Acquisition and Analysis of a Meme Corpus to Investigate Web Culture

Thomas Schmidt, Philipp Hartl, Dominik Ramsauer, Thomas Fischer, Andreas Hilzenthaler & Christian Wolff

Media Informatics Group, University of Regensburg, Germany

{firstname.lastname@ur.de}

15th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2020
Ottawa, Canada
July 20-25, 2020
Conference Abstracts

Keywords: Memes, Web Culture, Text Mining, Sentiment Analysis, Topic Modeling

Abstract.
Memes are a popular part of today's online culture reflecting current developments in pop-culture, politics or sports and are created and shared in large scale on a daily basis. We present first results of an ongoing project about the study of online-memes via computational Distant Reading methods. We focus on the meme type of image macros. Image macros memes consists of a reusable image template with a top and/or bottom text and are the most common and popular meme types. We gather a corpus for 16 of the most popular image macros memes by crawling the platform knowyourmeme.com thus creating a corpus consisting of 7840 memes incarnations and their corresponding metadata. Furthermore, we gather the text of the memes via OCR and make this corpus publicly available for the research community. We explore the application of various text mining methods like Topic Modeling and Sentiment Analysis to analyze the language, the topics and the moods expressed via online memes.

Link to version in the conference abstracts: https://dh2020.adho.org/wp-content/uploads/2020/07/590_AcquisitionandAnalysisofaMemeCorpustoInvestigateWebCulture.html
Link to the poster on Humanities Commons: http://dx.doi.org/10.17613/mw0s-0805

Please cite as:

Schmidt, T., Hartl, P., Ramsauer, D., Fischer, T., Hilzenthaler, A. & Wolff, C. (2020). Acquisition and Analysis of a Meme Corpus to Investigate Web Culture. In *15th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2020, Conference Abstracts*. Ottawa, Canada.

1 Introduction

Memes are a popular part of today's online culture, reflecting current developments in pop culture, politics or sports. That has led various scholars in the humanities and other research areas to examine the importance and role of memes (Shifman, 2014; Highfield & Leaver, 2016; McCulloch, 2019).

Bauckhage (2011) defines the term *Meme* as "contents or concepts that spread rapidly among Internet users". While memes with solely visual content are rising in popularity, one of the most common and historically important meme types is the "image macro" which consists of a reusable image template with a top and/or bottom text (figure 1).

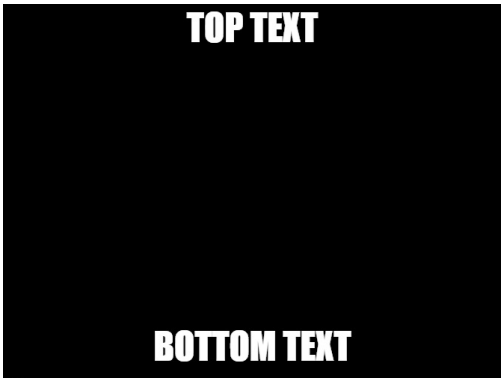

Figure 1: Typical format of an image macro

There are various established image templates (see figure 2 for an example) and with the growth of social media, new ones are constantly emerging. We differentiate between the meme template, which is basically just the image of a meme and the meme derivatives, which are the multiple manifestations of a meme template differing regarding the text of the meme.

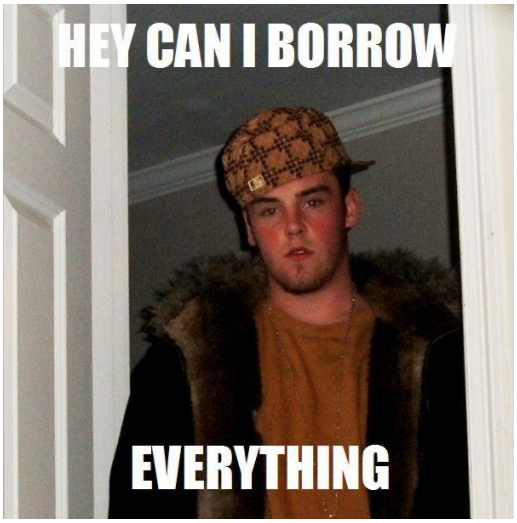

Figure 2: Example of "Scumbag Steve", a popular image macro meme

Although memes are distributed and shared in large quantities, the majority of current research on memes is qualitative, e.g. analyzing patterns and stylistic rules of a small number of memes (Shifman, 2012; 2014; Osterroth, 2015). Since image macros typically have a textual component, we want to use computational methods of *Distant Reading* (Moretti, 2013) to analyze memes in a large-scale approach. Our project aims to identify developments of the content and sentiment of memes in a diachronic way but is also based on image templates. In this paper we present first results on the corpus acquisition workflow we have developed as well as the application of general text analysis, topic modeling, and sentiment analysis on the overall corpus.

2 Corpus Creation

To create a corpus for our analysis we use the platform *knowyourmeme*[1]. It is one of the most popular platforms for uploading memes and offers the possibility to search for specific meme categories like image macros. Furthermore, the different derivatives of a meme template are collected under a single entry and are enriched with metadata. For our first analysis, we focus on 16 of the historically most popular templates and we have implemented a scraper to access the links to the meme derivatives and metadata. To get the text of the memes we use *Google Cloud OCR* on the images gathered. Our final dataset consists of 7.840 meme derivatives, metadata and the text (see table 1). In line with the conference theme of the open data movement, this corpus is publicly available for the research community to download and use[2]. Please note that we only include memes with English language since this is the language *knowyourmeme* is focused on.

#	Template	Amount of macros	Total tokens	Average tokens per macro
1	'Dat Ass	403	2122	5
2	Based God	129	950	7
3	[10] Guy	213	1930	9
4	Ancient Aliens	639	5494	9
5	Bad Luck Brian	674	7411	11
6	Ermahgerd	330	3538	11
7	Grumpy Cat	509	5654	11
8	Philosoraptor	649	7923	12
9	Scumbag Steve	474	5658	12
10	Ridiculously Photogenic Guy	322	4370	14
11	Annoying Facebook Girl	234	4056	17
12	Joseph Ducreux / Archaic Rap	930	16980	18
13	First World Problems	346	6584	19
14	Overly Attached Girlfriend	205	3862	19
15	Big Chungus	111	2661	24
16	Xzibit Yo Dawg	629	20175	32

Table 1: Corpus description

3 Corpus Analysis

For all approaches, we have implemented various preprocessing steps commonly used in text mining (e.g. lemmatization). Figure 3 shows a word cloud of the most frequent words of the entire corpus:

[1] https://knowyourmeme.com/
[2] https://github.com/lauchblatt/Memes_DH2020

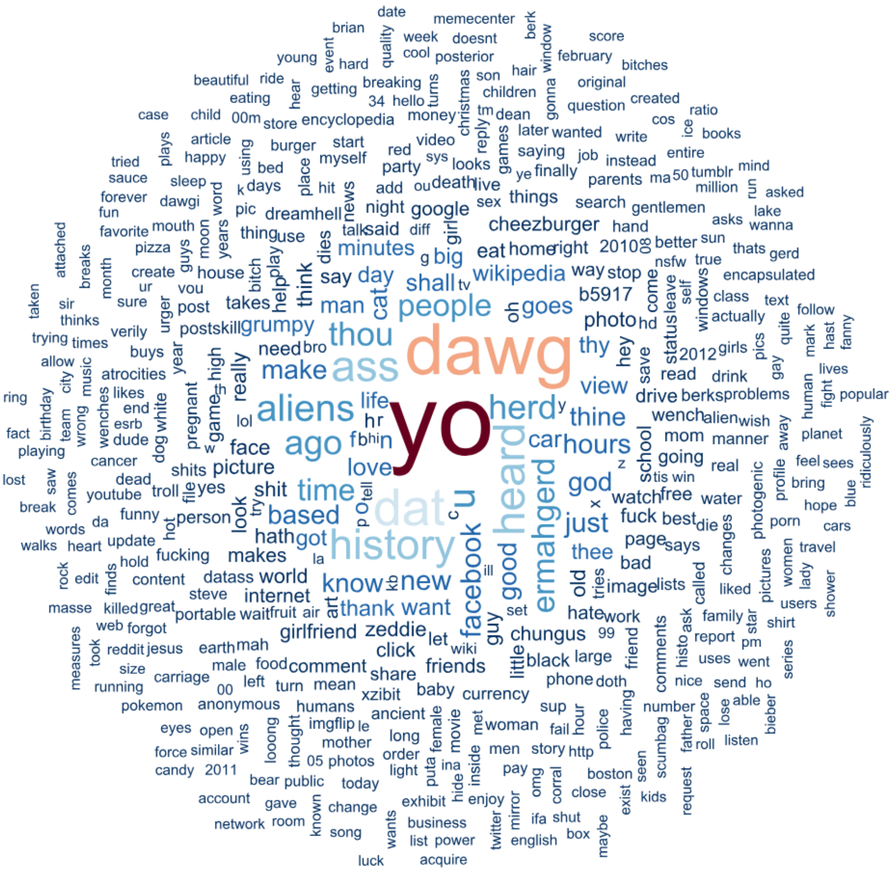

Figure 3: Word cloud of the most frequent words of the entire corpus

The word cloud illustrates the specifics of meme language like the dominance of slang. One can also identify some word patterns that are consistently used on some memes like e.g. “yo” and “dawg” being common words for the “Xzibit Yo Dawg”-meme template.

For topic modeling, we use *Latent Dirichlet Allocation* (LDA, Blei et al., 2003) to calculate 16 LDA topics. LDA topics are described by typical word clusters within documents (here: meme derivatives), thus topic modeling produces lists of words that appear frequently together in documents. Our assumption is that every meme template is equivalent to a topic, thus we chose the number of our image macros as topic number (16).

Figure 4 illustrates our results for the topic modeling analysis:

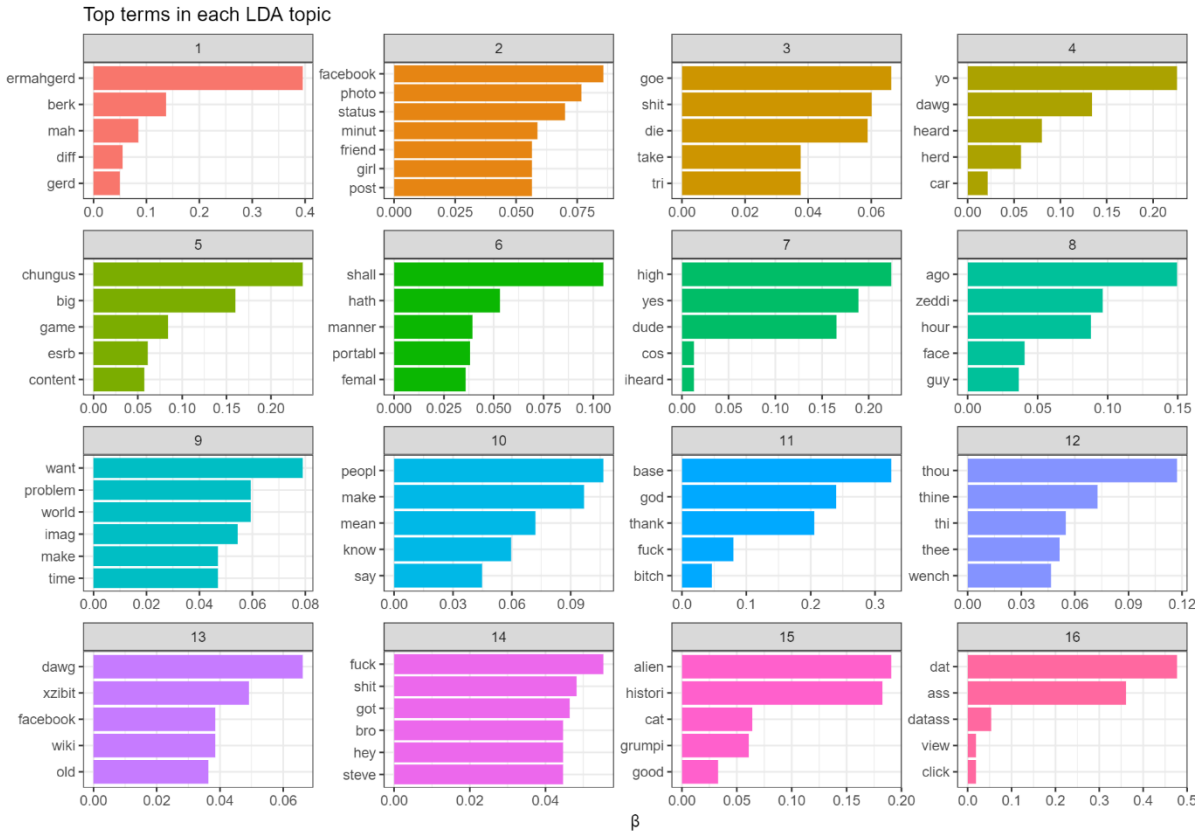

Figure 4: 16 LDA topics of the corpus; with the five most contributing tokens per topic

As expected, most of the topics are expressions of a single meme template (e.g. topic 1 for the "Ermahgered" or topic 3 for the "XZibit Yo Dawg" meme template) which shows that some memes consist of homogenous and reoccurring word patterns. However, there are some overlaps like topic 15, expressing words common in the "Ancient Alien" and "Grumpy Cat" meme. We plan to investigate these memes in future work in more detail to examine the similarities they have in more detail.

For the sentiment analysis, we use the sentiment lexicon "*Bing*" (Liu, 2012; Liu & Zhang, 2012) for polarity (positive, negative) and the *NRC Word-Emotion Association Lexicon* (Mohammad & Turney, 2013) for emotions. Figure 5 shows which words contribute the most to a specific overall sentiment:

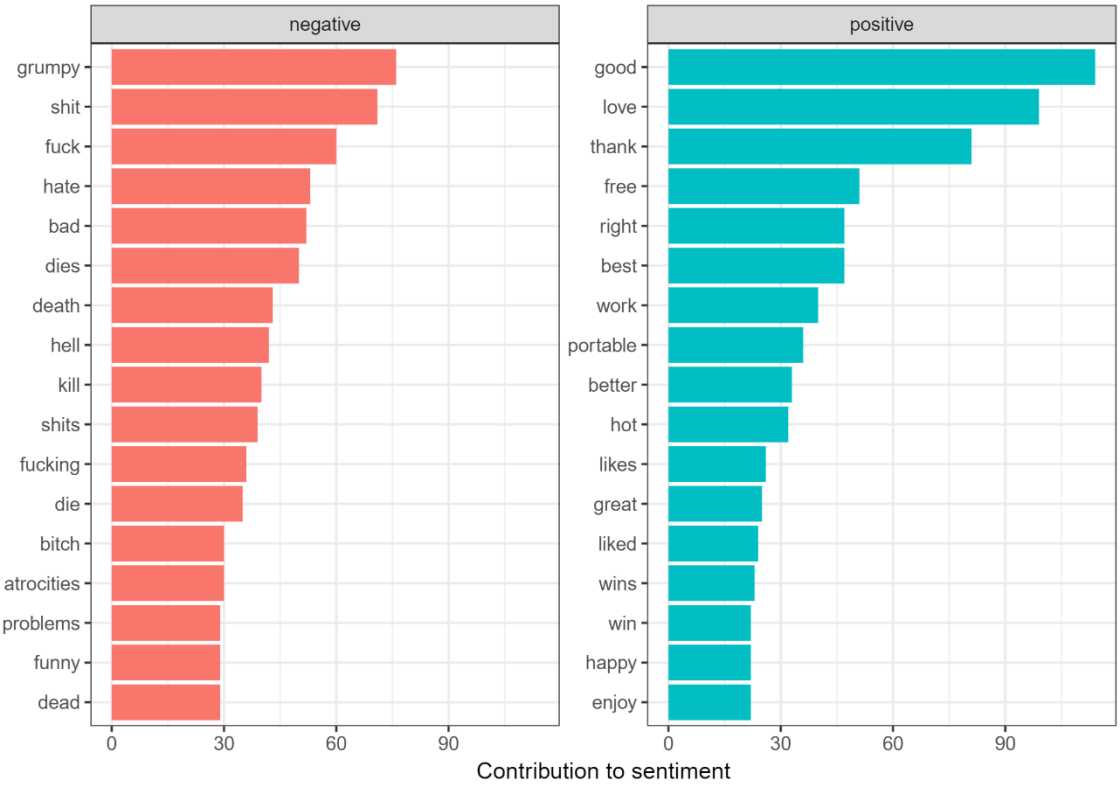

Figure 5: Most important tokens contributing to the overall sentiment in the corpus

Though we cannot report the results of the sentiment and emotion comparisons among the memes in detail, one outlier meme we want to highlight is the "Ancient Alien" meme. The "Ancient Alien" meme has the highest values for disgust and fear, which is a fitting result since those memes are often used in the context of conspiracy theories.

Currently, our research is at an early stage and exploratory. In future work, we want to continue our analysis by increasing our corpus, filtering out noise during the acquisition and gather more metadata to perform diachronic and meme based analysis and comparisons considering sentiments and topics.

References

Bauckhage, C. (2011, July). Insights into internet memes. In *Fifth International AAAI Conference on Weblogs and Social Media.*

Bauckhage, C., Kersting, K., & Hadiji, F. (2013, June). Mathematical models of fads explain the temporal dynamics of internet memes. In *Seventh International AAAI Conference on Weblogs and Social Media.*

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.

Davison, P. (2012). The language of internet memes. *The social media reader*, 120-134.

Highfield, T., & Leaver, T. (2016). Instagrammatics and digital methods: Studying visual social media, from selfies and GIFs to memes and emoji. *Communication Research and Practice*, 2(1), 47-62.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1), 1-167.

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data* (pp. 415-463). Springer, Boston, MA.

McCulloch, G. (2019). *Because Internet: Understanding the new rules of language*. Riverhead Books.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3), 436-465.

Moretti, F. (2013). *Distant reading.* Verso Books.

Osterroth, A. (2015). Das Internet-Meme als Sprache-Bild-Text. *Image*, 22, 26-46.

Shifman, L. (2012). An anatomy of a YouTube meme. *New media & society*, 14(2), 187-203.

Shifman, L. (2014). *Memes in digital culture.* MIT press.

Shifman, L. (2014). *The cultural logic of photo-based meme genres. Journal of Visual Culture*, 13(3), 340-358.