MULTIMODAL SENTIMENT ANALYSIS

Kabir Dhillon
Student# 1008229070
kabirs.dhillon@mail.utoronto.ca

Yousef Al Rawwash
Student# 1007684873
yousef.alrawwash@mail.utoronto.ca

Mariia Ostrenko
Student# 1008179436
mariiaostrenko@mail.utoronto.ca

Sparsh Kochhar
Student# 1009020220
sparsh.kochhar@mail.utoronto.ca

ABSTRACT

Abstract: In an era dominated by unprecedented growth in digital content and news media, this project seeks to develop a comprehensive sentiment detection system to aid in public image perception understanding. This system, built upon the power of advanced deep learning methodologies, targets multiple data types, not limited to news headlines, to discern sentiment attached to specific entities such as companies, individuals, and more. This innovative approach could have transformative implications across various sectors like finance, public relations, and policy-making, providing a nuanced understanding of public sentiment to inform critical decisions. With deep learning models' remarkable ability to process diverse and unstructured data types and detect intricate patterns, our system promises to outperform traditional methods of sentiment analysis. Equipped with the flexibility to adapt to evolving linguistic trends and data patterns, our project addresses the urgent need for automated sentiment analysis in our data-rich world. It aims to provide a robust, sophisticated tool that offers deeper insights into public sentiment, aiding in more informed and insightful decision-making processes. —-Total Pages: 9

1 APS360 PROJECT FINAL REPORT

1.1 BRIEF PROJECT DESCRIPTION

In a world where digital content and news media are growing exponentially, the sheer volume of data necessitates the creation of automated tools to navigate, interpret, and draw insightful conclusions from this vast information pool.

Driven by this pressing need, our focus lies in developing a sentiment detection system that leverages advanced deep learning techniques to get the analysis subject's public image. We're setting our sights on news headlines, audio recordings, tweets, etc., with the aim of creating a system that can determine the sentiment associated with a specific entity. The goal is to quantify sentiment in a manner that can enhance decision-making in various areas, including but not limited to finance, public relations, and policy-making.

The potential impact of this project is both multifaceted and fascinating. Consider the realm of finance, where a clear understanding of the sentiment surrounding a particular company could dramatically influence investment strategies with regard to its stock. This sentiment analysis provides insights into the public perception that might elude traditional financial metrics. Similarly, in public relations, sentiment analysis offers a powerful tool to evaluate public opinion and inform strategies.

Why deep learning, you may ask? The answer lies in its proven capacity to handle unstructured data like text and its ability to unravel complex patterns within this data. Traditional text analysis methods often struggle when tasked with capturing the subtleties and nuances of human language. But deep

learning models, such as RNN, have shown formidable ability in decoding context, sarcasm, and other linguistic intricacies.

The deep learning approach truly comes into its own when faced with a considerable volume of news data. As these models absorb more information and fine-tune their predictions, they grow progressively more skilled at identifying and interpreting sentiment in the text. They also display remarkable adaptability to shifting linguistic trends, ensuring that our sentiment detection system remains relevant and effective as language evolves with time.

Our project, at its core, responds to a modern need for automated sentiment analysis in a world that's brimming over with data. The significance of our work doesn't solely reside in its vast potential applications across different sectors, but it also lies in the adoption of deep learning technology. This methodology provides a strong and adaptable answer to the intricate challenge of text-based sentiment analysis. We're offering a smart tool to help navigate the intricate landscape of news media and make enlightened decisions rooted in public sentiment. Our mission isn't confined to detecting sentiment - we're about delving deeper to understand our world in a fresh, exhilarating way.

1.2 ILLUSTRATION

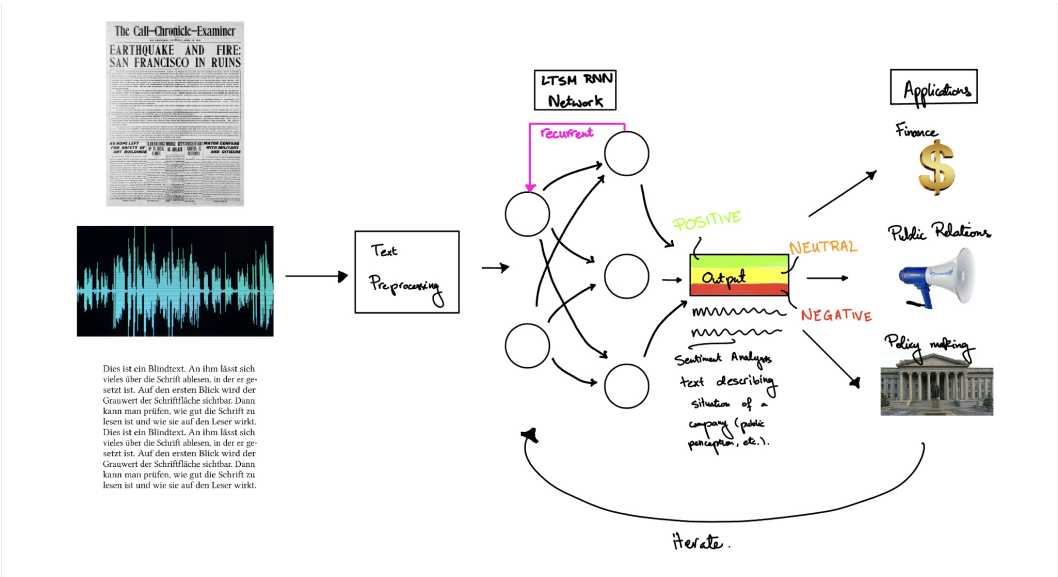

Figure 1: Illustration of the overall pipeline consisting of inputs, preprocessing, 3 layer RNN, output and applications

1.3 BACKGROUND AND RELATED WORK

Detect and measure human emotions: Affectiva employs computer vision and machine learning to derive emotion metrics from facial expressions and voice tones Affectiva (2021). Its vast dataset ensures robust detection across diverse facial features.

Sentiment analysis of movie reviews: Using an RNN, support vector machines yielded an accuracy of 82.9% Kaggle (2023). Combining unigrams and bigrams enhanced the precision. Another analysis found logistic regression to be the most effective with 88.89% accuracy.

Sentiment Analysis of COVID-19 Tweets: The project captured pandemic sentiment through tweet analysis Frontiers (2021). Leveraging natural language processing and the Conv1D-LSTM model, it offered key insights into public perceptions.

Airlines Tweets Sentiment Analysis: This study analyzed airline-specific tweets with LTSM, AdaBoost, and sklearn neural networks, categorizing sentiments Crowdflower (2023). The insights inform airlines on customer satisfaction and areas of improvement.

Sentiment Analysis using Amazon Product Reviews Dataset: The sentiment model trained on the Amazon dataset achieved a 93.5% accuracy using logistic regression, outpacing a naive Bayes approach Bittlingmayer (2023).

Financial news article sentiment: This analysis employed BERT to classify sentiments in financial articles Sbhatti (2023), serving as a valuable tool for industry stakeholders.

Drawing from these diverse applications in sentiment analysis, our project aims to synthesize the best of these techniques, addressing the evolving needs of today's digital landscape and pioneering a fresh approach to understanding public sentiment.

1.4 Data Processing

Data Collection

The data utilized in this project was collected from numerous sources, each providing distinct, valuable insights into our research focus. Textual datasets included a News Category Dataset Rmisra (2023) comprising headlines, authors, links, short descriptions, and categories, a Financial News Dataset Miguelaenlle (2023) comprising headlines, publishers, dates, stocks, and image URLs, a dataset with podcast transcriptions Patrickhallila1994 (2023) detailing episode names, transcription text, episode numbers, and speaker names, and a Job Interview Experiences dataset Rahulrrrd (2023) containing company names, experiences, upvotes, and URLs. These datasets in total had more than 2.1 million values. As our project features multimodal analysis, we also had three audio datasets of podcasts that we processed and utilized in our project. These datasets featured more than 50 hours of recorded audio Washingtongold (2023) along with 400+ audio fragments Sabahesaraki (2023) ranging from 2 minutes to 20 minutes in length.

Data Processing

The data processing began with converting the raw data into structured dataframes to facilitate the data cleaning and analysis process. From there, each dataset was examined for missing values to remove in order to prevent biases and errors in our analysis. Additionally, we identified and removed any duplicate values as these values can artificially inflate the importance of specific data points and lead to misleading results.

For textual data, we implemented several pre-processing steps. First, we removed special characters and punctuation as they don't carry significant meaning and could introduce noise, so this reduces computational complexity. Then, all words were converted to lowercase to establish consistency and ensure uniformity between capitalized and uncapitalized versions of the same word. Then, we removed stop words (e.g. "and", "the") as they are unnecessary for sentiment analysis and this helps decrease the computational cost for the model and improve its efficiency. Finally, to ensure we were only using relevant data for our project, we curated columns of interest from each dataset to create sub-datasets. These were then combined into a unified, comprehensive dataset, tailored for our model's training. This way, we got rid of the unnecessary publication dates, website links, author names, and other side information that was initially provided with the text. In the actual training process for the model, we utilized about 200,000 samples from this dataset due to the computational constraints of the machines we were running the model on.

For auditory data, we implemented similar pre-processing steps. We first transcribed the audio into textual form as this would allow us to apply text analysis techniques and allow our model to be trained on it. The audio was then segmented by silent intervals as auditory data often contains silent intervals that can artificially inflate the data's length and introduce noise into our analysis. This led to 162 text units derived from 600+ millisecond pauses. However, due to diverse speaking styles in the datasets, these units could represent roughly 130 to 200 actual sentences. From there, we lowercased all words to ensure uniformity and maintain data integrity. Then, we removed stop words to further reduce noise and computational redundancy. Finally, we consolidated the transcription into a single dataset with all text and audio in one column for streamlined training in the model. The first audio dataset with a 27-minute long podcast recording was fully transcribed and used to train

the model. The other two datasets only had 90 out of 400+ fragments uploaded to the drive, due to storage issues. We decided to only use half of that data directly for the model, as the training process was taking too long and was making it challenging to train the network while adjusting hyperparameters.

Data Statistics and Examples

As seen below in Table 1, each textual dataset we utilized was streamlined into two columns: one which captured the primary context (e.g. a "headline" or "title") and the other detailed supplementary information (e.g. "description" or "experience"). The auditory datasets had only one column with all transcribed text. The number of rows in each dataset signifies the distinct entries available for sentiment analysis. Table 2 demonstrates specific samples of cleaned and processed values from the datasets, picked out at random.

Dataset	Number of Rows	Number of Columns
Subdf1: News Categories	209,512	2
Subdf2: Financial News	1,845,559	2
Subdf3: Podcast Transcriptions	36,392	2
Subdf4: Job Experiences	7,768	2
Audio_df	162	1

Table 1: Statistics regarding individual sub-datasets used.

Textual Datasets Samples	Auditory Datasets Samples
Headline: 4 million americans roll sleeves omicrontargeted covid boosters. **Description**: health experts said early predict whether demand would match 171 million doses new boosters us ordered fall.	"Cnn opinion story by a former federal election commission general counsel larry noble he wrote a piece titled quote soliciting dirt on your opponents from a foreign government is a crime all that should have charged trump campaign officials with it"
Headline: american airlines flyer charged banned life punching flight attendant video. **Description**: subdued passengers crew fled back aircraft confrontation according us attorneys office los angeles.	"House speaker nancy pelosi did not call for impeachment proceedings today as she left a closed-door meeting with house democrats but she had a firm message and we believe that the president of the united states is engaged in a cover-up hours later in the rose garden president trump reacted."

Table 2: Samples of cleaned and processed values from textual and auditory datasets.

1.5 Architecture

1.5.1 Model Choice

Long Short-Term Memory (LSTM)

For our sentiment analysis project, we adopted the LSTM variant of the Recurrent Neural Network (RNN) architecture. The decision was based on the sequential nature of textual data, where context and word arrangements are essential for determining sentiment.

1.5.2 LSTM Rationale

LSTMs, compared to traditional RNNs, are better suited for retaining long-term dependencies in textual data. They employ a gating mechanism consisting of input, output, and forget gates, which helps in maintaining larger memory and handling long-range relationships in text. This makes them apt for sentiment analysis tasks.

1.5.3 Model Architecture

Layer 1 (Embedding Layer): Transforms words into vector representations that capture semantic nuances.

Layer 2 & 3 (LSTM Layers): Processes the word embeddings sequentially, understanding temporal dependencies.

Final Layer (Dense Layer): Predicts sentiment based on the information processed by the preceding layers.

In addition to that, we used padding and masking to ensure that the input of variable length fits into one computational graph and aids in efficient batch processing. Through numerous experiments, we have settled on having 30 hidden units, three epochs, a batch size of 512, and a learning rate of 0.01.

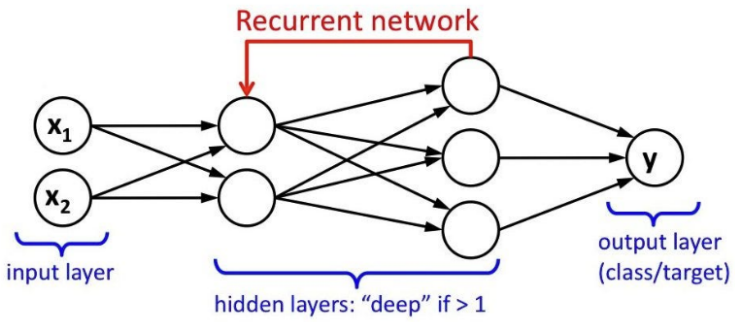

Figure 2: 3 layer recurrent neural network

1.6 Baseline Model

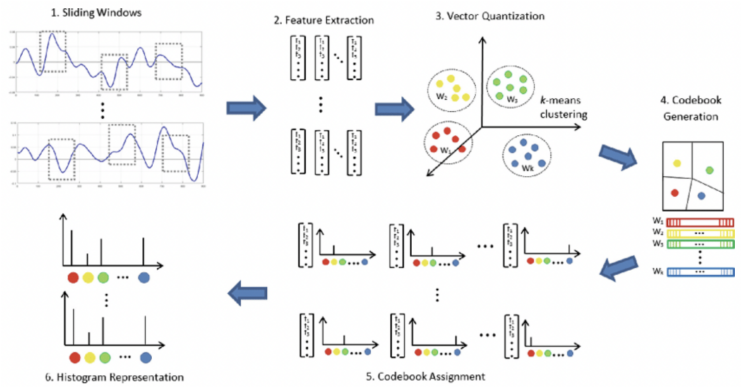

Figure 3: Bag-of-Words(BoW) model

In this sentiment analysis project, a Bag-of-Words (BoW) model along with a Logistic Regression classifier is used as a baseline model. This is a common and simple place to start for Natural Language Processing (NLP) projects.

Preprocessing: The raw news headlines are preprocessed in the same manner as in our main model by deleting superfluous characters and words, tokenizing, and then vectorizing. The BoW model does not, however, maintain the order of the words in contrast to deep learning models.

Feature Extraction: In the Bag-of-Words paradigm, which ignores syntax and even word order but maintains multiplicity, a text (such as a headline or an article) is represented as a bag (multiset) of its words.

	about	bird	heard	is	the	word	you
About the bird, the bird, bird bird bird	1	5	0	0	2	0	0
You heard about the bird	1	1	1	0	1	0	1
The bird is the word	0	1	0	1	2	1	0

Figure 4: Feature Extraction

Model Training (Logistic Regression): The output from the BoW feature extraction is then used to train a Logistic Regression model. This model is chosen for its simplicity and efficiency in handling high-dimensional data, and it works well for binary classification problems such as negative vs. positive sentiment.

Prediction and Evaluation: The trained model will be used to predict sentiment on the test dataset. The performance will be evaluated using standard metrics such as accuracy, precision, recall, and F1 score.

1.7 Quantitative Results

During the development phase of our machine learning model, we evaluated a variety of configurations differentiated by hyperparameters, such as epochs, batch size, learning rate, and hidden size. These hyperparameters are extremely important in deciding the model's performance.

We looked at the following configurations:

num_epochs=3, batch_size=512, learning_rate=0.01, hidden_size = 30

num_epochs=3, batch_size=512, learning_rate=0.1, hidden_size = 30

num_epochs=3, batch_size=512, learning_rate=0.1, hidden_size = 60

Following careful consideration, the configuration **(num_epochs=3, batch_size=512, learning_rate=0.01, hidden_size=30)** was chosen. It achieved the best possible balance of learning speed, model complexity, and overfitting risk, resulting in great accuracy on both the validation and testing datasets, as well as a noteworthy generalization to new data.

1.7.1 Training Results

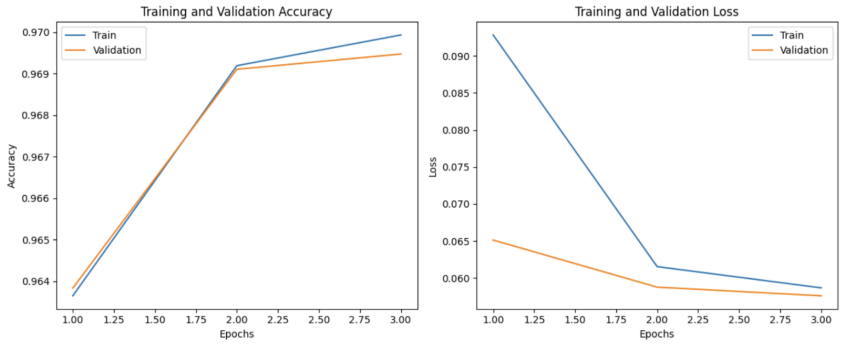

Figure 5: Training and Validation accuracy and loss plotted over epochs

Our model had a training accuracy of 0.9699, validation accuracy of 0.9694, and testing accuracy of 0.9689. These findings reveal that our model **properly predicted text sentiment in almost 97% of the occurrences** across each dataset. Such consistency in performance across the training, validation, and testing phases demonstrates the model's high generalization capabilities, which is critical for minimizing overfitting.

1.7.2 Confusion Matrix Insights

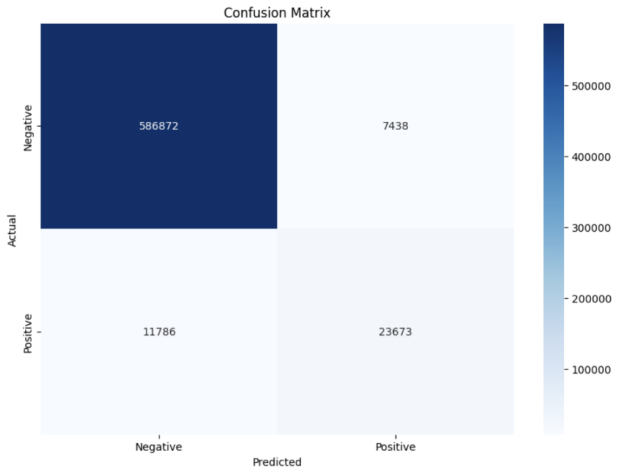

Figure 6: Confusion matrix to aid in visualizing the contribution of the classification model

Based on the confusion matrix, our model performed well in predicting negative attitudes but struggled with positive ones. This was demonstrated by a higher number of false negatives (Type II mistakes) than true positives. Concurrently, the model's low false positives (Type I errors) demonstrate its accuracy in recognizing negative thoughts. This observation corresponds to plots that show a negative skew between actual and projected attitudes.

1.8 Qualitative Results

Beyond the quantitative, the real-world applicability of our model is showcased through its adeptness in varied contexts, such as financial reports, podcast transcripts, and job interviews. Through the showcase of these results, the aim has been maintaining transparency while boasting the model's comprehensive performance through each phase of the machine learning workflow. This approach aligns with our objective of maintaining transparency and clarity.

1.8.1 Sample Outputs for Contextual Understanding

Textual Analysis: In the phrase "The local hospital gets generous donations, nurses are happy," our algorithm correctly interpreted a positive mood. This example demonstrates the model's capacity to grasp complex optimism.

Analysis of Audio Transcription: An audio clip transcribed as "Rising food prices make it difficult to buy basic products" was correctly classified as negative. This example emphasizes the model's consistency in interpreting sentiments from many sources.

The outputs chosen are intended to highlight the model's versatility in a variety of scenarios. Furthermore, the confusion matrix analysis and the negative skew in predictions reveal areas where the model might be improved, particularly in differentiating positive attitudes.

1.9 Evaluate Model on New Data

To ensure our sentiment analysis model's capacity to generalize beyond the training data, we took essential measures. Our dataset was deliberately partitioned into distinct training and testing sets, which would ensure that during the training phase, the model is kept unaware of a segment of unseen data, approximately 20% of the total data utilized for the model, which would prepare the model for an accurate test of its abilities. When the model was subjected to the testing set, it faced entirely new examples that it had never encountered during its training, thus providing insights into the model's real-world performance.

To gauge the model's performance on this unseen data, we employed a set of specific metrics. These included overall accuracy, which gave an overview of accurate predictions, as well as precision and recall, which went further to show how effectively it could classify sentiments. Overall, these indicators provided a detailed picture of how well-prepared our model is to deal with real-world situations.

However, to take it one step further and push the boundaries of our model's evaluations, our team also decided to actively look for recent news articles and create our own collection of samples that the model had never seen before. These were headlines we encountered in the media or ones we conceptualized to emulate the current news environment, thus ensuring that our model was put to the test in scenarios as close to genuine application as possible. To do this, we used a function that predicted sentiment using the previously trained Logistic Regression model. In the function, we transformed the input text to a vector using the vectorizer that was previously fit on our training data and then passed the text into the predict function of the trained Logistic Regression model to predict the sentiment. As seen in Table 3 below, our model effectively demonstrated commendable performance, even when tested against recent, real-world samples. Overall, we achieved a model consistency of 96.9% across the training, validation, and testing phases.

Positive Samples	Negative Samples
• "Promising new drug that could treat serious illnesses" ◦ **Prediction: Negative** • "Finally a perfect weather forecast for this weekend" ◦ **Prediction: Positive** • "The local hospital gets generous donations, nurses happy" ◦ **Prediction: Positive** • "Student robotics team wins the competition, the whole school is celebrating" ◦ **Prediction: Positive**	• "Serious floods destroy several cities, people are losing their homes" ◦ **Prediction: Negative** • "Toxic air quality becomes worse, more illnesses reported" ◦ **Prediction: Negative** • "Rising food prices make it difficult to buy basic products" ◦ **Prediction: Negative** • "20 people killed in the earthquake, 340 reported missing" ◦ **Prediction: Negative**

Table 3: Results of testing on unseen news headline samples.

1.10 Discussion

At its root, sentiment analysis is a complex issue. Decoding human language, which is inherently complex and full of nuance, idioms, slang, and cultural references, is the fundamental challenge. Several factors amplify the difficulty of this problem. For example, the issue of ambiguity as the meaning of many words can change based on the context or the introduction of sarcasm and irony in speech which can be challenging to detect in natural language processing.

Additionally, our project utilized diverse data sourced from multiple datasets like financial news, podcast transcripts, and job interviews, with the breadth and depth of the language structures involved in each becoming more complex. Even just transcribing spoken language for the audio datasets presents a challenge due to varying speech rates, accents, etc., that might not be present in written text. Each source, whether it be textual or auditory, can have its own distinct tone, style, and lexicon, which demands a robust model to generalize well across these datasets.

1.10.1 Results

Upon evaluating our results, the model exhibits noteworthy performance with consistent accuracy scores of 96.9% across training, validation, and testing datasets. Such consistency, especially given these difficulties and text-based machine learning, signifies a well-optimized model free from overfitting, enabling it to predict accurately on unseen data and manage the nuanced complexities of human language in real-world application.

Yet, examining the confusion matrix reveals its relative difficulty in identifying positive sentiments, evidenced by a higher rate of false negatives. This trend is intriguing and might be tied to our model's configuration, designed to reduce overfitting. The conservative configuration could have made the

model hesitant in classifying positive sentiments, which suggests that textual data's inherent nuance requires a careful balance between caution and assertiveness.

The model's ability to correctly identify sentiments across various contexts, from financial documents to podcast transcripts, without domain-specific tuning, stands out. This suggests that our dataset was diverse, encompassing a wide range of topics and sentiments.

1.10.2 Learnings and Takeaways

Balanced Configuration: A careful hyperparameter balance is critical. Our chosen configuration may have inadvertently restricted the model's capacity to confidently classify positive sentiments.

Dataset Diversity: The model's versatility underscores the importance of a comprehensive training dataset.

Beyond Accuracy: Relying solely on accuracy can overshadow specific prediction challenges, as revealed by our confusion matrix.

Iterative Refinement: The model's ability to identify positive sentiments can be enhanced in subsequent iterations.

In conclusion, our model exhibits adaptability and robustness. While its strengths are evident, recognizing and addressing its limitations is important. This endeavour has made clear the importance of a nunanced approach to machine learning that emphasizes comprehensive review, iterative refinement, and continous learning.

1.11 Ethical Considerations

Using sentiment analysis in domains like finance, public relations, and policy-making is a potent tool. Its power, however, necessitates ethical scrutiny.

Privacy: While data is sourced from public platforms, users may not have explicitly consented for their information to be repurposed. This treads a fine line between public and private data, raising privacy issues.

Representation and Bias: Online contributors don't represent the entire population. Without representation from all sections of society, the analysis might be skewed. Our training data, if primarily sourced from certain regions or periods, might lack global or historical perspectives.

Model Limitations: Despite the sophistication of neural networks, they aren't perfect. Challenges persist in accurately detecting nuances like sarcasm or irony. Misinterpretations can introduce inaccuracies.

Impact: The potential for model outputs to unduly influence public sentiment or decisions is real. Unrecognized biases in our model or data could inadvertently reinforce societal prejudices.

Transparency and Accountability: The clarity behind our model's decisions is crucial for ensuring responsible use, especially in high-stake areas like finance.

In advancing sentiment analysis, it's imperative to ensure that our techniques are not just technologically sound but also ethically responsible.

1.12 Colab Link

https://colab.research.google.com/drive/115evp9MLoX8AdPJ1JQW8lc5e5ihkj8Hi?usp=sharing

REFERENCES

Affectiva. Emotion ai 101: All about emotion detection and affectiva's emotion metrics, 2021. URL https://blog.affectiva.com/emotion-ai-101-all-about-emotion-detection-and-affectivas-emotion-metrics.

Bittlingmayer. Amazon reviews, 2023. URL https://www.kaggle.com/datasets/bittlingmayer/amazonreviews.

Crowdflower. Twitter airline sentiment, 2023. URL https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment.

Frontiers. Title. *Frontiers in Public Health*, 2021. doi: 10.3389/fpubh.2021.812735. URL https://www.frontiersin.org/articles/10.3389/fpubh.2021.812735/full.

Kaggle. Sentiment analysis on movie reviews, 2023. URL https://www.kaggle.com/competitions/sentiment-analysis-on-movie-reviews/discussion.

Miguelaenlle. Massive stock news analysis db for nlp backtests, 2023. URL https://www.kaggle.com/datasets/miguelaenlle/massive-stock-news-analysis-db-for-nlpbacktests.

Patrickhallila1994. Vox today explained audio, 2023. URL https://www.kaggle.com/datasets/patrickhallila1994/vox-today-explained-audio.

Rahulrrrd. Interviews dataset, 2023. URL https://www.kaggle.com/datasets/rahulrrrd/interviews.

Rmisra. News category dataset, 2023. URL https://www.kaggle.com/datasets/rmisra/news-category-dataset.

Sabahesaraki. Voxceleb-1 dataset, 2023. URL https://www.kaggle.com/datasets/sabahesaraki/voxceleb-1-dataset.

Sbhatti. Financial sentiment analysis, 2023. URL https://www.kaggle.com/datasets/sbhatti/financial-sentiment-analysis/code?datasetId=1918992.

Washingtongold. Voxconverse dataset, 2023. URL https://www.kaggle.com/datasets/washingtongold/voxconverse-dataset.