My Title

Faculty of Electrical Engineering and Computer Science
Institute of Data Science
Leibniz University Hannover

Master Thesis

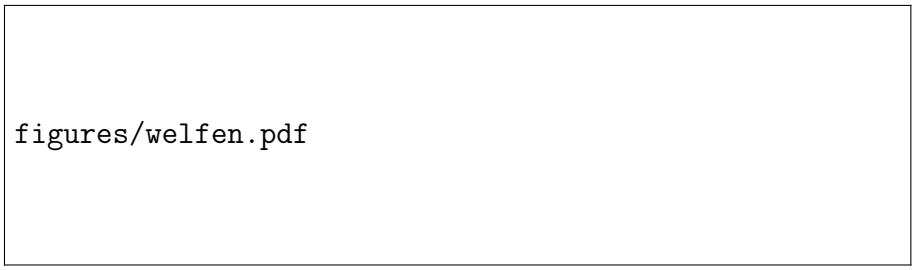
submitted for the degree of

Master of Science (M. Sc.)

by

Lea Rebecca Reinhart

Matriculation Number : 100



First Examiner: Prof. Dr. Ralph Ewerth

Second Examiner: Prof. Dr. Elias Examiner

Supervised By: M. Eng. Samuel Supervisor

January 1, 1970

Erklärung der Selbständigkeit

Hiermit versichere ich, dass ich diese Arbeit selbstständig verfasst habe, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt wurden, alle Stellen der Arbeit, die wörtlich oder sinngemäß aus anderen Quellen übernommen wurden, als solche kenntlich gemacht sind, und die Arbeit in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegt habe.

_____ Hannover, January 1, 1970

Lea Rebecca Reinhart

Abstract

Vivamus vehicula leo a justo. Quisque nec augue. Morbi mauris wisi, aliquet vitae, dignissim eget, sollicitudin molestie, ligula. In dictum enim sit amet risus. Curabitur vitae velit eu diam rhoncus hendrerit. Vivamus ut elit. Praesent mattis ipsum quis turpis. Curabitur rhoncus neque eu dui. Etiam vitae magna. Nam ullamcorper. Praesent interdum bibendum magna. Quisque auctor aliquam dolor. Morbi eu lorem et est porttitor fermentum. Nunc egestas arcu at tortor varius viverra. Fusce eu nulla ut nulla interdum consectetuer. Vestibulum gravida. Morbi mattis libero sed est.

Contents

List of Tables

List of Figures

1 Introduction

In today's interconnected world, individuals globally rely on news media to stay informed about both local and international events (Newman et al. [12]). However, news is more than just a source of information—it significantly shapes our understanding of the world (Happer and Philo [7]). This impact is particularly vital in democratic societies, where media is often deemed to be the unofficial fourth pillar of democracy because public beliefs can directly influence voting behavior (Badawy et al. [1]). Given this influence, news media has long been a political battleground (Miller and Krosnick [10]).

For consumers, it can be challenging to form balanced opinions in this battleground state of news (Newman et al. [11]). There is a growing trend of selective news avoidance; (Newman et al. [11]) reports that 39% of their sample at least occasionally avoid the news. One issue is the proliferation of fake news, which can continue influencing beliefs even after being debunked (Lewandowsky et al. [9]). Furthermore, regardless of their truthfulness, news stories often employ narrative strategies. While these strategies can facilitate engagement and learning (Dahlstrom [5] and Salvador and Cobos [14]), they can also influence the persuasiveness of messages (Braddock and Dillard [3]), making it harder to engage with news content objectively.

To combat dis- and misinformation in text-based media, various efforts such as fact-checking initiatives and techniques for detecting fake news through writing styles and narrative strategies have been researched (Hamby et al. [6]). [more examples/sources] However, as video increasingly becomes a dominant news medium, with 66% of surveyed individuals watching short-form news videos weekly and 51% engaging with longer formats (Newman et al. [11]), there is a relative scarcity of research on dis- and misinformation detection in this format. Current methods often depend solely on textual transcripts, missing the unique multimodal connections integral to video content.(SOURCE)

Specific temporal patterns, such as multimodal narrative strategies and filmic editing patterns, are distinctive to video content. An analysis of narrative patterns in news videos was conducted by John Bateman and Ciao-I Tseng in 2023, which revealed significant differences in pattern usage between state-owned and private media Bateman and Tseng [2]. However, this study relied on expert annotation. With the vast amount of news media published weekly, fueled even further by the emergence of AI-generated news, approaches reliant on recourse extensive human annotators, are merely a drop in the ocean of emerging news

content. To adequately address the current problems in the news landscape, automated approaches are vital Zellers et al. [15].

While AI can be utilized to generate news content, it can also be leveraged to analyze it. Early work in this field involved simple methods like hierarchical decision trees for narrative structure analysis (Phung et al. [13]) and hidden Markov models for classifying news types (Kolekar and Sengupta [8]). Contemporary research often focuses on complex domains like identifying speaker and situational contexts (Cheema et al. [4]). (find other projects with SOURCE). However, the automated analysis of temporal patterns such as narrative and filmic editing sequences remains largely underexplored.

This thesis aims to fill this research gap by evaluating various multimodal machine-learning approaches for their effectiveness in detecting temporal patterns in news videos. Additionally, we expand on existing knowledge by analyzing and exploring these patterns within our curated news video dataset.

Results

2 Related Work

This work falls into the area of digital humanities, an interdisciplinary field that bridges humanities and computer science. To establish the foundations of this research, we first examine key concepts from communication science, political science, and psychology, followed by an exploration of computational video analysis.

2.1 Media Background

News plays a pivotal role in shaping public discourse and informing society. Through agenda-setting, news organizations influence which topics dominate public discussion, determining societal priorities and framing public debate [(McCombs and Shaw, 1972)](add more recent sources). Beyond merely relaying information, news shapes perceptions of truth and societal beliefs about the topics it covers [(Happer and Philo, 2013); (Waisbord, 2018)].

Different formats of news like video, text, and audio—affect how individuals access and perceive information. Video formats, in particular, combine visual and auditory stimuli to create more engaging content, often enhancing emotional resonance and recall compared to textual formats [source]. However, this comes with risks; the emphasis on emotional engagement can sometimes distort factual precision and amplify biases. Recent trends underscore the rise of news videos as a primary source of information. According to the Reuters Digital News Reports (2021, 2024), two-thirds of surveyed individuals consume short-form news videos weekly, and over half engage with longer formats, emphasizing the growing influence of visual storytelling on audience perceptions and democratic discourse.

Despite its critical societal role, the news industry faces significant challenges, including a "post-truth" crisis characterized by the erosion of traditional journalistic standards and the rise of politically fragmented, agenda-driven news landscapes [(Waisbord, 2018)]. Often described as the unofficial fourth pillar of democracy, the integrity of news media is fundamental to informed public decision-making and democratic accountability. However, the prevalence of misinformation and disinformation poses substantial threats.

Misinformation refers to the unintentional spread of false or inaccurate information, such as reporting errors, while disinformation is deliberately fabricated information designed to mislead or advance specific agendas [(Fetzer, 2004)]. High-profile examples include Russian interference in the 2016 U.S. election, where disinformation campaigns sought to promote

narratives favoring Donald Trump [(Russian Influence)]. Addressing these issues is crucial, as the credibility of news media directly impacts public trust, social cohesion, and the integrity of democratic processes. [source]

The rise of misinformation and disinformation in news media presents profound challenges to societal trust and informed discourse. Psychological research highlights the persistence of false beliefs even after misinformation is corrected, a phenomenon known as the "continued influence effect." This effect is exacerbated by the emotionally engaging nature of video content, which can make misinformation more memorable and resistant to correction [(Lewandowsky et al., 2012)].

Compounding this issue is the growing trend of selective news avoidance, as reported by Newman et al. (2021, 2024). This behavior, where individuals deliberately disengage from news content, reflects a widening disconnect between media producers and consumers, reducing opportunities for audiences to access accurate and balanced information. These trends underscore the urgency of exploring how different aspects of news content, such as its temporal structure, impact audience engagement and trust. The basis for such works is being able to identify such patterns on a large scale.

Temporal Patterns

Filmic editing patterns are essential in shaping viewers' cognitive and emotional experiences, directly influencing their engagement with the narrative. Continuity editing, widely used in filmmaking, aligns with viewers' cognitive processes to segment and organize events, enhancing comprehension and memory. Techniques such as cuts and transitions guide the audience's understanding of spatial and temporal relationships, making complex narratives more accessible [(Magliano and Zacks, 2011)](https://doi.org/10.1111/j.1551-6709.2011.01202.x) [(Schwan et al., 2000)](https://doi.org/10.3758/BF03213801). Intensified editing styles, characterized by faster cuts and dynamic compositions, further amplify visual intensity and viewer immersion, offering more stimulating experiences [(Bordwell, 2002)](https://doi.org/10.1525/FQ.2002.55.3

Beyond cognitive organization, editing patterns influence emotional responses and narrative engagement. The Kuleshov Effect demonstrates how juxtaposed shots can manipulate emotional attributions by contextual framing, illustrating the psychological impact of editing [(Mobbs et al., 2006)](https://doi.org/10.1093/SCAN/NSL014). Similarly, narrative storytelling techniques, like character-driven plots and immersive structures, have been shown to enhance engagement, recall, and persuasion across various media contexts [(Braddock and Dillard, 2016)](https://doi.org/10.1017/CBO9781107589323).

The emotionally engaging nature of video formats, while beneficial for recall and engagement, also exacerbates challenges through misinformation. Narrative structures make false information more memorable, increasing its resistance to correction [(Lewandowsky et al., 2012)](https://doi.org/10.1080/10463283.2012.677998). By analyzing the temporal patterns in news videos, we can gain insights into how they shape audience perceptions, improve en-

gagement, and mitigate the spread of misinformation in an increasingly video-dominated news landscape.

Non-Automated Pattern Exploration

Narrative patterns have been identified as potential indicators of fake news, with research suggesting that such patterns are often more prevalent in fabricated content due to their focus on emotional engagement over factual accuracy [(Hamby et al.)]. This insight underlines the importance of studying narrative strategies in detecting misinformation and understanding its dissemination.

Research into Formalized Editing Patterns (FEPs) and narrative strategies in news media has traditionally relied on qualitative methods. Bateman and Tseng (2023) examined narrative strategies in news videos and uncovered significant differences between state-owned and privately-owned broadcasters. Their findings show how institutional goals—such as promoting neutrality versus prioritizing emotional engagement—manifest in editing styles and narrative choices. Similarly, Salvador and Cobos (2023) investigated narrative structures in educational videos, illustrating how these structures can enhance learning outcomes. Kolekar and Sengupta (2005) further contributed to the field by employing multimodal approaches to semantically index news video sequences, enabling better categorization and retrieval of video content.

While these studies provide valuable insights, the reliance on manual annotation and qualitative methods poses significant scalability challenges. Given the sheer volume of video content generated daily, this issue is exacerbated by the rise of AI-generated misinformation, which can rapidly produce sophisticated fake news [(Song et al., 2021)](https://consensus.app/papers/temp
evolving-graph-neural-network-for-fake-news-song-shu/75ef69f67dfd51288dcb77d737f0851f/?utm$_source$ =
$chatgpt$).$Recentadvancementsinvisualizingnarrativepatterns$, $suchasthe$"$time-sets$"$tooldevelopedbyK$
$NiemiandMasoodian$(2019), $offerpromisingsolutionsbyenablingtheanalysisofco-occurringvariablesi$
$NiemiandMasoodian$, 2019)](

Moreover, interdisciplinary approaches integrating psychological, linguistic, and computational methods highlight the potential for automated systems to address these limitations. For example, Zhou (2020) reviewed strategies for detecting fake news, emphasizing the need for scalable, explainable models that incorporate narrative and propagation patterns [(Zhou, 2020)](However, most of these systems are yet to fully incorporate temporal editing patterns, which play a crucial role in shaping audience engagement and perception. Bridging this gap represents a significant opportunity for advancing both theoretical understanding and practical applications in the fight against misinformation.

2.2 Multimodal Video Analysis

While qualitative methods have laid the groundwork for understanding FEPs and narrative strategies, they fall short in addressing the sheer scale of contemporary news production. Computational methods, leveraging advances in artificial intelligence and machine learning, offer a promising path forward.

2.2.1 General Video Analysis

High-level summary of past (CNN/RNN/GCNN) and nowadays a lot of transformer-based models (VideoMAE) Give some examples of popular video analysis tasks (action recognition, active speaker detection, ...) GAPS:

2.2.2 Videos Analysis using GenAI

... ...

2.2.3 News Video Analysis

More specific works that focos on multimodal analysis of news GAPS: typically aim to extract specific concepts but not temporal patterns

2.3 Temporal Pattern Detection

2.3.1 General Works

High-level summary GAPS

2.3.2 News Domain

More specific summary ... GAPS: although there are some works, most of them do not make use of state of the art models, they might not consider generalization issues given that data for such tasks is typically sparse, and do not focus on FEPs and narrative strategies

3 Foundations

Introduction

3.1 Communications Science for Video Analysis

Pick the 1-2 works that define the patterns you want to detect in the thesis

3.1.1 Filmic Editing Patterns

-definition -effect on viewers(might be in related work) -utilization in video content, specifically news videos

3.1.2 Narrative Patterns

-definitions, difficulties with definitions in videos - effect on consumers (might be in related work already) -utilization in video content, specifically news videos

3.2 Machine Learning

3.2.1 Random Forests

Random forests are a widely used machine learning technique that exemplifies the ensemble learning paradigm, where multiple models (in this case, decision trees) are combined to improve the overall predictive performance by reducing errors from individual models. By combining the predictions of multiple decision trees, random forests aim to improve overall predictive performance (Breiman, 2001). This method has proven effective for both classification and regression tasks and is valued for its robustness, versatility, and ability to handle high-dimensional data.

The construction of a random forest involves several key steps. First, a process called bootstrapping is employed, where random samples are drawn with replacement from the training dataset to create multiple subsets. Each subset serves as the basis for training an individual decision tree. To further enhance diversity among the trees, random forests incorporate feature randomization. At each decision node within a tree, only a randomly

selected subset of features is considered for splitting, rather than evaluating all available features. This deliberate introduction of randomness reduces the risk of overfitting and enhances the generalization capability of the ensemble. Once the decision trees are trained, their outputs are aggregated. For classification problems, the forest predicts the class based on majority voting, while for regression, the predictions of all trees are averaged (Ho, 1995).

An infographic illustrating the construction of a random forest—from bootstrapping to feature randomization and aggregation—would provide a clearer understanding of these steps. This visual could depict how individual trees are trained on different subsets of data and features, and how their outputs combine to form the final prediction.

The strengths of random forests make them a popular choice across various domains. Their robustness to overfitting is a standout feature, as the averaging of errors across trees mitigates the tendency of individual trees to overfit. This property also enabled random forests to outperform vision-language models (VLMs) in a news video generalization task, as demonstrated by the results reported in Section 2 (/cheemaetal).x20;

Additionally, random forests are scalable and can efficiently handle large datasets with numerous features. They are also inherently non-parametric, requiring no assumptions about the underlying data distribution, which contributes to their versatility. Another significant advantage is their ability to provide feature importance scores, offering insights into which variables most influence the predictions (Breiman, 2001).

However, random forests are not without limitations. The training process, involving the construction of multiple trees, can be computationally intensive, particularly for large datasets. Moreover, while feature importance scores enhance interpretability, the overall model is often regarded as a "black box" compared to simpler models like individual decision trees. There is also a potential trade-off between bias and variance due to the randomized selection of features, which may lead to underfitting in some cases.

Random forests have been applied in a wide range of fields, from fraud detection in finance (Breiman, 2001) and disease diagnosis in healthcare (e.g., prediction of diabetes) to text classification and sentiment analysis in natural language processing (Ho, 1995). In video analysis, Please write this part with citations.x20;

While random forests generally cannot handle time-series data directly, they can do so if combined with a sliding window and feature aggregation within the window. For instance, in predictive maintenance applications, sliding windows are used to segment sensor data into manageable intervals, and aggregated features like mean, variance, and frequency components are extracted for analysis (Zhang et al., 2019).x20;

In this thesis, random forests are utilized as a baseline method for detecting temporal patterns in news videos. By aggregating features extracted from video frames, such as color histograms and motion vectors, random forests provide an efficient mechanism to localize patterns. This foundational approach lays the groundwork for the deployment of more

sophisticated deep learning techniques, facilitating a systematic exploration of temporal patterns.

- Breiman, L. (2001). "Random Forests." *Machine Learning*, 45(1), 5–32. doi:10.1023/A:101093340432
- Ho, T. K. (1995). "Random Decision Forests." *Proceedings of the 3rd International Conference on Document Analysis and Recognition*.

3.2.2 Sequential Deep Learning Approaches

3.2.2.1 RNNs

Recurrent Neural Networks (RNNs) are a class of neural networks designed specifically for sequential data. Unlike traditional feedforward neural networks, RNNs incorporate a feedback loop that allows information from previous time steps to influence the current computation. This characteristic makes RNNs particularly suitable for tasks where temporal or sequential context is crucial, such as language modeling, speech recognition, and time-series prediction (Rumelhart et al., 1986).

The fundamental unit of an RNN is the recurrent layer, where a set of neurons processes sequential data step-by-step. At each time step, the input data is combined with the hidden state from the previous time step. This hidden state acts as a memory, capturing relevant information from past inputs. Mathematically, this can be expressed as:

$$h_t = \sigma(W_{hx}x_t + W_{hh}h_{t-1} + b_h)$$

Here, h_t represents the hidden state at time step t, x_t is the input, W_{hx} and W_{hh} are weight matrices, and b_h is the bias vector. The activation function σ, typically a non-linear function like tanh or ReLU, introduces non-linearity into the computation.

While this feedback mechanism enables RNNs to process sequential information effectively, it also introduces challenges such as the vanishing and exploding gradient problems. These issues arise during backpropagation when gradients are either excessively diminished or amplified, making it difficult to train the network on long sequences (Hochreiter, 1991).

Strengths and Limitations

RNNs offer several advantages in sequential data processing. Their ability to share weights across time steps reduces the number of parameters, making them computationally efficient compared to models that treat sequential data as independent instances. Furthermore, their inherent design allows them to model temporal dependencies, which are essential for understanding patterns in sequential tasks.

However, RNNs also exhibit notable limitations. The vanishing gradient problem significantly hampers their ability to learn long-term dependencies, limiting their effectiveness on datasets with extended temporal patterns. Moreover, RNNs can be computationally

expensive when applied to large datasets, as their sequential nature prevents parallelization during training.

In the domain of video analysis, RNNs have been employed to capture temporal relationships between frames, such as action recognition and event detection. By processing sequences of extracted features from video frames, RNNs can identify patterns that evolve over time. For instance, an RNN might analyze the sequence of motion vectors and keyframe features to detect transitions between different activities in a news broadcast. Although more advanced architectures like LSTMs and Transformers have largely superseded RNNs in recent years, they remain foundational to the study of sequential modeling.

In this thesis, RNNs are explored as a baseline method for detecting temporal patterns in news videos. Their relatively simple architecture provides a benchmark against which the performance of more sophisticated models, such as LSTMs and Transformers, can be compared.

- Rumelhart, D. E., Hinton, G. E., Williams, R. J. (1986). "Learning representations by back-propagating errors." *Nature*, 323(6088), 533–536. doi:10.1038/323533a0 - Hochreiter, S. (1991). "Untersuchungen zu dynamischen neuronalen Netzen." *Diploma, Technische Universität München*.

3.2.2.2 LSTMs

Long Short-Term Memory Networks (LSTMs)

Long Short-Term Memory (LSTM) networks are an extension of RNNs designed to overcome the challenges associated with learning long-term dependencies. By incorporating a sophisticated memory cell structure, LSTMs can selectively retain or discard information as needed, making them well-suited for tasks involving long sequential dependencies (Hochreiter Schmidhuber, 1997).

Architecture and Mechanisms

The core component of an LSTM is its memory cell, which is regulated by three gating mechanisms: the input gate, the forget gate, and the output gate. These gates control the flow of information into, within, and out of the memory cell, respectively:

Input Gate: Determines how much of the new input should be stored in the memory cell.

Forget Gate: Decides which information from the memory cell should be discarded.

Output Gate: Regulates the information passed from the memory cell to the next layer or time step.

Mathematically, these mechanisms can be expressed as:

Here, , , and are the forget, input, and output gates, respectively, is the memory cell state, and is the hidden state at time step . , , and represent weights and biases associated with the respective gates.

Advantages over RNNs

The gating mechanisms in LSTMs effectively address the vanishing gradient problem, allowing the network to capture long-term dependencies in sequential data. This capability has made LSTMs the architecture of choice for a wide range of tasks, including machine translation, speech recognition, and video analysis.

Applications in Video Analysis

In video analysis, LSTMs are often used to model temporal dependencies across frames. For instance, they have been employed to predict scene transitions, identify action sequences, and even detect subtle patterns in news video narratives. By leveraging their ability to remember context over extended periods, LSTMs excel in tasks where temporal coherence is essential.

In this thesis, LSTMs serve as a more advanced sequential modeling approach compared to RNNs. Their ability to manage long-term dependencies enables a deeper exploration of temporal patterns in news videos, providing insights that are critical for understanding complex narrative and filmic editing structures.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). "Learning representations by back-propagating errors." Nature, 323(6088), 533–536. doi:10.1038/323533a0

Hochreiter, S. (1991). "Untersuchungen zu dynamischen neuronalen Netzen." Diploma, Technische Universität München.

Hochreiter, S., & Schmidhuber, J. (1997). "Long short-term memory." Neural computation, 9(8), 1735-1780. doi:10.1162/neco.1997.9.8.1735

3.2.2.3 Transformer

From LSTMs to Transformers

Despite the advancements offered by LSTMs in addressing the limitations of RNNs, these architectures still struggle with certain challenges, such as inefficiency in parallelization and difficulties in capturing very long-range dependencies. Transformers, introduced by Vaswani et al. in 2017, represent a paradigm shift in sequential modeling by replacing recurrence with a self-attention mechanism. This innovation allows Transformers to process sequences in parallel and effectively model relationships across arbitrary distances within the data.

Transformers

Transformers are a neural network architecture that relies on self-attention mechanisms to model dependencies between all elements in a sequence simultaneously. Unlike RNNs and

LSTMs, which process data sequentially, Transformers analyze the entire sequence at once, making them highly efficient and capable of capturing long-range dependencies (Vaswani et al., 2017).

Architecture and Mechanisms

The core component of the Transformer is the self-attention mechanism, which computes a weighted representation of the entire input sequence for each element. This process is facilitated by three matrices: the Query (Q), Key (K), and Value (V). The self-attention operation can be expressed as:

Here, , , and are projections of the input sequence, and is the dimensionality of the Key vectors. The softmax function ensures that the attention weights sum to 1, enabling the model to focus on the most relevant parts of the sequence.

Transformers also employ positional encoding to account for the order of elements in the sequence, as they lack the inherent sequential structure of RNNs and LSTMs. Positional encoding adds unique embeddings to each input element based on its position, allowing the model to distinguish between different temporal locations.

Advantages over RNNs and LSTMs

Transformers overcome the limitations of RNNs and LSTMs in several ways. By processing sequences in parallel, they achieve significant improvements in training efficiency. Their self-attention mechanism enables them to capture global dependencies within the data, which is particularly beneficial for tasks requiring an understanding of long-range relationships. Additionally, Transformers are highly scalable and have become the foundation for state-of-the-art models in various domains, such as natural language processing and computer vision.

Applications in Video Analysis

In video analysis, Transformers have been used to model temporal patterns by treating sequences of frames or extracted features as input tokens. Their ability to capture complex dependencies across frames makes them ideal for tasks such as action recognition, scene segmentation, and multimodal analysis. For example, Vision Transformers (ViTs) and Temporal Attention models extend the Transformer architecture to process spatial and temporal information simultaneously.

In this thesis, Transformers represent the most advanced approach for analyzing temporal patterns in news videos. By leveraging their self-attention mechanism, Transformers provide a powerful framework for identifying and understanding intricate narrative and editing structures in video data.

References

Rumelhart, D. E., Hinton, G. E., Williams, R. J. (1986). "Learning representations by back-propagating errors." Nature, 323(6088), 533–536. doi:10.1038/323533a0

Hochreiter, S. (1991). "Untersuchungen zu dynamischen neuronalen Netzen." Diploma, Technische Universität München.

Hochreiter, S., Schmidhuber, J. (1997). "Long short-term memory." Neural computation, 9(8), 1735-1780. doi:10.1162/neco.1997.9.8.1735

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). "Attention is all you need." Advances in Neural Information Processing Systems, 30, 5998-6008.

3.2.3 GenAI

Vision-Language Models (VLMs): Building on Transformers

Vision-Language Models (VLMs) extend the Transformer architecture to handle multimodal data, integrating visual and textual information into a unified framework. By leveraging the self-attention mechanism and positional encodings of Transformers, VLMs can align visual features, such as object embeddings, with textual representations, such as captions or transcripts. This alignment enables VLMs to perform tasks that require a deep understanding of both modalities, including image captioning, visual question answering, and video understanding (Radford et al., 2021).

Architecture and Mechanisms

At the core of VLMs is a dual-encoder or unified architecture that processes visual and textual inputs. In the dual-encoder approach, separate encoders (e.g., Vision Transformers for images and Transformers for text) independently process each modality. Their representations are then aligned using similarity-based objectives, such as contrastive learning. In a unified architecture, a single Transformer model processes concatenated multimodal inputs, enabling joint attention across both modalities.

An example of a VLM is CLIP (Contrastive Language-Image Pretraining), which uses a dual-encoder architecture to learn visual and textual representations in a shared space. This approach allows for zero-shot classification and robust cross-modal retrieval capabilities (Radford et al., 2021).

Applications in Video Analysis

In video analysis, VLMs have become a cornerstone for tasks requiring multimodal integration. By aligning frame-level features with accompanying text, such as subtitles or metadata, VLMs enable tasks like temporal action localization, scene classification, and narrative structure analysis. For example, in news video analysis, VLMs can connect visual editing patterns with spoken or written narratives, providing insights into how content is structured and delivered.

In this thesis, VLMs are leveraged to explore the interplay between filmic editing patterns and narrative strategies in news videos. Their ability to integrate multimodal data offers

a powerful tool for detecting and analyzing temporal patterns in complex video datasets. (building on transformer)

4 Methodology

4.1 Data

4.1.1 Dataset Description

The dataset used in this study consists of annotated news videos from five German news channels. The different channels have been selected to cover a variety of editorial and journalistic practices:

- **BildTV**: A privately-owned news channel operated by Axel Springer SE, BildTV is an extension of the Bild tabloid, known for more sensationalist and emotionally engaging content. Germany's most successful commercial news? (SOURCE)
- **COMPACTTV**: Another privately-owned channel, COMPACTTV is known for its right-wing viewpoints. They were banned for being extremist in July 2024 but the ban was lifted in August 2024.
- **Tagesschau**: Germany's most widely watched news format, Tagesschau is produced by ARD, on of Germany's publicly financed channels.
- **HeuteJournal**: Produced by ZDF, Germany's second major public broadcaster. Longer format(?) (SOURCES and more info for channels, best would be views 2024 or for time where videos were produced)
- **Welt**: Description todo

The dataset's composition reflects these diverse reporting styles, enabling robust analysis of temporal and narrative patterns in public and private broadcasting contexts. All videos are in German, ensuring linguistic consistency, which is helpful for downstream tasks such as speech-to-text processing. Furthermore, since all channels cover similar topics (e.g., politics, society, and current events), comparisons across sources focus on their narrative and editorial techniques rather than content disparities.

This diverse dataset also allows for testing the generalization capabilities of machine learning models by training on three channels and evaluating performance on the fourth. Such an evaluation provides insight into the transferability of learned patterns across editorial and stylistic differences.

4.1.2 Annotation Process

Annotations in this dataset were conducted at the level of individual news stories, rather than entire broadcasts. This decision was caused by two key considerations:

1. **Boundary Independence**: Narrative patterns and filmic editing patterns rarely span multiple stories within the same broadcast. Segmenting the data into discrete stories ensures that each annotation captures a cohesive and self-contained unit of analysis.
2. **Machine Learning Preprocessing**: Story-level segmentation simplifies the preprocessing pipeline for machine learning models. This approach eliminates transitions or overlaps between stories, reducing noise and enabling the models to focus on well-defined temporal and narrative structures.

Annotations were carried out by communication scientists at the University of Bremen using the ELAN annotation format. ELAN allowed precise alignment of annotations with temporal segments, ensuring that patterns were tied directly to specific shots or timestamps(source to elan)

Both filmic editing patterns (FEPs) and narrative patterns were annotated using a standardized set of terms, covering concepts like shot types (e.g., "reverse-shot") and narrative strategies (e.g., "fragmentation" or "emotionalization"). The annotations were carried out at a fine-grained level, with exact timestamps marking the beginning and end of each occurrence.

Due to resource constraints, videos were annotated by a single annotator. While this streamlined the process, the lack of inter-annotator reliability checks is acknowledged as a potential limitation.

4.1.3 Dataset Statistics

The dataset contains **99 annotated stories**, totaling **X hours** of annotated video content. Each story has an average length of approximately **X minutes**, providing a manageable unit for analysis. The distribution of stories across the four channels is visualized in 4.1:

- **Tagesschau**: 53 stories
- **COMPACTTV**: 16 stories
- **BildTV**: 6 stories
- **HeuteJournal**: 19 stories
- **Welt**: 5 stories

Since public broadcasters (Tagesschau and HeuteJournal) tend to use fewer patterns, more stories from these sources were included to balance the overall number of pattern occurrences.

Pattern Occurrences

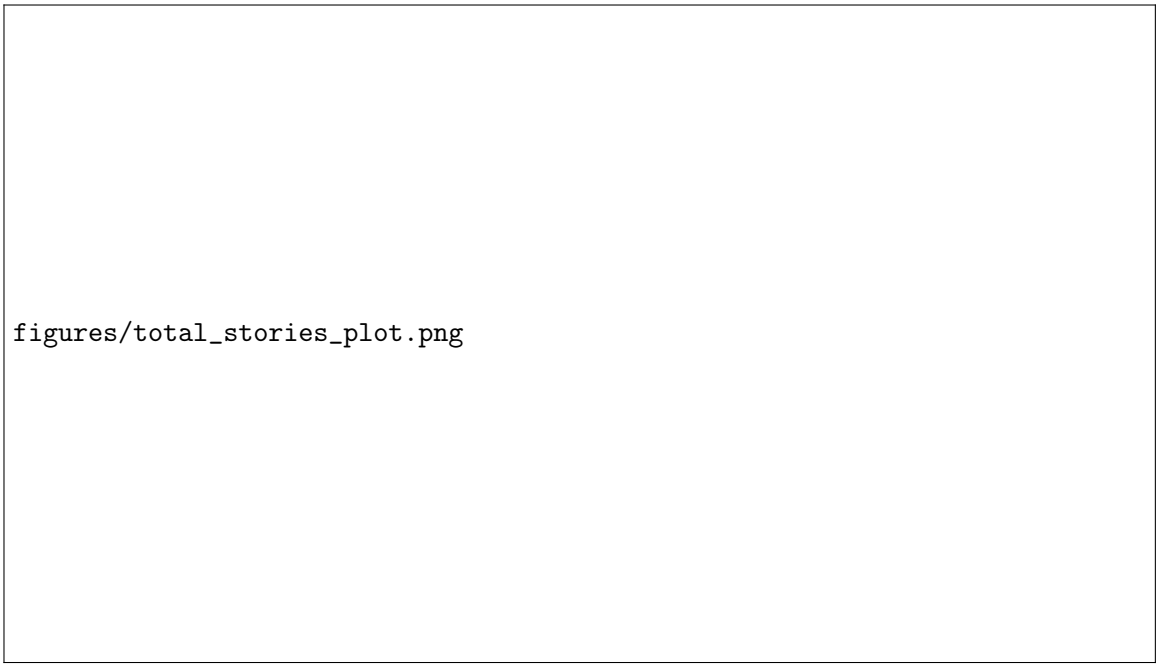


Figure 4.1: Distribution of news stories by channel

Distribution by Source The usage of patterns—both for Filmic Editing Patterns (FEPs) and Narrative Patterns—varies significantly between news sources. These distributions are shown in 4.2 and 4.3

- **FEPs by Source**: COMPACTTV demonstrates the highest use of FEPs, with **52 occurrences**, emphasizing its reliance on dynamic editing techniques. In contrast, Tagesschau employs **41 occurrences**, reflecting a more restrained editorial approach.
- **Narrative Patterns by Source**: Public channels like Tagesschau and HeuteJournal prioritize clarity-oriented patterns like "fragmentation," while private channels favor more emotionally charged strategies, such as "emotionalization-v2."

Overall Distribution Among the FEPs, the most frequent is **"fep:shot-reverse-shot"**, occurring **46 times**, followed by **"fep:intensify-v3"** (**28 occurrences**) and **"fep:alternating-shota"** (**25 occurrences**). Rare patterns like **"fep:fragmentation"** and **"fep:cut-away-v2"** occur only once each, highlighting significant class imbalance. These distributions are shown in **Figure 4**.

Similarly, narrative patterns like **"strategy:fragmentation-splitscreen"** (**17 occurrences**) dominate, while others, such as **"strategy:dramatization"**, appear only once.

Class Imbalance The dataset exhibits notable class imbalances. For instance:

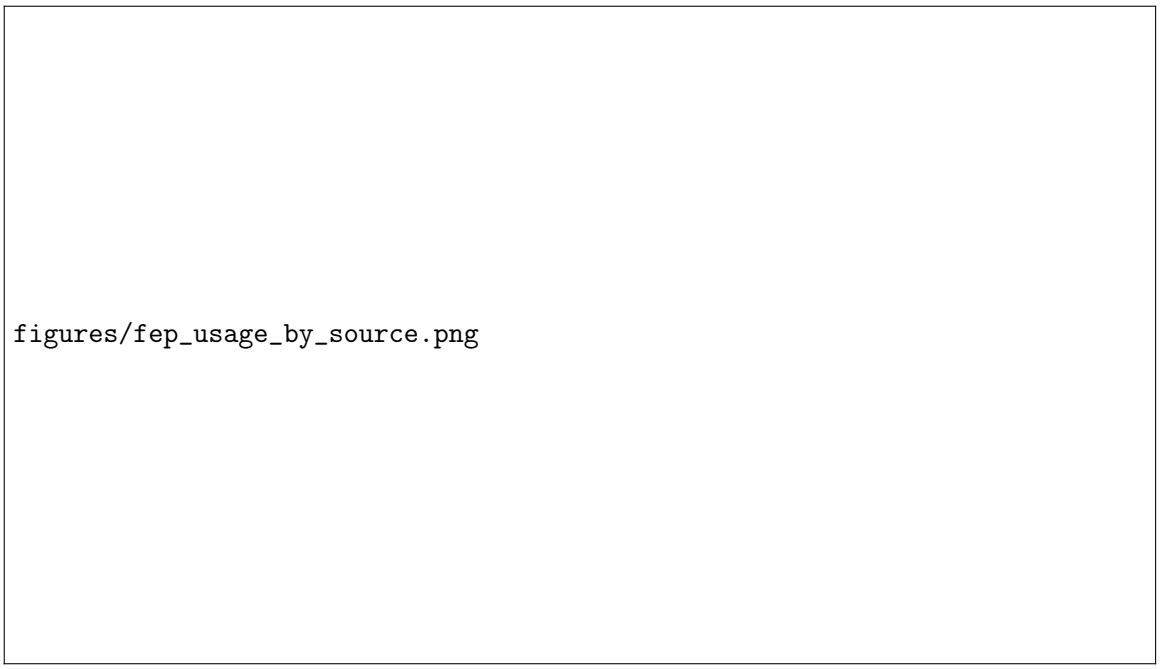


Figure 4.2: Filmic Editing Patterns by News source

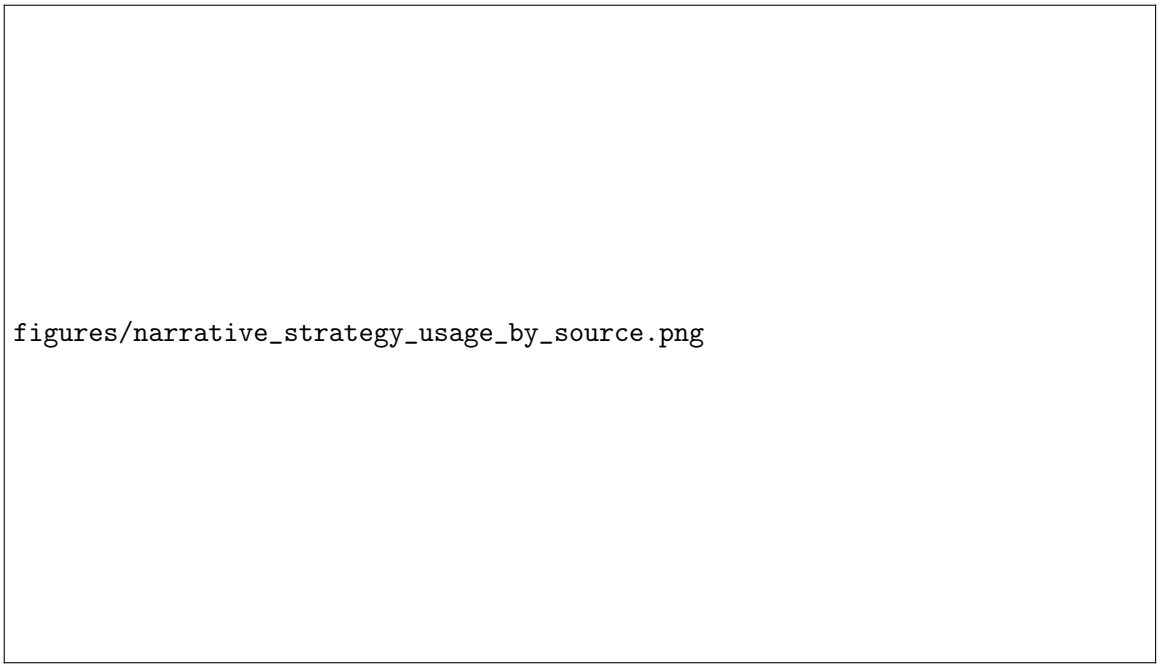


Figure 4.3: Enter Caption

- The most common FEP, **"fep:shot-reverse-shot"**, is 46 times more frequent than rare patterns like **"fep:fragmentation"**.
- Narrative patterns like **"strategy:fragmentation-splitscreen"** dominate over rare patterns such as **"strategy:dramatization."**

Addressing this imbalance during model training will require techniques like oversampling minority classes or employing weighted loss functions.

Preparation of Shot Data

1. **Data Preprocessing**
   - Conversion of raw video data into shots (explain the segmentation process briefly).
   - Overview of feature extraction methods:
     - Visual features: e.g., color histograms, motion vectors, keyframes.
     - Textual features: subtitles, speech-to-text.
     - Temporal features: shot lengths, transitions.
   - Mention if/how these features are derived from pretrained models or handcrafted algorithms.
2. **Normalization** How features are standardized and converted into a unified feature vector.

4.2 Machine Learning Approaches

4.2.1 Baseline: RF + Sliding Window

- **Overview**: Rationale for using a random forest (RF) classifier and sliding window for pattern detection.
- **Steps**:
  - Aggregate 2-3 consecutive shots into instances (include stride length).
  - Train an RF classifier:
    - Discuss architecture (number of trees, depth, etc.).
    - Describe the features used for classification.
  - Localize patterns using predicted probabilities per instance.
  - Highlight that the sliding window is non-trainable; it acts as a post-classification method.
- **Evaluation**: Discuss its pros and cons as a baseline, such as simplicity and interpretability.

4.2.2 Transformer-Based Approach

- **Overview**: Motivation for using transformers: ability to model long-term temporal dependencies.
- **Steps**:
  - Input: Use all instances of a video as a sequence (or batch them if required).
  - Transformer architecture: Specify details: attention mechanism, layers, positional encoding, etc.
  - Direct pattern detection: Unlike sliding window, patterns are detected end-to-end.
- **Optional Pretraining**:
  - Mention if pretraining on a similar task is considered.
  - Advantages of pretraining for temporal understanding.

4.2.3 Vision-Language Model (VLM)-Based Approach

- **Overview**: Leveraging multimodal inputs (images + text + feature vector) for pattern detection.
- **Steps**:
  - Input: VLM processes each shot (e.g., keyframe as image, corresponding text, and/or feature vector).
  - Prediction: Classify patterns at the shot level.
  - Aggregation: Combine predictions using a method similar to the sliding window.
- **Advantages**: Discuss the potential of VLMs in integrating diverse features.

4.3 2.3. Additional Pattern Mining

- Mention any postprocessing techniques for refining pattern detection.
- Example: clustering detected patterns for summarization or creating high-level abstractions.

4.4 Evaluation Framework

- If this fits in the chapter, outline how you will evaluate the methods:
  - Metrics: precision, recall, F1-score, IoU for pattern localization.
  - Baselines for comparison.

- Dataset splits: training, validation, testing.

4.5 Summary

- Recap of the key components of each approach.
- Highlight innovative aspects and challenges.

5 Experiments

6 Conclusion

Bibliography

[1] Adam Badawy, Emilio Ferrara, and Kristina Lerman. "Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign". In: *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (2018), pp. 258–265. DOI: 10.1109/ASONAM.2018.8508646.

[2] John A Bateman and Chiao-I Tseng. "Multimodal discourse analysis as a method for revealing narrative strategies in news videos". In: *Multimodal Communication* 12.3 (2023), pp. 261–285.

[3] Kurt Braddock and James Price Dillard. "Meta-analytic evidence for the persuasive effect of narratives on beliefs, attitudes, intentions, and behaviors". In: *Communication monographs* 83.4 (2016), pp. 446–467.

[4] Gullal S Cheema, Judi Arafat, Chiao-I Tseng, John A Bateman, Ralph Ewerth, and Eric Müller-Budack. "Identification of Speaker Roles and Situation Types in News Videos". In: *Proceedings of the 2024 International Conference on Multimedia Retrieval*. 2024, pp. 506–514.

[5] Michael F Dahlstrom. "Using narratives and storytelling to communicate science with nonexpert audiences". In: *Proceedings of the national academy of sciences* 111.supplement_4 (2014), pp. 13614–13620.

[6] Anne Hamby, Hongmin Kim, and Francesca Spezzano. "Sensational stories: The role of narrative characteristics in distinguishing real and fake news and predicting their spread". In: *Journal of Business Research* 170 (2024), p. 114289.

[7] C. Happer and Greg Philo. "The Role of the Media in the Construction of Public Belief and Social Change". In: *Journal of Social and Political Psychology* 1 (2013), pp. 321–336. DOI: 10.5964/JSPP.V1I1.96.

[8] Maheshkumar H Kolekar and S Sengupta. "Semantic indexing of news video sequences: a multimodal hierarchical approach based on hidden markov model". In: *TENCON 2005-2005 IEEE Region 10 Conference*. IEEE. 2005, pp. 1–6.

[9] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. "Misinformation and its correction: Continued influence and successful debiasing". In: *Psychological science in the public interest* 13.3 (2012), pp. 106–131.

[10] Joanne M. Miller and J. Krosnick. "News Media Impact on the Ingredients of Presidential Evaluations: Politically Knowledgeable Citizens Are Guided by a Trusted Source". In: *American Journal of Political Science* 44 (2000), pp. 301–315. DOI: 10.2307/2669312.

[11] Nic Newman, Richard Fletcher, Craig T Robertson, Amy Ross Arguedas, and Rasmus Kleis Nielsen. "Reuters Institute digital news report 2024". In: *Reuters Institute for the study of Journalism* (2024).

[12] Nic Newman, Richard Fletcher, Anne Schulz, Simge Andi, Craig T Robertson, and Rasmus Kleis Nielsen. "Reuters Institute digital news report 2021". In: *Reuters Institute for the study of Journalism* (2021).

[13] Quoc-Dinh Phung, Chitra Dorai, and Svetha Venkatesh. "Narrative structure analysis with education and training videos for e-learning". In: (2002).

[14] Patricia Salvador and Miguel Cobos. "Narrative as a Key Element for Learning Through Videos". In: *2023 IEEE Seventh Ecuador Technical Chapters Meeting (ECTM)*. IEEE. 2023, pp. 1–5.

[15] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. "Defending against neural fake news". In: *Advances in neural information processing systems* 32 (2019).

A Appendix Chapter 1

A.1 Example Section for Appendix