Ukrainian Catholic University

Master Thesis

Meme Generation for Social Media Audience Engagement

Author:
Andrew Kurochkin

Supervisor:
PhD. Kostiantyn Bokhan

A thesis submitted in fulfillment of the requirements
for the degree of Master of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2020

Declaration of Authorship

I, Andrew KUROCHKIN, declare that this thesis titled, "Meme Generation for Social Media Audience Engagement" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: ______________________________________

Date: ______________________________________

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Master of Science

Meme Generation for Social Media Audience Engagement

by Andrew KUROCHKIN[1]

Abstract

In digital marketing, memes have become an attractive tool for engaging an online audience. Memes have an impact on buyers and sellers online behavior and information spreading processes. Thus, the technology of generating memes is a significant tool for social media engagement.

In this study, we collected new memes dataset of ∼650K meme instances, applied state of the art Deep Learning technique - GPT-2 model [1] towards meme generation, and compared machine-generated memes with human-created.

We justified that MTurk workers can be used for the approximate estimating of users' behavior in a social network, more precisely to measure engagement.

Generated memes cause the same engagement as human memes, which didn't collect engagement in the social network (historically). Still, generated memes are less engaging then random memes created by humans.

[1] https://andrewkurochkin.com

Acknowledgements

I would like to thank my supervisor Kostiantyn Bokhan who provided a lot of useful advice and ideas, and for interesting discussions.

I want to thanks to Kateryna Kotliarova, Nikita Pankiv and Hanna Pylieva for help and support.

Contents

List of Figures

List of Tables

List of Abbreviations

CC	**C**omputational **C**reativity
CSS	**C**omputational **S**ocial **S**cience
NN	**N**eural **N**network
CNN	**C**onvolutional **N**eural **N**etwork
RNN	**R**ecurrent **N**eural **N**etwork
LSTM	**L**ong-**S**hort **T**erm **M**emory
TF	**T**erm **F**requency
IDF	**I**nverse **D**ocument **F**requency
OCR	**O**ptical **C**haracter **R**ecognition
ES	**E**ffect **S**ize
HIT	**H**uman **I**ntelligence **T**ask
MTurk	**M**echanical Turk

List of Symbols

μ mean of a sample
σ standard deviation of a sample
α significance level
H_0 null hypothesis
H_a alternative hypothesis
χ^2 Chi-square test
w Effect Size for Chi-square test

Chapter 1

Introduction

1.1 Key definitions

We adhered to the definition of word **engage** as a verb, which means to occupy, attract, or involve (someone’s interest or attention) [2].

Meme template - is the background image, which is common across many meme instances. Each template has its background, history, style of humor, and the sentiment. Context and sentiment of the memes aren’t always humoristic; it can be related to other emotions: confession, irony, sarcasm, appreciation, etc. An example of the meme template is shown in Figure 1.1 (left) - "Afraid to Ask" Andy. Memes which are based on this template are captioned with various confessions of one’s ignorance in current events or common knowledge followed by the phrase "...and at this point I’m afraid to ask." [3].

Image macro [1] - is a simple form of the internet **meme** shown in Figure 1.1 (right). A meme is a combination of the meme template, top and bottom captions.

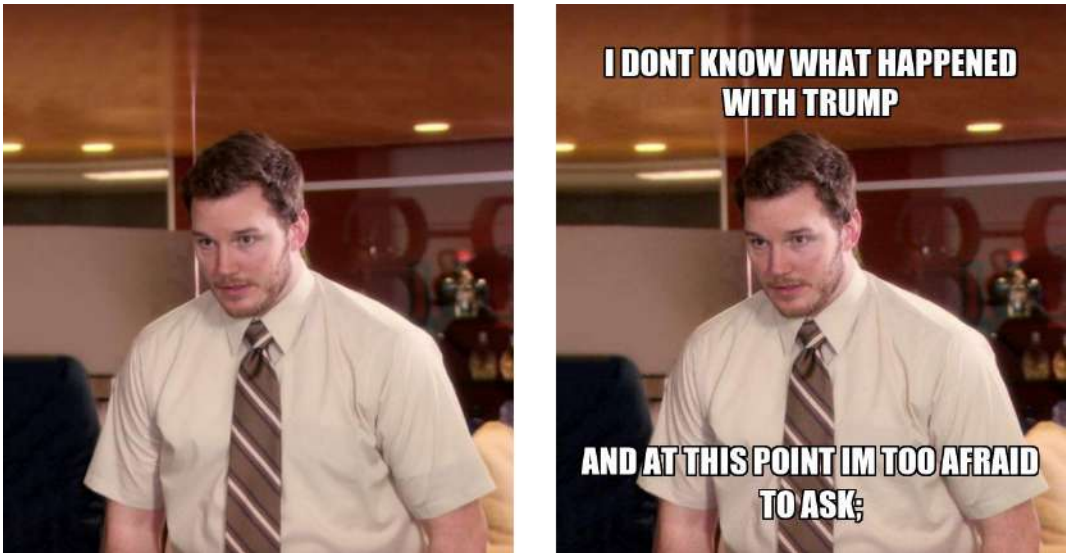

I DONT KNOW WHAT HAPPENED WITH TRUMP
AND AT THIS POINT IM TOO AFRAID TO ASK;

FIGURE 1.1: (left) Meme template named ’"Afraid to Ask" Andy’ and (right) its meme instance.
Meme instance was generated by the proposed system.

[1] We will use the phrase ‘Image macro’ and word ‘meme’ interchangeably in this work, as image macro is a form of a meme.

1.2 Background

This paper is related to the modeling of social media phenomena of the internet memes. Our research refers to the sub-disciplines named computational social science and computational creativity. Both CSS and CC are disciplines on intersection of a few fields, which makes this research multidisciplinary. We based our research on the social network Reddit. Generated memes (Figure C.1, Figure C.1) were assessed by expert evaluation method.

In 2018, digital consumers spent an average of 2 hours 22 minutes per day on social networks and messaging [4]. People use social networks to get news, professional information or consume content every day.

Research related to information gerrymandering shows that manipulation with information can be used to influence collective decision-making [5]. Social networks are mass media, they have impact on the society. Due to this fact, being present on social media is crucially important for all organizations which interact with their customers. Organizations put in great effort to be properly presented in social networks. Companies run massive information campaigns to promote their products or service. One of the primary purposes of this activity is to engage their audience and imply created emotional connection in the further buyer-seller relationship.

People who are involved in social media management track trending topics on a regular basis. Keeping an eye on the trends is only part of the work; another part is to create and post content which causes engagement in the audience.

1.3 Motivation

Different kinds and forms of information spread in social networks. Information can be in the form of text, video, audio or image. Image superimposed with sarcastic or humoristic text is one of the most common form of the internet meme [6].

To create a meme on the relevant topic, an author of the meme has to come up with a caption which will cause emotions in the audience, as well as select the image to supplement the meme. Once meme has been composed, an author creates a title or description (it depends on the specifics of a social network). When the post is ready to be published the right choice of the posting time is essential. This whole process is time-consuming.

Due to this fact we designed a solution to automate the creation of posts with the image macro to engage the audience. Engagement is a widely used to measure success for content in social media. Different actions can be used for the measurement of people engagement and its power [7]: views, likes, comments, shares, and reposts.

Generation of memes which engage the audience in the social media using image superimposed with English sentences is problem of computational creativity. The goal of computational creativity is to model, simulate or enhance creativity using computational methods [8].

In this paper we investigated how modern Deep Learning approaches for natural language processing cope with task of meme generation. In particularly, we applied a technique for natural language modeling - GPT-2 model [1] in solving a creativity problem which traditionally is a prerogative of a human.

1.4 Thesis structure

In Chapter 2 we reviewed some of related works. Proposed approach and methodology to collect dataset are defined in Chapter 3. Solution is described in Chapter 4. Evaluation of generated memes is presented in Chapter 5. Conclusions and vision of future work are written in the Chapter 6.

Chapter 2

Related work

Our paper relates mainly to three research topics: story generation and image captioning, meme generation, engagement and virality in social networks. They are briefly reviewed in this chapter.

2.1 Story Generation and Image Captioning

The problem of generating memes caption can be approached as a task to produce a short story based on the tags which set the storyline. In [9] authors approached the problem of hierarchical story generation where the model first generates a premise and then transforms it into a passage of text [9]. Researchers used sequence-to-sequence(seq2seq) models [10] with the usage of a fusion mechanism [11], as it had been shown that fusion mechanisms could help seq2seq models build dependencies between their input and output [9]. In the scope of this work an open-source sequence modeling toolkit was used FAIRSEQ [12].

The problem of generating natural language descriptions from the image has been studied in [13]. The approach to encode images with Convolutional Neural Network into vector embeddings was proposed. Decoder uses embeddings to generate sentence based on the Long-Short Term Memory network. The LSTM was chosen due to its ability to deal with vanishing and exploding gradients, which are a common problem of the Recurrent Neural Networks [14].

In other work, the authors concentrated on generating captions for images and videos with different styles [15]. In this work the authors utilized the FlickrStyle10K dataset and aimed it at the generation of humoristic or romantic image caption. The solution architecture was based on the encoder-decoder design with the modifications. The most valuable of which is the factored LSTM, it automatically distills the style factors in the monolingual text corpus [15]. A meme image can be the image with a penguin, but the main message or subject of the joke can be related to the awkward social situation [16]. Since the scene presented on the image can have different meaning than the meme template with its cultural background, an image caption does not solve our problem as the image is not the right source of information for memes caption.

2.2 Meme Generation

The language of Internet memes was modeled in [17], where an approach which is common in the economic modeling - copula methods [18][19] was applied. The authors claim that the predictive power of copula models could be used for joint modeling of raw images, text descriptions, and popular votes [17]. They employed reverse image search to get text information about the input image.

In [20], the results from [13] were adopted, however, with ResNet-152 [21] replaced CNN as a feature extraction method. In this work, authors proposed Funny Score, which was used as a loss function. Funny Score metric is based on the stars from the BoketeDB which display the degree of funniness of a caption evaluated by users of the Bokete [22].

The authors of [23] based their solution on [13] approach. In order to create image encoding, the system utilized a pre-trained Inception-v3 network. An important contribution of the work was a new beam search implementation in order to encourage diversity in the captions [23]. For the evaluation, perplexity and human assessment were used. Images or a combination of image and its name served as input data. The same image template can have various memes text related to it. Due to this fact, we claim that memes names have insufficient descriptive power. The authors mention that the separator between the text at the top and bottom can improve training results, so we take this into account in our work.

2.3 Engagement and Virality in Social Networks

In the [7] 4-level system of engagement classification based on human actions was proposed: from Level 1 - views, less public and more private expressions of engagement, Level 2 is "like" action, Level 3 - comment or share, to Level 4 - external posting, the most public level of engagement. The model for predicting Level 4 engagement was provided.

The study of memes propagation, evolution, and influence across the Web was done in the [24]. The authors used a processing pipeline based on perceptual hashing, clustering techniques, and a dataset of 160M images from 2.6B posts [24]. The researchers performed collection of the memes description based on the site Know Your Meme [25], which gives information about the memes concepts. This information was used for cluster analysis of the memes and to create their embeddings.

In [26] the authors analyzed how post popularity depends on the way the content is presented (the title), the community it was posted to, whether it has been seen before, and the time it was posted. The unique contribution of this work is the dataset which contains 132K submissions, only 16.7K of which were unique, whereas the others were resubmission. The mentioned circumstances make it possible to determine for the submission the influence of title, community, and posting time. Community and language models which help target social media audience were developed in [26]. In this paper, research interest was on the viral content, in the form of republished submissions.

In [27], the phenomenon of the image virality investigated from a computer vision perspective. Virality score based on the image resubmission was proposed. The neural network for the image virality prediction was created. The results show that in the task of image virality prediction based on the high-level image description (capturing semantic information), machine performs better than human. The model showed 68.10% accuracy relative to 60.12% of human performance.

Chapter 3

Approach

3.1 Dataset collection

As the aim of the project is to generate memes we required a large memes dataset which was unavailable. We collected dataset with combinations of meme template, image captions for the top section and bottom section, post title, score, and comments.

We decided to generate content in the English language since English is the most common language on the internet; it was in use by 25.2% among internet users for April 30, 2019 [28].

Social network Reddit is used as a source of data, since according to the official blog [29], it has 330 million users who generate 58 million votes, 2.8 million comments daily, and has 80% of the content in English. It is common practice in computational social science and social network analysis to use this platform for scientific researches.

We built our dataset on the data collected by Jason Baumgartner [30], which includes all Reddit posts and comments since 2005. Reddit has 850K communities related to different topics, but we were interested only in the posts which contain meme image. Due to this fact, a limited number of submissions were used, as we need only posts that contain meme images. However, the proposed pipeline can be applied to extract information from the Reddit for any period since 2005.

We faced the problem of meme submissions recognizing among the hundred of millions of others posts. This problem was solved by using subreddit (community) AdviceAnimals [31], which is dedicated only for posts with the image macro, memes that we were searching for. The community was founded in 2010 and is one of the most popular places for people to share their image macro. Usage of only one community can be concerned as its content can be biased, but as this community has 8 million members, we assumed that it can be a reliable source of memes for the research.

Data collection pipeline included next steps:

1. Download Reddit submissions and comments data for the subreddit [31].

   Dataset collected by Jason Baumgartner [30] is available in the Google Big Query [32], so data is publicly available and can be queried with no need to been previously downloaded. We selected all submissions and comments from the target subreddit [31] since the end of 2010 (time when it was created) till June of 2019.

2. Download images from the submissions were it was possible.

3. Recognize a meme template, as we utilize template id as part of context during generative model training.

Method proposed in the study [24] was applied:

(a) Embed each image as a vector with 64 elements based on its Perceptual Hash (pHash) [33]. Basically, pHash represents an image as vectors in the space, such that images that look similar for a human eye will be close to each other.

(b) Hamming distance was used to calculate distances between pHashes.

(c) Cluster images using DBSCAN [34], so memes with same or very similar image template will be related to the same cluster. We tuned original code [35] from the mentioned paper [24], our version can be found in the repository [**repo_template_recognition**].

We removed irrelevant images (not memes) or very rarely used templates, which can be observed less than 100 times for period.

4. Perform optical character recognition (OCR) to extract top and bottom pieces of text from the meme image.

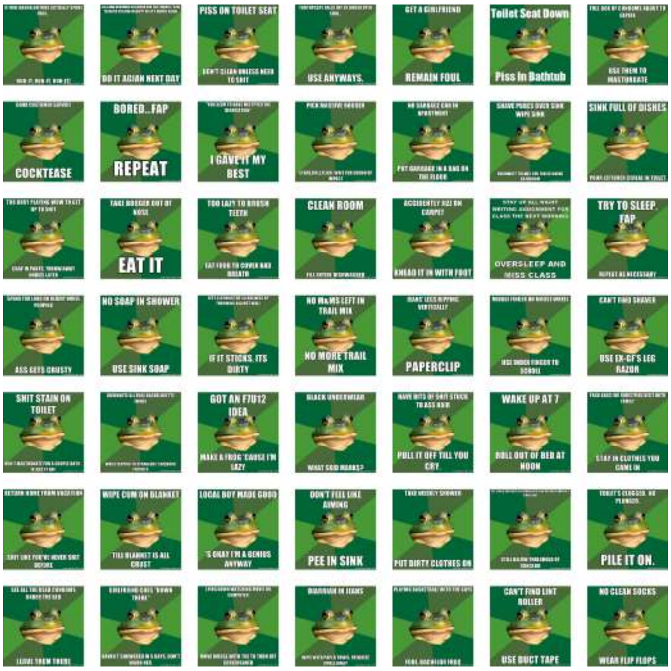


FIGURE 3.1: Memes grid prepared to OCR.

We tried a few approaches for this problem: Text Detection [36] with further Text Recognition [37], Tesseract OCR [38] and Microsoft Azure Computer Vision [39]. Based on the manual result evaluation Azure showed the best accuracy on the OCR task, so we used it in our research.

Major concern of chosen approach is price, Azure Recognize Text feature costs \$2.50 per 1000 transactions [40]. We had ∼650K images to be processed, the total price for OCR at this point was estimated in \$1625, which didn't fit our research budget. We found that Microsoft Azure defines a transaction as one JPEG, PNG, or BMP file with size ≤4Mb, and image dimensions must be at least 50 x 50, at most 4200 x 4200 [41]. We performed preprocessing for downloaded memes to standardize image dimensions to the average for the meme template; after that, we merged as many memes in one image as were possible in 4200px. We created grids with memes related to one template with whitespaces between images (it was found that paddings improve the accuracy of ORC). Example of such grid is in Figure 3.1.

The number of images after "gridding" became ∼25 times less; with Azure's student initial credit ($100), it was free to perform OCR for ∼650K images. Finally, we mapped text with coordinates on the grid to its meme image.

5. Merge memes and comments data in the dataset.

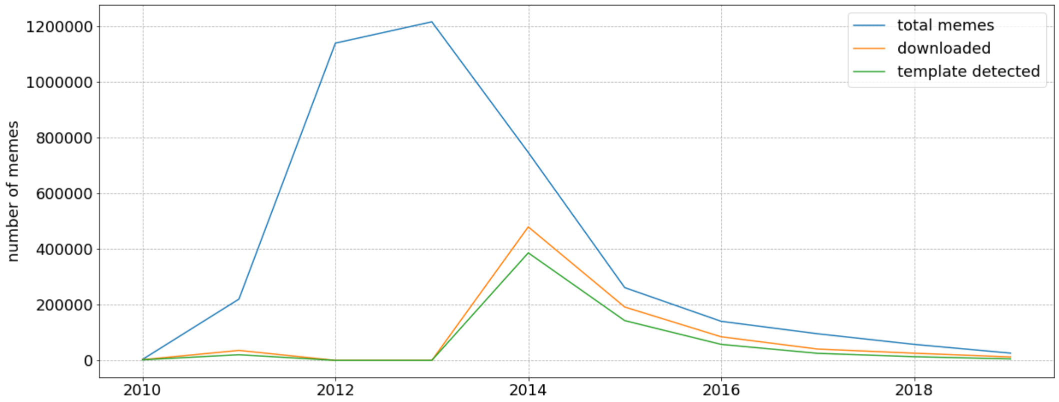


FIGURE 3.2: Dataset characteristics.

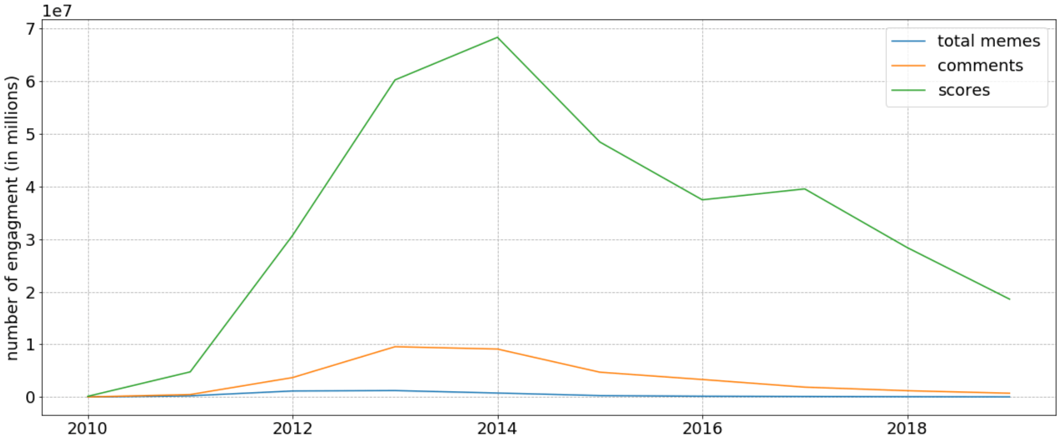


FIGURE 3.3: Volumes of engagement in dataset.

The final dataset size is ∼650K memes, other dataset properties shown in Figure 3.2. We didn't process submissions for 2012 and 2013 years, as they have more than 2 million submissions, and it was time-consuming to process such amount of data. We decided that the volume of the dataset of ∼650K will be enough to achieve stated objectives.

Volumes of engagement in the dataset shown in Figure 3.3. It is an interesting inference that in 2012 and 2013 were submitted more than 2 million of the memes, which is more than all other years combined together. Also, after 2013 we can observe a downward trend on the number of submissions, which can mean that this community became less popular, or users are more interested in making other types of submissions, or they use other social media to spread such content.

3.2 Engaging content creation

Our approach is to create a system to generate posts in social media, and each post should contain image macro (meme) and title. Created memes should cause feedback from the user in the form of the score (upvote in the Reddit); to do so, we use context information as model input. High-level overview of whole content generation pipeline can be seen in Figure 3.4.

Since the number of comments is a good indicator of content engagement level [7], we decided to use information from the comments. We assumed that comments from the most discussed submissions from the Reddit communities can be a source for an input context for the generator (Figure 3.4 (A)). The advantage of this source is that information can be extracted relatively easy.

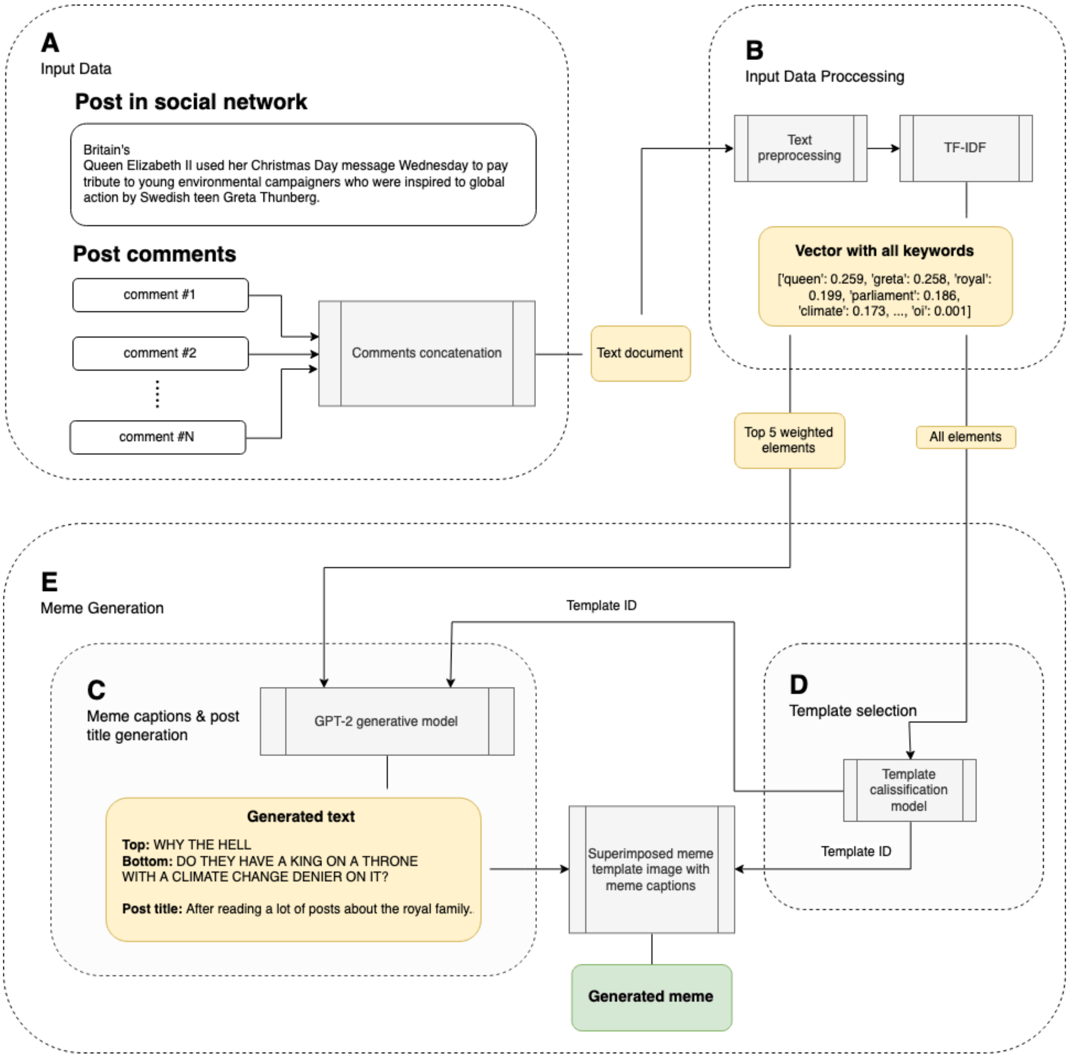

A
Input Data
Post in social network
Britain's
Queen Elizabeth II used her Christmas Day message Wednesday to pay tribute to young environmental campaigners who were inspired to global action by Swedish teen Greta Thunberg.
Post comments
comment #1
comment #2
comment #N
Comments concatenation
Text document
B
Input Data Proccessing
Text preprocessing
TF-IDF
Vector with all keywords
['queen': 0.259, 'greta': 0.258, 'royal': 0.199, 'parliament': 0.186, 'climate': 0.173, ..., 'oi': 0.001]
Top 5 weighted elements
All elements
E
Meme Generation
Template ID
C
Meme captions & post title generation
GPT-2 generative model
Generated text
Top: WHY THE HELL
Bottom: DO THEY HAVE A KING ON A THRONE WITH A CLIMATE CHANGE DENIER ON IT?
Post title: After reading a lot of posts about the royal family.
D
Template selection
Template calissification model
Template ID
Superimposed meme template image with meme captions
Generated meme

FIGURE 3.4: High-level overview of content generation pipeline.

One of our main assumptions is that comments written to post strongly reflect the idea of the post, as comments have explanatory power for the topic of the submission. Our approach is to represent comments with several the most relevant keywords and use them as input for our model for meme caption generation.

From all collected memes, we used only posts that got a number of comments is greater 1. This approach is similar to the approach used for dataset collection for GPT-2 training [1], where links from the posts with a number of the score more than 3 were used for model training. Each submission is related to one or a few topics, so we extracted information from the submission's comments. We used this information during GPT-2 finetuning to emphasize the ability to generate content (memes) related to a specific topic.

We decomposed the system into two models:

1. Model to choose the most appropriate image template based on text (comments) can be seen in Figure 3.4 (D).

2. Model to generate meme captions and post title based on the template and keywords (Figure 3.4 (C)).

3.3 Meme template selection

Meme template is crucial for setting-up the context of a joke. There are ∼23K of memes described in the biggest internet meme encyclopedia [25]. However, for our system, we used limited number of templates that were presented in our dataset.

An image (meme template) that reflects the idea of further posts based on comments should be chosen. We approached this problem as a multinomial classification based on the comments text (documents).

We transformed documents with TF-IDF [42] to present them as vectors of weighted words; weight depends on the importance of the word in the document (Figure 3.4 (B)).

$$tf(t,d) = \log(1 + freq(t,d)) \tag{3.1}$$

$$idf(t,D) = \log\left(\frac{N}{count(d \in D : t \in d)}\right) \tag{3.2}$$

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D) \tag{3.3}$$

TF-IDF (term frequency-inverse document frequency) - is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus [43], it is commonly used for the task of text information retrieval. In the essence this algorithm: calculates term frequency of a word t in a document d 3.1, calculates the inverse document frequency of the word t across a set of documents D 3.2, and final weight of word t for the document d is a product of two previous matrices 3.3.

Each meme was presented as target class (template) and related vector(based on comments). Model selection and training is described in the Section 4.1.

3.4 Memes and post title generation

The key problem in our work was to generate memes text and post title; we treated this problem as a language modeling task. Language modeling is usually framed as unsupervised distribution estimation from a set of examples $(x_1, x_2, ..., x_n)$ each composed of variable length sequences of symbols $(s_1, s_2, ..., s_n)$ [1].

In study [1] - GPT-2 text-generation model was presented. GPT-2 is based on multi-layer Transformer decoder [44] which is a variation of the transformer architecture [45] shown in Figure 3.5. GPT-2 beat state of the art results on various tasks without finetuning; it was trained on a huge variety of Internet texts, including Reddit. Due to this fact, we used GPT-2 as NN for our approach regarding its ability to catch text nature and generate coherent text. This model gives us with structured memory for handling long-term dependencies in text which [46] are important to generate meme captions based on the input text.

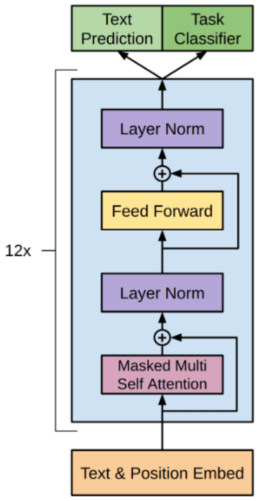


FIGURE 3.5: The Transformer architecture used in the GPT-2.
Source: [46].

We used the smallest GPT-2 117M which has 117 million of parameters, as this model is less computationally expensive than other GPT-2 modifications (355M, 774M, 1558M) and can be trained on the free of charge Google Colab [47] machines with GPUs in relatively short terms. Our approach is to finetune the GPT-2 117M model on our domain-specific data. Model input contains context in the form of template id and keywords, and outcome in the form of memes caption and posts title (Figure 3.4 (C)). Final data preparation and model training is described in the Section 4.2.

Chapter 4

Solution

4.1 Meme template selection

4.1.1 Data preparation

We defined template selection as a multinominal classification task, where each class represents a meme template.

The dataset has 650K meme instances; we left only memes that collected more than 1 comment, as we use comments for model training. After this manipulations we had 347 000 memes that were based on the 72 unique templates. We manually checked a few meme instances from each template to defined how many meme instances were detected incorrectly for this template, as our template recognition solution sometimes made errors. We found templates that have a high error rate and excluded them from the dataset. This led to noise reduction in the input data for model training.

A number of memes related to each template in the dataset are shown in Figure 4.1(a). It can be observed that templates are imbalanced. We filtered out templates that had less than 1000 memes related to it and made undersampling for templates to a maximum number of observations 3000 per template. After these manipulations, templates had become more balanced, which is shown in Figure 4.1(b). The final total amount of observations was 90194 meme instances based on the 38 unique templates. We used vector embeddings of concatenated submission comments as input for a classifier. We represented comments using sklearn implementation of TF-IDF [48].

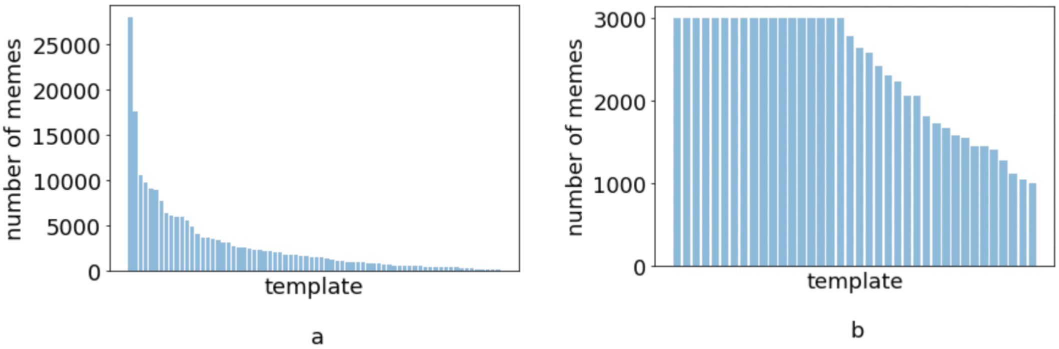


FIGURE 4.1: Memes count distribution across templates.

a - meme templates prior resampling (imbalanced), b - meme templates after resampling (balanced).
One bar is related to one meme template.

TABLE 4.1: Classifier models metrics.

Model	Metric			
	accuracy	precision	recall	F_1 score
LogisticRegression	**0.4998**	0.5181	**0.4998**	**0.4983**
LinearSVC	0.4899	0.4895	0.4899	0.4876
KNeighborsClassifier	0.3597	0.4066	0.3597	0.3662
RandomForestClassifier	0.3405	0.5309	0.3405	0.3502
AdaBoostClassifier	0.3015	0.5292	0.3015	0.3403
MultinomialNB	0.2946	**0.5929**	0.2946	0.3029

4.1.2 Model selection

We split data on train and test samples with a proportion of 80% - train, 20% - validation. We trained 6 models to approach classification problem. Models were trained with default hyperparameters and $random\_state = 0$, with two exceptions: RandomForestClassifier with hyperparameters $n\_estimators = 1000$, $max\_depth = 5$; and KNeighborsClassifier with hyperparameter $n\_neighbors = 38$.

Models metrics shown in Table 4.1. Based on the models performance, we chose multinomial logistic regression [49] as a system component for template selection. It gets a text as input and predicts the most probable meme template. We used logistic regression with accuracy 0.499% for the task of classification between 38 categories. Even though this accuracy is acceptable for our task, it can be improved as part of future work. In the worst case, when our model misclassifies and chooses the wrong meme template, it wouldn't be a problem as a result still can be used for the generative model. The only concern is that in approximately every second case person would use different meme template then our system predicts; however, this is not critical.

4.2 Meme generation

4.2.1 Data preparation

We finetuned GPT-2 with domain-specific memes data. Each meme had a context in the form of keywords and meme template and result in the form of memes captions and post title. To take context into account, we applied a method for keywords encoding [50] proposed by Max Woolf [51]. We built text corpus where each record presents 1 meme with a unique combination of template id, submission keywords, submission title, meme top caption, and meme bottom caption:

```
<|startoftext|>~‘
001~^
photographer catdog mediocre said maybe~@
As a photography student...~}
just because you own a camera~{
it does not make you a. photographer
<|endoftext|>
```

1. `<|startoftext|>` - start tokens.

2. `001` - template id.

3. `photographer catdog mediocre said maybe` - keywords.
4. `As a photography student...` - submission title.
5. `just because you own a camera` - top meme caption (lowercased).
6. `it does not make you a.  photographer` - bottom caption (lowercased).
7. `<|endoftext|>` - end tokens.
8. `~‘, ~^, ~@, ~}, ~{` - sections separators.

Each meme has top-10 most weighted keywords, extracted from submission comments. Each meme was presented 3 times with 5 randomly chosen keywords in the training corpus; we randomized the order of keywords to avoid overfitting based on the sequence of the keywords. We split 80% data into training set, and 20% in validation set.

4.2.2 Generative model training

We trained a model on Google Colab as they provide access to free of charge machines with GPUs. We utilized framework [52] which is wrapper written by Max Woolf on top of original GPT-2. We trained network for 30K epochs with hyperparameters shown in Table A.1 for ∼19 hours. It can be seen in Figure 4.2 that the model loss decreased on the train set during 30K epochs. However, validation loss decreased until 17K epochs and started to grow after. This indicates model overfitting after the 17K-th step. We were saving models each 5K epochs, so the closets saved model was saved on the 15K-th step. We used this instance of the model for our final system.

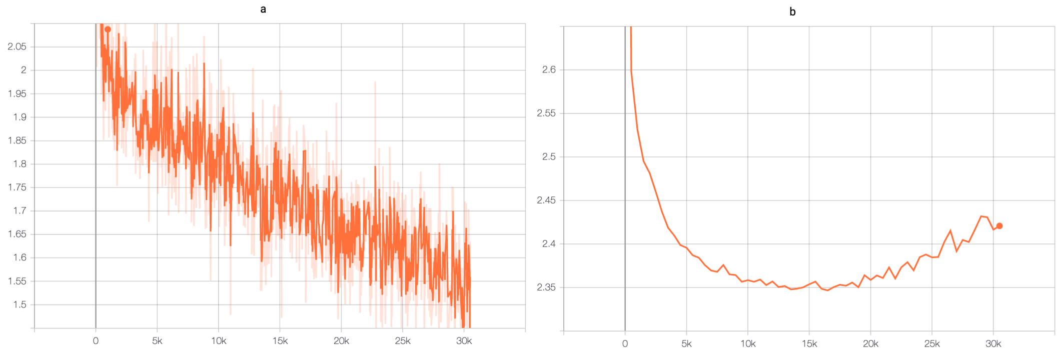


FIGURE 4.2: GPT-2 model losses.
a - training loss, b - validation loss.

We did a side experiment, which was a fail. We trained the network in 3 stages and provided more engaging memes as training data on each next stage. We split training data into 3 groups: memes with no scores, memes with number of comments less than the median (5 comments), and greater than the median. Our idea was to improve model output quality with an improvement of input data quality, as the last two batches will content memes which caused engagement. We measured model validation loss on data from the same group on each stage. We concluded experiment failure as a final model metric (Figure B.2) was not interpretable enough to evaluate model instance trained with this approach. However, it doesn't mean that this approach is not valid; it can be a theme for another research.

4.3 Continuous engaging content generation

When core models for template selection and text generation were trained and evaluated, we merged them in the final system shown in Figure 4.3.

Final pipeline to generate engaging content is next:

1. Select top-scored submissions from subreddit; for this purpose, any subreddit can be used.

   We used community with world news [53], which frequently updates (a few submissions every hour). We used PRAW [54] - Reddit client wrapper, to get new data from the Reddit. We select top 30 the most scored submissions for the day.

2. We perform TF-IDF on top of the comments from each post.

3. Define the most probable meme template for this comments.

4. Run our generative model with template id and top-5 keywords extracted from the comments as input. Generative model hyperparameters are described in Table A.2. On this stage, we generate the title, top, and bottom captions for the meme.

5. We superimpose the template image with text.

Generated memes can be found in Figure C.1 and Figure C.2.

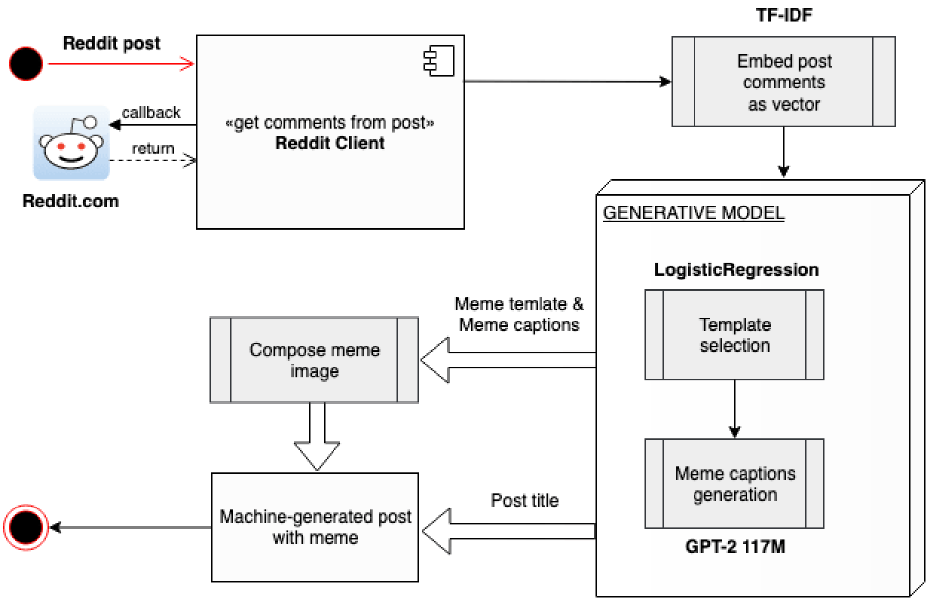


FIGURE 4.3: System for engaging content generation.

Chapter 5

Evaluation

Perfect evaluation should be done in the wild environment, where content should be posted in social media without information that the author is NN (to avoid bias). Big Reddit communities can be a destination for posts since a lot of memes are posted there every minute. The number of scores and comments from the social network can be used to compared engagement from human created and machine-generated content. Significant evaluation requires the number of submitted memes (observations) more than 1000; derivation of this number is described in Subsection 5.3.1. It is a time-consuming process to post such number of submissions with respect to Reddit rules named Reddiquette [55].

To evaluate memes in a limited time, we set up a series of controlled experiments with crowdsourcing expert evaluation. We decided to use MTurk as there are a few studies [56][57][58] where best practices and MTurk specific were described. We targeted our audience to US citizens only, as Reddit users are mostly from the USA (55% [59]).

MTurk uses term **HIT** - a Human Intelligence Task, a single, self-contained task that a Worker can work on, submit an answer, and collect a reward for completing [60]. In our case, **HIT** is a single action (like/don't like) for a meme from one worker.

Our budget for the evaluation was - $150.

For the evaluation, we used the number of likes (analog of upvotes) to measure the engagement level. We compared how generated memes engage the audience compared to the real memes based on the scores given by MTurk workers.

5.1 Applied statistics

We utilized a few statistics to compare two groups:

1. $\boldsymbol{\chi^2}$ **test**, we used Pearson's Chi-Square test [61] for statistical hypothesis testing as our data has a categorical nature (Like/Not Like). This method is commonly used to calculate the probability to get difference which had been observed between two groups. For each test, we used a contingency table and calculated χ^2 statistics. Contingency tables are used in statistics to summarize the relationship between several categorical variables. χ^2 is in equation 5.1, where O_i is the number of observations of type i, E_i is the expected count of type i, and n is the number of cells in the contingency table.

   The same $\boldsymbol{H_0}$ *and* $\boldsymbol{H_a}$ *were utilized for all tests.* Null hypothesis - that there is no significant difference between group A and group B, $H_0 : p_A = p_B$; and alternative hypothesis - groups have significant difference, $H_a : p_A \neq p_B$.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{5.1}$$

2. **Effect size w** - the minimum difference between groups that we consider as significant [62]. ES for Chi-Square test was used to measure the difference between groups. Pre-calculated w for small, medium, and large effect sizes can be seen in Table D.1.

   It was mentioned that effective size w can be more significant metric than test p-value, as it quantifiy the difference between two groups [63]. Equation 5.2 was used to calculate w, where N is the total count in all the cells.

$$w = \sqrt{\frac{\chi^2}{N}} \tag{5.2}$$

3. **Test power** for Chi-Square test as part of the meta-analysis. We used Cohen Power Tables [62] to calculate test power.

5.2 MTurk workers evaluation

5.2.1 Experiment settings

The main task of this evaluation was to define whether number of likes given by workers are related to the number of likes (upvotes) from the real world. We done an experiment to approve that MTurk can be used to measure engagement, and it is related to engagement in social media.

We have memes from the social network, they have already been scored and commented. For the evaluation we selected 2 groups of memes: good memes (group A) - top 10 the most scored memes, and bad memes (group B) - random 10 images that didn't collect any engagement (their total score was 0), we mixed memes from both groups for the experiment.

We built an MTurk task in which workers were asked to choose how they would react on the meme if they had seen them in social media. UI of this task can be seen in Figure E.1. We consider that group A and group B have a significantly different nature, as one group is "engaging memes" and the other is "not engaging memes".

During our MTurk meta-analysis we checked whether difference in collected engagement will be significant between group A and group B, as it is in the real world. We performed Chi-square test to analyse the difference.

We specified that all workers should be citizens of the USA. We didn't apply any filters on the workers' qualifications, which we used for the next evaluations.

5.2.2 Experiment analysis

Collected engagement can be seen in Table 5.1.

TABLE 5.1: Meme samples characteristics.

Group	**Description**	**Memes num.**	**Workers num.**	**Hits num.**	μ	$\sigma(x)$
A	*good memes*	10	10	100	0.41	0.49
B	*bad memes*	10	10	100	0.21	0.41

Based on collected results we calculated statistics described in Section 5.1 which can be seen in Table 5.2. Statistics show that $H_a : p_A \neq p_B$ is true - MTurk workers provided significantly more engagement to memes from group A, those collected a lot of scores in the social network. Based on this, we claim that MTurk can be used to estimate amount of engagement caused by the meme, which is relatively similar to the amount caused by Reddit users.

TABLE 5.2: Chi-square test statistics.

Test #	$\mu_A - \mu_B$	H_0	H_a	χ^2	**p-value**	w	α	power
1	0.2	$p_A = p_B$	$\boldsymbol{p_A \neq p_B}$	8.4385	0.0037	0.2054	0.05	0.82

5.3 Memes evaluation

During this evaluation, we compared machine-generated image macros with random human and bad human memes (which didn't collect scores). We applied the Chi-square test to calculate the significance of engagement difference between two samples and used these numbers to evaluate generated memes.

5.3.1 Chi-square test preliminary calculations

We had defined effect size, and minimum number of observations for the statistically significant test with power ≥ 0.9, before the test was performed.

We used $w = 0.1$, as this value was proposed by Cohen for a small difference [62]. ES values for the Chi-square test can be found in Table D.1.

We used Cohen Power Tables [62] to find minimum number of needed observations for predefined values $w = 0.1$, $\alpha = 0.05$, degree of freedom $df = 1$ and $power = 0.9$. We needed ~1000 total observations to conduct test with defined characteristics.

The number of HITs that we needed to collect in order to have enough observations to detect a difference with ES ≥ 0.1 was 1000. We wanted to compare two groups in each test, so we needed a minimum of 500 observations for each group.

We collected 900 HITs for each group of memes; for each test, ~1800 total observations were used.

5.3.2 Evaluation setting

We created 3 groups for comparison: machine-generated memes, random human-created and bad human-created image macros. Generated memes were selected randomly; however, before sending it to the MTurk workers, they were manually moderated to exclude violent content. In the real community, this procedure is called moderation, when moderators delete content that is violent, harmful or brakes community rules. Both human-created memes groups were selected from the 2019 year memes, as our system creates image macros on topics related to 2019. Only memes based on the templates that were used in the sample with machine-generated memes (to minimize the difference between samples). We had 270 memes in our final evaluation.

To collect the needed amount of observations, we collected HITs from 10 workers for 90 memes in each sample, 900 HITs per sample.

We targeted workers with the next characteristics:

- Worker is MTurk Master.
- Location is the United States.
- Hit approval rate(%) $\geq$ 95%.

- Number of HITs approved ≥ 5000.

We collected 2740 HITs from MTurk; additional 40 HITs were collected during UI testing. To clean data from HITs that could be poorly executed, we filtered out HITs, which were done faster or slower than most other HITs. We defined anomaly as 5% of observations that were on the extreme left and right sides of the distribution. In other words, we removed (2.5%) of the HITs, which had execution time higher than other 97.5%, as well as (2.5%) of HITs that were done slower than other 97.5%. When anomaly exclusion had been done, we got 2603 observations to calculate test statistics.

Characteristics of HITs and collected engagement are in Table 5.3.

TABLE 5.3: Characteristics of engagement for memes groups.

Group	**Description**	**Memes num.**	**Hits num.**	μ	$\sigma(x)$
A	*human, random*	90	846	0.35	0.48
B	*human, bad*	90	881	0.27	0.44
C	*machine*	90	876	0.24	0.43

TABLE 5.4: Memes samples comparison.

Test #	**Groups**	$\mu_1 - \mu_2$	H_0	H_1	χ^2	**p-value**	w	α	**power**
1	A vs B	0.08	$p_A = p_B$	$\boldsymbol{p_A \neq p_B}$	11.06	<0.001	0.08	0.05	0.91
2	A vs C	0.09	$p_A = p_C$	$\boldsymbol{p_A \neq p_C}$	21.29	<0.001	0.11	0.05	0.99
3	B vs C	0.03	$\boldsymbol{p_B = p_C}$	$p_B \neq p_C$	1.4	0.236	0.02	0.05	0.22

5.3.3 Random human vs bad human memes

During this test, we compared random human memes (A) with bad human memes (B). Test statistics can be found in Table 5.4 Test #1.

Conducted Chi-square test had p-value < 0.001, which is less then our a, so based on p-value, we can reject the null hypothesis that group B caused the same engagement as a group A. $H_a : p_A \neq p_C$ is true.

The conclusion is that bad human memes caused less engagement in workers than random image macros with significant difference.

5.3.4 Random human vs machine-generated memes

We compared random human image macros (A) with memes generated using our system (C), test statistics can be found in Table 5.4 Test #2. ES $w = 0.11$ is greater than small, which can be observed in Table D.1.

Chi-square test with the null hypothesis that two samples cause the same volumes of engagement was conducted. This test had p-value< 0.001, based on p-value, we accept alternative hypothesis - $p_A \neq p_C$.

The test confirmed that group A caused significantly more engagement than group C. Random human memes cause more engagement than image macros generated by the machine, which is a very realistic result.

5.3.5 Bad human memes vs machine generated memes

We want to mention that small amount of engagement is not always a reliable metric of a bad meme; it can be a matter of time when the submission was published, wrong title [26], or just a random. It means that there are submissions that were similar to regular human memes, but they didn't collect engagement because of time when they were posted (for example, at late night when community members are asleep). So in the bad memes sample should be image macros that cause the same engagement as random memes.

Test statistics are shown in Table 5.4 Test #3. The null hypothesis is that bad human memes (B) don't have a significant difference in engagement level compared to machine-generated memes (C). This test has p-value statistic equals 0.236, which means that we can not reject the null hypothesis, $H_0 : p_A = p_B$. We can't make decisions based only on test results as test power is 0.22.

However, $w = 0.02$, which is 5 times smaller than value 0.1 which we considered as a significant difference between groups. Based on $w = 0.02$, which is a very small difference, we stated that there is no significant difference between group B and C.

We concluded that bad human-created memes have the same engaging power as memes generated by our system.

5.4 Evaluation results

To conclude all inferences from tests described above:

1. MTurk workers expert evaluation was approved as an appropriate tool for engagement measurement.
2. Machine-generated memes caused the same engagement as bad human-created memes. Difference is not significant, $w = 0.02$.
3. Machine-generated image macros caused less engagement than random human memes. Based on Chi-square test results.
4. Random human memes caused more engagement than bad human memes. Difference is significant.

Chapter 6

Conclusions

6.1 Contribution and Summary

Results of the project is successful as main objective was achieved. We proposed a unique method in order to generate memes to engage an audience in social media using the GPT-2 model. Generated memes causes the same engagement as unsuccessful memes (which didn't collect scores in the Reddit) created by human. Machine-generated image macros can be found in Figure C.1 and Figure C.2.

The problem of content generation to engage an audience is relatively new, and it involves different disciplines and scientific areas. There have been studies on the engagement analyses, social media influence, modeling of information spreading, even meme generation already have been done. However, achieved results make the contribution of this work valuable.

In this study, we proposed an approach for meme generation, created a dataset to justify our method, described implementation details such that our results can be reproduced, and evaluated engagement caused by model-generated memes. The proposed pipeline can be used as a base for generating more complex scenes then image macro. We did a statistical evaluation of engagement that machine-generated memes cause in the audience compared to memes created by humans.

The evaluation showed that machine-generated memes engage people with the same power as human memes that didn't collect engagement in social media. However, generated memes cause less engagement than random memes created by humans. Our solution can be used as a baseline to create system for producing content that will be close to content created by a human.

A mixed approach that combines current Deep Learning state-of-the-art techniques and established statistical methods can be used for solving computational creativity problems. Data with content from social media is publicly available, and the GPT-2 model as well, so system which is similar to ours, can be built by enthusiasts.

6.2 Future work

List of improvements which can increase engagement caused in the audience by machine-generated memes (content):

1. Advanced template selection. We achieve 0.499% accuracy on classification task with 38 templates using logistic regression. However, advanced techniques can be applied. Thus improvement can increase content diversity.

2. We used text from all comments as input for the system, however only comments that collected positive scores can be used to improve the quality of input for generator.

3. Detect and filter out offensive generated content. The rule-based system of offensive words vocabulary can be used as baseline. The more complicated way is to train a model for binary classification task for distinguishing violent memes from ones that can be published.
4. Make an advance selection for training data, clean input dataset from violent and harmful content, make it more divorced, and finetune GPT-2 on top of it. A few sources of memes can be used to increase generated content variety.
5. Publish generated submissions in the significant public communities to evaluate engagement in the wild. We recommend to do it only when content filtering will be implemented.
6. Investigate which word sequences are commonly used in engaging memes based on the NN weights.
7. Use different communities as a source for input data and evaluate memes generated based on their posts.

Appendix A

GPT-2 hyperparameters

TABLE A.1: GPT-2 model training hyperparameters

hyperparameter	value
batch_size	1
learning_rate	0.0001
accumulate_gradients	5
sample_every	100
sample_length	1023
sample_num	1
use_memory_saving_gradients	FALSE
only_train_transformer_layers	FALSE
optimizer	adam
val_every	500
val_batch_size	2
val_batch_count	40

TABLE A.2: GPT-2 meme generation hyperparameters

hyperparameter	value
temperature	0.7
top_k	40
nsamples	1
batch_size	1
truncate	`<|endoftext|>`
include_prefix	FALSE

Appendix B

Fail GPT-2 training

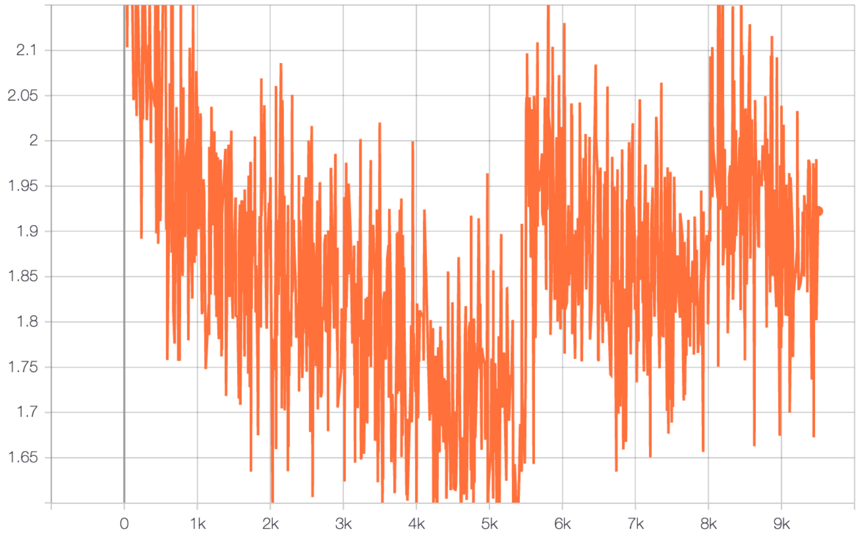


FIGURE B.1: Fail model training loss.

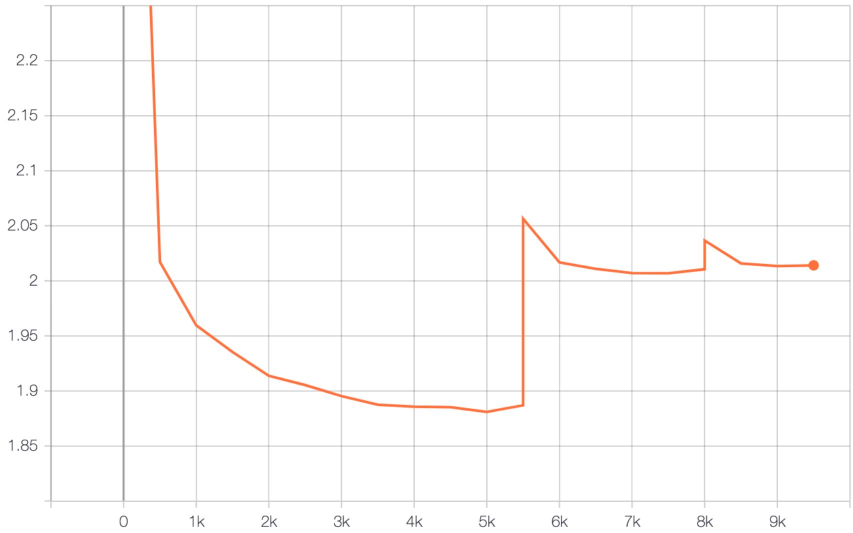


FIGURE B.2: Fail model validation loss.

Appendix C

Machine-generated memes

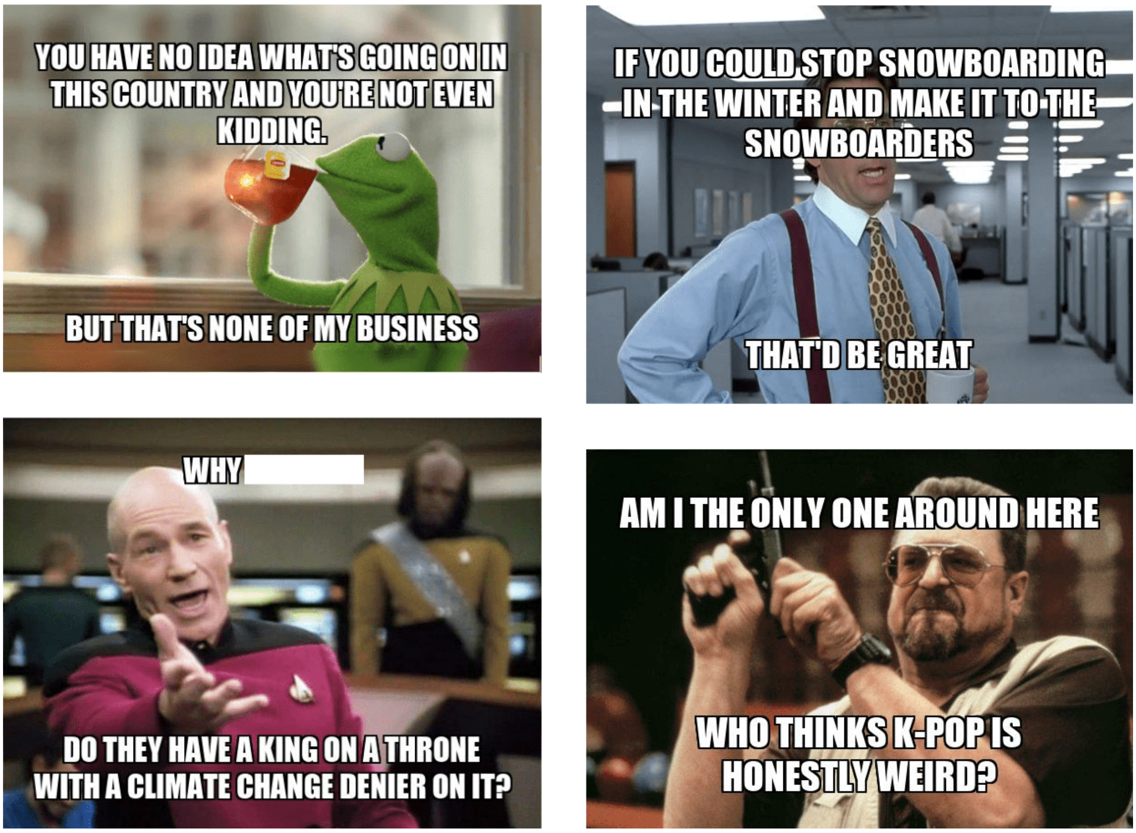

YOU HAVE NO IDEA WHAT'S GOING ON IN THIS COUNTRY AND YOU'RE NOT EVEN KIDDING.
BUT THAT'S NONE OF MY BUSINESS
IF YOU COULD STOP SNOWBOARDING IN THE WINTER AND MAKE IT TO THE SNOWBOARDERS
THAT'D BE GREAT
WHY
DO THEY HAVE A KING ON A THRONE WITH A CLIMATE CHANGE DENIER ON IT?
AM I THE ONLY ONE AROUND HERE
WHO THINKS K-POP IS HONESTLY WEIRD?

FIGURE C.1: Generated memes example #1.

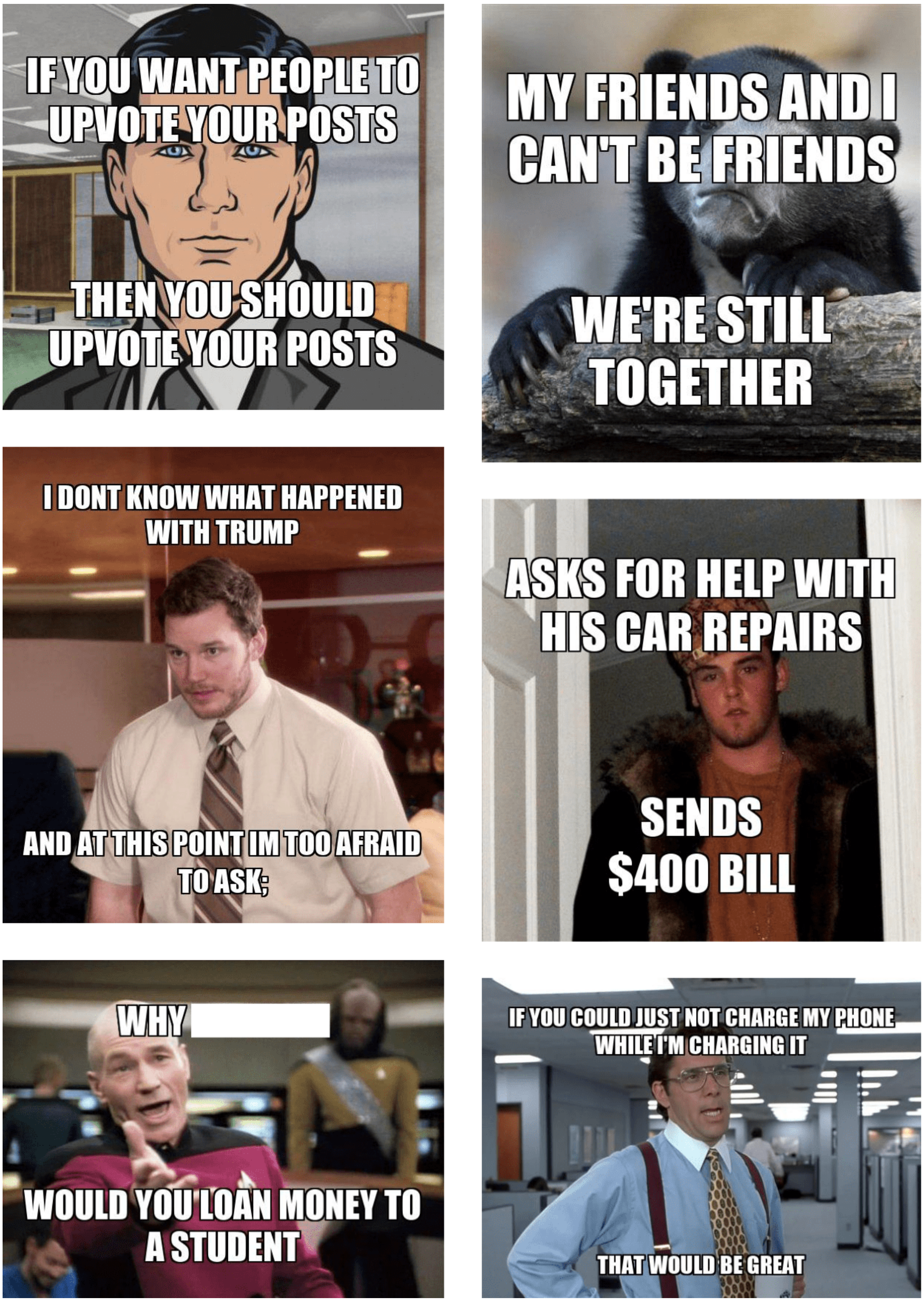

IF YOU WANT PEOPLE TO UPVOTE YOUR POSTS
THEN YOU SHOULD UPVOTE YOUR POSTS
MY FRIENDS AND I CAN'T BE FRIENDS
WE'RE STILL TOGETHER
I DONT KNOW WHAT HAPPENED WITH TRUMP
AND AT THIS POINT IM TOO AFRAID TO ASK;
ASKS FOR HELP WITH HIS CAR REPAIRS
SENDS $400 BILL
WHY
WOULD YOU LOAN MONEY TO A STUDENT
IF YOU COULD JUST NOT CHARGE MY PHONE WHILE I'M CHARGING IT
THAT WOULD BE GREAT

FIGURE C.2: Generated memes examples#2.

Appendix D

Cohen effective size table

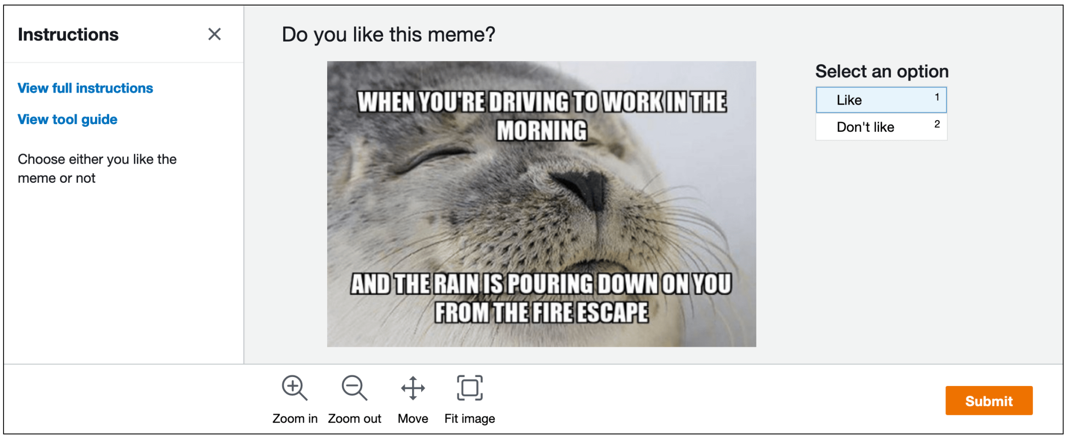
TABLE D.1: Cohen effect size w.

Effect size w		
small	**medium**	**large**
0.1	0.3	0.5

Source: [62].

Appendix E

Task for engagement collecting



Instructions

Choose either you like the meme or not

Do you like this meme?

Select an option

View full instructions
View tool guide
WHEN YOU'RE DRIVING TO WORK IN THE MORNING
AND THE RAIN IS POURING DOWN ON YOU FROM THE FIRE ESCAPE
Like 1
Don't like 2
Zoom in
Zoom out
Move
Fit image
Submit

FIGURE E.1: MTurk HIT to collect engagement.

Bibliography

[1] A. Radford et al. "Language models are unsupervised multitask learners". In: *OpenAI Blog* 1.8 (2019).

[2] *Engage | Definition of Engage by Lexico*. lexico.com/en/definition/engage. Accessed: 2020-01-02.

[3] *"Afraid to Ask" Andy*. https://knowyourmeme.com/memes/afraid-to-ask-andy. Accessed: 2020-01-07.

[4] N. Bayindir and D. Kavanagh. *GlobalWebIndex's flagship report on the latest trends in social media*. Flagship Report 2018. 2018.

[5] C. T. Bergstrom and J. B. Bak-Coleman. "Information gerrymandering in social networks skews collective decision-making". en. In: *Nature* 573.7772 (Sept. 2019), pp. 40–41.

[6] M. Knobel and C. Lankshear. "Online memes, affinities, and cultural production". In: *A new literacies sampler* 29 (2007), pp. 199–227.

[7] K. K. Aldous, J. An, and B. J. Jansen. "View, Like, Comment, Post: Analyzing User Engagement by Topic at 4 Levels across 5 Social Media Platforms for 53 News Organizations". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 13. 2019, pp. 47–57.

[8] H. Toivonen and O. Gross. "Data mining and machine learning in computational creativity". In: *WIREs Data Mining Knowl Discov*. Lecture Notes in Artificial Intelligence (LNAI) 5.6 (Nov. 2015), pp. 265–275.

[9] A. Fan, M. Lewis, and Y. Dauphin. "Hierarchical Neural Story Generation". In: (May 2018). arXiv: 1805.04833 [cs.CL].

[10] I. Sutskever, O. Vinyals, and Q. V. Le. "Sequence to Sequence Learning with Neural Networks". In: *Advances in Neural Information Processing Systems 27*. Ed. by Z Ghahramani et al. Curran Associates, Inc., 2014, pp. 3104–3112.

[11] A. Sriram, H. Jun, S. Satheesh, and A. Coates. "Cold Fusion: Training Seq2Seq Models Together with Language Models". In: (Aug. 2017). arXiv: 1708.06426 [cs.CL].

[12] M. Ott et al. "fairseq: A Fast, Extensible Toolkit for Sequence Modeling". In: (Apr. 2019). arXiv: 1904.01038 [cs.CL].

[13] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. "Show and tell: A neural image caption generator". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3156–3164.

[14] Y Bengio, P Simard, and P Frasconi. "Learning long-term dependencies with gradient descent is difficult". en. In: *IEEE Trans. Neural Netw.* 5.2 (1994), pp. 157–166.

[15] C. Gan et al. "Stylenet: Generating attractive visual captions with styles". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 3137–3146.

[16] *Socially Awkward Penguin | Know Your Meme*. https://knowyourmeme.com/memes/socially-awkward-penguin. Accessed: 2019-09-30.

[17] W. Y. Wang and M. Wen. “I can has cheezburger? a nonparanormal approach to combining textual and visual information for predicting and generating popular meme descriptions”. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2015, pp. 355–365.

[18] B Schweizer and A Sklar. *Probabilistic Metric Spaces*. en. Courier Corporation, Oct. 2011.

[19] B. Rayens and R. B. Nelsen. *An Introduction to Copulas*. 2000.

[20] K. Yoshida et al. “Neural Joking Machine : Humorous image captioning”. In: (May 2018). arXiv: 1805.11850 [cs.CV].

[21] K. He, X. Zhang, S. Ren, and J. Sun. “Deep Residual Learning for Image Recognition”. In: (Dec. 2015). arXiv: 1512.03385 [cs.CV].

[22] *Bokete*. https://bokete.jp. Accessed: 2019-09-30.

[23] A. L. Peirson V and E Meltem Tolunay. “Dank Learning: Generating Memes Using Deep Neural Networks”. In: (June 2018). arXiv: 1806.04510 [cs.CL].

[24] S. Zannettou et al. “On the Origins of Memes by Means of Fringe Web Communities”. In: *Proceedings of the Internet Measurement Conference 2018*. IMC ’18. Boston, MA, USA: ACM, 2018, pp. 188–202.

[25] *Know Your Meme: Internet Meme Database*. https://knowyourmeme.com. Accessed: 2019-09-30.

[26] H. Lakkaraju, J. McAuley, and J. Leskovec. “What’s in a name? understanding the interplay between titles, content, and communities in social media”. In: *Seventh International AAAI Conference on Weblogs and Social Media*. 2013.

[27] A Deza and D Parikh. “Understanding image virality”. In: *Proceedings of the IEEE conference on* (2015).

[28] *Top Ten Internet Languages in The World - Internet Statistics*. https://www.internetworldstats.com/stats7.htm. Accessed: 2019-09-30.

[29] *Upvoted – The official Reddit blog*. https://redditblog.com/. Accessed: 2019-09-30.

[30] *Reddit Statistics - pushshift.io*. https://pushshift.io. Accessed: 2019-09-30.

[31] *r/AdviceAnimals*. https://www.reddit.com/r/AdviceAnimals/. Accessed: 2019-11-22.

[32] *Google BigQuery*. https://bigquery.cloud.google.com/dataset/fh-bigquery:reddit_posts. Accessed: 2020-01-01.

[33] V. Monga and B. L. Evans. “Perceptual image hashing via feature points: performance evaluation and tradeoffs”. en. In: *IEEE Trans. Image Process.* 15.11 (Nov. 2006), pp. 3452–3465.

[34] M. Ester et al. “A density-based algorithm for discovering clusters in large spatial databases with noise”. In: *Kdd*. Vol. 96. 1996, pp. 226–231.

[35] *zsavvas/memes_pipeline*. https://github.com/zsavvas/memes_pipeline. Accessed: 2020-01-08.

[36] J. Baek et al. “What Is Wrong With Scene Text Recognition Model Comparisons? Dataset and Model Analysis”. In: (Apr. 2019). arXiv: 1904.01906.

[37] X. Zhou et al. “EAST: an efficient and accurate scene text detector”. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2017, pp. 5551–5560.

[38] *tesseract-ocr/tesseract: Tesseract Open Source OCR Engine (main repository)*. https://github.com/tesseract-ocr/tesseract. Accessed: 2020-01-01.

[39] *Computer Vision | Microsoft Azure*. https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/. Accessed: 2020-01-01.

[40] *Pricing - Computer Vision API | Microsoft Azure*. https://azure.microsoft.com/en-us/pricing/details/cognitive-services/computer-vision/. Accessed: 2020-01-01.

[41] *Cognitive Services APIs Reference*. https://bit.ly/3algRH2. Accessed: 2020-01-01.

[42] S. J. Karen. “A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL”. In: *Journal of Documentation* 28.1 (Jan. 1972), pp. 11–21.

[43] A. Rajaraman and J. D. Ullman. *Mining of massive datasets*. Cambridge University Press, 2011.

[44] P. J. Liu et al. “Generating Wikipedia by Summarizing Long Sequences”. In: (Jan. 2018). arXiv: 1801.10198 [cs.CL].

[45] A. Vaswani et al. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems 30*. Ed. by I Guyon et al. Curran Associates, Inc., 2017, pp. 5998–6008.

[46] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. “Improving language understanding by generative pre-training”. In: *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf* (2018).

[47] *Welcome To Colaboratory - Colaboratory*. https://colab.research.google.com/. Accessed: 2019-09-30.

[48] *sklearn.feature_extraction.text.TfidfTransformer*. https://bit.ly/2RtsCSS. Accessed: 2020-01-01.

[49] *sklearn.linear$_m$odel.LogisticRegression*. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html. Accessed: 2019-12-29.

[50] *gminimaxir/gpt-2-keyword-generation*. https://github.com/minimaxir/gpt-2-keyword-generation. Accessed: 2020-01-02.

[51] *Max Woolf’s Blog*. https://minimaxir.com/. Accessed: 2020-01-02.

[52] *minimaxir/gpt-2-simple*. https://github.com/minimaxir/gpt-2-simple. Accessed: 2020-01-02.

[53] *World News*. https://www.reddit.com/r/worldnews/. Accessed: 2020-01-01.

[54] *praw-dev/praw*. https://github.com/praw-dev/praw. Accessed: 2020-01-02.

[55] *reddit: the front page of the internet*. https://www.reddit.com/wiki/reddiquette. Accessed: 2020-01-06.

[56] D. Archambault, H. Purchase, and T. Hoßfeld. *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments: Dagstuhl Seminar 15481, Dagstuhl Castle, Germany, November 22 – 27, 2015, Revised Contributions*. en. Springer, Sept. 2017.

[57] K Hara et al. “A data-driven analysis of workers’ earnings on amazon mechanical turk”. In: *Proceedings of the* (2018).

[58] A. M. Brawley and C. L. S. Pury. “Work experiences on MTurk: Job satisfaction, turnover, and information sharing”. In: *Comput. Human Behav.* 54 (Jan. 2016), pp. 531–546.

[59] *reddit.com Competitive Analysis, Marketing Mix and Traffic - Alexa*. https://www.alexa.com/siteinfo/reddit.com. Accessed: 2020-01-06.

[60] *Amazon Mechanical Turk*. https://www.mturk.com/worker/help. Accessed: 2020-01-06.

[61] K. Pearson. “X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50.302 (1900), pp. 157–175.

[62] J Cohn. “Statistical power analysis for the behavioral sciences”. In: *Technometrics* 31.4 (1988), pp. 499–500.

[63] R Coe. “It’s the effect size, stupid: What effect size is and why it is important”. In: (2002).