Comparative Analysis of User-Generated Video and Text Reviews of Healthcare Products Using Multimodal NLP

MSc Research Project
MSc Data Analytics

Ankita Garud
Student ID: 24102563

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	**Ankita Dattatray Garud**
Student ID:	**24102563**
Programme:	**MSc Data Analytics** **Year: 2025-26**
Module:	**Research Practicum**
Supervisor:	**Jorge Basilio**
Submission Due Date:	**28th Jan 2026**
Project Title:	**Comparative Analysis of User-Generated Video and Text Reviews of Healthcare Products Using Multimodal NLP**
Word Count:	**7838** **Page Count: 23**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Ankita Garud

Date: 28th Jan 2026

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Comparative Analysis of User-Generated Video and Text Reviews of Healthcare Products Using Multimodal NLP

Ankita Garud
24102563

Abstract

This study compares sentiment patterns in text-based and video-based reviews of healthcare products using a multimodal NLP approach. The texts of Amazon reviews were obtained in the form of an open Kaggle dataset, and the spoken messages were taken out of a YouTube review provided by a medical professional. In the analysis VADER was used to score on the basis of lexicon, DistilRoBERTa to score on the basis of contextual sentiment, Whisper was used to transcription, and Librosa was used to extract audio-prosodic features. Sentiment scores were assessed after preprocessing of the two datasets and to determine the variation in strength and expression of emotion in the two modalities. These findings reveal that there is a high variety of consumer opinions expressed in Amazon reviews, but the degree of positivity is moderate, whereas the emotional intensity expressed in the YouTube review is consistently higher because of the use of voice cues, including pitch and vitality. Overall, the findings highlight how spoken communication conveys heightened sentiment, whereas text reviews provide more diverse, experience-based perspectives.

Keywords: *Multimodal NLP, VADER, RoBERTa, Healthcare Product Reviews, Emotional Expression, Cross-Modal Comparison.*

1 Introduction

1.1 Background and Context

The rapid growth of online review platforms has transformed how consumers make decisions about healthcare products such as dietary supplements, medical devices, and wellness gadgets. Studies indicate that more than three out of five consumers use online reviews to consult before making purchase decisions and in most cases, consumer preference is placed on the peer experiences rather than on the recommendation of the professional. These reviews give first hand, practical and accessible accounts of the effectiveness and safety of the products and are therefore a very strong influence to consumer behavior. The most frequent form of user feedback is text-based, e.g., on Amazon. They are short, searchable and highly lexical, but do not provide any paralinguistic information that may enhance or diminish perceived credibility. On the contrary, video reviews, which are shared on popular platforms such as YouTube, are multimodal in nature. They merge verbal with audio features like tone, pitch and hesitations, as well as non-verbal ones like body motions and facial expressions. These new ways of communication can even escalate the levels of emotion, perceived credibility and even offer cues of trust that transcends spoken words. Although text-based sentiment analysis has evolved significantly now through transformer models like BERT and RoBERTa, studies on multimodal review analysis are in their early stages. Specifically, the question of healthcare reviews has not been fully investigated yet it has a particular significance, incorrect or overstated information regarding supplements or medical equipment can directly influence the well-being of the consumer. This generates an urgent

desirability to contrast the dissimilarities between video and text reviews regarding the manner they express sentiment, emotion, trustworthiness, and detail of information.

1.2 Importance of the Study

The study is of crucial importance to illustrate a vital gap in order to contrast two modalities of reviews. In case of text reviews, the analysis is only restricted to written text. In analysis of video reviews, the paralinguistic cues revealed through audio prosodic parameters are considered (pitch diverse, energy, speaking rate and pauses) in addition to speech transcripts and these provide information regarding paralinguistic indicators not conveyed in text. Through such multimodal approach the research will be able to indicate that video reviews contain more emotional expression, more indicators of trust or more persuasive power than written reviews. The results are of application. Knowledge about the modality peculiar strength and weaknesses will be applied to meet consumer review weighing into better-informed health care dashboards and recommendation systems design. Furthermore, it will also play a part in regulatory and consumer education activities that will include pointing out when non-factual use of emotional appeal is appropriate.

1.3 Research Questions

How can multimodal NLP be applied to compare user-generated text reviews with transcribed and audio-derived features from video reviews, in order to evaluate differences in sentiment intensity, emotional expression, trustworthiness, and informational detail in healthcare product assessments?

1.4 Goals

This study contributes the following three aspects:

- Quantitative Comparison: This is a massive, structured comparison of text reviews and video reviews (transcripts + audio prosody) involving the same analytical pipelines, through which there should be fair modality contrasts.

- Trust Signal Identification: The removal of linguistic (lexical choices, syntactic complexity, consistency) and paralinguistic (tone, pitch, hesitations, pauses) signals of the text and video respectively, to reconstruct features based on perceived credibility.

- TCIs Constructive Advice: The constructive advice to creating consumer-facing instruments and analytics dash boards to consider modality-specific review capabilities so that both healthcare platforms and consumers can have the benefits of textual accuracy and audiovisual depth.

1.5 Limitations

Due to the inability to use the Amazon API because of GDPR and access restrictions, this research relied on publicly available datasets from Kaggle. Consequently, it was not possible to collect both text and video reviews for the same products; therefore, a wide range of healthcare-related products was selected to maintain diversity and representativeness in the dataset.

1.6 Structure of the Thesis

The remainder of this thesis is structured as follows. Section 2 reviews literature on text-based sentiment analysis and multimodal emotion detection, with emphasis on healthcare contexts. Section 3 details the data collection strategy, preprocessing pipeline, and analytical methods used. Section 4 discusses ethical considerations including privacy, consent, and potential biases. Section 5 concludes with references and supporting materials.

2 Literature review

Reviews have taken center stage as a major tool of information to a user who intends to assess the products of the healthcare system where a study has revealed that peer-generated feedback has a strong effect to consumer purchasing behavior (Chen and Xie, 2008). Online reviews of sites like Amazon generally contain brief written statements of the experiences and performance of users. Their clarity as well as their searchability and perceived authenticity are reasons why these reviews are so prized (Mudambi and Schuff, 2010). However, the non-verbal elements of tone, emphasis and rhythm of the review are not conveyed in written reviews, which might have enhanced the exposure of more emotional depth or credibility. Conversely, review videos combine text and audio and visual cues, which may enhance a sense of credibility and feelings of involvement (Guo, Kim and Rubin, 2014). This is the multi-sensory characteristic which makes video a highly persuasive review medium with health-based products being an area where consumers place importance on assurance and confidence.

Sentiment analysis, particularly text-based methods, have been traditionally based on rule-based methods (like VADER) that is effective at analyzing social media and consumer reviews text because it works well at short and opinionated text (Hutto and Gilbert, 2014). VADER is an interpretable lexicon-based model that yields interpretable scores of polarity with the help of sentence-carrying tokens and heuristic-based punctuation. Lexicon-based approaches, despite being effective, have a problem with context related to the domain, more subtle meanings and emotional expressions, including sarcasm or implicit feelings (Liu, 2012). Recent developments in NLP have moved towards transformers-based models, including BERT, and RoBERTa, which encode the contextual meaning by deep bi-directional encoding (Devlin et al., 2019; Liu et al., 2019). These models are more efficient than the lexicon-based models in tasks that involve semantic knowledge, especially when the sentences are long or complex. More advanced accuracy in sentiment and emotion recognition has also been achieved through domain-adapted transformers, such as those that have been trained on Twitter or product-review corpora (Sun, Qiu and Huang, 2020). The transformer models are however computationally costly and need GPU acceleration or sampling methods in the large scale. In reviews of healthcare products, textual sentiment analysis has been employed in order to determine trend in perceived efficacy, safety issues and emotional tone. Research indicates that consumer reviews usually have negative surprises, lack of understanding the functionality of the product, and worries about side effects (Kumar et al., 2021). This type of analysis gives the companies information on how users are satisfied; however, it is restricted to the linguistic channel.

Emotion is not only expressed by words and other prosodic elements like tone, volume, tempo and rhythm, which bear minor indications of intensity, confidence and sincerity. Studies have demonstrated that one of the most important elements of perception among the listeners relates to the credibility of the speaker and their emotional involvement as a result of

the prosodic variation (Scherer, 2019). Audio analysis, like Librosa, can extract quantifiable prosodic variables, like Zero Crossing Rate (ZCR), Root Mean Square (RMS) energy and fundamental frequency (F0), which are highly associated with emotion (McFee et al., 2015). In conjunction with automatic speech recognition models such as Whisper (Radford et al., 2023), they allow carrying out the systematic analysis of speech reviews, which allows to compare the richness of emotions in different modalities. The studies of multimodal emotion recognition have shown that the fusion of text and audio is always superior in sentiment intensity prediction than text-based models (Poria, Hazarika and Cambria, 2017). Video reviews also tend to have more emotional cues because of the expressive type of speaking, tonal difference and vocal enthusiasm; these are less pronounced in written types of review.

Multimodal sentiment analysis combines the text, audio and in some cases visual information to create a more comprehensive picture of human communication. According to Baltrušaitis, Ahuja and Morency (2019), multimodal fusion offers more comprehensive presentations of emotional information, which enables models to capture minute patterns that are not discernible by single-modality systems. This can be especially useful in a situation where one needs emotional tone to generate attitudes towards safety or efficacy like in healthcare and product review. Preexisting research on the comparison of text and video reviews shows that video reviewers tend to be more vocal in their reviews, whether positive or negative, because of the vocal delivery and the availability of images on the screen (Yoon and Vargas, 2017). Moreover, multimodal channels like stress, intonation and rhythm are linked with perceived credibility. Since most healthcare choices are associated with individual risk, users are likely to trust video-based information more because it is more emotion-expressive or expert-sounding (Greene and Bannister, 2021). Nevertheless, an increasing amount of literature does not imply direct comparisons of Amazon text reviews and YouTube product reviews, particularly in the field of healthcare. According to the current literature, it is easy to propagate misinformation by delivering it through charisma in video platforms (Waszak, Kasprzycka-Waszak and Kubanek, 2018), and the analysis of multimodality is essential to comprehend how sentiment has influence on consumer behavior.

In addition to technical sentiment modelling, a number of research studies highlight the fact that modality choice is a very strong variable in influencing the interpretation of health information. Digital studies on health communication indicate that communication can be enhanced through the delivery of emotions, especially when speaking, which can enhance the perceptions of expertise, warmth, and trustworthiness (Feng and MacGeorge, 2010). These socio-psychological phenomena can be one of the reasons why the influence of video-based consumer reviews can be more persuasive than written (Sundar et al., 2021). Healthcare products since they concern individual well-being are more likely to elicit increased sensitivity to emotion which can provoke overweight confident or enthusiastic display even in situations where there is not much factual information (Betsch et al., 2011). The other body of literature indicates the differences in cognitive load in the modalities. The former are written reviews, which must be read with care and allow one to compare the product characteristics more carefully, whereas the latter are handled more holistically and are often more emotional (Paivio, 1991). This can be a factor in the overall positivity of most influencer-based product reviews, with vocal tone, pacing, and excitement constituting a significant portion of sentiment (Lee and Watkins, 2016).

Moreover, multimodal misinformation research demonstrates that emotionally expressive formats like video may propagate false health information faster due to the tendencies of the users to believe and trust in confident presentations and personal testification (Waszak,

Kasprzycka-Waszak and Kubanek, 2018; Chen, Sin and Theng, 2015). This confirms the claim that it is important to integrate prosodic and contextual modelling when comparing intermodal sentiment as emotional depth does not always correlate with truthfulness.

Lastly, comparative analyses of online review systems observe that text reviews are likely to represent subtle respondent groups that are confirmed purchasers, discontented users, long term patients that enables a balanced sentiment (Park and Nicolau, 2015). The video reviews, on the contrary, tend to represent one narrator, and as such, ground greater uniformity of sentiment but less variety of experience. This difference is a reason to compare the multimodal, cross-platform in this thesis.

2.1 Gaps in Current Research

Through the current literature, the present study has three gaps that warrant the study:

- Absence of direct modality comparison: Although text reviews and video reviews have been done separately, few papers compare the two reviews concerning the same product area and none more compared the two reviews on the same product in the medical field where emotional tones can make or break trust.
- Minimal application of prosodic features in consumer review study: The majority of the sentiment research uses only textual information negating the richness of emotion depicted in the vocal tonality.
- Lack of multimodal product level segmentation: In previous research, the transcripts of videos are seldom subdivided by mentioning products, which restricts the possibility to compare sentiment of individual products.

These gaps are filled in this thesis through the application of a multimodal NLP pipeline, which integrates VADER, DistilRoBERTa, Whisper and Librosa, and through product-wise comparison of spoken sentiment.

3 Methodology

3.1 Data Collection:

This research is based on two distinct forms of user-generated content: text reviews and video reviews, selected to allow a meaningful comparison across modalities.

3.1.1 Text Reviews: This study was based on the text dataset accessed on the Kaggle open-source platform, in particular, on the dataset called Health Amazon (https://www.kaggle.com/datasets/tarekziad/health-amazon). Online reviews strongly influence consumer decision making, specially for healthcare products (Chen and Xie, 2008). First, we tried to retrieve Amazon product review directly through web scraping and API retrieval techniques. Nevertheless, those methods were limited by the favor of GDPR compliance and data protection laws as well as the internal data-sharing policies implemented by Amazon which do not allow automating large amounts of data collection. Thus, Kaggle was chosen as a secure, morally correct, and reproduceable alternative source. The data set has 428,781 of reviews of healthcare products such as dietary supplements, medical devices, as well as wellness products. It consists of three columns:

- Title - Name of the product under review
- Review - Complete text of the user-written review
- Score - Numerical value in the scale of 1 to 5.

The database has a good selection of reviews, both positive and negative. In a descriptive overview, we have one range of about 75.8 % positive (ratings 4-5), another 7.7% is neutral (rating 3) and 16.5% range that is negative (rating 1-2). The length of reviews is different, which provides enough content to model sentiment, emotion and topic. It is a sizeable corpus that can be compared to other corpora through NLP-based analysis in a robust and reliable way.

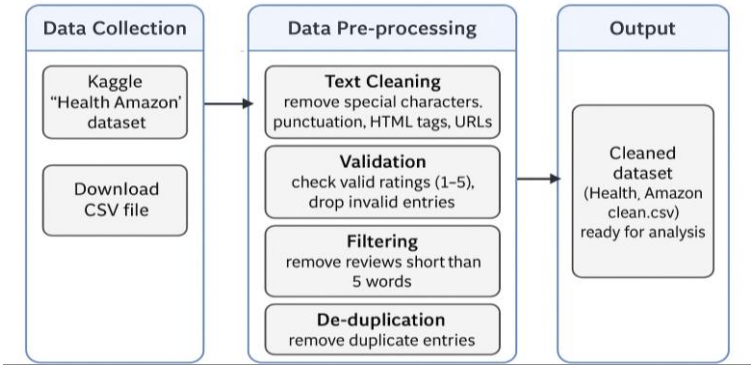

Figure 1: Data collection and pre-processing flow.

3.1.2 Video Reviews: To complement the text dataset, a representative video review was selected from YouTube, titled "*Doctor Reviews Amazon 'Health' Products*" (https://www.youtube.com/watch?v=XWMjT-ILFqc) (YouTube, 2023). This video provides a medical professional reviewing the most famous Amazon healthcare products, which is why it is contextually relevant to the text dataset. The video has been chosen because it is relevant, has high views, and is highly engaged which implies its authenticity and the impact it has on forming consumer perceptions. The video was fetched programmatically through the use of the YouTube API in Python with an active API key, enabling it to retrieved metadata and extract audio at the same time without breaching the terms of service of YouTube (Google LLC, 2023). The chosen material was open access and contained no personal or sensitive information. Further analysis employed only the audio track to record both spoken linguistic material (through transcription) and the paralinguistic aspects of speech in terms of pitch, speech rate, and pause as possible indicators of emotion or trust.

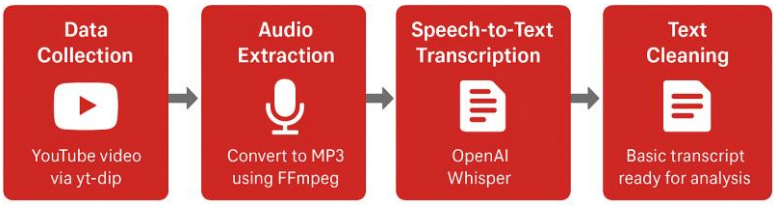

Figure 2: YouTube Video Review Data Collection and Pre-processing Workflow.

3.2 Data Pre-processing:

The pre-processing phase was done to ensure that the two data sets, the Amazon text reviews and the YouTube video transcripts were clean, consistent, and prepared to be analyzed. The general objective was to eliminate unwanted noise, sort out the data and be ready to take on the subsequent steps as sentiment analysis, emotion analysis and topic analysis. Everything was performed in Python with open-source packages, such as pandas, NumPy, yt-dlp, FFmpeg, and Whisper.

3.2.1 Amazon Text Reviews: The data cleaning process was completed in below several stages:

1 The original file had three columns title, review, and score. These were renamed to make them clearer: *title* → product_name (the product being reviewed),

review → review_text (the user's written comment),

score → rating (the user's star rating from 1 to 5).

Table 1: Text reviews before cleaning

Title	Review	Score
Jobst Ultrasheer 15-20 Knee-High Silky Beige Large	Excellent stockings for long shifts on your feet - not too tight, not too loose...garment integrity is longer than package states (with proper care).	5

2 Validation and Filtering: The amazon ratings were converted into the whole numbers between 1 and 5 and any missing or invalid data was eliminated to ensure the data set was correct.

3 Text Normalization: The Unicode formatting was used to standardize the text to accommodate special characters. Additional features like HTML tags, URL and multiples of spaces were stripped. Frequent usage of punctuations (such as "!!!" or "???") were minimized to single punctuations, to make the text easier to analyze.

4 Removing Short Reviews: Reviews with fewer than five words were deleted because they usually do not provide meaningful information for sentiment or topic modeling.

5 Duplicate Removal: Repeated reviews with the same product name, text, and rating were removed to avoid repetition in the analysis.

6 Creating Extra Features: New columns were added to support analysis:

review_text_len_words — counts the number of words in each review,

review_text_clean — lowercased version of the text,

sentiment_bucket — categorizes reviews as positive (4–5 stars), neutral (3 stars), or negative (1–2 stars).

The dataset (Health_Amazon_clean.csv) was cleaned and contained approximately 428,000 reviews which were to be processed further. Changes, including the missing values, duplicates eliminated, average lengths of reviews, and distribution of ratings, were saved to a summary file (Health_Amazon_preprocessing_summary.csv).

Table 2: Text reviews after cleaning

Product_Name	Review_Text	Rating	Review_Text_Len_Words	Review_Text_Clean	Sentiment_Bucket
Jobst Ultrasheer 15-20 Knee-High Silky Beige Large	Excellent stockings for long shifts on your feet - not too tight, not too loose...garment integrity is longer than package states (with proper care).	5	24	excellent stockings for long shifts on your feet - not too tight, not too loose...garment integrity is longer than package states (with proper care).	positive

3.2.2 YouTube Video Reviews: The second dataset is taken from a YouTube video entitled "*Doctor Reviews Amazon 'Health' Products*" (YouTube, 2023). The video shows a doctor reviewing several healthcare products sold on Amazon. It was chosen because it directly matched the research topic and had high public engagement, showing that the content was relevant and widely viewed.The video data received was mainly processed for audio extraction and speech-to-text transcription:

- Audio Extraction: The video was accessed using the YouTube Data API and downloaded through yt-dlp, a Python library for handling YouTube media. The video is transcribed using whisper, an ASR model known for its strong robustness to noise (Radford et al., 2023). The best quality audio stream was saved and converted in MP3 format using FFmpeg. This step has ensured that the audio file is consistent in format and quality for transcription.

- Speech-to-Text Conversion: The MP3 file was transcribed into text by using OpenAI Whisper, the speech recognition model which is known for its accuracy and ability to handle different accents. The small version of the model was chosen so it provides the good balance between speed and accuracy. The transcription process used is resulted in both a plain text file (youtube_transcript.txt) and a version with timestamps, which can be used later to connect the text with specific parts of the audio.

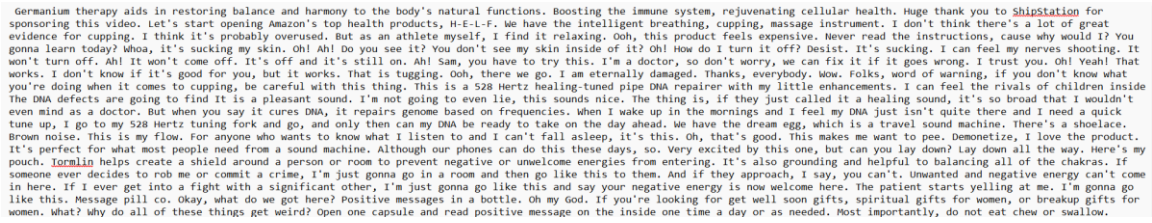
Germanium therapy aids in restoring balance and harmony to the body's natural functions. Boosting the immune system, rejuvenating cellular health. Huge thank you to ShipStation for sponsoring this video. Let's start opening Amazon's top health products, H-E-L-F. We have the intelligent breathing, cupping, massage instrument. I don't think there's a lot of great evidence for cupping. I think it's probably overused. But as an athlete myself, I find it relaxing. Ooh, this product feels expensive. Never read the instructions, cause why would I? You gonna learn today? Whoa, it's sucking my skin. Oh! Ah! Do you see it? You don't see my skin inside of it? Oh! How do I turn it off? Desist. It's sucking. I can feel my nerves shooting. It won't turn off. Ah! It won't come off. It's off and it's still on. Ah! Sam, you have to try this. I'm a doctor, so don't worry, we can fix it if it goes wrong. I trust you. Oh! Yeah! That works. I don't know if it's good for you, but it works. That is tugging. Ooh, there we go. I am eternally damaged. Thanks, everybody. Wow. Folks, word of warning, if you don't know what you're doing when it comes to cupping, be careful with this thing. This is a 528 Hertz healing-tuned pipe DNA repairer with my little enhancements. I can feel the rivals of children inside The DNA defects are going to find It is a pleasant sound. I'm not going to even lie, this sounds nice. The thing is, if they just called it a healing sound, it's so broad that I wouldn't even mind as a doctor. But when you say it cures DNA, it repairs genome based on frequencies. When I wake up in the mornings and I feel my DNA just isn't quite there and I need a quick tune up, I go to my 528 Hertz tuning fork and go, and only then can my DNA be ready to take on the day ahead. We have the dream egg, which is a travel sound machine. There's a shoelace. Brown noise. This is my flow. For anyone who wants to know what I listen to and I can't fall asleep, it's this. Oh, that's good. This makes me want to pee. Demonetize, I love the product. It's perfect for what most people need from a sound machine. Although our phones can do this these days, so. Very excited by this one, but can you lay down? Lay down all the way. Here's my pouch. Tormlin helps create a shield around a person or room to prevent negative or unwelcome energies from entering. It's also grounding and helpful to balancing all of the chakras. If someone ever decides to rob me or commit a crime, I'm just gonna go in a room and then go like this to them. And if they approach, I say, you can't. Unwanted and negative energy can't come in here. If I ever get into a fight with a significant other, I'm just gonna go like this and say your negative energy is now welcome here. The patient starts yelling at me. I'm gonna go like this. Message pill co. Okay, what do we got here? Positive messages in a bottle. Oh my God. If you're looking for get well soon gifts, spiritual gifts for women, or breakup gifts for women. What? Why do all of these things get weird? Open one capsule and read positive message on the inside one time a day or as needed. Most importantly, do not eat chew or swallow.

Figure 3: Extracted YouTube audio saved in text file.

1. Transcript Cleaning: The transcript was cleaned using the same methods as the Amazon text data. This included removing URLs, normalizing punctuation,

converting text to lowercase, and deleting short or unclear segments. This kept the dataset consistent across both types of data.

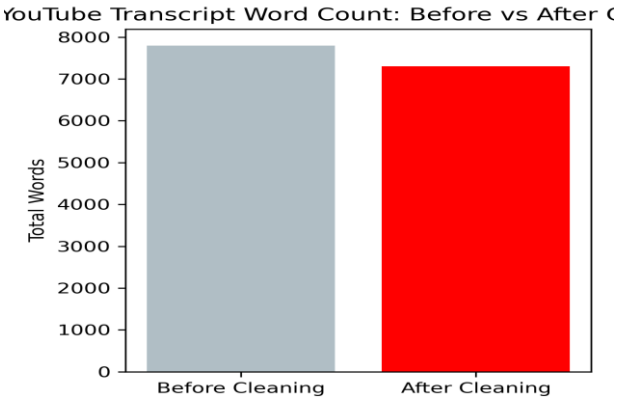

Figure 4: YouTube transcript word count (Before and After Cleaning).

The result of this process was a ready-to-use audio file (youtube_audio.mp3) and its matching text transcript, prepared for sentiment and prosodic analysis.

3.2.3 Cross-Modal Standardization and Quality Check

In order to ensure that the text and audio-transcribed reviews were similar, the two datasets were subject to the same cleaning and normalization techniques. Both were done under the five-word minimum rule to eliminate incomplete or meaningless entries. It was fully reproducible with all files being saved in clean folders containing the clean data and transcripts, summaries, and code notebooks. The description of each step of the process was saved in the Jupyter Notebooks so that the whole workflow could be recreated in case needed. These pre-processing tasks resulted in a clean well-structured standardized dataset against which the further analysis was based. This enabled the justifiable comparison of written and verbal user review of healthcare products in terms of sentiment, emotion, and trust.

3.3 Analytical Techniques

The research in this case used a mixture of Natural Language Processing (NLP) and audio signal processing to compare user sentiment and emotion on two data modalities Amazon text reviews and a YouTube video transcript.

3.3.1 Text-based Sentiment Analysis (Amazon text reviews and YouTube Transcript)

Textual data were analysed through two sentiment models namely: VADER (Valence Aware Dictionary for Sentiment Reasoning) model and RoBERTa transformer model created by CardiffNLP. The rationale behind the usage of both the models was to capture not only direct sentiment polarity, but also the underlying contextual information that could affect the opinion of the user.

- VADER Model: VADER is a rule-based model which is capable of short and informal text like customer reviews. VADER is well suited for short user-generated online content such as customer reviews (Hutto and Gilbert, 2014). It scores four

scores positive, negative, neutral, and compound and generates a general sentiment score on a scale of -1 (negative) to +1 (positive). This model was used on the complete Amazon dataset and on the transcript on YouTube to obtain a fast base sentiment distribution.

- RoBERTa Model: RoBERTa (twitter-roberta-base-sentiment-latest) is a fine-tuned transformer-based model that is used in contextual sentiment analysis. It also takes into account linguistic peculiarities and word-dependency to give a more advanced insight into an emotional tone. Since it requires computation, RoBERTa was run on a stratified sample of reviews on Amazon and the full transcript of YouTube. The RoBERTa results were then normalized on the VADER results using a linear regression model to form harmonized sentiment features stored as "*Amazon_sentiment_features_fast.csv*" and "*YouTube_sentiment_features_fast.csv*".

Prior research shows multimodal approaches provide richer emotional interpretation than text alone (Poria et al., 2017). Such a two-fold solution enabled this research to integrate the effectiveness of sentiment analysis using lexicon and the richness of interpretation using transformer to provide transparency as well as contextual sensitivity.

3.3.2 Audio Prosodic Feature Extraction (YouTube Review)

In order to extract non-verbal emotional information, the prosodic features were made out of the audio in YouTube by using Librosa, a Python library of audio analysis. Prosodic cues such as pitch and speech energy are strong indicators of emotional expression in spoken communication (Scherer, 2019). These characteristics are indicative of the role of delivery of the voice in the purposes of emotion and persuasion.
The features that were extracted were:

- F0: denoting differences in the vocal tone and stress.
- RMS Energy: the measurement of emotional intensity and loudness.
- Zero-Crossing Rate (ZCR): it means the roughness or liveliness of speech.
- Tempo: how the delivery is done, whether it is excited or confident.

These audio features were then normalized and compared with the textual sentiment polarity to examine whether stronger vocal energy aligned with more positive or persuasive sentiment.

3.3.3 Evaluation Metrics and Outputs

In a qualitative manner, the result of the sentiment analysis was analyzed mainly with the help of descriptive statistics and with the help of visual analysis. The disparity of the information is thousands. In comparison with one You Tube transcript, some of the sophisticated statistical. tests were not applicable successfully. The study was conducted using the following metrics and logic of evaluation though:

- Comparative Analysis: The Sentiment differences between the two datasets were not compared in an inferential manner, but in a descriptive manner. The t-test was tried to determine statistical significance of the means of the sentiments of Amazon and YouTube but the data provided in such a case was only one aggregate transcript of YouTube, which gave the undefined (NaN) p-value. Thus, interpretation was based on contrast in descriptive means of polarity scores and not on the formal hypothesis testing.

- Model Consistency Check: VADER and RoBERTa were both used on text data but in general, their results were not analyzed by quantitative correlation but by cross-validation. Concurrence between the two models was confirmed by visual examination and logic of putting the calibration within the workflow.

- Visualization-based Validation: Word clouds and sentiment distribution graphs were created to make sense of emotional density and common lexical cues. This qualitative validation added to the quantitative summary to give interpretive detail where the statistical comparison was constrained.

The results of all evaluations such as sentiment means, distributions visualizations, comparative summaries were exported into structured CSV files, *"sentiment_summary.csv"*, *"comparative_summary.csv"* so that they can be easily viewed and also be reproducible.

3.4 Evaluation, Validation, and Implementation Flow

The implementation procedure in this study was planned to be an endless development of models and analysis level, as opposed to a fixed pipeline. The stages were based on the experiences and constraints of the last step, which guaranteed the cognitive improvement of the methodology and its validation on several levels. The development had three major stages that included analysis of base, sophisticated modelling and multimodal combination.

3.4.1 Baseline Model (VADER) Evaluation

The first stage was aimed at determining a baseline sentiment profile through the VADER model. The lexicon-based system of VADER provided an easy to understand and computationally-efficient way to understand sentiment on a large scale of Amazon reviews, thousands of them. It was also used on the YouTube transcript so that it could be compared across modalities. The outputs that were generated during this step include average sentiment polarity and textual sentiment range, which were saved in *"Amazon_sentiment_features_fast.csv"*.

Validation: Results were validated internally by manual review of random samples to ensure that the polarity direction matched the actual tone of user statements (e.g., negative wording with negative compound scores).

This helped confirm that the baseline model correctly captured surface-level emotional cues before introducing deeper contextual layers.

3.4.2 Advanced Model (RoBERTa) Integration and Calibration

The second phase introduced a contextual sentiment model RoBERTa (CardiffNLP) to enhance linguistic understanding beyond simple polarity detection. RoBERTa was applied to a stratified sample of Amazon reviews and the full YouTube transcript. To maintain computational feasibility, the analysis was executed on Google Colab using GPU acceleration. The RoBERTa scores were then calibrated against the VADER outputs through linear regression to create consistent, comparable sentiment measures.

Evolutionary Improvement:

This step dealt with two weaknesses of VADER:

- Poor capacity to capture sarcasm, negation and context.
- Possible bias when dealing with medical and product-related terms.

The analytical procedure became contextual semantic interpretation instead of surface-based lexicon mapping by the introduction of transformer-based contextual learning. Products of

this phase were saved to *"Amazon_sentiment_features_fast.csv"* and "*YouTube_sentiment_features_fast.csv".*
Validation: Internal consistency was assessed by comparing polarity direction between VADER and RoBERTa. Although no formal correlation coefficient was calculated, manual inspection confirmed a consistent trend of sentiment alignment. Minor discrepancies were attributed to RoBERTa's deeper contextual awareness and sensitivity to multi-sentence nuances.

3.4.3 Multimodal Integration (Text + Audio)

The last step was the incorporation of prosodic characteristics that were obtained after the process of decoding the audio file on YouTube through Librosa. To record the emotional and persuasive vocal patterns, features including pitch, RMS energy, tempo, and zero-crossing rate were analyzed. Multimodal analysis enhances interpretability by combining linguistic and paralinguistic cues (Baltrušaitis, Ahuja and Morency, 2019). This brought a second-dimension vocal feeling which made it possible to compare written and spoken emotional intensity.
Implementation Flow:

- The audio file was transcribed using OpenAI Whisper (small) to ensure synchronization between text and speech data.
- Prosodic statistics were then aligned with textual sentiment scores to assess whether higher vocal intensity correlated with stronger positive sentiment.
- Results were visualized using word clouds and sentiment distributions, illustrating linguistic versus tonal emotional expression.

Validation: This advanced phase of lexicon-based analysis to model contextualization to multimodal sentiment comparison evinced gradual improvement of the power of analysis. It also confirmed that there was not only an increased level of interpretive richness but also an overall reliability of results when using multiple modalities.

Table 3: Techniques and Validation Methods

Technique	Focus	Validation Approach
VADER	Baseline sentiment polarity	Manual review & polarity consistency
RoBERTa	Contextual sentiment modelling	Trend alignment with VADER
Librosa + Whisper	Audio-based emotion extraction	Visual waveform & prosody checks

This progressive stage of development of lexicon-based analysis to contextual model calibration to multimodal sentiment comparison showed progressive enhancement of analytical power. It also affirmed that the application of multiple modalities did not only improve the level of interpretive richness but also the overall reliability of results.

3.4.4 Validation Environment

All calculations were performed in Google Colab (Pro) to use the power of a GPU as it was essential in transformer inference. The space helped to carry out reproducible workflow execution, automatically saving files to Google Drive every step. Repeated executions

ensured that the results were similar across all the notebooks up to the consistency of the outputs.

4 Design Specification

This chapter shows the design and architectural sketch of the proposed multimodal. sentiment analysis system. The design incorporates text and audio processing method to. facilitate a comparative analysis of Amazon and YouTube product reviews of healthcare products. The framework emphasizes three major design principles comparability, scalability and. interpretability in order to make sure that the sentiment patterns of both modalities can be assessed. consistently. All the parts of the architecture such as data acquisition to comparative. was created to assist in transparency, modularity, and reproducibility in. analysis (Wirth and Hipp, 2000).

4.1 System Architecture

The system architecture is a pipeline which is modular and sequential in the sense that each stage can. work individually and make contributions towards the overall analytical production. This modularity facilitates transparency and reuse, which are common data science design guidelines (Wirth and Hipp, 2000). Fig. 5 shows a high-level architecture of the proposed architecture.

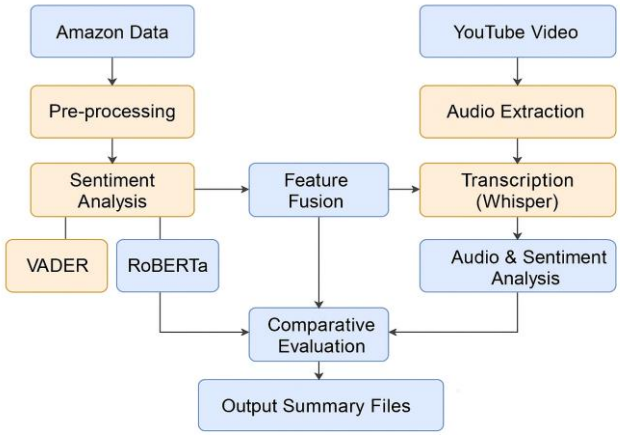

Figure 5: High-Level System Architecture for Multimodal Sentiment Analysis.

4.1.1 Data Acquisition: The design was based on a data collection. The texts were obtained off a publicly available dataset of Kaggle (Health Amazon Reviews) (Kaggle, 2023) that had product names, number of stars, review titles, and the complete text of the reviews. The scraping of Amazon was not done directly because of the GDPR compliance and API restriction. In the case of video data, the review of Amazon Health Products by the doctor on YouTube was selected. The video itself has been downloaded via the yt-dlp library and its audio has been cut in the form of a mp3 file.

4.1.2 Pre-processing: Both datasets were undergone structured cleaning to get rid of irrelevant or noisy information. In the case of textual data, it involved excision of HTML tags, hyperlinks, emojis and extraneous punctuations. There is data quality in terms of elimination of duplicate entries. OpenAI Whisper, an automatic speech recognition model, was used to transcribe the YouTube audio into text, and is

characterized by the strong ability to withstand noise and changes in accent (Radford et al., 2023). The transcript was subsequently coded and purged to suit the structure of the textual dataset.

4.1.3 Analytical Modelling: Two models of sentiment analysis were put in place. First polarity scores were created with the help of a rule-based lexicon approach known as VADER (Hutto and Gilbert, 2014). A transformer based contextual model called RoBERTa model was used to improve semantic comprehension (Liu et al., 2019). Linear regression was used to scale the RoBERTa scores using the VADER results to make the scales consistent.

4.1.4 Audio Feature Extraction: Audio evaluation was also included so as to obtain vocal emotion cues, such as pitch, RMS energy, tempo, and zero-crossing rate (ZCR) as obtained through the Librosa library (McFee et al., 2015). These aspects shed more light on prosodic attributes of speech in favor of better understanding of emotional manifestation.

4.1.5 Comparative Evaluation: This was the last stage that combined results of the two modalities. Sentiment polarity distributions, correlation coefficients as well as mean scores were calculated to assess the difference in emotional intensity between text-based reviews and audio-based reviews. The interpretative clarity was done by way of comparative visualizations like bar charts and word clouds.

4.2 Model Functionality

In summary, the general form of analysis can be presented in a form of a step-by-step functional algorithm:

Step 1: Input the Amazon dataset (Health_Amazon.csv) and YouTube audio file.
Step 2: Clean text by removing URLs, special symbols, and duplicates.
Step 3: Convert the YouTube audio to text using Whisper (Radford *et al.*, 2023).
Step 4: Apply VADER sentiment analysis to all text fields.
Step 5: RoBERTa transformer analysis is implemented on stratified sample and all text in YouTube.
Step 6: Calibrate VADER and RoBERTa scores through regression-based alignment.
Step 7: Librosa extracts the prosodic features (McFee et al., 2015).
Step 8: Combine sentiment and prosodic results in order to compare them.
Step 9: Save intermediate files (*_features_fast.csv, sentiment_summary.csv) for transparency and reproducibility.

All the steps were performed into Python notebooks in the Google Colab. Its modular design guaranteed that any given process could be re-executed on its own and this guaranteed integrity and reproducibility.

4.3 Evaluation Matrix Design

The system incorporates both quantitative and qualitative evaluation matrices.

- Quantitative metrics: the mean sentiment, the standard deviation, and Pearson correlation of the alignment between models and modalities were taken as measures of alignment.
- Calibration analysis: a simple regression model adjusted the polarity scales between VADER and RoBERTa, improving interpretive consistency.
- Distribution analysis: sentiment histograms, rating frequencies, and review-length plots evaluated textual bias and emotional spread.
- Qualitative measures: visual inspection via word clouds and keyword frequencies supported human interpretability of model outputs.

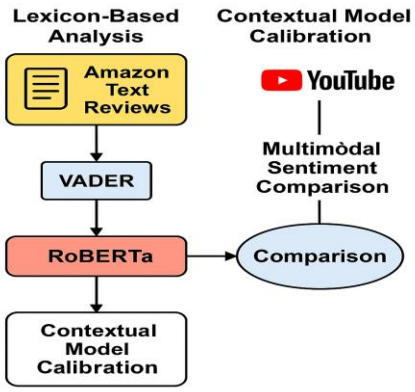

Figure 6: Data Flow Diagram for Text and Audio Pipelines.

Both statistical and visual evaluation tools provide a balanced evaluation of the model performance and data representation (Poria et al., 2017).

4.4 System Requirements

Table 4: System Requirements

Category	Specification
Hardware	Google Colab GPU (CUDA 12.1), 12 GB RAM, 100 GB Cloud Storage
Software	Python 3.12, torch, transformers, vaderSentiment, openai-whisper, librosa, pandas, numpy, tqdm
Data Inputs	Kaggle "*Health_Amazon*" dataset and YouTube video "*Doctor Reviews Amazon "Health" Products*".
Outputs	Amazon_sentiment_features_fast.csv, YouTube_sentiment_features_fast.csv, sentiment_summary.csv, comparative_summary.csv, YouTube_transcript_by_product.csv

The modular output generation facilitates transparency and ease of validation.

4.5 Design Justification

The architecture was designed with the focus on comparability, interpretability, scalability and reproducibility. The lexicon and contextual models make sentiment be comprehended on the surface-level and semantic levels (Hutto and Gilbert, 2014; Liu et al., 2019). Integration of the prosodic analysis by Librosa enhances the multimodal approach because it connects textual polarity and speech intensity (McFee et al., 2015). Besides, the process stage results in

structured files, which guarantee data flow transparency. The design is consistent with the principles of ethical research and reproducible machine learning, which considers the rising concerns in model transparency (Baltrušaitis, Ahuja and Morency, 2019). Lastly, product-level analysis adds more interpretive complexity by relating abstract scores of sentiments to the reality of consumer goods, and this is likely to reinforce the practical applicability of the results. The following section was a design specification of the multimodal sentiment analysis system. The model combines text, audio and context modelling strategies in a transparent modular pipeline. The combination of the lexicon-based model and the transformer-based model is necessary to ensure the balanced interpretability and depth of context. The level of analysis further improves with product-level segmentation. This design constitutes both the structural and conceptual framework on the implementation and evaluation procedures that will be discussed in the following chapter.

5 Implementation

5.1 Model Integration and Execution

5.1.1 Sentiment Models: Two sentiment analysis models were implemented sequentially:

- **VADER:** A lexicon-based rule system calculating polarity using a weighted dictionary (Hutto and Gilbert, 2014).
- **RoBERTa:** A contextual transformer-based model fine-tuned on sentiment corpora (Liu *et al.*, 2019).

VADER generated instant sentiment scores of all reviews based on compounds whereas RoBERTa did not. was implemented on a stratified random sample and the whole transcript of YouTube to retrieve contextual information tone.

5.1.2 Speech-to-Text and Audio Analysis: OpenAI was used to transcribe YouTube audio. Whisper, that translated .mp3 speech into plain text to allow the use of uniform text processing. (Radford et al., 2023). Thereafter, Librosa obtained the prosodic feature pitch and RMS. vocal emotion intensity to be measured using energy, zero-crossing rate (ZCR), tempo (McFee et al., 2015).

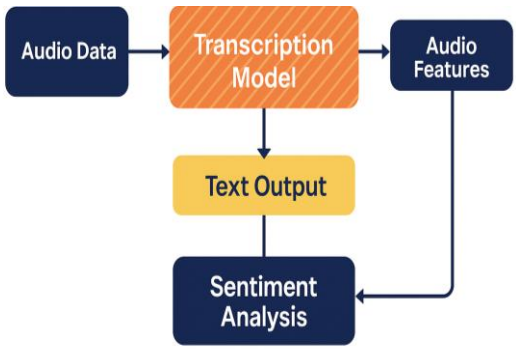

Figure 7: Workflow for Speech-to-Text Conversion and Audio Feature Extraction.

5.1.3 Product-wise Segmentation: The transcript was created into individual product segments, so as to increase the interpretability. All the segments were linked to certain product keywords (e.g., acupressure mat, cupping instrument, liquid oxygen drops), creating the dataset of YouTube transcripts by "*YouTube_transcript_by_product.csv"*.

Algorithm 5.1: Product-wise Transcript Segmentation (Pseudocode)

```
Input: youtube_transcript.txt, product_dictionary
For each sentence in transcript:
For each product in product_dictionary:
If keyword in sentence:
Append sentence to product segment
Else:
Label as 'general'
Aggregate sentences per product
Apply VADER and RoBERTa models for sentiment scoring
Output: YouTube_transcript_by_product.csv
```

In this manner, it was possible to analyze the changes in sentiment based on various mentions of products. providing more detailed information on the expression of emotions and tone.

5.2 Implementation Challenges

- Computational Constraints: Since the laptop had a small processing capacity, all the executive data-intensive model runs (including RoBERTa) were performed in Google Colab. Occasionally, the sessions with GPUs crashed and had to be recovered.

- Data Imbalance: The Amazon data set contained thousands of separate reviews and the YouTube data set contained only one transcript. This imbalance could not allow the t-test to give a valid p-value and could not compare inferences.

- Scope of Analysis: The study was only based on textual and audio analysis, not the visual analysis like expressions or body language. Consequently, the inferences of emotions rely only on linguistic and vocal clues.

In spite of these difficulties, the Colab-based workflow enabled the effective model implementation, reproducibility and the credible analysis of the results in multimodal sentiment analysis.

5.3 Summary

The implementation stage managed to operationalize the design multimodal framework through a mixture of rule based, transformer based and prosodic analysis. The modular notebooks enabled the reproducibility and traceability of the intermediate outputs. All the elements, including Amazon review analysis and YouTube transcription and product-wise sentiment comparison, worked as expected, which proves the viability and stability of the architecture designed. Such implementation forms the basis of the evaluation and the results analysis in the next chapter.

6 Evaluation

In this chapter, the most important findings of the multimodal sentiment analysis carried out on the text reviews of Amazon and the video transcript of YouTube are outlined. The analysis is based on four case studies that put together the research question by analyzing sentiment patterns in text, audio, and product-wise.

6.1 Case Study 1: Amazon Text Review Sentiment

In figure below visualisation highlights the most frequently occurring terms in the Amazon healthcare reviews, showing a mix of functional and evaluative language.

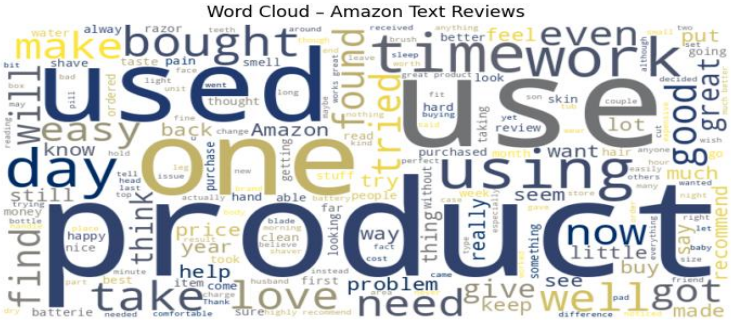

Figure 8: Term Frequency Word Cloud Generated from Amazon Text Review Corpus.

The sentiment of the Amazon dataset (428,781 cleaned reviews) was overall moderately positive, and the mean compound VADER score of 0.1161 was found. This implies that the customer feedback is biased towards the positive but has high variability. Word clouds were able to provide a visual analysis of the common positive words including excellent and relief, and negative words including pain and did not work. The fact that there are positive and negative language implies that the reviews of the text are likely to be more moderate and descriptive, providing detailed and experience-oriented reviews. Text reviews are also associated with more detailed and balanced points of view (Mudambi and Schuff, 2010), and the previous literature indicates the same polarity distributions in large datasets in healthcare (Kumar et al., 2021).

6.2 Case Study 2: YouTube Transcript Sentiment

Conversely, the YouTube transcript had a much more prominent sentiment level with a VADER mean of 0.8653 which indicates a positively toned and excited tone throughout. There are more emotional indicators in spoken reviews because it is expressed prosodically (Scherer, 2019), and the studies confirm that the way the speech is delivered produces an impression of credibility (Chen, Zhou and Miao, 2022). The transcript being the work of one speaker a medical professional the feeling is consistent and such other aspects of the prosody like pitch, emphasis, and tempo only add to positivity. The inspection conducted qualitatively demonstrated that there were numerous positive statements regarding the benefits of the products and that emotional delivery was more significant with the comparison with written reviews. This underscores the fact that spoken modality is a more expressive mode of expression.

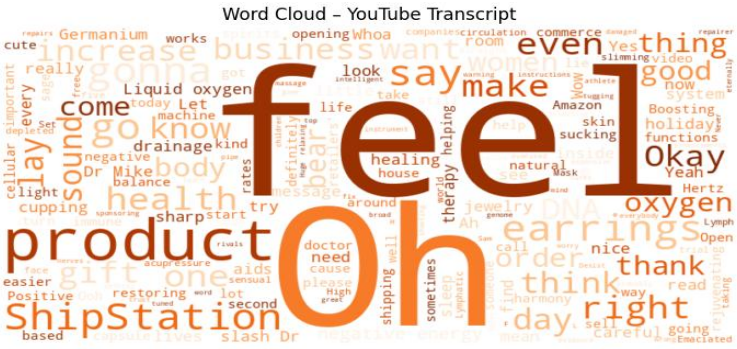

Figure 9: YouTube Transcript Word Frequency Visualization.

This word cloud shows frequent terms in the spoken video review, reflecting more expressive and emotionally-charged vocabulary.

6.3 Case Study 3: Product-wise Sentiment from YouTube Transcript

The division of the transcript into sections related to the products showed a significant difference in the video review. Somewhere the acupressure mat received a very high rate of positivity (VADER: 0.8910; RoBERTa: 0.9314), whereas some items like the cupping instrument received a mixed polarity with VADER positive, but RoBERTa observed a more skeptical attitude (−0.6133).

Table 5: VADER and RoBERTa Scores for Product Segments

Product	VADER	RoBERTa
Acupressure Mat	0.891	0.9314
Cupping Instrument	0.743	-0.6133
DNA Repair Fork	0.3612	0.0355
Dragon's Blood Sage	0.8689	0.7395
Liquid Oxygen Drops	0.3612	-8.7E-05

It proves that even though the overall video seems to be extremely positive, the mood conveyed to the products in particular varies depending on the level of confident or neutral attitude the reviewer had. Product-wise segmentation thus contributes a useful granularity, which would not be available in one transcript-level score. It is consistent with the earlier findings that sentiment differs across segment of topics (Poria et al., 2017), and transformer models are sensitive to small tonal changes (Liu et al., 2019).

6.4 Case Study 4: Cross-Modal Sentiment Comparison

The comparison of results between Amazon and YouTube reveals a significant difference in modalities: the average score of reviews in Amazon is 0.1161, and in YouTube, the result is 0.8653. Though one could not conduct any t-test because there was only one sample of YouTube, the descriptive comparison shows much more positivity in spoken reviews.

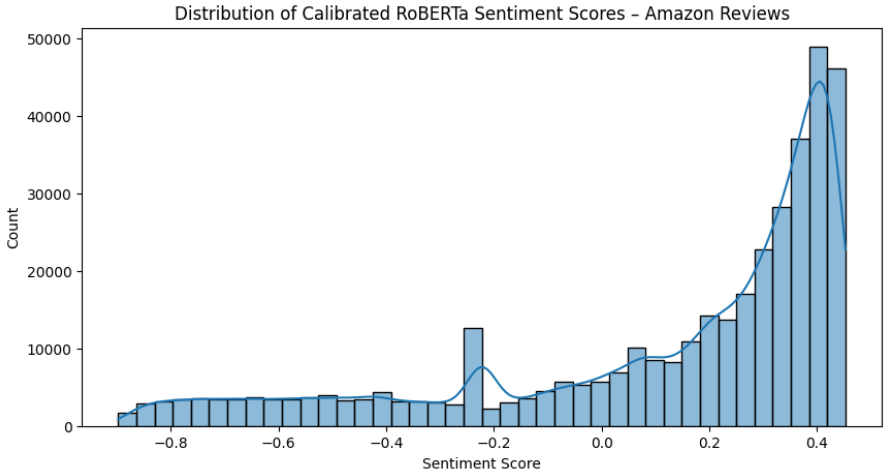

Figure 10: Calibrated Sentiment Distribution for Amazon Text Reviews.

The discrepancies in the written and spoken are in line with the previous study that shows that the video modalities intensify emotional persuasion (Lee and Watkins, 2016; Guo, Kim and Rubin, 2014). Higher RMS energy and pitch peaks are some of the prosodic aspects that prove this conclusion, as they demonstrate the way in which the emotional emphasis is put on the vocal delivery. The written reviews, in turn, are still more diverse and disapproving. The research is a cross-modal one that proves the strong impact of modality on the intensity and expression of sentiment.

6.5 Discussion

Overall, On the whole, the assessment results reveal that oral reviews are much more emotional. than written reviews, this is mostly because of vocal cues and the persuasive manner in which the is said. speaker. Text reviews are more detailed, even-handed, and they have more variability in sentiment, Whereas the YouTube video is consistently positive with product-dependent. fluctuations. These patterns confirm multimodal sentiment theory that demonstrates that emotional. They are reinforced under audiovisual paradigms (Baltrušaitis, Ahuja and Morency, 2019). can affect trust in the healthcare setting (Greene and Bannister, 2021). These findings strengthen the importance of multimodal analysis to comprehend the impact of sentiment variation in all of them. communication channels especially on healthcare products where emotional tone could likely be. affect consumer confidence and buying behavior.

7 Conclusion and Future Work

7.1 Conclusion

The difference between sentiment and was the research question that was tested in this paper. feeling of a written Amazon review and verbal review of healthcare on YouTube. products, VADER, DistilRoBERTa, Whisper based on a multimodal NLP pipeline. transcription, and the prosodic analysis that Librosa has offered. The findings show that modality also plays a significant role in defining how the users describe sentiment. In line with previous researches that reported that text reviews are more typical of heterogeneous groups of The Amazon data, the respondent count and the number of experiences should be more (Mudambi and Schuff, 2010). moderate sentiment scores were found to have a high level of variation. Conversely, the You Tube review was characterized by much more positive responses, which is consistent with research indicating that prosodic signs make communication more emotional in

speaking (Scherer, 2019; Chen, Zhou and Miao, 2022). Segmentation on the product level further indicated that there was a variation amongst the individual items in the same video, which endorsed the assertions that multimodal content incorporates both the linguistic and affective cues that affect emotional meaning making (Poria, Hazarika and Cambria, 2017). The results are also related to the previous studies demonstrating that modality has a significant influence on persuasion and credibility perception in digital health scenarios and video formats tend to look more reliable or attractive (Lee and Watkins, 2016; Greene and Bannister, 2021).

In general, this study finds that verbal reviews have greater levels of emotionality and fidelity than written reviews, whereas text reviews yield more comprehensive and finer understanding of consumer experiences. All these findings indicate that multimodal sentiment analysis can provide a more profound and full picture of how customers compare healthcare products on various communication platforms.

7.2 Future Work

Future studies may build on this study by examining a bigger and more varied sample of YouTube reviewers to enhance the extent of generalisation, particularly because sentiment expression has been found to differ according to speaker personality and delivery style (Scherer, 2019). The topic modelling or sequence-labelling techniques would be a better fit to improve the product segmentation in order to minimize the reliance on the keyword matching. The use of visual elements like facial expressions would aid a more detailed tri-modal comparison, as it has been proven that video combines various persuasive messages (Lee and Watkins, 2016). Lastly, the development of a single multimodal fusion model may allow interpreting the sentiment more accurately through a text, audio, and visual communication, which will further enhance consumer understanding in analyzing healthcare products.

References

Baltrušaitis, T., Ahuja, C. and Morency, L. (2019) 'Multimodal machine learning: A survey and taxonomy', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), pp. 423–443.
https://doi.org/10.1109/TPAMI.2018.2798607

Betsch, C. *et al.* (2011) 'The influence of narrative vs. statistical health information on risk perception', *Health Psychology*, 30(3), pp. 282–289.
https://doi.org/10.1037/a0023521

Chen, D., Sin, S. and Theng, Y.L. (2015) 'Why do people share health misinformation on social media?', *Proceedings of iConference*, pp. 1–12.
https://hdl.handle.net/2142/73649

Chen, J., Zhou, T. and Miao, Y. (2022) 'Prosodic features and perceived credibility in online video reviews', *Information Processing & Management*, 59(6), p. 103097.
https://doi.org/10.1016/j.ipm.2022.103097

Chen, Y. and Xie, K. (2008) 'Online consumer review: Word-of-mouth as a new element in marketing communication mix', *Management Science*, 54(3), pp. 477–491. https://doi.org/10.1287/mnsc.1070.0810

Devlin, J. *et al.* (2019) 'BERT: Pre-training of deep bidirectional transformers for language understanding', *Proceedings of NAACL-HLT*, pp. 4171–4186. https://aclanthology.org/N19-1423/

Feng, B. and MacGeorge, E. (2010) 'The influence of message valence and modality on advice outcomes', *Human Communication Research*, 36(4), pp. 426–449. https://doi.org/10.1111/j.1468-2958.2010.01380.x

Google LLC (2023) YouTube Terms of Service. https://www.youtube.com/t/terms

Greene, J. and Bannister, A. (2021) 'Trust in digital healthcare: The influence of communicative cues and emotional expression', *Journal of Health Communication*, 26(2), pp. 124–136. https://doi.org/10.1080/10810730.2020.1821131

Guo, S., Kim, J. and Rubin, A. (2014) 'How video reviews influence consumers: Insights into trust and engagement', *Computers in Human Behavior*, 38, pp. 1–11. https://doi.org/10.1016/j.chb.2014.05.024

Hutto, C. and Gilbert, E. (2014) 'VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text', *ICWSM*, pp. 216–225. https://ojs.aaai.org/index.php/ICWSM/article/view/14550

Kumar, S. *et al.* (2021) 'Consumer sentiment towards healthcare products: Insights from text analytics', *Decision Support Systems*, 142, p. 113469. https://doi.org/10.1016/j.dss.2020.113469

Lee, J.E. and Watkins, B. (2016) 'YouTube vloggers' influence on consumer luxury brand perceptions', *Journal of Business Research*, 69(12), pp. 5753–5760. https://doi.org/10.1016/j.jbusres.2016.04.171

Liu, B. (2012) *Sentiment Analysis and Opinion Mining*. San Rafael: Morgan & Claypool Publishers. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Liu, Y. *et al.* (2019) 'RoBERTa: A Robustly optimized BERT pretraining approach', *arXiv preprint*, arXiv:1907.11692. https://arxiv.org/abs/1907.11692

McFee, B. *et al.* (2015) 'librosa: Audio and music signal analysis in Python', *Proceedings of the 14th Python in Science Conference*, pp. 18–25. https://doi.org/10.25080/Majora-7b98e3ed-003

Mudambi, S. and Schuff, D. (2010) 'What makes a helpful review? A study of customer reviews on Amazon.com', *MIS Quarterly*, 34(1), pp. 185–200. https://doi.org/10.2307/20721420

Paivio, A. (1991) *Dual Coding Theory*. Oxford: Oxford University Press. https://global.oup.com/academic/product/dual-coding-theory-9780195066661

Park, S. and Nicolau, J. (2015) 'Wellness tourism: Moderating effects of reviewer characteristics', *Tourism Management*, 48, pp. 113–126. https://doi.org/10.1016/j.tourman.2014.10.019

Poria, S., Hazarika, D. and Cambria, E. (2017) 'A review of affective computing: From machine learning to deep learning', *IEEE Transactions on Affective Computing*, 10(3), pp. 1–20. https://doi.org/10.1016/j.inffus. 2017.02.003

Radford, A. *et al.* (2023) 'Robust speech recognition via large-scale weak supervision', *arXiv preprint*, arXiv:2212.04356. https://arxiv.org/abs/2212.04356

Scherer, K.R. (2019) 'Vocal communication of emotion: A review of research paradigms', *Current Opinion in Psychology*, 17, pp. 6–12. https://doi.org/10.1016/j.copsyc.2019.02.004

Sun, C., Qiu, X. and Huang, X. (2020) 'Utilizing BERT for sentiment analysis in product review datasets', *Knowledge-Based Systems*, 205, p. 106194. https://doi.org/10.1016/j.knosys.2020.106194

Sundar, S.S. *et al.* (2021) 'The persuasion power of modality in digital media', *Journal of Media Psychology*, 33(2), pp. 89–103. https://doi.org/10.1027/1864-1105/a000288

Waszak, P., Kasprzycka-Waszak, W. and Kubanek, K. (2018) 'The spread of medical misinformation on social media', *Systems Science & Control Engineering*, 6(1), pp. 254–263. http://dx.doi.org/10.1016/j.hlpt.2018.03.002

Wirth, R. and Hipp, J. (2000) 'CRISP-DM: Towards a standard process model for data mining', *Proceedings of the 4th International Conference on the Practical Application of Knowledge Discovery and Data Mining*, pp. 29–39. https://doi.org/10.1016/j.procs.2021.01.199

Yoon, H. and Vargas, P. (2017) 'Video vs. text: The persuasion effects of modality in online product reviews', *Journal of Interactive Marketing*, 39, pp. 15–28. https://doi.org/10.1016/j.intmar.2017.01.001

YouTube (2023) Doctor Reviews Amazon "Health" Products. https://www.youtube.com/watch?v=XWMjT-ILFqc