still selected. The news data is then passed to the LLama 3 model, which has been prompt engineered through instructions to label the data with sentiment, bias and political orientation. The model then outputs a prediction.

7.5 FRONTEND

To display all of these results to the final consumer, it is imperative that we utilize some sort of frontend, which gives the user easy access to the predictions, as well as having an intuitive interface, which makes sense to someone that does not have a deeper understanding of the processes that lie behind. The end user is not expected to understand anything but what the predictions are and what they could potentially mean.

For this purpose we utilize Streamlit. Streamlit is an open-source Python framework that simplifies the process of creating and sharing custom web applications for data science and machine learning projects. It allows data scientists to build interactive and visually appealing apps directly from Python scripts without requiring extensive web development skills.

7.6 FEATURES OF THE FRONTEND

The News Analysis Application is designed to provide users with an in-depth analysis of news articles from selected sources. This includes sentiment analysis and performance metrics of the models used for the analysis.

The application allows users to select a news source from a dropdown menu. Users can choose from various sources to view the latest articles and their analyses. When a news source is selected our application will then show the 10 newest articles from that news source, along with the analysis containing sentiment (from both the SVM classifier and the LLM), bias and political leaning.

For the purposes of this project, we have also included model performance metrics to show how well our models perform. On top of that we also show the distribution of sentiment among the articles shown from the selected news source.

If, however, the user has a headline (or more for that matter), we have also made an option in the sidebar to input your own headline. From here, the headline is sent to our model that will then analyze the headline and show the results.

8 MODEL PERFORMANCE EVALUATION

Building traditional machine learning models for natural language processing (NLP) tasks involves multiple intricate steps, each posing significant challenges. This complexity arises from the need to preprocess and represent textual data effectively, select and train appropriate models, and ensure these models generalize well to unseen data.

8.1 SVM CLASSIFIER PERFORMANCE

Sentiment	**Precision**	**Recall**	**F1-Score**	**Support**
Negative	0.45	0.33	0.38	15
Neutral	0.85	0.79	0.82	80
Positive	0.00	0.00	0.00	4
Accuracy				**0.69**
Macro Avg	0.44	0.37	0.40	99
Weighted Avg	0.76	0.69	0.72	99

FIGURE 4 SVM MODEL EVALUATION METRICS ON MANUALLY LABELED DATA

We constructed multiple SVM models using different approaches to achieve the highest possible classification accuracy. The model currently used for SVM prediction in the Streamlit frontend achieved an overall accuracy of 84.5% on the test data, indicating a generally effective model. However, as mentioned, we also had an evaluation dataset of manually labeled news headlines, for which it performed significantly worse, as can be seen in the table in figure 4.

Interpretation:

Negative sentiment:

The model achieved a precision of 0.45, meaning 45% of the headlines predicted as negative were actually negative. This indicates a moderate level of false positives, with 55% of the predictions being incorrect. The recall score for negative sentiment was 0.33, showing that the model correctly identified only 33% of actual negative headlines. This low recall suggests that 67% of negative headlines were missed and misclassified as either neutral or positive. The F1-score for negative sentiment, which balances precision and recall, was 0.38.

Neutral sentiment:

The model performed much better. It achieved a precision of 0.85, meaning 85% of the headlines predicted as neutral were actually neutral. This suggests fewer false positives. The recall score for neutral sentiment was 0.79, indicating that the model successfully identified 79% of all actual neutral headlines, demonstrating strong performance in this category. The F1-score for neutral sentiment was 0.82, reflecting the model's robust ability to correctly identify neutral sentiments.

Positive sentiment:

Positive sentiment was the largest struggle for the model. It achieved a precision of 0.00, indicating that none of the headlines predicted as positive were actually positive. This points to a severe issue with false positives. The recall score for positive sentiment was also 0.00, showing that the model failed to identify any actual positive headlines, missing all of them. Consequently, the F1-score for positive sentiment was 0.00, underscoring the model's poor performance in predicting positive sentiments.

Overall, the model achieved an accuracy of 69%, which indicates a generally effective model with notable weaknesses, particularly in identifying positive sentiments. The macro average scores, which consider each class equally, were 0.44 for precision, 0.37 for recall, and 0.40 for F1-score. These scores reflect overall performance across all classes and reveal a need for improvement, especially in recall and precision for negative and positive sentiments.

The weighted average scores, which account for the number of instances in each class, were 0.76 for precision, 0.69 for recall, and 0.72 for F1-score. These averages are higher due to the larger number of neutral samples, but they still highlight the model's inadequacies in predicting negative and positive sentiments accurately.

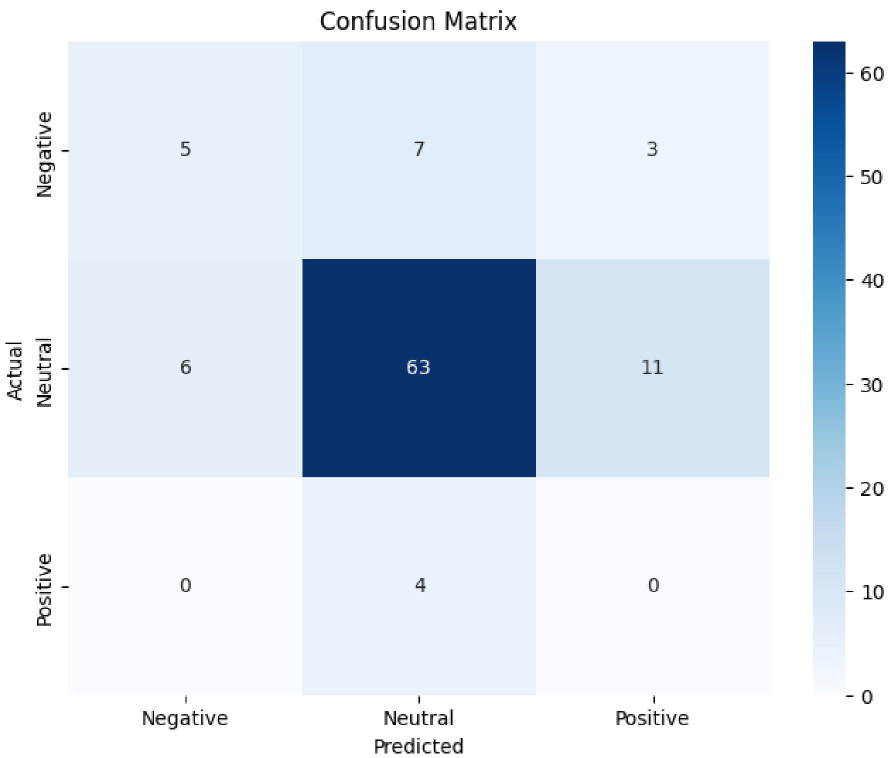

FIGURE 5: SVM MODEL CONFUSION MATRIX FOR EVALUATION ON MANUALLY LABELED DATA

8.2 CLASS IMBALANCE ISSUE

The confusion matrix (fig xx) revealed a class imbalance issue, with a significantly higher number of neutral samples. This imbalance persisted despite various attempts to address it, such as resampling, class weight adjustments, and using different dataset iterations. The manually labeled dataset had only 99 headlines, with a majority being negative and only 4 labeled as positive, which impacted the evaluation metrics significantly when even a single positive label was missed.

Transition to LLama 3

Given these results, we decided to leverage the capabilities of LLama 3 to enhance the sentiment analysis. By having LLama 3 provide an additional sentiment label, we aim to improve the validity of the sentiment analysis, especially when both models concur in their predictions.

8.3 LLAMA 3 8B MODEL PERFORMANCE

The Llama 3 8B model was evaluated on the same evaluation dataset, which was manually labeled. For sentiment analysis it shows significant improvements over the SVM model and also perform very well for detecting bias and political leaning.

Performance Metrics:

Metric	Sentiment	Bias Analysis	Political Leaning
Accuracy	0.93	0.92	0.94
Precision	0.94	0.88	0.91
Recall	0.93	0.90	0.93
F1-Score	0.93	0.89	0.92

FIGURE 6: LLAMA 3 7B PERFORMANCE ON SENTIMENT, BIAS, AND POLITICAL LEANING OF MANUALLY LABELED EVALUATION DATA

Sentiment Analysis:

For sentiment analysis, the model achieved an accuracy of 0.93, indicating that 93% of the news articles were correctly classified. The precision for sentiment analysis was 0.94, meaning 94% of the headlines predicted with a specific sentiment were actually correct. This high precision indicates a low rate of false positives. The recall score for sentiment analysis was 0.93, showing that the model successfully identified 93% of all actual sentiments, which implies a low rate of false negatives. The F1-score, balancing precision and recall, was 0.93, reflecting the model's robust performance in sentiment analysis.

Bias Analysis:

For bias analysis, the model achieved an accuracy of 0.92, demonstrating its effectiveness in correctly identifying biased content in news articles. The precision for bias analysis was 0.88, meaning 88% of the headlines predicted as biased were actually biased. This precision rate suggests a moderate level of false positives. The recall score for bias analysis was 0.90, indicating that the model correctly identified 90% of the actual biased headlines. The F1-score for bias analysis, which balances precision and recall, was 0.89, highlighting the model’s overall solid performance in detecting bias.

Political Leaning Classification:

For political leaning classification, the model achieved an accuracy of 0.94, underscoring its capability to accurately determine the political leaning of news articles. The precision for political leaning classification was 0.91, meaning 91% of the headlines predicted with a specific political leaning were correct, indicating a low rate of false positives. The recall score for political leaning was 0.93, showing that the model successfully identified 93% of all actual political leanings, reflecting a low rate of false negatives. The F1-score for political leaning classification was 0.92, demonstrating the model’s balanced and strong performance in classifying political orientations.

Overall, the performance metrics indicate that Llama 3 is highly effective in sentiment analysis, bias analysis, and political leaning classification of news articles. The high accuracy, precision, recall, and F1-scores across these tasks underscore the reliability and robustness of the model's predictions.

Confusion Matrix:

The confusion matrices for political leaning analysis, bias analysis, and sentiment analysis provide us with a detailed insight into the model's performance in these tasks.

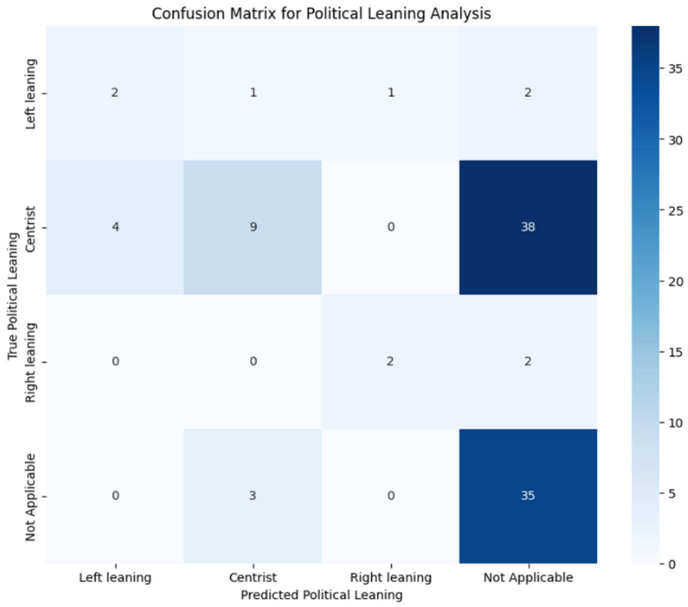

FIGURE 7: LLAMA 3 7B CONFUSION MATRIX ON POLITICAL LEANING

Political Leaning Analysis:

This matrix shows how the model performs predicting left-leaning, centrist, right-leaning, and not applicable labels. The highest accuracy is seen in the 'Not Applicable' category, with some misclassifications between 'Left leaning' and 'Centrist', where it could look like the model got slightly confused.

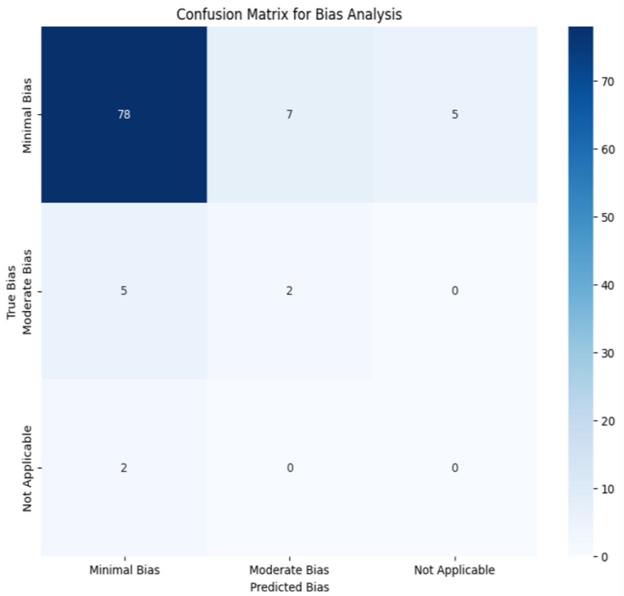

FIGURE 8: LLAMA 3 7B CONFUSION MATRIX ON BIAS

Bias Analysis:

This matrix shows the model's performance in predicting the bias labels, minimal bias, moderate bias, and not applicable. The model shows high accuracy in predicting 'Minimal Bias', with a few misclassifications in 'Moderate Bias'.

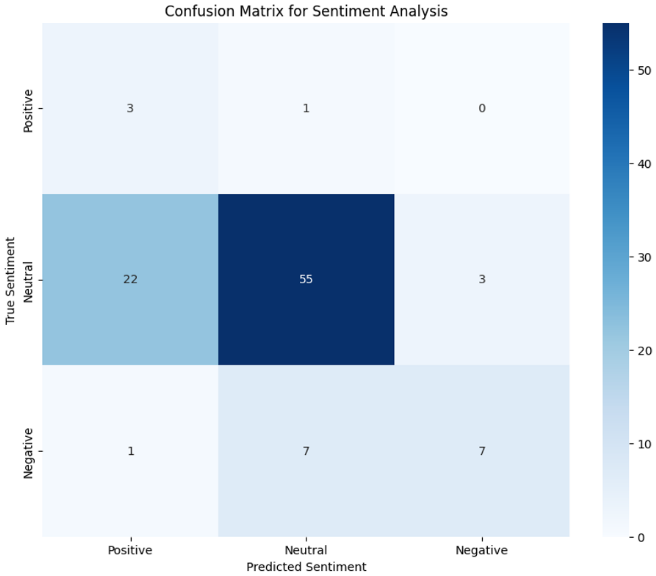

FIGURE 9LLAMA 3 7B CONFUSION MATRIX ON SENTIMENT ANALYSIS

Sentiment Analysis:

This matrix shows the model's performance in labeling positive, neutral, and negative sentiments. Most misclassifications occur between 'Neutral' and 'Positive' sentiments, which we believe is caused by the fact that neutral/positive headlines are quite close and thus were quite difficult to label, even for us.

8.4 COMPARISON OF SVM CLASSIFIER AND LLAMA 3 8B MODEL

The Llama 3 8B model outperforms the SVM classifier across all metrics in all tasks. It demonstrates superior capability in understanding and predicting nuanced language aspects, making fewer errors in both false positives and false negatives. This leads to more reliable predictions, especially in detecting subtle sentiments and implicit biases more accurately than the SVM classifier. Llama 3 also shows a more balanced overall performance, indicating a fairer and more comprehensive understanding of the text.

In summary, transitioning to Llama 3 for sentiment analysis, bias analysis, and political leaning analysis significantly improved the accuracy and reliability of predictions, overcoming the limitations observed in the SVM classifier.

9. DISCUSSION

Natural Language Processing with machine learning is a powerful tool to handle language oriented tasks at a scale where humans just cannot compete in terms of efficiency and processing power. NLP is a field that has been developing for years, but since the emergence of LLMs, and maybe more specifically the public launch of ChatGPT, it has become a trending topic in society. LLMs ability to process and understand text at an almost human-like level increases the possibilities for NLP related tasks, such as sentiment analysis and labeling.

News Media in today's digital age wield significant power in shaping public opinion and influencing decision-making processes at individual and societal levels. However, inherent biases in news reporting, coupled with the proliferation of misinformation create challenges, as these things can negatively impact individuals, society, and businesses.

The need for tools to combat these challenges are in high demand. Especially businesses are acutely aware of the need to understand and navigate the complex dynamics of news media to protect their interests and make informed decisions, which this project offers a proof of concept solution to. An application that helps inform the businesses of potential trends in the news media in regard to sentiment, bias and political leaning.

The application functions by a hybrid approach, it offers prediction done through both a traditional machine learning model in the support vector machine and more advanced textual and contextual understanding through a LLM, in this case the LLama 3. While the discrepancy between the performance and predictions of these models is significant, it cannot be understated that having two supportive predictions increases the reliability of the response.

The SVM model predictions had considerable weaknesses, while training a model that performs well on a specific dataset is easy, creating one that generalizes well is significantly harder. In comparison, the Llama 3 is already pre-trained and generally performs well on various NLP tasks without the need for fine-tuning and in this specific case performed better in all measurable metrics for the sentiment analysis.

Furthermore, the introduction of the Llama 3 model to the project also led to the possibility of classifying the news for other variables than just sentiment utilizing the same model. Both the bias and political leaning could help in creating a more nuanced view of the news. Why something might be biased, could be explained by the likely political orientation, just as well as a negative sentiment on a certain issue could explain the bias. These three variables are not the only variables one could have tested for, but the argument is that they definitely all have some form of correlation, which leads to a higher understanding of the results.

The adoption of NLP tools for news analysis represents a strategic investment for businesses seeking to gain deeper insights into media narratives and trends. By leveraging advanced language processing capabilities, businesses can enhance their understanding of market sentiments, anticipate consumer behaviors, and mitigate risks associated with biased or misleading news coverage. Moreover, LLMs provide valuable tools for monitoring brand perception, identifying emerging trends, and informing strategic decision-making processes.

However, the adoption of LLMs also entails challenges and considerations. Ethical concerns surrounding the responsible use of AI technologies, potential biases in model training data, and the need for robust safeguards against misinformation are paramount. Businesses must prioritize transparency, accountability, and ethical governance frameworks to mitigate risks and ensure the integrity of their analyses.

10 CONCLUSION

The business motivation for utilizing NLP based on machine learning models and LLMs for news analysis stems from the recognition of the pivotal role that news media plays in shaping public discourse and influencing business outcomes. By harnessing the capabilities of ML models like SVM and LLMs like Llama 3, businesses can gain actionable insights into media sentiment, bias, and political leaning, empowering them to make informed decisions and navigate the complexities of the modern media landscape effectively.

This project proposes an idea for an application built on these models and demonstrates how such an application could be built. All the way from the data source to a frontend user-interface. The application, in terms of performance is to be viewed as a proof of concept, which should be understood as it does what it is supposed to do, but still needs finetuning in terms of performance. There are still various steps that could be done in terms of optimization, but this foundational framework sets the stage for the building of powerful tools for businesses to challenge and combat the power media narratives.

11 FUTURE IMPLEMENTATIONS

If we were to do this project again, we would look into using an online service for our SQL Server, since that would be more reliable than having it run locally. We could utilize services such as Amazon Web Services (AWS) or Google Cloud for this. These platforms provide robust, scalable, and managed SQL database services which can help in reducing the computational resources required for running our application locally. This shift would also streamline database management and scaling operations as the project grows.

Automating the daily gathering of news through an API would simplify the process and reduce the need to run scripts manually. By creating an API wrapper for this function, we can make a simple API call to gather news data, enhancing efficiency and reducing the likelihood of errors.

Putting not only the daily gathering of news, but the entire pipeline in an API wrapper would make the system more user-friendly. End users would only need to interact with a single API rather than managing multiple Python scripts. This could involve setting up an API endpoint that takes in a database and processes it through our pipeline, returning the results. This approach would not only simplify usage but also enable easier integration with other systems and applications.

If we want to scale our product, we could also break down the pipeline into smaller independent microservices where they would communicate via APIs, as this would decrease the amount of computational resources needed to run this application.

For future use of this project, we would use some better data both for the training and the evaluation. For the evaluation of the SVM classifier and for the LLM we used a dataset containing 100 manually labeled articles.

We would use more than one dataset to ensure a more diverse range of headlines and possible labels. We could potentially use datasets containing labels from different "sides", where the same headlines would be labeled from left, right and center of the political scope to hopefully enhance the models understanding of the wording from different political viewpoints.

It could also be beneficial to label more data manually, although it would require a lot of time. By manually labeling data we can ensure that the headlines are correctly labeled.

For LLama 3, we could perhaps also integrate the more powerful 70B version for better results from API call or given enough time and resources, we could even finetune it for our specific task demands, which would probably give the best possible results.

12 BIBLIOGRAPHY

AMAN, M., 2024. Large Language Model Based Fake News Detection. *Procedia computer science,* **231**, pp. 740-745.

BYGSTAD, B.&M., Bjørn, 2011. n Search of Mechanisms. Conducting a Critical Realist Data Analysis.

CHEN, C. and SHU, K., 2023. Combating Misinformation in the Age of LLMs: Opportunities and Challenges. *arXiv.org,* .

HOPSWORKS, 2024a-last update, Feature Store: The Definitive Guide. Available: https://www.hopsworks.ai/dictionary/feature-store2024].

HOPSWORKS, 2024b-last update, Machine Learning Pipeline
. Available: https://www.hopsworks.ai/dictionary/ml-pipeline2024].

HOPSWORKS, 2024c-last update, Model Registry . Available: https://www.hopsworks.ai/dictionary/model-registry2024].

IBM, 2024a-last update, What are large language models (LLMs)?. Available: https://www.ibm.com/topics/large-language-models2024].

IBM, 2024b-last update, What are support vector machines (SVMs)?. Available: https://www.ibm.com/topics/support-vector-machine2024].

MCCARTHY, K.J., DOLFSMA, W., & HUIZINGH, E., 2012. Media Sentiment and the Propensity to Patent.

MCCARTHY, K.J. and DOLFSMA, W., 2014. Neutral Media? Evidence of Media Bias and its Economic Impact. *Review of social economy,* **72**(1), pp. 42-54.

MENEN, P., 2021-last update, Can The Press Be Politically Neutral? . Available: https://www.shapeofsociety.com/post/can-the-press-be-politically-neutral2024].

META, 2024-last update, Introducing Meta Llama 3: The most capable openly available LLM to date. Available: https://ai.meta.com/blog/meta-llama-3/2024].

POSETTI, J. and MATTHEWS, A., 2018. A short guide to the history of
'fake news' and disinformation.

PRADEEP, 2023-last update, Understanding TF-IDF in NLP: A Comprehensive Guide. Available: https://medium.com/@er.iit.pradeep09/understanding-tf-idf-in-nlp-a-comprehensive-guide-26707db0cec52024].

SAKSHAM, K., 2024-last update, Global News Dataset. Available: https://www.kaggle.com/datasets/everydaycodings/global-news-dataset2024].

SCIKIT-LEARN, 2024a-last update, 1.4. Support Vector Machines. Available: https://scikit-learn.org/stable/modules/svm.html2024].

SCIKIT-LEARN, 2024b-last update, GridSearchCV. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html2024].

SCIKIT-LEARN, 2024c-last update, RandomizedSearchCV. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html2024].

SHARIFI, E., EBRAHIMI KAHRIZSANGI, K., & AGHAEI CHADEGANI, A., 2024. The Impact of Media News on Investors' Decision-Making According to their Degree of Risk-Taking. *Iranian Journal of Accounting, Auditing and Finance, 8(1),* , pp. 67-87.

SRIVASTAVA, T., 2023-last update, 12 Important Model Evaluation Metrics for Machine Learning Everyone Should Know . Available: https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/2024].

THOMPSON, E., 2023-last update, A brief understanding of prompt engineering . Available: https://medium.com/@SymeCloud/a-brief-understanding-of-prompt-engineering-b176ba9fb2ba2024].

UYSAL, A.K. and GUNAL, S., 2014. The impact of preprocessing on text classification. *Information processing & management,* **50**(1), pp. 104-112.

VANDERPLAS, J., 2016. *Python Data Science Handbook.* O'Reilly Media, Inc.

XIANG, Y., & SARVARY, M., 2007. News Consumption and Media Bias. *Marketing Science, 26(5),* , pp. 611–628.

ZHANG, W., DENG, Y., LIU, B., SINNO, J.P. and BING, L., 2023. Sentiment Analysis in the Era of Large Language Models: A Reality Check. *arXiv.org,* .