trivadis

Sparx S1E3: Article

Using artificial intelligence to detect hate on the internet

By Maximilian Janisch

Watch Episode

The easy and fast spread of hateful content online is a big problem of our internet age. Unsurprisingly, there is thus great interest in systems that can detect and delete such content in automated fashion. This is a particularly difficult task for memes, which are a combination of image and text. Thanks to progress in the field of multimodal deep learning and the development of more powerful machine learning methods, hateful memes can already be identified automatically with a very low error rate.

Information in the real world often comes in multiple modalities. We see objects, hear sounds, feel texture, smell odors, and taste flavors. A prime example of multimodal information are memes, which are a combination of an image with text created most often for humor. According to YPulse, 30% of 13-35-year-olds share more than 7 memes per week.[1] According to Google Trends, memes are about as popular as Jesus.[2] It is no wonder thus that memes are used increasingly for hateful intentions online. For example, Pepe the Frog became a big part of the American alt-right movement. Automatic detection of hateful memes is hard because the image and text are often related in intricate ways based on societal conventions that are hard to understand for a computer.

Great interest from facebook

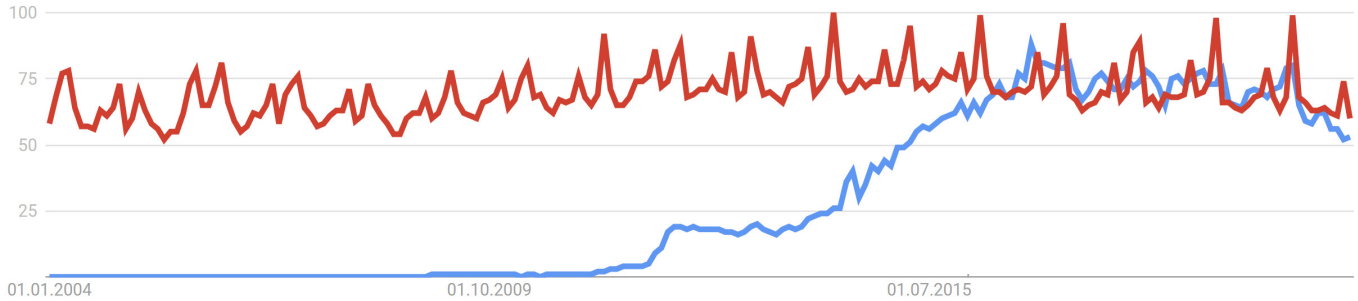
Less than 5 years ago, automatic detection of hateful memes was completely impossible for computers. However, interest for this problem was renewed in recent times, in part because companies such as Facebook have great interest in automatic tools that can reliably detect hateful memes. For example, in May 2020, Facebook released the Hateful memes challenge[3] with a cash prize of 100'000 US-dollars for whoever can detect hateful memes the best.

The goal of AI in this area

In the future, we would like AI to improve its understanding of sophisticated content that is currently understood only by humans. This requires that the meaning of words and images can be recognised not only independently of each other, but also in combination. Researchers are working in particular on systems for identifying multimodal hate speech. Facebook provides AI researchers with a dataset of 10'000 multimodal examples for this purpose.

Error rate of only 10%

Thanks to advances in the field of multimodal deep learning and the development of more powerful methodology in machine learning, current state-of-the-art algorithms can detect hateful memes correctly with an error rate of about 10%. It should be noted that, since it is often disputable even for humans whether a given meme is hateful or not, an error rate of less than 10% is nearly impossible.



Why are memes such an important topic? Since 2016, about as many people search for memes (blue) on Google as for the term Jesus (red).

Where research is already successful

Multimodal deep learning, a highly active area of research, has also led to the solving of other "hard" problems, such as automatic image captioning or the automatic answering of questions about the content of an image. Some further areas where multimodal deep learning plays a big role are audio-visual speech recognition, summarization and indexing of videos, robotics (a robot has to "understand" its surroundings based on input from many different types of sensors) and medicine (for instance the automatic detection of organ issues is based on different forms of input data).

In conclusion, I believe that the advancement of multimodal deep learning is yet another demonstration that we are able to solve more and more sophisticated tasks with the help of mathematics and computer science, as was the quintessence of my talk on black holes and how to take pictures of them (watch the video with the QR code above).

[1] https://www.ypulse.com/article/2019/03/05/3-stats-that-show-what-memes-mean-to-gen-z-millennials/

[2] https://trends.google.com/trends/explore?date=all&q=memes,jesus

[3] https://ai.facebook.com/blog/hateful-memes-challenge-and-data-set/

About the author

Maximilian Janisch was born on August 8 2003 in Zurich, Switzerland. The high school Gymnasium Immensee accepted him as a student when he was 8 years old. At the age of 9 years, he passed the final high school exams in mathematics with top marks. In the summer 2018, he finished all high school exams at the age of 15 and got matriculated as a regular student at the University of Zurich where he studies mathematics. He has an IQ of 149+.

His interest in mathematics was joined early on by an interest in computer science. He is doing Machine Learning under the instruction of the Alpnach location of the Centre Suisse d'Electronique et de Microtechnique. He is particularly concerned with the classification, segmentation and generation of images using artificial neural networks. Together with his team, he took part in Facebook's Hateful memes challenge and ranked in the top 6% of over 3000 participants in the overall public leaderboard of phase 1 of the challenge.