Studying Web Culture:

Distant Reading of Online Memes

Note: This contribution is intended to be a poster contribution

Memes are a popular part of today's online culture, reflecting current developments in pop-culture, politics or sports. Bauckhage (2011) defines the term *Meme* as "contents or concepts that spread rapidly among Internet users". One of the most common and popular meme types is the "image macro" which consists of a reusable image template with a top and/or bottom text (figure 1).
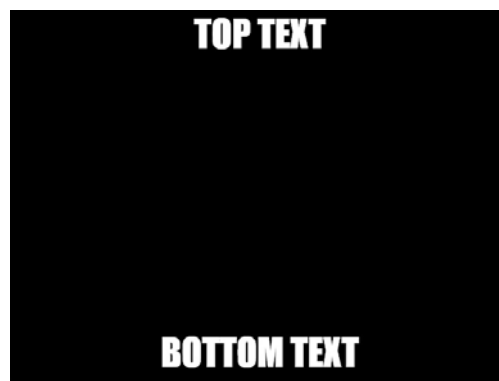

Figure 1: Typical format of an image macro

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 in regards of 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 is of qualitative research, e.g. analyzing patterns and stylistic rules of a small number of memes (Shifman, 2012; 2014; Osterroth, 2015). Since memes typically have a textual component, we want to use computational methods of *Distant Reading* (Moretti, 2013) to analyze memes in a large-scale approach to gain insights about the language and the content of this specific internet phenomenon.

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 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 7840 meme derivatives, metadata and the text (see table 1). This corpus is publicly available for the research community[2].

---

1   https://knowyourmeme.com/
2
    https://docs.google.com/spreadsheets/d/1ESecAyzfWvXTELfIbWwZEXWE_tGKq8wmTTpOTS
    SrDJY/edit?usp=sharing

#	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

Besides general text analysis, we focus on topic modeling to identify the topics of various memes and sentiment analysis to analyze the expressed moods of the memes. 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:

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. We select 16 topics since this fits well with our number of image macros. 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. Figure 4 illustrates our results for the topic modeling:
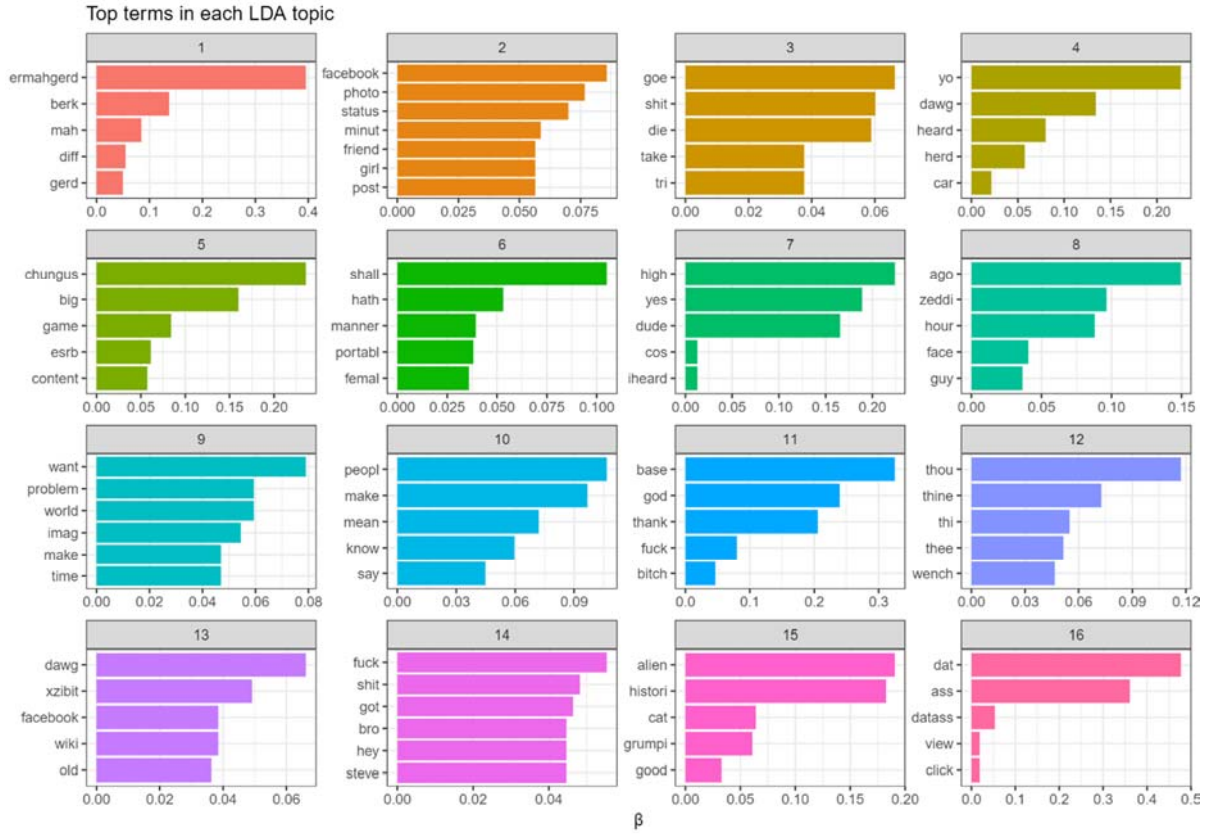
Figure 4: 16 LDA topics of the corpus; with the five most contributing tokens per topic

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.

For the sentiment analysis, we use the sentiment lexicon *"Bing"* (Liu, 2012; Liu & Zhang, 2012) for polarity (positive, negative) and the *NRC* (Mohammad & Turney, 2013) for emotions. Figure 5 shows which words contribute the most to a specific overall sentiment:
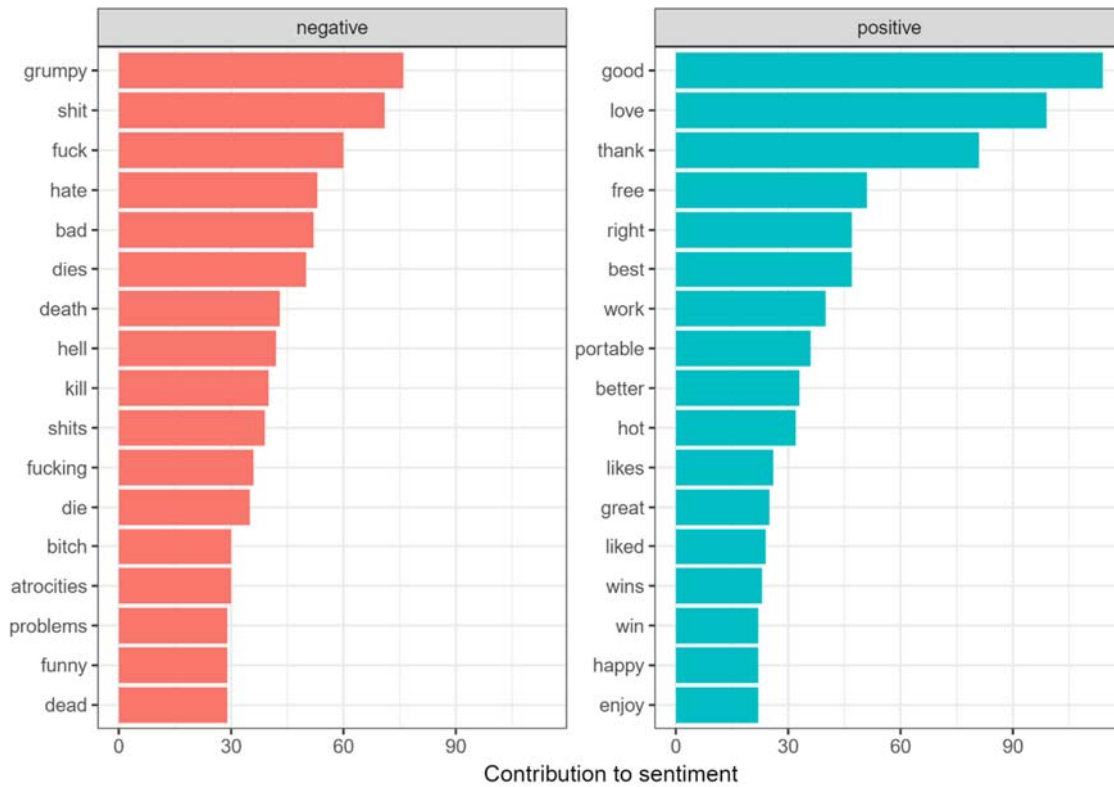
Figure 5: Most important tokens contributing to the overall sentiment in the corpus

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.

In future work, we want to continue our analysis by increasing our corpus, filtering out noise during the acquisition and using other *Distant Reading* visualization techniques like diachronic analysis to explore the linguistic manifestations and developments of internet memes in more detail.

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.

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.

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