Burghardt, M. & Weiß, C. (Hrsg.): 4th Workshop InfDH,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023 11

A Corpus of Memes from Reddit: Acquisition, Preparation and First Case Studies

Thomas Schmidt[1], Fabian Schiller[2], Mathias Götz[3] and Christian Wolff[4]

Abstract: We present a corpus of memes and their textual components that were acquired from the popular meme platform r\memes, a subreddit of Reddit and one of the major outlets of online meme culture. The corpus consists of the most popular memes from 2013-2021 on the platform and we acquired 11,701 memes and 280,351 text tokens. We conduct several case studies focused on diachronic analysis to highlight the possibilities of the corpus for research in internet studies and online culture. We examine the general activity on the platform throughout the years and identify a significant increase in meme production beginning 2017. Results of sentiment analysis show a tendency towards memes with positively classified texts. The analysis of most frequent words per half-year spotlights the importance of certain cultural events for meme culture (e.g. the 2016 US election). Using the LIWC to analyze swear and sexual words shows an overall decrease in the usage of these words pointing to an increased moderation of the platform. The corpus is publicly available for the research community for further studies.

Keywords: memes; internet studies; corpus; natural language processing; sentiment analysis; Reddit

1 Introduction

Memes have become a popular media type in today's internet culture. Bauckhage [Ba11] defines a meme as "content or concepts that spreads rapidly among internet users" and Davison [Da12] extends this definition to memes as "a piece of culture, typically a joke, which gains influence through online transmission". One of the most popular meme types is the usage of an image with a text, usually consisting of one or two lines that formulate a joke or punch line (see figure 1 for a conceptual template, figure 2 for an example). The image is referred to as image macro or meme template and they are usually reused with different text changing based on context, goal, and situation of the overall meme. We refer to the different forms based on differing texts of the image macros as meme incarnations.

Due to the large popularity and widespread usage of memes in social media, memes have become a topic of interest for various humanities branches like internet and cultural studies. From a methodological point of view, research is focused on qualitative / hermeneutic methods with a rather low number of memes in qualitative case studies [Sh12, Sh14, Sh13,

[1] Media Informatics Group, University of Regensburg, thomas.schmidt@ur.de
[2] Media Informatics Group, University of Regensburg, fabian.schiller@ur.de
[3] Media Informatics Group, University of Regensburg, mathias-christian.goetz@ur.de
[4] Media Informatics Group, University of Regensburg, christian.wolff@ur.de

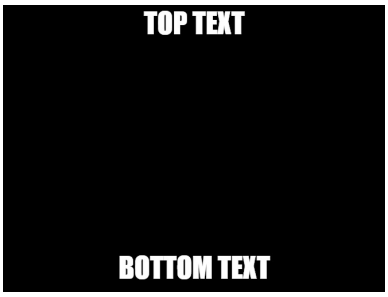

Fig. 1: Typical format of an image macro meme.

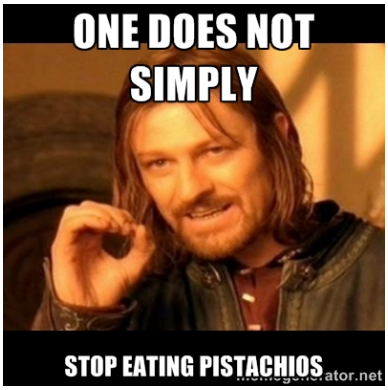

Fig. 2: Example of a meme of the corpus (posted in the second half-year of 2015).

Os15, HL16, Mc19]. The application of computational methods for meme analysis in digital humanities (DH) is rather rare [Sc20b]. In natural language processing (NLP), memes often serve as rich data source in the context of other primary tasks e.g. to evaluate algorithms for hate language detection [Ki21], AI-based meme generation [AT18, Sa20] or sentiment analysis [Sh20]. Sherratt [Sh22] proposes the focus on multimodal methods of online memes but also argues that there is still a lack of public well-structured corpora for the creation of necessary knowledge bases. First developments towards that goal can be seen in the data science and machine learning community with the publication of datasets on open data platforms[5] but they often lack structured metadata, text captions or point to deleted online content. Schmidt et al. [Sc20b] acquired and analyzed around 8,000 meme incarnations of 16 image macros from the platform Know Your Meme. Vyalla and Udandarao [VU20] created a corpus of over 1 million meme text captions but limited to 128 image macro classes and use this to develop a meme generation system.

In this paper, we present the methodological approach and the results of a project for the acquisition and analysis of a corpus of memes. Our goal is to publish a multimodal

[5] https://www.kaggle.com/datasets/electron0zero/memegenerator-dataset

(images and text) corpus with rich metadata of this unique media type to support research in media, internet, and cultural studies. Furthermore, we present first case studies with various methods of computational text analysis to showcase the potential of this corpus for further studies. We primarily examine the development of textual meme content across time to explore influence and interactions with general culture.

2 Corpus

We refer to the subreddit r\memes[6] as our source of acquisition. The platform, which started in 2008, has more than 26 million subscribers, 22 million posts and is the 10th largest subreddit on Reddit.[7] r\memes is one of the major curated outlets for memes of the image macro type in the English language. Users can post memes as submissions on the platform and receive comments, upvotes and downvotes. As one of the largest Reddit communities a strict moderation is in place deleting non-meme content or content violating the terms of use (e.g. racist content).

We have built a Python script using the Reddit pushshift fast-API[8] to scrape the top 1,000 memes (as measured by the highest number of upvotes) for each half-year (or less than 1,000 if less memes were posted) in the timespan of 2013-2021. We decided for 2013 as a starting point since from 2008 to 2012, this subreddit was not very active or popular.[9] We scraped the memes as image files from the corresponding image url (see figure 2 for an example) including all additional metadata as json-data: upvotes, upvote ratio, number of comments, submission time, full link to the meme submission) and created based on this data a corpus in csv-format. However, we removed meme submissions of these top 1,000 per half-year if they contained no links to images or links to deleted content (noise). The remaining data was extended with the textual content of the memes using the optical character recognition (OCR) of the Google Cloud Vision API.[10] The OCR returns valid results as analysis showed since the clear and large text poses no challenge to modern OCR. However, the OCR also included non-caption text like words in the background or the url of the meme creation tool used. While we tried to filter these in the preprocessing, certain OCR artefacts will certainly remain in the final corpus. Overall, we acquired a corpus of 11,701 meme-images, metadata, and the corresponding textual content.

Table 1 illustrates general corpus statistics. Token and sentence-based analysis was performed via Spacy.[11] The 11,701 memes sum up to 280,351 tokens of text. On average a meme consists of 24.5 tokens and 1.8 sentences. Figure 3 illustrates the distribution among memes in the corpus per half-year.

[6] https://www.reddit.com/r/memes/
[7] All data as of May 5th 2023.
[8] https://github.com/pushshift/api
[9] See https://knowyourmeme.com/memes/sites/rmemes for more information.
[10] https://cloud.google.com/vision/docs/ocr
[11] https://spacy.io/

Type / Metric	absolute	AVG	SD	Min	Max
Tokens	280,351	24.5	37.7	1	2,120
Sentences	20,340	1.8	4.0	1	397
Punctuation	23,773	2.0	10.4	0	800

Tab. 1: General corpus statistics.

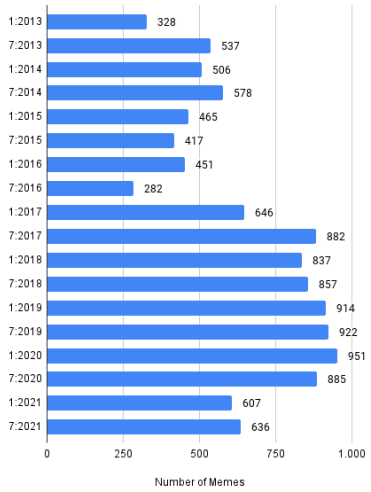

Fig. 3: Distribution of memes among each half-year.

3 Case Studies

To highlight the possibilities of this corpus, we have performed a first exploratory analysis with a focus on diachronic analysis of the textual content of the memes based on half-year sub-corpora of this corpus. Additional data relating to these analyses as well as the corpus itself are publicly available for the research community.[12]

3.1 Most Frequent Words

To gain an overview of the topics that are discussed, and the terms used on memes, we analyzed most frequent words per half-year and most frequent words in the corpus as a whole using word clouds after stop word removal. The overall word cloud of all memes points to swear words and general terms as being very frequent (figure 4).

However, word clouds for specific half-years do mirror important topics in society (see figure 5 and 6 for examples). More word clouds can be examined on our GitHub repository.

[12] https://github.com/lauchblatt/reddit_memes

Fig. 4: Word cloud of the overall meme corpus.

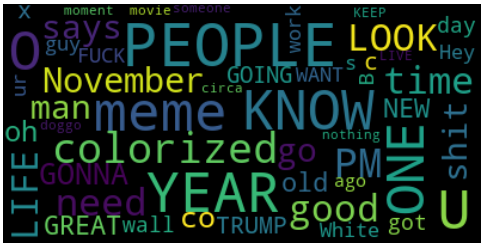

Fig. 5: Word cloud for all memes of the second half-year of 2016 highlighting the 2016 presidential election as important topic of meme culture that year (“wall”, “Trump”, “November”).

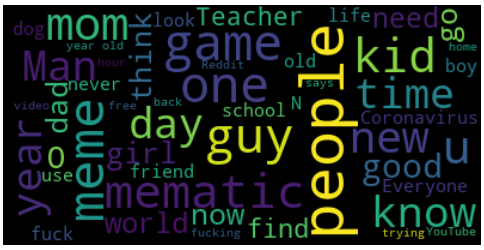

Fig. 6: Word cloud for all memes of the first half-year of 2020 including “Coronavirus” as one of the most frequent words.

3.2 Sentiment Analysis

We also explored sentiment analysis using the Python toolkit Vader[13] [HG14] which produces an average sentiment value for each text of each meme and is a recommended lexicon-based sentiment analysis often used for the analysis of social media content [SKW20, Mo20]. Following the recommendations of Vader, we regard an average sentiment value higher than 0.05 as positive and lower than -0.05 as negative. The final classification shows that 36% (3,866 memes) of the memes are positive, 33% (4,178) negative and 31% (3,674) neutral. The higher percentage of positively classified memes is in line with the general idea of the humoristic often light-hearted content of the memes. As an example, figure 7

[13] https://github.com/cjhutto/vaderSentiment

shows the meme with the highest sentiment score (the most positive meme according to the classification).

Fig. 7: Meme with the highest sentiment score of the overall corpus (0.983).

We also calculated the average sentiment score per half-year to get an understanding of the sentiment progression over time. We identified that the half-year overall values of the textual content of the memes sums up to a rather neutral value (between -0.05 and 0.05). However, as figure 8 illustrates, there is a development towards more positive sentiment beginning in 2016. This is in line with the general rise of popularity of r\memes but also a more serious moderation and professionalization of the platform including a more controlled removal of racist and trolling content that is likely to be classified as negative.

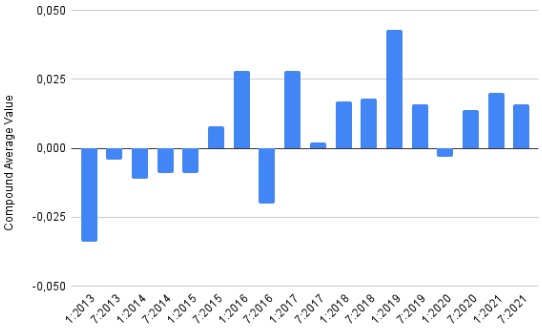

Fig. 8: The average sentiment score per half-year across time.

3.3 LIWC Analysis

Finally, we used the English version of the Linguistic Inquiry Word Count lexical resource[14] (LIWC; [Bo22]) to investigate the development of certain word categories across time. The LIWC is a psychologically validated lexical resource that offers lists of words for certain

[14] https://www.liwc.app/.

topics and while it is primarily used in psychology, it has also shown to be beneficial in the context of language analysis in social media by analyzing the overall proportions of word categories [Mo20]. The word lists consist of a fixed list of words including their inflections and words extended with regular expressions. We focus on the word categories swear (130 words, regex-forms not included, like bullshit, sucks, dumb and other slurs) and sexual (131 words like erotic, naked and sexual slurs) as those pose interesting insights in the context of the previously performed sentiment analysis. Please note however that this dictionary approach is limited due to the increased moderation of the subreddit (deleting and adjusting memes with explicit content) and also approaches by the community to avoid swear words by using placeholders (e.g. b*llsh*t for bullshit). Nevertheless, calculating the average proportion of these word categories among all memes results in an average of 0.19 for swear words and 0.07 for sexual words.

Looking at the proportion of these words across the time confirms previous findings of the sentiment analysis and shows the importance of swear words for this media type with average proportion values up to 0.42 in the second half-year of 2014 (figure 9). We also identified a moderate decrease from the peaks in 2014 to 2021 that is in line with the sentiment development and might be a sign of the increased professionalization and moderation of the platform deleting and managing memes with content like racial slurs that are part of the swear word category. Our results show that words with sexual connotations are not as important and frequent compared to swear words and rather rare in certain years (e.g. 2018, 2019). However, they do peak just recently in 2021.

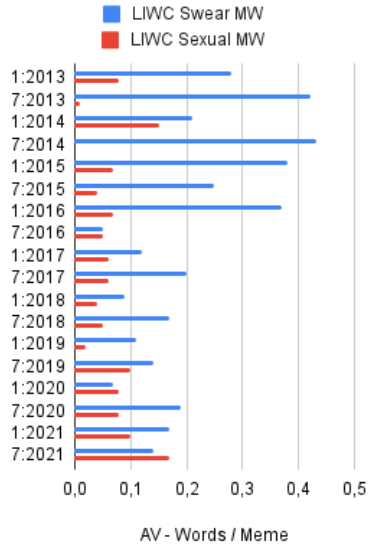

Fig. 9: Distribution of swear and sexual terms across the time. X-axis is the average proportion of word categories across all memes.

4 Limitations & Future Work

While these analyses are exploratory at the moment, they show the potential of this corpus and possible research ideas. We want to extend these cases studies by addressing several limitations:

We used Spacy, a standard NLP tool and general purpose solution to create tokens and sentences. However, we noticed that the special vocabulary and syntax used on memes pose challenges to this standard approach. We evaluated the functionality of the token and sentence segmentation with some random samples and while we got the impression that it works well most of the time, we also found problems dealing with slang, complex punctuation and URLs. In many cases, specific sentence segmentation decisions are open to interpretation. We plan to further analyze this problem by defining syntactical and token based units for our meme corpus and examining results of multiple segmentation tools in more detail.

The applied sentiment analysis as lexicon-based approach is rather limited and we did not evaluate the approach. We intend to annotate sub-corpora of the meme corpus with sentiment to evaluate methods and apply more sophisticated machine learning methods including more multifaceted emotion prediction [SDW21]. Furthermore, we intend to perform more fine-grained sentiment analysis considering the time spans e.g. sentiment progressions per month to get a more in-depth view of the data. We also argue that we can gain further insights of meme culture by analyzing other word categories of the LIWC. The analysis of the gender-based categories female and male offer potential for in-depth analysis in the context of gender studies [Sc20a]. We plan to investigate the method of topic modeling to examine the diachronic development of important topics discussed on memes in more depth. Additionally, we did not perform yet any major analysis into the interaction of certain metadata with sentiment or LIWC analysis e.g. the correlation of sentiment and popularity. We intend to uncover various analysis ideas of that branch in future work.

Furthermore, we agree with Sherratt [Sh22] that memes are a multimodal medium and we want to include methods of computer vision (e.g. image classification, object detection) to further analyze our corpora. This is further supported by the increased usage of video snippets, GIFs and similar content in recent meme culture. Currently, we are in the process to increase the overall corpus with more memes from r\memes but also other subreddits to get a broader view of the phenomenon and extend the corpus also with video based memes. Recent applications of computer vision methods applied in DH fields like theatre studies [SBW19, SW21], film studies [Sc21, Ho19, SK22, EKSW22] and internet studies [Sc20c] offer interesting possibilities in this context of meme analysis.

To our knowledge, only few structured meme corpora are publicly available [Sc20b] and we were able to gain interesting insights about the Reddit meme culture in regard to sentiment, topics and platform moderation that have implication for internet, media and social media studies.

Bibliography

[AT18] AbelL.Peirson, V.; Tolunay, E. M.: Dank Learning: Generating Memes Using Deep Neural Networks. ArXiv, June 2018.

[Ba11] Bauckhage, Christian: Insights into Internet Memes. Proceedings of the International AAAI Conference on Web and Social Media, 5(1):42–49, 2011. Number: 1.

[Bo22] Boyd, Ryan; Ashokkumar, Ashwini; Seraj, Sarah; Pennebaker, James: The Development and Psychometric Properties of LIWC-22. February 2022.

[Da12] Davison, Patrick: 9. The Language of Internet Memes. In: 9. The Language of Internet Memes, pp. 120–134. New York University Press, March 2012.

[EKSW22] El-Keilany, Alina; Schmidt, Thomas; Wolff, Christian: Distant Viewing of the Harry Potter Movies via Computer Vision. In: Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022). Uppsala, Sweden, pp. 33–49, 2022.

[HG14] Hutto, C.; Gilbert, Eric: VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Proceedings of the International AAAI Conference on Web and Social Media, 8(1):216–225, May 2014. Number: 1.

[HL16] Highfield, Tim; Leaver, Tama: Instagrammatics and digital methods: studying visual social media, from selfies and GIFs to memes and emoji. Communication Research and Practice, 2(1):47–62, January 2016. Publisher: Routledge _eprint: https://doi.org/10.1080/22041451.2016.1155332.

[Ho19] Howanitz, Gernot; Bermeitinger, Bernhard; Radisch, Erik; Gassner, Sebastian; Rehbein, Malte; Handschuh, Siegfried: Deep Watching - Towards New Methods of Analyzing Visual Media in Cultural Studies. July 2019.

[Ki21] Kiela, Douwe; Firooz, Hamed; Mohan, Aravind; Goswami, Vedanuj; Singh, Amanpreet; Fitzpatrick, Casey A.; et al.: The Hateful Memes Challenge: Competition Report. In (Escalante, Hugo Jair; Hofmann, Katja, eds): Proceedings of the NeurIPS 2020 Competition and Demonstration Track. volume 133 of Proceedings of Machine Learning Research. PMLR, pp. 344–360, December 2021.

[Mc19] McCulloch, Gretchen: Because Internet: Understanding the New Rules of Language. Penguin, July 2019. Google-Books-ID: Tfo5DwAAQBAJ.

[Mo20] Moßburger, Luis; Wende, Felix; Brinkmann, Kay; Schmidt, Thomas: Exploring Online Depression Forums via Text Mining: A Comparison of Reddit and a Curated Online Forum. In: Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task. Association for Computational Linguistics, Barcelona, Spain (Online), pp. 70–81, December 2020.

[Os15] Osterroth, Andreas: Das Internet-Meme als Sprache-Bild-Text. IMAGE, 22:26–46, 2015.

[Sa20] Sadasivam, Aadhavan; Gunasekar, Kausic; Davulcu, Hasan; Yang, Yezhou: , memeBot: Towards Automatic Image Meme Generation, April 2020. arXiv:2004.14571 [cs].

[SBW19] Schmidt, Thomas; Burghardt, Manuel; Wolff, Christian: Toward Multimodal Sentiment Analysis of Historic Plays: A Case Study with Text and Audio for Lessing's Emilia Galotti. In (Navarretta, Costanza; Agirrezabal, Manex; Maegaard, Bente, eds): Proceedings of the Digital Humanities in the Nordic Countries 4th Conference (DHN 2019). volume 2364 of CEUR Workshop Proceedings, CEUR-WS.org, Copenhagen, Denmark, pp. 405–414, March 2019.

[Sc20a] Schmidt, Thomas; Engl, Isabella; Herzog, Juliane; Judisch, Lisa: Towards an Analysis of Gender in Video Game Culture: Exploring Gender specific Vocabulary in Video Game Magazines. In: Proceedings of the Digital Humanities in the Nordic Countries 5th Conference (DHN 2020). Riga, Latvia, pp. 333–341, 2020.

[Sc20b] Schmidt, Thomas; Hartl, Philip; Ramsauer, Dominik; Fischer, Thomas; Hilzenthaler, Andreas; Wolff, Christian: Acquisition and Analysis of a Meme Corpus to Investigate Web Culture. In (Estill, Laura; Guiliano, Jennifer, eds): 15th Annual International Conference of the Alliance of Digital Humanities Organizations (DH 2020), Conference Abstracts. Ottawa, Canada, July 2020.

[Sc20c] Schmidt, Thomas; Mosiienko, Anastasiia; Faber, Raffaela; Herzog, Juliane; Wolff, Christian: Utilizing HTML-analysis and computer vision on a corpus of website screenshots to investigate design developments on the web. Proceedings of the Association for Information Science and Technology, 57(1):e392, 2020.

[Sc21] Schmidt, Thomas; El-Keilany, Alina; Eger, Johannes; Kurek, Sarah: Exploring Computer Vision for Film Analysis: A Case Study for Five Canonical Movies. In: 2nd International Conference of the European Association for Digital Humanities (EADH 2021). Krasnoyarsk, Russia, September 2021.

[SDW21] Schmidt, Thomas; Dennerlein, Katrin; Wolff, Christian: Emotion Classification in German Plays with Transformer-based Language Models Pretrained on Historical and Contemporary Language. In: Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. Association for Computational Linguistics, Punta Cana, Dominican Republic (online), pp. 67–79, November 2021.

[Sh12] Shifman, Limor: An anatomy of a YouTube meme. New Media & Society, 14(2):187–203, March 2012. Publisher: SAGE Publications.

[Sh13] Shifman, Limor: Memes in Digital Culture. MIT Press, October 2013.

[Sh14] Shifman, Limor: The Cultural Logic of Photo-Based Meme Genres. Journal of Visual Culture, 13(3):340–358, December 2014. Publisher: SAGE Publications.

[Sh20] Sharma, Chhavi; Bhageria, Deepesh; Scott, William; PYKL, Srinivas; Das, Amitava; Chakraborty, Tanmoy; Pulabaigari, Viswanath; Gambäck, Björn: SemEval-2020 Task 8: Memotion Analysis- the Visuo-Lingual Metaphor! In: Proceedings of the Fourteenth Workshop on Semantic Evaluation. International Committee for Computational Linguistics, Barcelona (online), pp. 759–773, December 2020.

[Sh22] Sherratt, Victoria: Towards Contextually Sensitive Analysis of Memes: Meme Genealogy and Knowledge Base. In (Raedt, Lud De, ed.): Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22. International Joint Conferences on Artificial Intelligence Organization, pp. 5871–5872, July 2022.

[SK22] Schmidt, Thomas; Kurek, Sarah: Der Einsatz von Computer Vision-Methoden für Filme - Eine Fallanalyse für die Kriminalfilm-Reihe Tatort. In: DHd 2022 Kulturen des digitalen Gedächtnisses. 8. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum"(DHd 2022). Potsdam, Germany, March 2022.

[SKW20] Schmidt, Thomas; Kaindl, Florian; Wolff, Christian: Distant Reading of Religious Online Communities: A Case Study for Three Religious Forums on Reddit. In: Proceedings of the Digital Humanities in the Nordic Countries 5th Conference (DHN 2020). Riga, Latvia, pp. 157–172, 2020.

[SW21] Schmidt, Thomas; Wolff, Christian: Exploring Multimodal Sentiment Analysis in Plays: A Case Study for a Theater Recording of Emilia Galotti. In: Proceedings of the Conference on Computational Humanities Research 2021 (CHR 2021). Amsterdam, The Netherlands, pp. 392–404, 2021.

[VU20] Vyalla, Suryatej Reddy; Udandarao, Vishaal: Memeify: A Large-Scale Meme Generation System. In: Proceedings of the 7th ACM IKDD CoDS and 25th COMAD. CoDS COMAD 2020, Association for Computing Machinery, New York, NY, USA, pp. 307–311, January 2020.