Emotion-aware Multimodal Meme Retrieval

Bachelorarbeit

Bachelor Angewandte Informatik

Aleksandr Litvin

May 16, 2025

Supervisor/Betreuer:

Christopher Bagdon

Reviewer/Prüfer:

Prof. Dr. Roman Klinger

Lehrstuhl für Grundlagen der Sprachverarbeitung
Fakultät Wirtschaftsinformatik und Angewandte Informatik
Otto-Friedrich-Universität Bamberg

Abstract

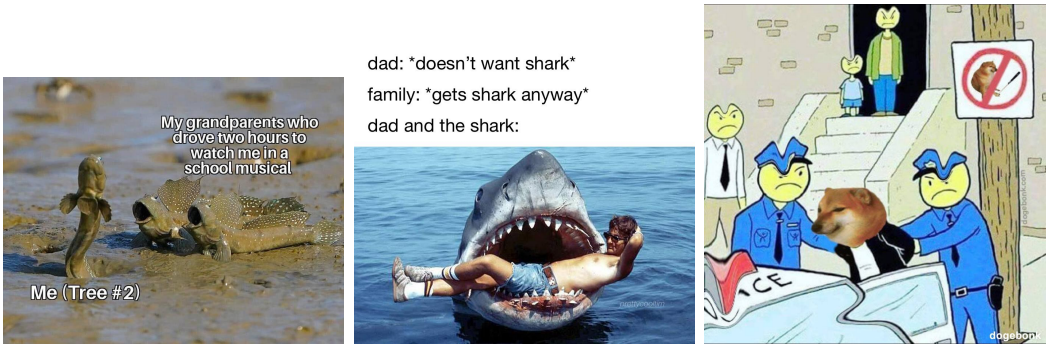
As an important part of modern communication, memes are socially relevant and allow a better understanding of social processes. However, existing research focuses primarily on the binary classification task without considering more general approaches. Work dedicated to emotion analysis considers memes only in a specific context or moves away from the task of classification towards a more specialized extraction of metaphors. Also, research on retrieval and recommendation systems often does not consider memes, preferring more general approaches. This thesis aims to fill the research gaps in these areas. With a focus on the social sciences, we propose a framework that does not require human resources for the labeling task and allows each meme to be processed as a graphical image with text. We investigate the influence of the emotion component in an iterative multimodal retrieval system in terms of user satisfaction by comparing two approaches: a multimodal approach (considering textual and visual information) and a multimodal approach with an emotion component (considering textual, visual and emotion components). We obtain features with the Transformer-based models, CLIP for visual, textual and emotion features from images and RoBERTa for emotion features from text. Furthermore, we cluster the obtained features using the Affinity Propagation algorithm. We survey 40 participants of varying ages and genders after their interaction with our iterative recommendation system system. Analysis of eleven quantitative metrics in the categories of user satisfaction, overall experience, quality of recommendations, system responsiveness, and interaction costs does not indicate significant differences in user satisfaction between settings with and without an emotional component. The developed framework can serve as a platform for further computer science and social research. It can be extended to cover more modalities (e.g. audio), adjust the system's parameters and configuration (e.g. using a predefined number of clusters), and be implemented in more natural setups (e.g. social media).

Contents

1 Introduction

Along with the entry of the Internet into everyday life, meme culture has emerged (Shifman, 2013). Although memes do not necessarily require digital tools for their creation and the first instances could be found before the invention of the Internet and even the personal computer (Wagener, 2024), it was only with the advent and spread of social media that the meme culture began to truly evolve (Shifman, 2013). Both social media and Internet culture have changed and continue to change and under their influence, so do the people who create that culture (Shifman, 2013) (Wagener, 2024). Despite their reputation as "funny pictures", memes can be a useful source for understanding society as a living system, its moods, opinions, and feelings (Wagener, 2024)(Shifman, 2013)(Milosavljević, 2020). Memes not only reflect individual opinions but also capture collective moods, attitudes, and reactions to societal events. Given their viral nature and rapid evolution, memes are crucial in understanding public opinion and the collective consciousness (Shifman, 2013).

Despite the social importance of memes, there is a lack of research on memes. Existing studies tend to focus on binary classification tasks, such as identifying whether a meme is hateful or non-hateful (Afridi et al., 2021). This narrow focus overlooks the broader potential of memes to inform us about society as a living system. Limiting analysis to simple categories fails to capture the layers of humor, irony or satire embedded in memes. Memes function as a nuanced "visual language" with cultural and emotion components that reflect society's diversity and complexity (Shifman, 2013). By categorizing them in a binary way, we lose the opportunity to explore their deeper meaning and limit our insights into how people use digital humor to communicate beliefs, foster connections and share feelings about current events. In



(a) A meme containing text (b) A meme with text as description (c) A meme without textual component

Figure 1: Examples of meme formats we consider

this thesis we address the gap by moving beyond binary classifications and examine the influence of emotion component. We only consider memes in the form of a graphic image, most often containing text. However, memes that rely solely on visual metaphor (without text) can also be found in the dataset we use. Examples of the memes we are examining can be found in Figure 1. We aim to investigate how

the effect of emotion labels for memes by creating a meme retrieval system based on user selection and comparing a multimodal approach and a multimodal approach with an emotion component based on a user experiment. Such a system can change the way researchers and social scientists approach understanding social trends and digital culture. By creating and analyzing the model that uses a multimodal approach combined with an emotion component, we provide a platform for a more sophisticated understanding of meme content that is more suited to its role in modern digital society and aims to close the existing research gap in this area. Although memes are firmly embedded in modern culture as markers of social attitudes, there is still room for research we aim to address. Hence, the research question of this thesis: is there a significant difference in user satisfaction while using multimodal retrieval system and while using multimodal retrieval system with emotion component?

Author's note: As a non-native English speaker, I used DeepL for grammatical and stylistic refinement in this thesis.

2 Background

This section covers the basic concepts and frameworks that are used in this thesis that are essential to understanding the Related work section. Here we focus mainly on the technical part of building a multimodal retrieval system. Firstly, we introduce retrieval and recommendation systems, then discuss multimodal machine learning, and finally the topic of emotion analysis.

2.1 Recommendation systems

To answer our research question we am to build a retrieval system. Recommendation and retrieval systems are systems and algorithms designed to help users find information or content in large data sets or databases based on certain preferences, patterns, or context (Li et al., 2024). While retrieval systems focus on extracting relevant data from large collections based on a query (e.g., Google search), recommendation systems go further by suggesting content that a user may not have explicitly searched for but may find interesting based on their previous interactions (Li et al., 2024) (e.g. online shop). Since the system we examine contains features of both retrieval and recommendation systems, we can speak of a content-based recommendation system with retrieval characteristics or recommendation system driven by a retrieval process.

2.2 Multimodal Machine Learning

Tadas Baltrušaitis et al. (2019) explain that multimodal machine learning aims to create models that can process and relate information from multiple modalities, where modality refers to the way something happens or is experienced (e.g. we can hear a dog and see it barking). To do so we firstly need to extract information from these modalities or perform multimodal feature extraction.

2.2.1 Transformer architecture

The Transformer architecture, introduced by Ashish Vaswani et al. (2017), is state-of-the-art in many applications at the time of writing this thesis (Sajun et al., 2024). It relies entirely on a self-attention mechanism and process input sequences in parallel, leading to faster training (Vaswani et al., 2017). A Transformer-based model weighs the importance of different parts of the input sequence when making predictions, which enables it to capture long-range dependencies in the data more effectively than more traditional Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN). Despite its limitations, such as computational complexity and hallucinations, (Sajun et al., 2024) (Gao et al., 2021) (Vaswani et al., 2017) (Peng et al., 2024) the Transformer-based models outperform more traditional CNN and RNN in several task (Sajun et al., 2024), enable more fast computations

(Vaswani et al., 2017) and can be adapted to different modalities. During their evolution, Transformers have been applied to the tasks like image recognition (e.g., Vision Transformer (ViT) (Dosovitskiy et al., 2020)), speech processing (Dong et al., 2018), and multimodal learning (Radford et al., 2021). Such flexibility is provided by their ability to model relationships between any parts of the input, regardless of modality which makes Transformers a powerful tool for complex feature extraction and alignment tasks such as multimodal analysis we do in this thesis.

2.2.2 Multimodal feature extraction

Feature extraction is one of the first steps in using machine learning. This is the process of transforming raw data (e.g., text) into a set of representative features that can be used by machine learning algorithms (LeCun et al., 2015). That helps portray the underlying patterns or characteristics required for a specific task (LeCun et al., 2015).

Multimodal feature extraction is the process of extracting meaningful representations from each of multiple modalities (e.g. text and image combined together), so that the essential information from each modality is preserved (Baltrušaitis et al., 2019). However, since each modality can represent information encoded in a particular way, it is important to combine these features into a common or joint embedding space (Tsai et al., 2019). Such feature alignment aims to ensure that features from different modalities correspond to the same concepts or objects, which allows the model to reveal relationships across modalities (Tsai et al., 2019). In addition, Principal Component Analysis (PCA) (Jolliffe and Cadima, 2016) is a general approach to reduce the dimensionality of the embeddings (Hinton and Salakhutdinov, 2006). Reducing the number of dimensions simplifies the model and reduces computation time, helping to prevent overfitting(Ajibade et al., 2024).

2.2.3 Multimodal analysis

After extracting features, we can use a Transformer-based model to perform a task, which in our case is multimodal analysis. As proposed by Limor Shifman (2013), memes consist of content (the specific ideas, jokes, or messages conveyed), form (the physical and structural form like image, video, etc.), position (the way the meme positions itself, often reflecting humor, satire, or criticism) (Shifman, 2013). A multimodal approach when working with memes is therefore necessary. Both textual and visual data need to be analyzed together to fully understand the content and context being transmitted. Multimodal analysis is integration and interpretation of different forms of data (such as text, images, audio or video) simultaneously, while the unimodal approach considers only one modality. Due to the nature of memes, the multimodal approach is the one considered in this thesis, on the assumption that it is more accurate than the unimodal approach (e.g. only visual analysis) (Kiela et al., 2020). However, it is also worth remembering that memes often use

emotional resonance to convey cultural, ideological, or informational content (Shifman, 2013)(Wagener, 2024). Focusing on objective features such as the image (what objects or subjects are displayed in the image) and the text (the text in the meme) may miss the essence of the humor contained in the meme. In contrast, extracting emotions and using them as an additional label (which we can define as emotion component) may help to retain the necessary information and, as a result, classify memes more accurately based on their similarities. In order to do this, we need to perform an emotion analysis.

2.3 Emotion analysis

Emotion analysis natural language processing (NLP) is the process of identifying and categorizing emotions expressed in data (Pang and Lee, 2008). It focuses on identifying and interpreting the emotional tone conveyed by text, voice or images (Huang et al., 2019). This is particularly important for social media content, where emotions can be central to the message (Kiela et al., 2020). Emotion analysis takes a deeper look at the specific emotional states, such as happiness, anger, sadness or fear, embedded in the content (Huang et al., 2019).

We can emphasize three main approaches in emotion analysis: classification with singular emotion label, multiple emotion labels, and continuous value labels. For singular emotion label classification, the goal is to assign one dominant emotion to a text (Plaza-del Arco et al., 2024). For this task a discrete emotion model can be used, such as Ekman's taxonomy which identifies 6 basic emotions (happiness, sadness, anger, fear, disgust and surprise) (Ekman, 1992) or Plutchik's Wheel of Emotions (Plutchik, 1980). To perform the singular emotion label classification we can use rule-based or lexicon-driven systems, like the Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015) or VADER (Hutto and Gilbert, 2015), which map words to predefined emotion categories. More traditional machine learning algorithms (e.g. Support Vector Machine (SVM)) and deep learning architectures like BERT (Devlin et al., 2019) are trained on labeled datasets to predict single emotions.

Multiple emotion label classification enables texts to be annotated with emotions that occur alongside each other. It can be useful in case of mixed emotions (e.g. a funny text with sarcasm and fear in it). To extract multiple emotion label we can also use neural networks which identify overlapping emotions.

Continuous value labeling predicts not the discrete classes, but the emotion intensity score indicating how strongly an emotion is conveyed. For this task we also can use more regression models (e.g., Random Forest (Fisher et al., 2025)) or BERT (Devlin et al., 2019) to predict numerical scores from text features capturing emotional shifts such as escalating anger.

3 Related work

Recent research in meme classification and retrieval has been conducted in a variety of areas, including multimodal processing and emotion recognition. The section reviews related work in the areas, which can be extended by this thesis. We discuss the topics presented in the Background in relation to our research question.

3.1 Recommendation systems in social media

The use of recommendation systems is an important aspect of online information processing and, consequently, of improving the user experience when using various services, including social networks (Li et al., 2024). These systems are designed to provide users with personalized recommendations, for example on online products or, in the case of social networks, to set up an intelligent news feed (Li et al., 2024). The same applies to memes or any images with text. As we state before, in the context of this thesis we can talk about a content-based recommendation system with retrieval characteristics, or a recommendation system driven by a retrieval process.

Yang Li et al. (2024) write that one of the most widely used approaches in recommender systems is a system based on Collaborative Filtering (CF). This method is based on the idea that users with similar preferences and behaviors will have similar opinions and therefore be interested in similar recommendations. CF is driven by user profiles built from their past interactions, such as their shopping history or the movies they have watched. However, the disadvantage of the approach is the "cold start" problem, where the user and/or the information item have no previous interaction history (Li et al., 2024). Without accumulated knowledge of user behavior, CF-based recommendations are random. It is also relevant to our work, as it shows that we need to establish the initial user preference first in order to have a starting point for future recommendations. Tao Chen et al. (2016 address this problem and present a CITING framework for dealing with tweets (microblogging context). Unlike like existing approaches that focus on images or text in isolation (unimodal approach), Tao Chen et al. target the unique nature of image tweets by combining multimodal and social cues, which is also relevant for our thesis. In order to improve recommendations the authors fuse traditional collaborative filtering with contextual information to better predict which tweets a user will find relevant (Chen et al., 2016). This approach is more sophisticated than simply using user-object interaction data and outperforms the standard recommendation model. In our thesis, we move away from the objective context (such as a historical event that triggers the reaction in an image tweet (Chen et al., 2016)) and aim to create a system that analyses the user's emotions rather than the context.

This thesis comes closest to addressing the problem of recommender systems for memes as images with text. Existing work either only considers one modality, and this applies to both recommendation (Won et al., 2019) and retrieval systems (Stathopoulos et al., 2023), or relies on context that's often unavailable to memes

such as the external web page linked to by the tweet's embedded URL (Chen et al., 2016). As Yang Li et al. point out, images in social media are more semantic but diverse and need to be understood in the context of their mention (Li et al., 2024). This is also true for memes which can be based on the context of many social phenomena (Shifman, 2013). The lack of researches in recommendation and retrieval systems area that focuses on memes points to a research gap and the importance of this thesis.

3.2 Meme Classification

Moving from the recommender system in general to the method in particular, it is important to examine existing work in the field of meme classification. The work on meme classification or categorization has largely focused on identifying specific types of content, such as hate speech, and has not paid sufficient attention to flexible, user-centered search that takes into account the nuances of meme content. For instance, Tariq Habib Afridi et al. (2021) propose a general framework for visual-linguistic multimodal challenges and discuss the limitations of existing methods in addressing meme classification, particularly in the context of automatic content targeted on misinformation and hate speech on social media platforms. Authors identify several open research questions and presents a roadmap for future developments in machine learning for visual-linguistic tasks (Afridi et al., 2021). This study provides valuable insights due to their attention to multimodal nature of mems, but focuses primarily on hateful/non-hateful classification, i.e., binary classification. The authors discuss neural network architectures like CNNs which are often used to process image data and RNNs for sequential data like text.

Tariq Habib Afridi et al. also highlight (2021) the rise of Transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) which use attention mechanisms to weigh contextual relationships in text more effectively than traditional RNNs (Afridi et al., 2021). The Transformer-based approach has become widespread, resulting in models such as BERT (Devlin et al., 2019), EmotionBERT (Huang et al., 2019), RoBERTa (Zhuang et al., 2021), and several others. Because of the advantages of Transformer-based models, namely relatively fast computation (which is important when interacting with users), high accuracy and multimodal usability, they are often used for classification tasks (Zhuang et al., 2021)(Huang et al., 2019)(Dong et al., 2018). The main Transformer-based model we consider in this thesis is CLIP, a multimodal Transformer-based approach developed by OpenAI (Radford et al., 2021). CLIP focuses heavily on contrastive learning and aligns visual and textual features in a common embedding space, however lacks a fusion mechanism (Radford et al., 2021). The model aligns text and images without a deeper common representation of the two (Radford et al., 2021). This can be useful for some classification tasks, but does not provide ready-made solutions with the result of already combined multimodal features.

Afridi et al.'s survey highlights the field's emphasis on binary classification tasks, such as differentiating hateful from non-hateful memes (Afridi et al., 2021). However, meme retrieval systems benefit from more nuanced understanding and diversity of content. Going beyond basic categorization, the aim of this thesis is to extend the scope of meme retrieval through sophisticated multimodal clustering that combines visual, textual and emotion features for potentially more accurate classification and improved retrieval and recommendation capabilities. A more specific example of a multimodal approach and its comparison with a unimodal approach is discussed by Douwe Kiela et al. (2020). Researchers use a multimodal approach to detect (to classify binary) hate speech (Kiela et al., 2020). The work shows that current state-of-the-art multimodal models perform relatively poorly compared to human performance. However, Douwe Kiela et al. demonstrates that the multimodal approach outperformed the unimodal approach (Kiela et al., 2020). The work indicates that there is still room for improvement in multimodal classification, even in the case of a binary task, and highlights the challenges associated with the topic (Kiela et al., 2020). Therefore, we do not seek to replicate the classification task, but build our own framework on the knowledge of the need to use a multimodal approach when dealing with memes.

Vasiliki Kougia and John Pavlopoulos, who participated in the same challenge (Kiela et al., 2020), demonstrate that a multimodal approach based on BERT outperforms unimodal approaches (in this case the text modality) (Kougia and Pavlopoulos, 2021). The authors again emphasize the need for a multimodal approach.

Lanyu Shang et al. (2021) address the same issue. The authors present a deep learning-based Analogy-aware Offensive Meme Detection (AOMD) framework that uses visual, textual, and contextual features from the content of meme posts as input (Shang et al., 2021). However, like the aforementioned studies, it considers a specific binary classification problem rather than a more generalized approach which we address in this thesis.

Although there are several works that work with memes in a multimodal context (Kiela et al., 2020)(Shang et al., 2021), they focus primary on binary classification and are constructed in the context of defense or filtering. This thesis aims at the opposite situation: finding memes that can be shown to the user. A multimodal approach alone, however, is not sufficient and does not answer the research question of this paper. To examine the emotion component, we need to look at related work in the field of emotion analysis.

3.3 Emotion analysis

When defining the emotion component, it is important to clarify the classification of emotions that can be used. One of the basic approaches is Paul Ekman's approach which provides a clear structure for such research. For instance, Shivam Sharma et al. (2024), who focus on improving the ability to identify emotions conveyed by memes, present the framework ALFRED (Sharma et al., 2024) for emotion-based classification using the Ekman's taxonomy. It shows quite high accuracy

in the task of classifying memes into basic emotions analyzing both multimodal and unimodal approaches. As in the work of D. Kiela et al. (2020), the multimodal approach (especially the framework created by the authors themselves) shows better accuracy compared to the unimodal approach which again address the necessity of a multimodal approach.

However, the basic emotions classification can be not fully suitable in the case of memes. Memes are often sarcastic, ironic, or ambiguous, which may not be well captured by a strictly categorical model like Ekman's (Ekman, 1992). A meme that appears 'happy' may actually be intended to express sarcasm or irony, requiring more nuance. In an attempt to capture the more fine-grained meanings of memes, EunJeong Hwang et al. (2023) address the task of captioning memes and hey publish a dataset containing the visual metaphors of memes which is used in this thesis (Hwang and Shwartz, 2023). Each dataset picture is labeled with the literal image caption and the visual metaphors. However, such a detailed approach is also insufficient for the task of categorization. The visual metaphors highlighted in the paper add an additional description closer to human perception, but they are not an emotion component in themselves.

As an intermediate option, a taxonomy of 27 emotions (and neutral) proposed by D. Demszky et al. (2020), can be considered. In contrast to Ekman's system, the emotions presented by the researchers are not only positive, but also negative and ambiguous (Demszky et al., 2020). This classification can allow for more nuanced interpretations, which can be important in multi-level social messages such as memes while remaining within a defined classification with a limited number of labels.

It is worth noting that visual memes tend to contain very little textual data. Without a lot of information, it can be difficult to identify the emotion contained in such texts. Ming Chen (2022) propose a deep learning-based method to analyze emotions in short texts, with applications in tracking cultural trends in Western literature and media (Chen, 2022). Like in several previous papers, the researcher uses a Transformer-based (in this case BERT-based) approach that first converts text into context-aware word vectors, captures nuanced emotion cues, and then processes sequential dependencies in text to improve semantic representation by considering bidirectional context (Chen, 2022). Nevertheless, despite more accurate prediction compared to traditional deep learning models (Chen, 2022), the proposed approach has its limitations: first of all, it requires large, labeled emotion datasets for training (Chen, 2022). It also focuses on Western texts (Chen, 2022) and therefore may not generalize to non-Western contexts without adaptation. Ming Chen's work is not in line with the social science orientation of the thesis due its focus only text and only in the Western setting, but it demonstrates the success of the BERT-based model in the context of emotion extraction.

Thus, for this work, the most appropriate emotion classification that maintains a balance between the social and emotional subtleties of memes, while being a classification with a defined number of labels, is the GoEmotions dataset (Demszky et al., 2020). Since the BERT-based approach has shown its validity for emotion

analysis, we next use the BERT-based model (Zhuang et al., 2021) trained on this dataset for the task of emotion feature extraction.

3.4 Summary

In summary, despite recent improvements in multimodal analysis, existing models still face challenges in understanding the humor as a derivative of emotion of memes. Existing research has mainly focused on the task of identifying hate speech using a binary classification (Afridi et al., 2021)(Kiela et al., 2020)(Shang et al., 2021)(Kougia and Pavlopoulos, 2021). This thesis can provide a more theoretical basis for a sophisticated understanding of similarity classification that can be applicable in future research in this area, including the social sciences. In addition, the existing works specifically in the field of memes do not focus on a recommendation system, which is a relevant issue aimed more at practical applications (e.g. in social media). Some studies have looked at the emotion component of memes (Hwang and Shwartz, 2023)(Sharma et al., 2024), but this has not been used as a separate modality or label in classification tasks, indicating a research gap in this area. Some studies focus on only one aspect, such as short texts (Chen, 2022), rather than memes as a whole, and have limitations in terms of the need to label the data. The lack of research on meme recommendation systems also underlines the relevance of this thesis.

4 Dataset

The dataset MemeCap (Hwang and Shwartz, 2023) created by EunJeong Hwang et al. and mentioned earlier is well suited for use in this thesis. The dataset includes 6.3K memes from Reddit, primarily the /r/memes subreddit, but also posts with a meme in the post title (Hwang and Shwartz, 2023). These memes are filtered to exclude those with no text, sexual imagery, excessive number of characters or profanity (Hwang and Shwartz, 2023). Despite this, some of the memes remaining in the dataset can be considered ambiguous and even offensive. This requires additional filtering, which is discussed in section 5.1.1. The dataset also contains diverse captions, image information and retrieved metaphors (Hwang and Shwartz, 2023). However, the provided information is not necessary for the thesis. The aforementioned focus on the social sciences requires as little data pre-training as possible. Therefore, methods that require manual image labeling are not suitable for such an approach. In this regard, although there is additional information, only the dataset's images are used in the thesis. The flexibility of our approach allows us to replace the dataset with any dataset containing memes in a visual format.

We should also discuss the structure of the dataset. According to authors there are four types of memes in the dataset: text dominant, image dominant, complementary and without visual metaphor (Hwang and Shwartz, 2023). A metaphor vehicle type can be person or character, object or property, facial expression or gesture or action (Hwang and Shwartz, 2023). Additionally, metaphor target type can be behavior or stance, meme poster, approach or concept, another person and desire vs reality type (Hwang and Shwartz, 2023). An even distribution of these categories across the training and test samples makes the final dataset more balanced (Hwang and Shwartz, 2023). However, they do not represent the final classes for the classification task when it comes to human preferences in humor and we cannot build our groups for retrieval system based on these categories alone. From person to person, memes that seem funny may fall into different categories. Taking this into account, the classification task in this thesis becomes a clustering task which we discuss further in the section Clustering approach.

5 Methods

The section provides an explanation of the methods we use in this thesis. We first explain the multimodal approach, then introduce and discuss clustering and retrieval approaches. Figure 2 shows a pipeline of our framework, from the raw data to the retrieval algorithm.

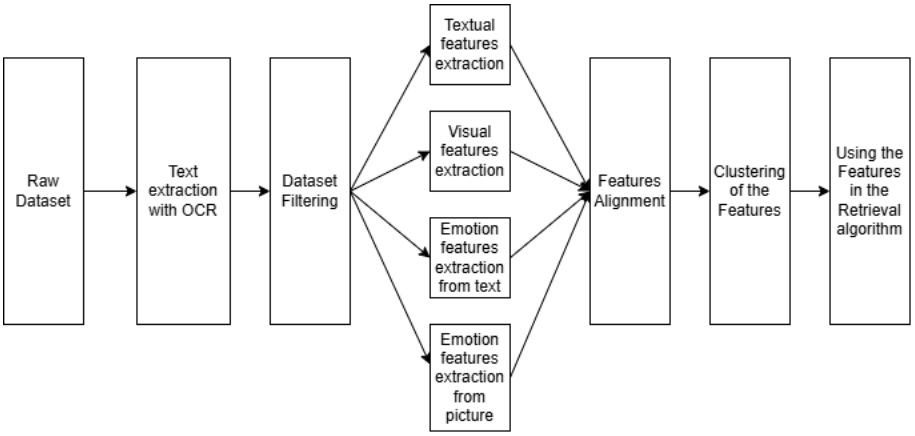


Figure 2: Pipeline of the system

5.1 Multimodal approach

To obtain embeddings for each individual meme, we use the CLIP model (Radford et al., 2021) and RoBERTa (Zhuang et al., 2021) models. We implement the framework that can be divided into several parts:

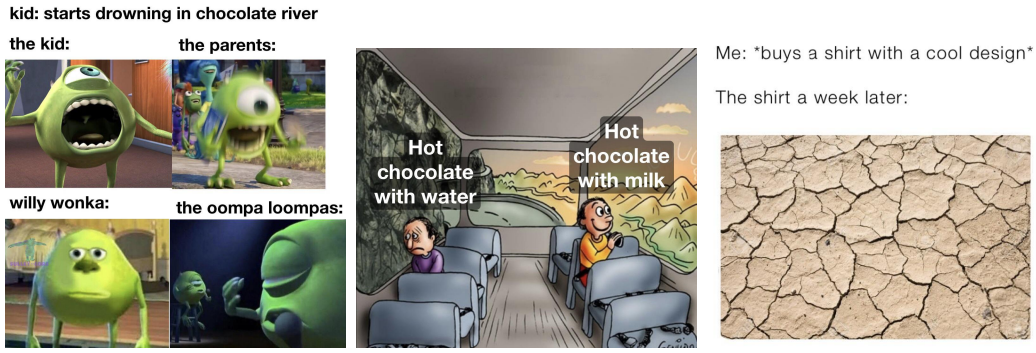
- Extraction of textual modality features;
- Extraction of emotion component features;
- Extraction of visual modality features;
- Alignment of all the extracted features;
- Clustering of aligned features;
- Retrieval based on the user input.

To obtain multimodal information, we first need to extract features from each meme. We extract each type of feature (textual, visual and emotion components) and combine them into a common feature. For the multimodal approach, these features include visual and textual features, and for the emotion component approach, they also include the emotion component. We use these features to form different groups or clusters of memes with similar characteristics, which are then used for the retrieval system. Our user experiment is designed to test the difference between the two approaches: when the user deals with a system that works with features containing only textual and visual information, and a system that also includes an emotion component.

5.1.1 Retrieval of the Textual Modality

In order to find textual features of each meme, the first step is to extract the textual component or text. Since the used dataset (Hwang and Shwartz, 2023) initially

doesn't not contain the text in the image as additional information and we aim to create a procedure that does not require labels or the need to create them, text extraction is one of the phases we need to conduct. As the practice shows, this step proves to be the most computationally and time consuming. For this purpose, we can use Optical Character Recognition (OCR) (Sharma, 2023). OCR is a technology that converts images or scanned documents into a digital text (Sharma, 2023)(Smith, 2007). To do so the OCR-tools use pattern recognition to identify characters and words. An important disadvantage of OCR-based tool in this task is the order of the sentences. Since memes can be in different formats (e.g. a comic in 4 parts, a meme with captions for different visual entities, an image with a caption, a picture with a text message as shown in Figure 3), the most accurate text capture still would be manual captioning. However, we still get enough text to analyze the textual content of memes without using human resources.



(a) A meme containing 4 parts

(b) A meme with captions for different visual entities

(c) An image with a caption

Figure 3: Example of different meme formats

We store text of each meme and then perform a filtering: images containing certain words (e.g. obscene language or nationality) are removed from the dataset. However, given the complexity of visual and textual metaphors in memes, this approach does not completely eliminate all content that could be taken offensively.

After that, we use the CLIP model generates a text embedding by processing both short and long input texts. Even in case of memes we face the problem of large texts. The model that we apply uses 77 tokens during training. If short texts (less than 77 tokens) can be processed directly by CLIP to produce an embedding without any problems, the longer texts are normally truncated, resulting in a loss of information and context. We propose to solve this problem by breaking long texts into overlapping parts in order to save the contextual information. Each chunk of the longer text is embedded independently by the model, the overlapping chunks provide continuity between words or phrases softening abrupt breaks. The resulting embeddings for all fragments are merged into the final embedding using an averaged embedding merging strategy. This approach allows CLIP to handle text of any length, balancing model's input constraints and semantic consistency. It also reduces

the amount of processing required. However, it is a less accurate approach compared to full text processing in the case of long texts.

5.1.2 Retrieval of the Emotion Component

As an emotion component, we consider the emotion embeddings derived from textual and visual input from each image. First, we must identify the emotion from the text. Since CLIP does not provide such functionality in the case of text, we use BERT-based (RoBERTa) model which is trained on the previously mentioned GoEmotions dataset (Demszky et al., 2020) with 28 labels. The model analyses the text and returns list of emotions of this text.

We convert the retrieved text in a previous step into a 28-dimensional emotion probability vector using the emotion classifier. The classifier identifies probable emotions (e.g. "happiness", "anger") with their corresponding confidence scores. The final vector represents the emotion composition of the text as a probability distribution over 28 predefined categories. However, we need to note that a model which is trained specifically on memes may be more accurate.

Only extracting emotions from text would be a unimodal rather than a multimodal approach. To make the prediction, we use the list of 28 emotion labels mentioned earlier. In this case CLIP can predict emotion or emotions from an image by matching visual features with textual emotion labels. The model computes similarity scores between the image and each emotion label, which indicates how well the image matches each textual descriptor of the emotion. The scores are converted to probabilities using softmax, resulting in a distribution across all emotion labels, so that the full probability distribution is smoothed into a vector to preserve nuanced confidence in all emotions. With this approach we capitalize on the multimodal capabilities of CLIP to combine visual content and emotion semantics, as stated before, without the need for a training phase.

5.1.3 Retrieval of the Visual Modality

To extract the visual modality features we again use the CLIP model. Each input image is pre-processed by the CLIP processor and then converted into a tensor compatible with the model. As result we get high-dimensional visual features. This produces a compact, fixed-dimensional vector representation of the meme.

5.1.4 Multimodal embedding fusion

After retrieving the embeddings from each modality (text, visual and both emotion embeddings), we need to process them to create fused representations. Because CLIP lacks a fusion mechanism we need to develop such an mechanism ourselves when working with the model. First, we normalise each embedding type using Z-score standardisation to ensure consistent scales across features. The normalised

embeddings are then fused into two variants: one with emotion features (combining text, visual and both emotion vectors) and one without emotion features (text and visual only). We store all these concatenated embeddings and apply Principal Component Analysis (PCA) (Jolliffe and Cadima, 2016) to reduce the dimensionality of the embeddings. In our case, we aim for 128 dimensions. The number of dimensions is chosen to strike a balance between having space to encode complex concepts hidden in memes, and having operational data small enough for fast computation. After dimension reduction, the obtained embeddings are ready for clustering algorithm to cluster or group our embeddings for both approaches, with and without the emotion component.

5.2 Clustering approach

To form a clustering algorithm, we should note that since we are talking about user satisfaction, the calculation after user input should be as fast as possible without losing much accuracy. For this purpose, we perform all computationally and time consuming calculations in the model preparation and training phase. During this time, the embeddings are grouped into clusters based on similarity.

A simpler approach is using an algorithm that requires the number of clusters to be known in advance, such as the K-Nearest Neighbor (KNN) (Cunningham and Delany, 2021) algorithm. However, such approaches are less flexible and lose some of their representation power. More accurate solution is Affinity Propagation algorithm. Affinity Propagation (AP) is a clustering algorithm that identifies representative data points exemplars and groups similar data around them (Frey and Dueck, 2005). During the clustering phase we process the embeddings obtained before. In this step, we not only group the memes into clusters, but also try to identify the relationships between these clusters. In this way, we aim to identify semantic similarities between different groups of memes. The algorithm groups similar images into clusters using the Affinity Propagation algorithm. The algorithm automatically determines the number of clusters based on inherent data patterns using Euclidean distances between points to build the similarity matrix for identifying the exemplars. Once clusters are defined, the we compute pairwise similarities between cluster centers using cosine similarity to determine how closely related different clusters are. We store these similarities as edge weights ranging from -1 (dissimilar) to 1 (identical) reflecting the angular orientation of the cluster centroids in the embedding space. We retain these weights in order to recognize the relationships between clusters, which will be further useful for the retrieval tasks and explained further in the Retrieval approach section. The priority of the algorithm is to preserve important metadata while ensuring computational efficiency. We obtain 386 clusters for the multimodal approach with emotions and 381 for the approach without emotion component.

5.3 Retrieval approach

When working on the recommendation system, it is important to make it iterative. That means that user of the system should be able not only to retrieve the most similar memes, but also to evaluate the degree of similarity in order to get the most. This gives users the ability to not rely on a single query, but to provide feedback to get the most satisfying result. We propose a framework that retrieves images by combining visual similarity (direct embedding similarity) and semantic cluster relationships (inter-cluster relationships), dynamically adapting to user input. As we discuss in the Recommendation systems in social media section we need some initial set of pictures for the purpose of avoiding the "cold start" problem. As such, the user can select a few memes from this initial set to serve as a starting point for follow-up recommendations.

After determining the starting point, we need to construct an algorithm to find memes that are similar to the query. The simpler approach would be to return a meme from the same cluster as the one selected by the user. However, this approach is not flexible. The user can select memes from different clusters, united by characteristics common to those individual instances, but not to the clusters as a whole. That means that the 1:1 ratio is not respected when changing the number of memes selected (e.g. from 5 to 3) and the number of memes retrieved (e.g. from 5 to 8).

A more flexible approach is to measure the average similarity of the user-selected memes to each meme in the dataset. Then the N (5 in this experiment) memes with the highest similarity value to the query embeddings are output. Comparing over the whole dataset allows the algorithm to be more flexible and to focus on the individual characteristics of the user input rather than the data unit. A disadvantage of this approach, however, is the use case where the memes selected by the user are in clusters that are far apart from each other. In this situation, the embedding of these memes may not accurately reflect any of them, as the suggested memes may lie in a "middle ground" between these different clusters, which can make them less relevant to the user. Furthermore, by averaging, the embedding may dilute the specific features that made each selected meme unique. As a result, the retrieved memes may not capture the strong individual similarities to any of the selected memes, but may only be "moderately close" to all three.

To avoid this issue, a more careful approach to calculating similarity than the mean is needed. One possible solution is to have a threshold that cuts off a part of the samples that are exactly inappropriate or, on the contrary, too appropriate. Similarly, to avoid median values, only memes that are maximally similar to some of the memes selected by the user can be selected. As a variation of this approach, different similarity coefficients can be weighted and used to select the most similar samples.

However, the above-mentioned solutions are simplified. In addition to the difficulty of choosing an individual strategy (e.g. for thresholding) for different systems (respectively different clustering), such approaches do not guarantee the same behavior for different user inputs. More advanced the method is of assigning weights

not depending on the similarity coefficient, but on the cluster in which the meme is located. In this case clusters are represented as a weighted connected graph, where the numerical value of an edge decreases the coefficient or weight that affects the final similarity coefficient between memes from different clusters. In this way, clusters closest to cluster X are weighted 1, clusters further away have lower value, completely dissimilar -1. Therefore, the data source we obtain in the Clustering approach stores not only parameters for each mems (e.g. reference number and cluster), but also the relation between all clusters.

For an initial query without feedback, the system constructs a query by averaging feature vectors of selected images. This averaged embedding represents the "middle ground" of the selected examples in the feature space. The system also identifies the most frequent cluster among the selected images (e.g. if 3 selected images belong to cluster A and the other 2 to cluster B, cluster A becomes the semantic anchor of the query). When feedback scores are provided, the system refines the query and filters the results. Positive feedback (score is greater than 0) increases the influence of the corresponding image's features in the query average, pulling results towards its features. Negative feedback (score is less than 0) subtracts the image's features from the query. It pushes away results from its features and excludes it from the output entirely. The query's cluster context is dynamically updated: as feedback introduces images from new clusters, the dominant cluster is recalculated to reflect all user input. During the retrieval, at each feedback iteration, the selected and retrieved images are evaluated (how good or relevant each retrieved image is) using a cosine similarity between their features and the query vector, enhanced by the pre-computed affinity between their cluster and the dominant cluster of the query. For example, an image in cluster C (which is historically aligned with cluster A) will receive an enhanced score even if its raw visual similarity is slightly lower than an image in a less related cluster. The result priorities proximity to the query vector, while favoring clusters with contextual links to the user's initial selections. The duality of the approach allows the system to balance perceptual matches with broader conceptual relationships learned during clustering. The method is also flexible enough to vary the number of memes to be selected and that the user receives as output, as long as this number is greater than 2. However, in this thesis we stick to 5 memes as input and output.

6 User experiment

We answer our research question by formulating the following null and alternative hypotheses:

Null Hypothesis (**H0**): There is no significant difference in user satisfaction while using multimodal retrieval system and while using multimodal retrieval system with emotion component.

Alternative Hypothesis (**H1**): There is significant difference in user satisfaction while using multimodal retrieval system and while using multimodal retrieval system with emotion component.

To test our hypothesis we compare two approaches: multimodal approach with emotion component and without it. The experiment requires that each participant receives a retrieval system that is tailored to a particular approach and rates their satisfaction with the system. We can test our hypothesis by comparing these user satisfaction reports.

To facilitate participation in the experiment, we develop a website with the recommendation system accessible from the Internet. It contains the retrieval algorithm described in the section 5.3.

We use two data sources with memes and clusters information, one for method without the emotion component and one with it. After opening the website, the user has a 50/50 chance of getting a group of 3 or 8. The group depends the data source: 3 corresponds to the method with an emotion component and 8 without. This approach also provides an opportunity to screen out participants who do not read the questionnaire properly and answer incorrectly or randomly. The survey requires users to be at least 18 years old, have advanced English skills and not use neural networks.

The website firstly provides user 20 random memes and asks to select 5 memes keeping in mind that the system will search for similar memes. The user interface of the website is shown in the Figure 4.

The system then retrieves or recommends 5 new memes for users to rate on a scale of -5 to 5. The screenshot of such recommendations with feedback input is shown in Figure 5. One of the memes shown on the screenshot is hovered over with the mouse to zoom in and improve readability. Each iteration, including the very first one, is counted in an iteration counter, which is further asked in the questionnaire. The users' goal is to select and rate the images in order to get a satisfactory result, if possible. Regardless of whether or not a satisfactory result is obtained, the user has to continue filling in the questionnaire.

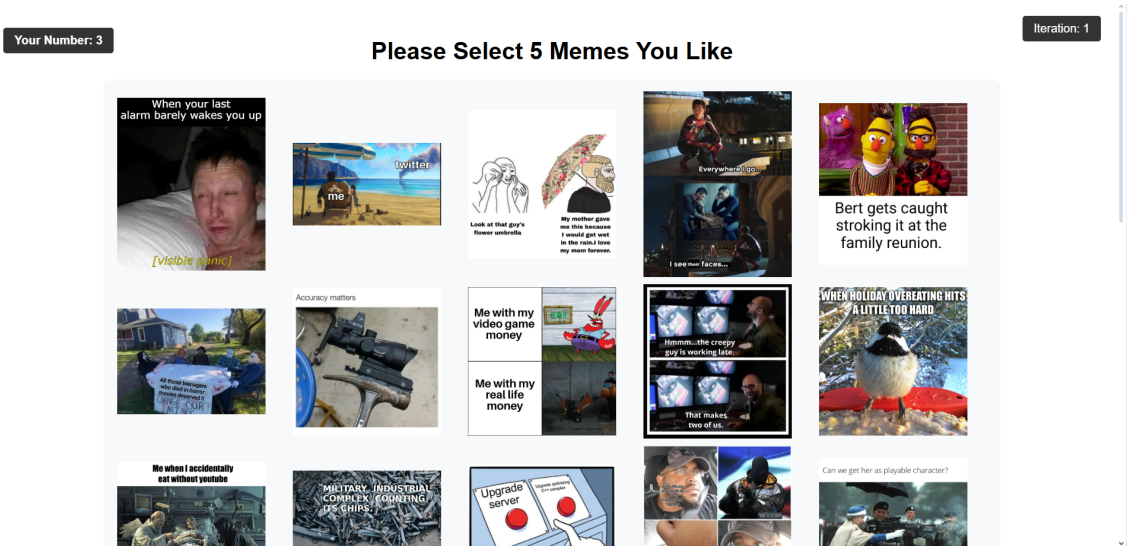


Figure 4: Interface of the website showing the initial set of random memes

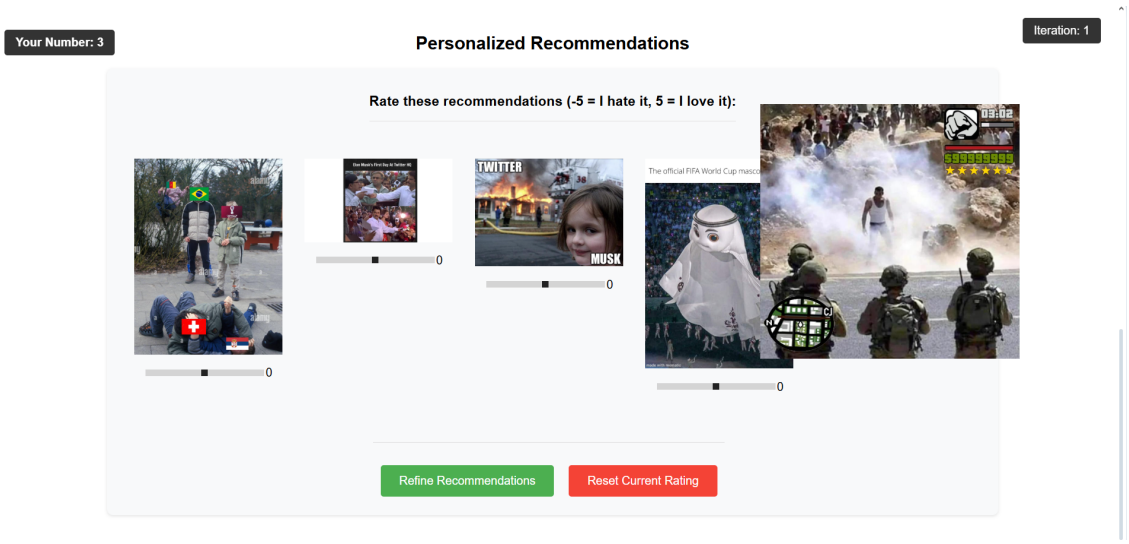


Figure 5: Interface of the website showing the recommended memes with feedback sliders

We recruit users to participate in the experiment both through private channels (e.g., asking friends and asking them to distribute the survey to friends of friends) and via the university forums. The message can vary to be more personalised or to target a wider audience. The text of a standard message is shown below:

Message when looking for participants

Hello! As part of my thesis, I am conducting a user experiment to test a meme recommendation system.
The experiment will take about 10 minutes, the questionnaire and the memes are in English, and all data will be processed anonymously. You don't need a Google account to participate.
Link to the study: *a link to the questionnaire form*
The website with the system itself is specified on the second page of the form. In countries outside the EU, images may take longer to load. Please use a PC to open the website. The mobile version is unfortunately not supported.
Feel free to contact me with any (technical) questions. I can also provide the results at the end of the experiment.
The site is written using the http protocol, so it may appear insecure in modern browsers. However, the site is secure and does not require any data entry. If there are security concerns, we can have a quick meeting in Zoom or Teams so that you can participate in the experiment remotely from my computer.
Thank you for your time and participation in this survey. Please share it with friends if it's possible.

The first page of the questionnaire contains the following text:

Text from the first page of the questionnaire

Dear participant, thank you for your interest in the experiment to test a meme recommendation system. Please be assured that your responses will remain confidential and will be used solely for research purposes. Your identity will be anonymized in any reports or publications resulting from this study.
Your task is to test the recommendation system according to the instructions on page 2 and answer some questions about your experience as truthfully as possible. You will also be asked to provide some demographic and personal information. Please be aware that this study may contain material that is intended for individuals over the age of 18. The content may include mature topics or language that could be deemed sensitive. You should also be able to understand English at an advanced level.
The approximate time required to complete the survey is 10 minutes. You can only take the survey once. However, if you would like to explore the system further, you can do so after you have taken the survey. Feel free to quit at any time without giving a reason, in this case your answer won't be saved.
Please do not use AI tools such as ChatGPT, any other kind of AI assistant software, or any other external source.
This study is run by Aleksandr Litvin, and overseen by Christopher Bagdon and Roman Klinger. If you have any questions regarding the experiment or the study you can write to the following email address: aleksandr.litvin@stud.uni-bamberg.de

The first page of the questionnaire contains the following text:

Text from the second page of the questionnaire

To get started, please go to this website: *website of hosting provided by university*
You need to copy and paste the entire link into your browser. The link is only accessible with a computer.
5000 must be left in the link. The site is made on an older protocol, so an insecurity message may appear, the example of it is below. Please ignore the message and follow the link anyway.
The site is safe to visit, no user data is requested."
We also provide the clarification what the security error is and screenshot of possible appearance. The following instruction are given on the page: "Depending on your location, the time it takes for images to load may vary. When you see the text "Loading... Please wait.", the loading of pictures is still in progress.
After :5000 in URL you will see a number. For example, 5000/4. Please do not correct the number. You will need this number in one of the questions.
Please do not reload the website until you have completed the questionnaire.
The instruction and the goal of experiment
You will see 20 random memes. Please select exactly 5 of them, keeping in mind that the system will search for similar memes. Once you have selected 5 images, click on "Find similar memes".
You will be presented with 5 new images. You can rate the recommended memes from -5 to 5 and click on "Refine Recommendations". If you want to change the rating, move the slider to a different number or click "Reset Current Rating" to change the rating for all 5 current images.
Your goal is to select and evaluate the pictures in order to get a satisfactory result.
Once you have achieved a satisfactory result (you are completely satisfied with the memes you have received), please don't close the website and complete the questions below.
If you can't achieve the satisfactory result after many attempts, that please don't close the website and complete the questions below.

Users are asked about several categories of questions: user satisfaction and overall experience, quality of recommendations, system responsiveness and interaction costs, and baseline questions. A list of all questions can be found in the appendix A. The study includes 40 participants of different ages and genders, with 20 participants in each group. The demographic composition of the participants is shown in Figure 6.

We also ask the participants how often they share memes in their everyday life. The distribution is shown in Figure 7. The vast majority actively interact with memes in their daily lives, indicating that the participants are familiar with the subject

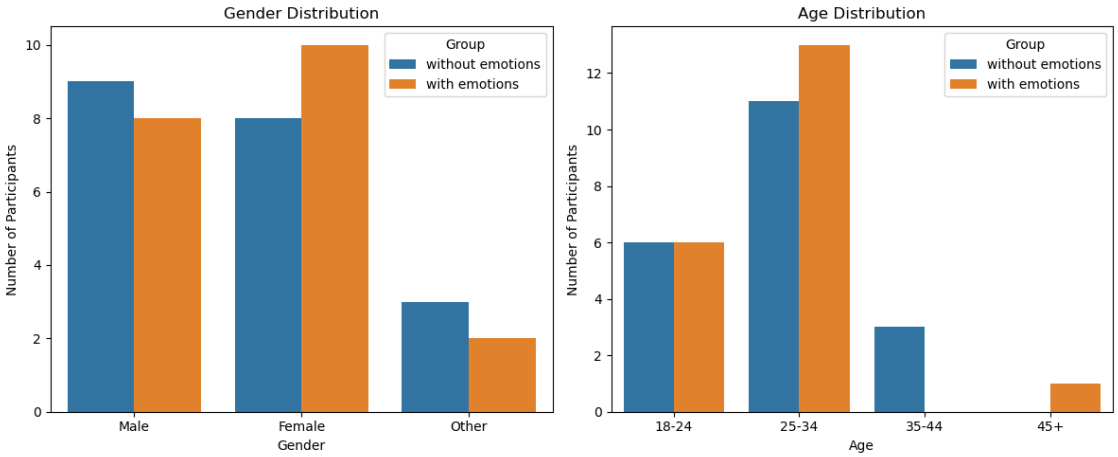


Figure 6: The demographic composition of the participants

matter of the study. The implementation details of the experiment can be found in A.2

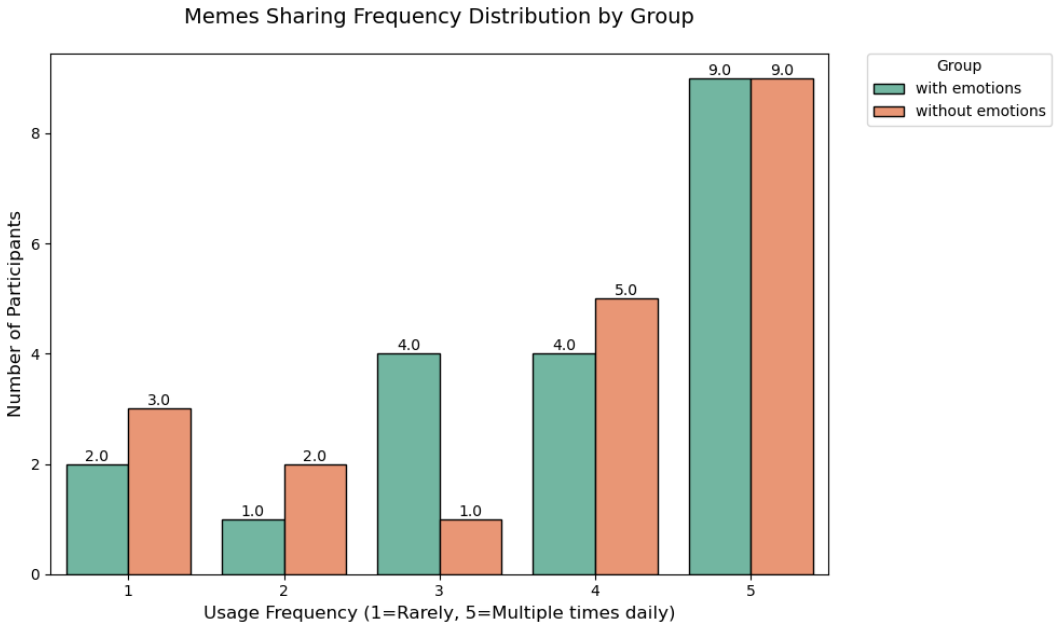


Figure 7: Memes Sharing Frequency Distribution by Group

7 Evaluation

7.1 Statistical Procedure

The Mann-Whitney U test is a non-parametric statistical procedure designed to evaluate two independent groups originate from populations with identical distributions (Mann and Whitney, 1947). The test is suitable for ordinal data, such as Likert-scale responses, where numerical values represent ordered categories but do not assume equal intervals between adjacent points (Mann and Whitney, 1947). Parametric alternatives, such as the independent t-test, rely on assumptions of normality and interval-level measurement that may be violated in ordinal data, especially when responses are skewed or have ceiling/floor effects (Delacre et al., 2017).

In the context of our experiment which compares two groups using a 1–5 ordinal scale, the Mann-Whitney U test addresses the primary research question: is there a significant difference in user satisfaction between group 3 (multmodal approach with emotion component) and group 8 (multimodal approach only)?

7.2 Statistical Power Analysis

The required sample size depends on three key parameters (Cohen, 1988):

- **Effect Size (d)**: For ordinal data, effect size is often expressed as the probability of superiority (P), defined as $P = \Pr(X > Y) + 0.5 \cdot \Pr(X = Y)$, where X and Y are observations from Groups A and B.
- **Power ($1-\beta$)**: Typically set to 80%, ensuring an 80% probability of detecting a true effect (Hoenig and Heisey, 2001).
- **Significance Level (α)**: Conventionally $\alpha = 0.05$ in two-tailed test.

Given a significance level of $\alpha = 0.05$ in two-tailed test and equal group sizes of $n = 20$, a power analysis shows that with this sample size, the study has approximately 80% power ($1 - \beta = 0.80$) to detect a large effect size (Cohen's $d = 0.8$), based on established power calculation methods (Cohen, 1988)(Faul et al., 2007). However, the power to detect medium ($d = 0.5$) or small ($d = 0.2$) effects is limited under the current sample size.

Therefore, although our results are robust enough to detect large effects in user satisfaction and behaviour between the two recommendation systems, we should be cautious in interpreting the null results, as the test may not be robust enough to detect more subtle effects.

8 Results

As we mention before, we want to test our hypothesis that there is no significant difference in user satisfaction when using a multimodal retrieval system and when using a multimodal retrieval system with an emotion component.

We use Mann-Whitney U tests which reveals no statistically significant differences (all $p > 0.05$) in user satisfaction between the the multimodal approach with emotion component and the multimodal approach without it while iterating with retrieval system. The experimental conditions for each of the 11 metrics are evaluated, the consistent null results suggest that under the current experimental conditions, the emotion component presence has no measurable effects on the user experience dimensions.

For 7 of 11 metrics (e.g., satisfaction, improvement, enjoyment) median scores are identical. Small differences ($\Delta \leq 0.5$ Likert points) are observed in relevance, responsiveness, and consistency, but none reaches statistical significance. Figure 8

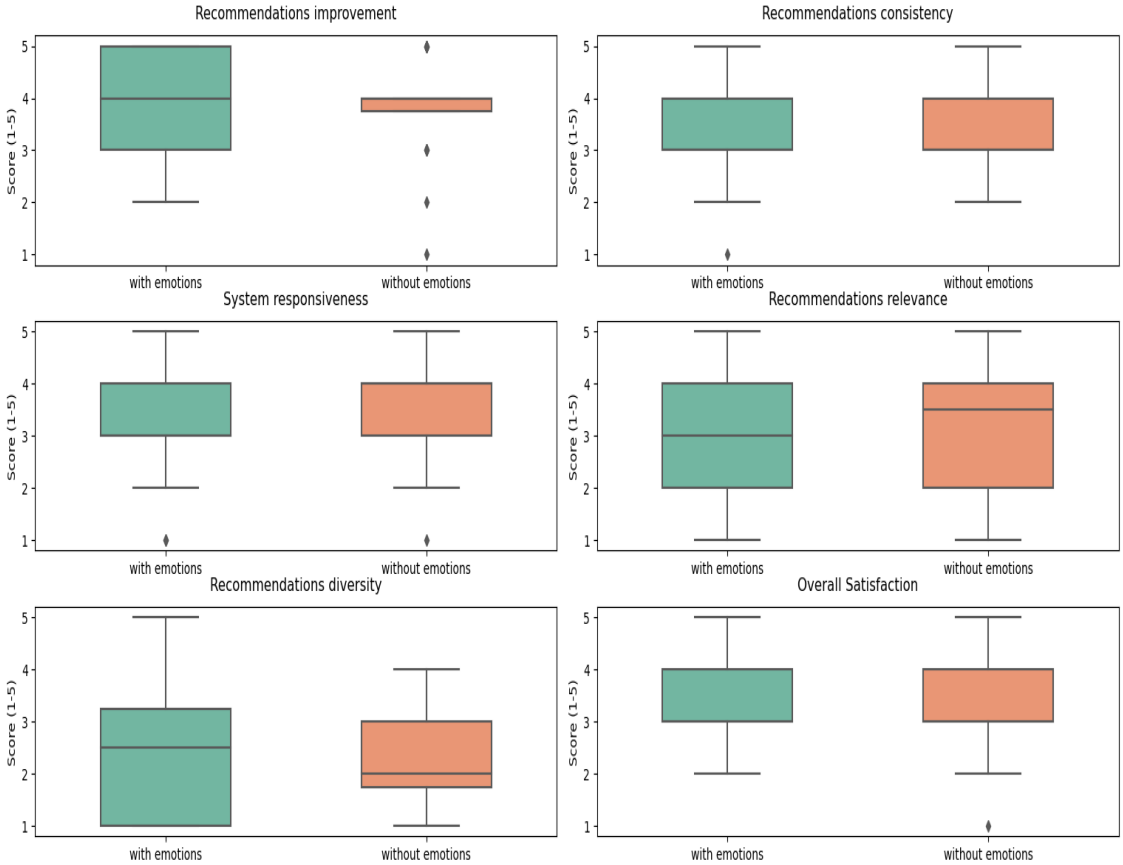


Figure 8: Iterations Metrics and Overall Satisfaction

shows box plots of user satisfaction associated with the system's iterative approach (how well each iteration brought experimental participants closer to a satisfactory outcome for them) and overall satisfaction. As can be seen from the box plots of

Recommendation improvement and Recommendation diversity, the group with emotions has a wider spread than the group without emotions. However, as we mention above, when looking at the medians, there is no statistically significant difference.

Figure 9 shows a comparison of the box plots of overall satisfaction, satisfaction with the system, willingness to use the system in the future, and willingness to recommend the system to friends.

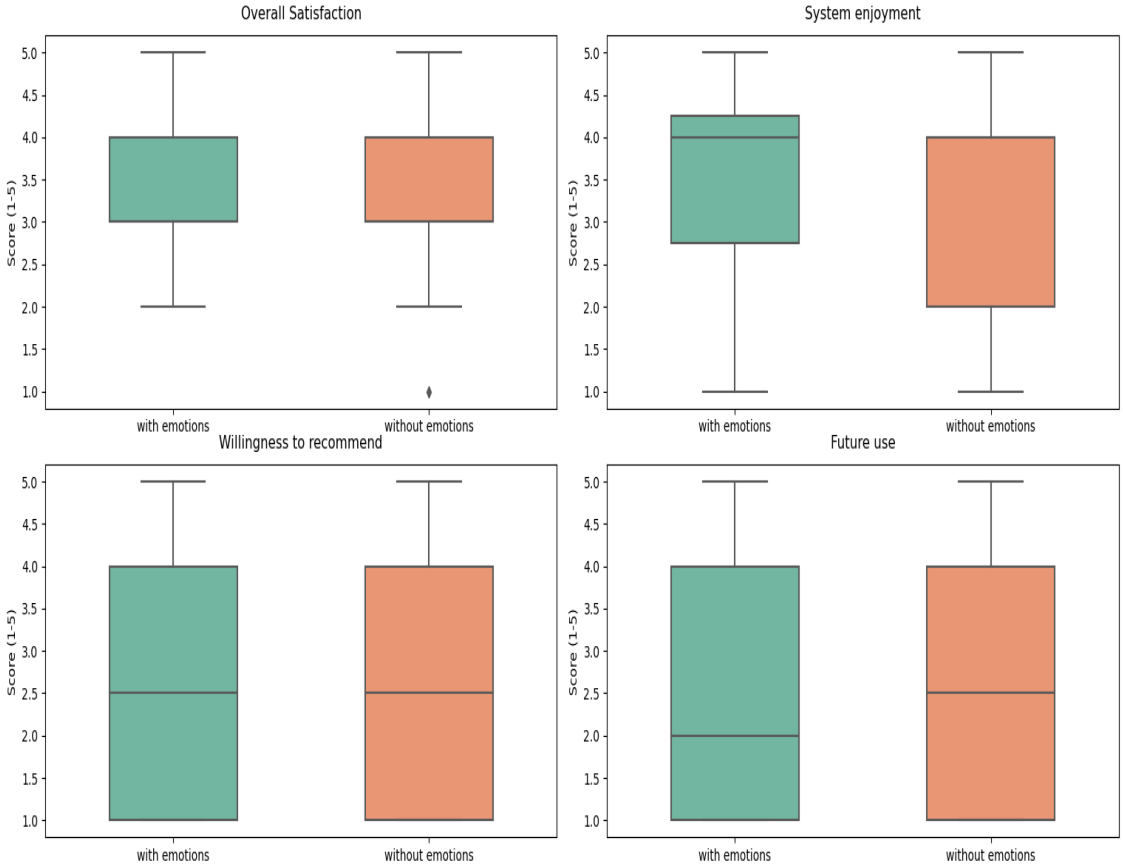


Figure 9: Overall Satisfaction and System Satisfaction

There is no gender difference in the comparison of user satisfaction. The comparison shows that for all metrics **p** values > 0.05 (range: 0.61-1.00) there is no evidence that emotion labels affect the two genders differently. For the comparison we exclude participants with the selected gender "Other" due to the small number (only 5 participants). For men, the improvement and enjoyment metrics slightly favor the approach without emotion labels, for women all metrics (satisfaction, improvement and enjoyment) favor the emotion labels approach (but trivial effects: $\mathbf{r} \leq 0.13$). The observed effects and sample size (8-10 participants in each group) are too small to draw any conclusions. The comparison box plots are shown in the Figure 10.

We also compare the groups with high (group which evaluates the relevance of initial mems higher than median) and lower (the relevance is below or equal median) interest in initial memes. For both interest groups: $\mathbf{p} > 0.77$ with negligible effect

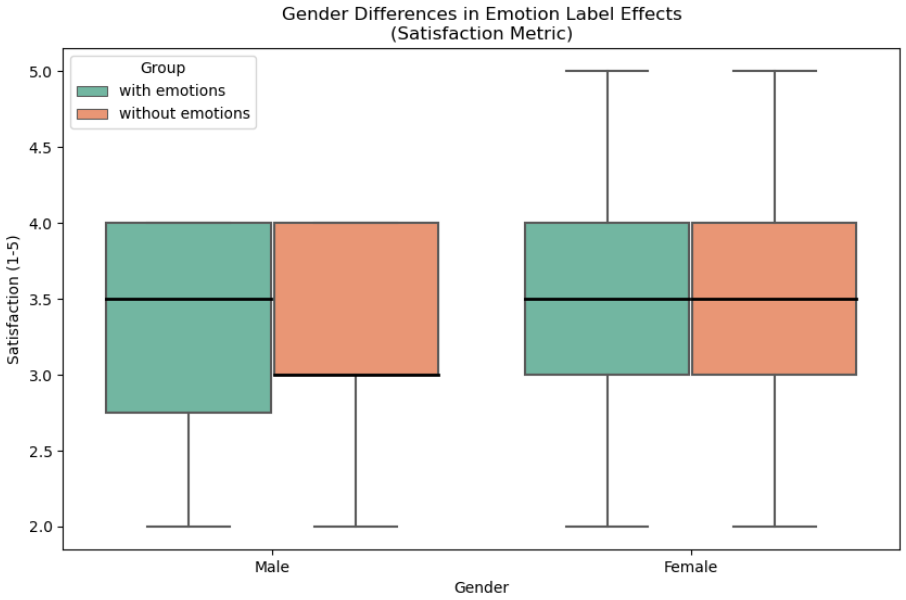


Figure 10: Gender Comparison in Satisfaction Metric

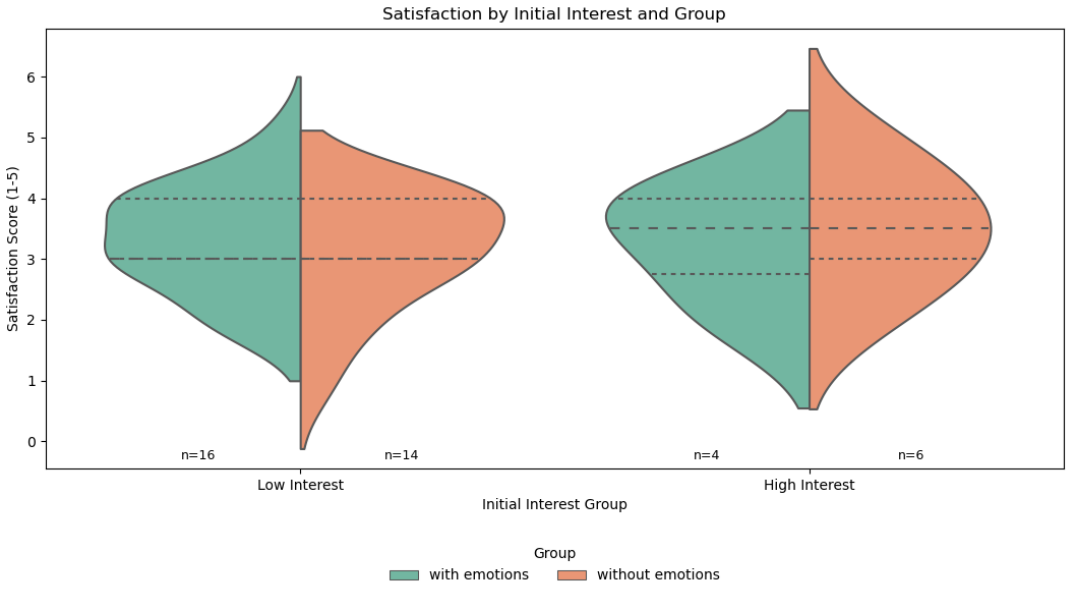


Figure 11: Violin plot showing User Satisfaction based on Relevance of Initial Set of Mems

sizes ($\mathbf{r} = -0.06\ to + 0.13$) High-interest group slightly trends toward preferring without emotion labels trends toward preferring without emotion labels, low-interest group shows no meaningful difference. The comparison can be found in the Figure 11.

In addition, we analyze whether there is a difference in satisfaction between participants who share memes (low frequency of sharing, less than median, above median) and those who share memes frequently (high frequency of sharing) in everyday life. The comparison is shown in the Figure 12. The analysis reveals no statistically significant differences between the groups across all metrics with all $\mathbf{p} - values > 0.05$ (range: 0.18–0.97), the largest effect ($\mathbf{r} = 0.24$ for responsiveness metric). Therefore, we cannot confirm that emotion labels make a significant difference when comparing participants who share memes often and those who do not. But we can't prove that there is no difference either, because of the sample size and the small effects.

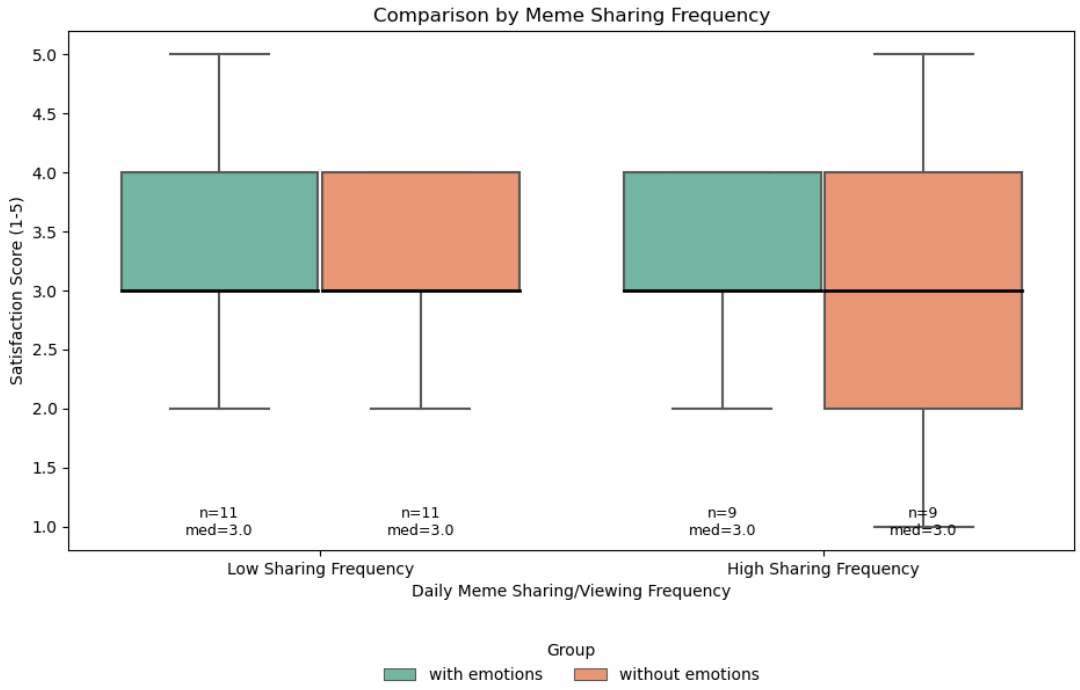


Figure 12: Comparison based on the Memes Sharing Frequency

Finally, we want to mention the iteration counter metric. Seven participants in the group with emotions marked 100 as iterations, signaling dissatisfaction with the sample. In the group without emotions, there were four such participants. As previously mentioned, there are 20 participants in each group. Since the other metrics, including overall satisfaction, do not show a statistically significant difference, we cannot draw conclusions based only on the iterations metric. We also want to note that some of the participants indicated 100 iterations despite a score of 3 in overall satisfaction with the system.

9 Discussion

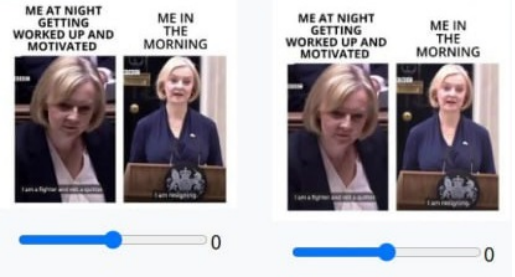
The lack of significant differences between the two methods may be due to several factors:

1. There are indeed no significant differences between the approaches, multimodal analysis of memes is not exactly different from multimodal analysis with the addition of an emotion component;
2. The conducted experiment and the created framework are not precise enough to reveal the difference between the approaches.

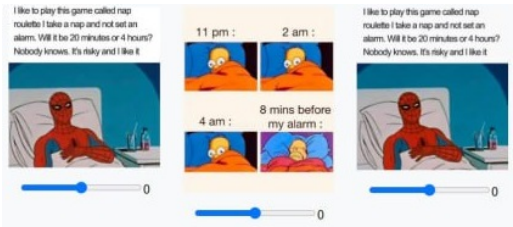
We would also like to point out that the perception of the whole framework can have been influenced by factors such as:

- technical issues (some participants in the experiment complain about the inaccessibility or long response time of the website, while others have no such problems);
- lack of understanding of how the concrete recommendation system works due to the unfriendly user interface;
- lack of familiarity with a particular dataset. The dataset we use only includes memes shared on Reddit. Such humor may influence the perception of recommended memes and the desire to use the system in the future or recommend it to friends.

In addition, some participants give feedback about the repetition of memes. This problem could have occurred at the dataset creation stage: the same memes could have been taken from different sources, but since the differences in the images are insignificant (they are cropped differently or have slightly different colors), this was not noticed by the dataset authors (Hwang and Shwartz, 2023). Screenshots of such system's behavior can be found in the Figure 13.



(a) An example of a duplicate with a difference in color



(b) An example of a duplicate with a difference in cropping

Figure 13: The examples of duplicate memes

An important aspect of the experiment is the sample size: 40 participants is a powered sample for detecting only large differences (large effects). As our analysis shows, we cannot reject the null hypothesis, which means that we do not have enough evidence to conclude that there are statistically significant differences between the groups. This does not confirm that there are no differences, but rather that any true effects are either smaller than our study could detect, or would require a larger sample to detect reliably. These limitations mean that the results should be interpreted as exploratory rather than conclusive, as the risk of type II error (we fail to detect true differences) is considerable. The mixed directionality (which group has a higher metric median) of the effects across metrics may reflect the noise inherent in small samples rather than meaningful patterns.

10 Future Work

Further work can be done in several directions. First, the accuracy of each of the steps can be improved to produce a more accurate final result. Second, qualitative interviews can provide more insight into users' cognitive processes during meme selection in order to improve the retrieval system. Using the system in more natural settings, such as social networking platforms, can test its scalability and validity. Future work could also experiment with the parameters of the retrieval system: initial sampling, number of memes to select, and number of memes recommended. Using different clustering algorithms that favor fewer or a fixed number of clusters can also provide valuable insights.

The framework we develop can also be applied to the field of emotion analysis and the development of more accurate recommendation systems than the one we present, unimodal or multimodal. The work can be extended to other modalities, such as video or sound, to cover more forms of memes and human interactions with them. For each of these studies, the number of participants should be increased.

11 Conclusion

The comprehensive null results across all user experience metrics suggest that the emotion component in multimodal iterative retrieval system does not significantly impact subjective satisfaction metrics, perceived system performance and behavioral engagement measures under current experimental conditions. While we observe small effects, the study has limited power (sample size = 40) and therefore we cannot reject the null hypothesis. We lack sufficient evidence to conclude there are statistically significant differences between the groups with and without emotion component. Also technical limitations in the system's implementation (e.g., emotion extraction, interaction design) may lead to constrained observable outcomes. These findings highlight that there is still room for improvement in both experimental design and technical execution. This work can serve as a foundation for future technical and social research. The work can be continued in several directions, including implementing the system with more natural settings (e.g. social media) and increasing/changing modality (e.g. by adding audio modality).

A Appendix

A.1 Questions

1. Which number do you see on the page?
2. How relevant or interesting was the initial set of memes?
    - 1 — Not at all
    - 5 — Very much
3. How satisfied are you with the final recommendations?
    - 1 — Very dissatisfied
    - 5 — Very satisfied
4. How much did recommendations improve from your initial selection?
    - 1 — Got worse
    - 5 — Improved dramatically
5. How relevant were the recommendations to your preferences?
    - 1 — Not relevant
    - 5 — Extremely relevant
6. How many iterations did you need to be completely satisfied? If you weren't satisfied, please enter 100.
7. How diverse were the recommendations?
    - 1 — Too repetitive
    - 5 — Very diverse
8. How consistently did the recommendations improve with each refinement?
    - 1 — Got worse each time
    - 5 — Improved every time
9. How quickly did the system adapt to your feedback?
    - 1 — Too slow
    - 5 — Instantly responsive
10. Would you recommend the system to friends?
    - 1 — Absolutely no
    - 5 — Absolutely yes

11. How likely are you to use such a recommendation system in the future?
    - 1 — Very unlikely
    - 5 — Very likely
12. How enjoyable was you experience with the system?
    - 1 — Not enjoyable
    - 5 — Very enjoyable
13. Please select your gender.
    - Male
    - Female
    - Other
14. Please select your age.
    - 18-24
    - 25-34
    - 35-44
    - 45+

A.2 Implementation details

In this section we will briefly discuss the implementation details of the framework. Due to the academic limitations of this thesis, we consider only open-source models. For the extraction of visual, textual and emotion embeddings from text we use *openai/clip-vit-base-patch32* as a CLIP model (Face, a). At the time of writing, the model is available for import through the transformers (Face, c) library and does not require additional installations. As a RoBERTa we use *SamLowe/roberta-base-go_emotions* which is also available via the transformers library (Face, b). As an OCR-tool we utilize EasyOCR library (Pallets, a) which shows more accurate text recognition in comparison with more widely known Tesseract (Smith, 2007) in the case of memes. The backend of the website is written in Python using the Flask library (Pallets, b). It contains the retrieval algorithm we describe in the section 5.3. For the Frontend we use JavaScript, HTML and CSS, the screenshots of user interface can be found in section User experiment. Our code is available in a public repository on GitHub.

Bibliography

Tariq Habib Afridi, Aftab Alam, Muhammad Numan Khan, Jawad Khan, and Young-Koo Lee. *A Multimodal Memes Classification: A Survey and Open Research Issues*. Springer International Publishing, 2021. URL https://link.springer.com/chapter/10.1007/978-3-030-66840-2_109.

Sure Ajibade, Fatima Sunoloso, and Abiodun Okunola. The impact of dimensionality reduction techniques on machine learning algorithm efficiency. *International Journal of Advanced Computer Science and Applications*, 2024. URL http://dx.doi.org/10.14569/IJACSA.2021.0120480.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. URL https://ieeexplore.ieee.org/document/8269806.

Ming Chen. Emotion analysis based on deep learning with application to research on development of western culture. *Frontiers in Psychology*, 2022. URL https://pubmed.ncbi.nlm.nih.gov/36186353/.

Tao Chen, Xiangnan He, and Min-Yen Kan. Context-aware image tweet modelling and recommendation, 2016. URL https://dl.acm.org/doi/10.1145/2964284.2964291.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 1988. URL https://api.pageplace.de/preview/DT0400.9781134742707_A24419622/preview-9781134742707_A24419622.pdf.

Pádraig Cunningham and Sarah Jane Delany. k-nearest neighbour classifiers - a tutorial. *ACM Comput. Surv.*, 2021. URL https://dl.acm.org/doi/abs/10.1145/3459665.

Marie Delacre, Christophe Leys, Yuhan Mora, and Daniël Lakens. Why psychologists should by default use Welch's t-test instead of Student's t-test. *International Review of Social Psychology*, 2017. URL https://rips-irsp.com/articles/10.5334/irsp.82.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020. URL https://www.researchgate.net/publication/343300732_GoEmotions_A_Dataset_of_Fine-Grained_Emotions.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*

(Long and Short Papers). Association for Computational Linguistics, 2019. URL https://aclanthology.org/N19-1423/.

Linhao Dong, Shuang Xu, and Bo Xu. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018. URL https://ieeexplore.ieee.org/document/8462506.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, 2020. URL https://iclr.cc/virtual/2021/oral/3458.

Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 1992. URL https://psycnet.apa.org/record/1993-00392-001.

Hugging Face. openai/clip-vit-base-patch32 · hugging face, a. URL https://huggingface.co/openai/clip-vit-base-patch32.

Hugging Face. Samlowe/roberta-base-go_emotions · hugging face, b. URL https://huggingface.co/SamLowe/roberta-base-go_emotions.

Hugging Face. Transformers, c. URL https://huggingface.co/docs/transformers/v4.17.0/en/index.

Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. G*Power 3: A flexible statistical power analysis program. *Behavior Research Methods*, 2007. URL https://link.springer.com/article/10.3758/BF03193146.

Hadar Fisher, Nigel Jaffe, Kristina Pidvirny, and et al. Using natural language processing to track negative emotions in the daily lives of adolescents. *Research Square*, 2025. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC12047991/.

Brendan J. Frey and Delbert Dueck. Mixture modeling by affinity propagation. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*. MIT Press, 2005. URL https://www.researchgate.net/publication/221619627_Mixture_Modeling_by_Affinity_Propagation.

Shang Gao, Mohammed Alawad, M. Todd Young, John Gounley, Noah Schaefferkoetter, Hong Jun Yoon, Xiao-Cheng Wu, Eric B. Durbin, Jennifer Doherty, Antoinette Stroup, Linda Coyle, and Georgia Tourassi. Limitations of transformers on clinical text classification. *IEEE Journal of Biomedical and Health Informatics*, 2021. URL https://ieeexplore.ieee.org/document/9364676.

Geoffrey Everest Hinton and Ruslan Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 2006. URL https://www.science.org/doi/abs/10.1126/science.1127647.

John M. Hoenig and Dennis M. Heisey. The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, 2001. URL https://www.vims.edu/people/hoenig_jm/pubs/hoenig2.pdf.

Yen-Hao Huang, Ssu-Rui Lee, Mau-Yun Ma, Yi-Hsin Chen, Ya-Wen Yu, and Yi-Shin Chen. Emotionx-idea: Emotion bert – an affectional model for conversation, 2019. URL https://www.researchgate.net/publication/335258426_EmotionX-IDEA_Emotion_BERT_--_an_Affectional_Model_for_Conversation.

CJ Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, 2015. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14550.

EunJeong Hwang and Vered Shwartz. MemeCap: A dataset for captioning and interpreting memes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023. URL https://www.researchgate.net/publication/370982120_MemeCap_A_Dataset_for_Captioning_and_Interpreting_Memes.

Ian T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2016. URL https://pubmed.ncbi.nlm.nih.gov/26953178/.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: detecting hate speech in multimodal memes. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020. Curran Associates Inc. URL https://www.researchgate.net/publication/341311100_The_Hateful_Memes_Challenge_Detecting_Hate_Speech_in_Multimodal_Memes.

Vasiliki Kougia and John Pavlopoulos. Multimodal or text? retrieval or BERT? benchmarking classifiers for the shared task on hateful memes. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*. Association for Computational Linguistics, 2021. URL https://www.researchgate.net/publication/353491550_Multimodal_or_Text_Retrieval_or_BERT_Benchmarking_Classifiers_for_the_Shared_Task_on_Hateful_Memes.

Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton. Deep learning. *Nature*, 2015. URL https://www.nature.com/articles/nature14539.

Yang Li, Kangbo Liu, Ranjan Satapathy, Suhang Wang, and Erik Cambria. Recent developments in recommender systems: A survey [review article], 2024. URL https://ieeexplore.ieee.org/document/10494051.

Henry B. Mann and Donald R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 1947. URL https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-18/issue-1/On-a-Test-of-Whether-one-of-Two-Random-Variables/10.1214/aoms/1177730491.full.

Ilija Milosavljević. *THE PHENOMENON OF THE INTERNET MEMES AS A MANIFESTATION OF COMMUNICATION OF VISUAL SOCIETY - RESEARCH OF THE MOST POPULAR AND THE MOST COMMON TYPES*. MEDIA STUDIES AND APPLIED ETHICS, 2020. URL https://www.researchgate.net/publication/340051035_THE_PHENOMENON_OF_THE_INTERNET_MEMES_AS_A_MANIFESTATION_OF_COMMUNICATION_OF_VISUAL_SOCIETY_-RESEARCH_OF_THE_MOST_POPULAR_AND_THE_MOST_COMMON_TYPES.

Pallets. Easyocr, a. URL https://github.com/JaidedAI/EasyOCR.

Pallets. Welcome to flask — flask documentation, b. URL ttps://flask.palletsprojects.com/en/3.0.x/.

Bo Pang and Lillian Lee. *Opinion Mining and Sentiment Analysis*. Now Foundations and Trends, 2008. URL OpinionMiningandSentimentAnalysis.

Binghui Peng, Srini Narayanan, and Christos Papadimitriou. On limitations of the transformer architecture, 2024. URL https://api.semanticscholar.org/CorpusID:267636545.

James Pennebaker, Ryan Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015, 2015. URL https://www.researchgate.net/publication/282124505_The_Development_and_Psychometric_Properties_of_LIWC2015.

Flor Miriam Plaza-del Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. ELRA and ICCL, 2024. URL https://aclanthology.org/2024.lrec-main.506/.

Robert Plutchik. *Emotion: A Psychoevolutionary Synthesis*. Harper & Row, 1980.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021. URL https://icml.cc/virtual/2021/oral/9194.

Ali Reza Sajun, Imran Zualkernan, and Donthi Sankalpa. A historical survey of advances in transformer architectures. *Applied Sciences*, 2024. URL https://www.mdpi.com/2076-3417/14/10/4316.

Lanyu Shang, Yang Zhang, Yuheng Zha, Yingxi Chen, Christina Youn, and Dong Wang. Aomd: An analogy-aware approach to offensive meme detection on social media. *Information Processing Management*, 2021. URL https://www.sciencedirect.com/science/article/abs/pii/S0306457321001527.

Parikshit Sharma. Advancements in ocr: A deep learning algorithm for enhanced text recognition. *International Journal of Inventive Engineering and Sciences*, 2023. URL https://www.ijies.org/portfolio-item/f42630812623/.

Shivam Sharma, Ramaneswaran Selvakumar, Md Shad Akhtar, and Tanmoy Chakraborty. Emotion-aware multimodal fusion for meme emotion detection. *IEEE Transactions on Affective Computing*, 2024. URL https://ieeexplore.ieee.org/abstract/document/10475492.

Limor Shifman. Memes in a digital world: Reconciling with a conceptual troublemaker. *Journal of Computer-Mediated Communication*, 2013. URL https://doi.org/10.1111/jcc4.12013.

Ray Smith. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, 2007. URL https://ieeexplore.ieee.org/document/4376991.

Evangelos A. Stathopoulos, Anastasios I. Karageorgiadis, Alexandros Kokkalas, Sotiris Diplaris, Stefanos Vrochidis, and Ioannis Kompatsiaris. A query expansion benchmark on social media information retrieval: Which methodology performs best and aligns with semantics? *Computers*, 2023. URL https://www.mdpi.com/2073-431X/12/6/119.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019. URL https://aclanthology.org/P19-1656/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. URL https://dl.acm.org/doi/10.5555/3295222.3295349.

Albin Wagener. *Memes, Emotional Engagement and Politics*. De Gruyter, 2024. URL https://www.researchgate.net/publication/382912279_Memes_Emotional_Engagement_and_Politics.

Ha-Ram Won, Yunju Lee, Jae-Seung Shim, and Hyunchul Ahn. A hybrid collaborative filtering model using customer search keyword data for product recommendation. In *2019 18th IEEE International Conference On Machine Learning And*

Applications (ICMLA), 2019. URL `https://ieeexplore.ieee.org/document/8999276`.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*. Chinese Information Processing Society of China, 2021. URL `https://link.springer.com/chapter/10.1007/978-3-030-84186-7_31`.

Declaration of Authorship

Ich erkläre hiermit gemäß §9 Abs. 12 APO, dass ich die vorstehende Bachelorarbeit selbstständig verfasst bzw. erbracht habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt worden sind. Ferner, dass die digitale Fassung der gedruckten Ausfertigung ausnahmslos in Inhalt und Wortlaut entspricht und dass zur Kenntnis genommen wurde, dass die digitale Fassung einer durch Software unterstützten, anonymisierten Prüfung auf Plagiate unterzogen werden kann.

Ort, Datum

Unterschrift