UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Computer Science Curriculum

Tarmo Pungas

Towards a Semantically Rich Meme Dataset

Bachelor's Thesis (9 ECTS)

Supervisors: Riccardo Tommasini, PhD
Radwa El Shawi, PhD

Tartu 2022

Towards a Semantically Rich Meme Dataset

Abstract:
This thesis presents the idea of a system capable of recommending internet memes based on textual context. The goal of the thesis is to help realize this system by preparing a semantically rich dataset. Methods for exploring, cleaning, and analyzing data are presented. Different metrics for assessing inter-annotator agreement are discussed. A pipeline is created to enable the preprocessing and enrichment of an existing meme dataset. As part of the preprocessing, the data are cleaned, filtered, and integrated with other datasets. For semantic enrichment, crowdsourcing is used to collect information for 50 memes. The resulting semantic annotations are analyzed and evaluated, revealing that the collected data are of mixed quality and that further data collection is necessary. Finally, future improvements to the pipeline are suggested.

Keywords:
Internet memes, memes, semantic analysis

CERCS:
P175 - Informatics, systems theory

Semantiliselt rikkaliku meemiandmestiku suunas

Lühikokkuvõte:
Selles lõputöös esitletakse ideed süsteemist, mis oleks võimeline tekstilise konteksti põhjal soovitama internetimeeme. Töö eesmärk on aidata luua selline süsteem, valmistades ette semantiliselt rikkalik andmestik. Töös esitletakse meetodeid andmete uurimiseks, puhastamiseks ja analüüsimiseks. Arutatakse erinevaid mõõdikuid, millega hinnata annoteerijate vahelist nõustumist. Luuakse konveier, mis võimaldab olemasoleva meemiandmestiku eeltöötlemist ja rikastamist. Eeltöötluse osana andmed puhastatakse, filtreeritakse ja integreeritakse teiste andmestikega. Andmete rikastamiseks kogutakse ühisloome abil 50 meemi kohta semantilist infot. Saadud semantilisi annotatsioone analüüsitakse ja hinnatakse ning jõutakse järeldusele, et kogutud andmed on erineva kvaliteediga ja andmete kogumist tuleb jätkata. Lõpetuseks pakutakse välja soovitusi tulevikus konveieri parendamiseks.

Võtmesõnad:
Internetimeemid, meemid, semantiline analüüs

CERCS:
P175 - Informaatika, süsteemiteooria

Contents

1 Introduction

Memes are cultural features or types of behavior passed from one generation to another [1]. The term, a portmanteau of *mimesis* and *gene*, was initially coined by Richard Dawkins in 1976 in his book The Selfish Gene [2]. The study of memes and their spread has become known as memetics.

Memes are prevalent in human society and are especially recognizable in the internet culture as *internet memes*. An internet meme is a piece of culture, typically a joke, which gains influence through online transmission [3]. Although widespread and rich with information, internet memes remain relatively unexplored in scientific literature. The concept has been investigated semiotically [4], but few data-driven applications exist.

The main obstacle that prevents a systematic semantic analysis of memes is the difficulty of acquiring a well-annotated dataset. Some reasons contributing to this difficulty are the (1) great variety among memes, (2) large number of memes being spread across multiple platforms, and (3) constant evolving and changing of memes. Recently, work has been done to create a meme knowledge graph [5] which will be explored in this work to gain insights into memes and determine which features best characterize a meme (see Figure 1 for meme characteristics).

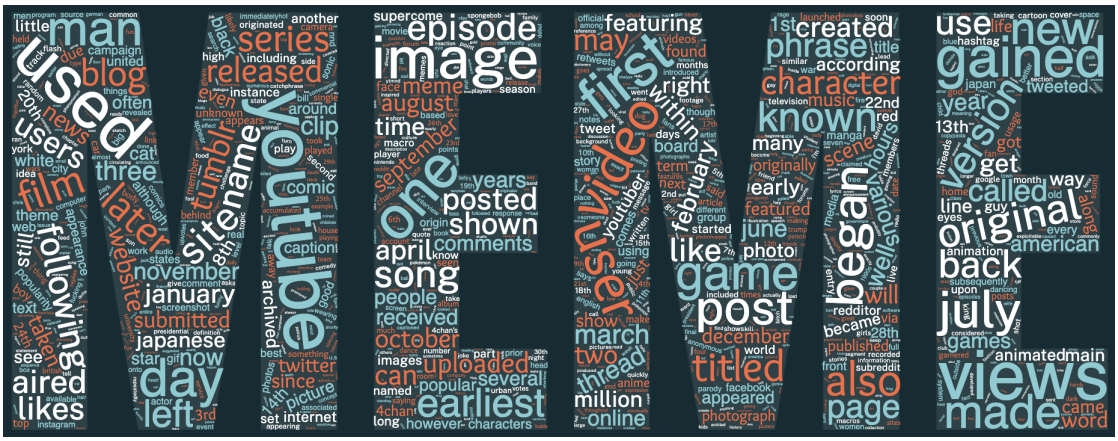

Figure 1. Word cloud of textual data describing internet memes.

Memes transgress social and cultural boundaries and have the potential to aid communication. This work positions itself within the vision of a system for meme understanding. We can imagine a system capable of understanding meme semantics and recommending them memes based on some external context, e.g., an article. In order to create this system, three components are required. One is a text summarization tool that can take an article and output its summary. The second part is a collection of meme embeddings acquired from a database. The final component combines these two pieces, comparing meme embeddings with text summaries and generating a ranked output of captioned

memes.

This work contributes to the creation of the second component. The goal is to prepare a semantically rich dataset that can be used to create meme embeddings. Based on that dataset, we hope to understand the best way to represent memes – either metadata or semantics – which will decide the focus of future research. Figure 2 visualizes the contribution of this thesis and the design of the recommendation system. The contributions can be broken down into four parts:

1. Exploring the existing data to understand the context.
2. Creating a preprocessing pipeline based on received insights.
3. Semantically annotating memes manually and through crowdsourcing.
4. Analyzing the annotation results.

The thesis is structured as follows. The second chapter provides an overview of the existing dataset, crowdsourcing platform used for annotations, preprocessing methods, and inter-annotator agreement. Chapter three explores the data in detail, presents the preprocessing pipeline, and, finally, describes and analyzes the semantic enrichment of data.

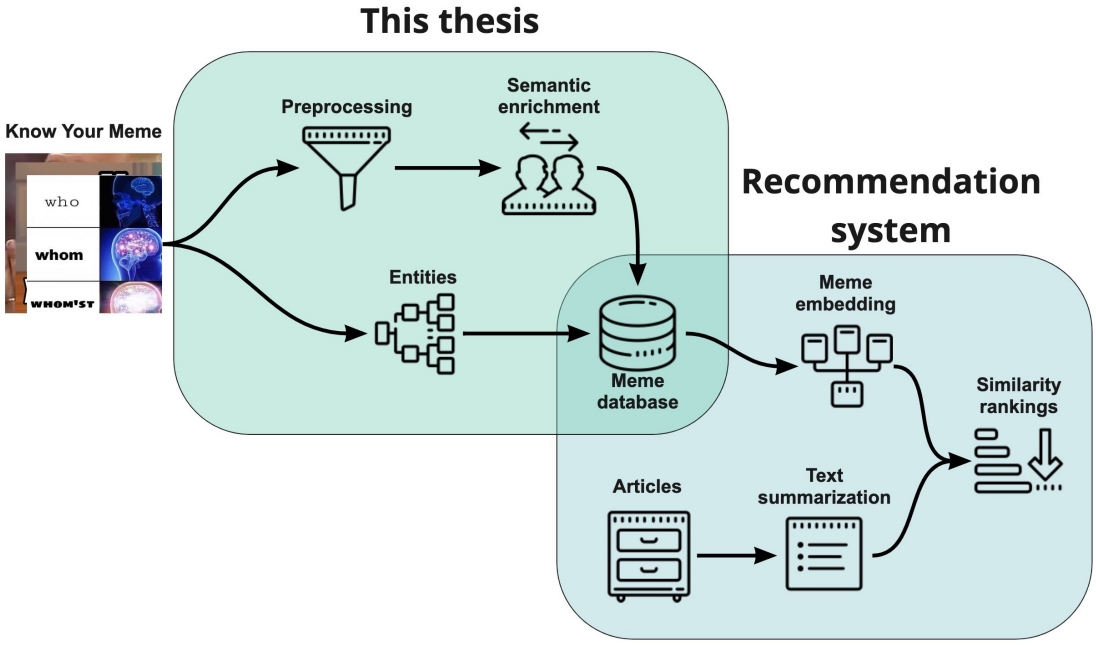

Figure 2. Overview of the thesis and recommendation system.

2 Background

This section introduces prior knowledge needed to understand the content of the thesis. Additionally, it provides an overview of relevant literature and reasoning behind the choices made in section 3.

2.1 Meme dataset

This study was made possible by a preceding bachelor's thesis, which designed a pipeline to facilitate the construction of an internet meme knowledge graph, covering data ingestion, processing, storage, and application [5]. The pipeline includes scraping the website Know Your Meme[1] and enriching the collected data by extracting image and text entities using Vision AI[2] and DBpedia Spotlight[3] respectively. This results in three distinct but linked datasets described below.

Know Your Meme data. The data is scraped from Know Your Meme and contains 28799 entries, out of which 15406 are memes. Technically, the data contains meme templates as opposed to meme instances. Each meme template can have multiple or even thousand instances. A meme template attempts to describe its instances' semantics, origin, and spread. However, for the sake of simplicity, we will use *meme* to refer to both memes and meme templates in this work. There are 18 features in the Know Your Meme dataset:

1. Title
2. URL
3. Date when the meme was last updated
4. Category, e.g., memes, viral videos, events
5. Template image URL
6. Meta, e.g., site descriptions for social media
7. JSON-LD
8. Date when the meme was added
9. Status, e.g., submission, confirmed, rejected
10. Origin, e.g., a specific movie or website
11. Year – first occurrence of the meme
12. Type, e.g., image macro, exploitable
13. Content
    (a) About
    (b) Origin
    (c) Spread
    (d) Notable examples
    (e) Search interest (iFrame to Google Trends[4])
14. Tags
15. Additional references
16. Parent – presented as "Part of a series on"
17. Siblings (related entries)
18. Children (related subentries)

[1] https://knowyourmeme.com/
[2] https://cloud.google.com/vision
[3] https://www.dbpedia-spotlight.org/
[4] https://trends.google.com/trends/

The most relevant features for this study were the title, URL, category, year, type, and content. From the *categories*, only memes are relevant for this study. The *year* can be used to sort memes based on recency. Since more than 100 *types* are listed on Know Your Meme, focusing on a small subset can be beneficial. *Content* includes the least structured but most crucial information, possibly including semantics.

Image entities. The image entities were extracted from the original dataset with Vision AI, a service provided by Google. For each image, it performs label detection, explicit content detection, and web entity detection. The resulting dataset consists of 419,002 labels (4,722 distinct) and 197,582 entities (42,746 distinct). The most frequent labels are *font*, *art*, *happy*, and *gesture*; the most common entities are *image*, *internet meme*, *Know Your Meme*, and *meme*.

Text entities. The text entities were extracted from the About sections of the original dataset using DBpedia Spotlight, a tool that provides entities (with confidence ratings) associated with the provided text. The dataset includes 38,801 entities, out of which 10,262 are distinct. The most common entities are *image macro*, *Japanese language*, *internet meme*, and *YouTube*[5].

2.2 Toloka

Toloka[6] is a crowdsourcing and microwork[7] platform – similar to Amazon Mechanical Turk[8] – owned and operated by Yandex N.V. It is directed toward two types of users: (1) people who need to label large amounts of data, called *requesters*, and (2) people who perform the labeling, called *performers*. The former group can use the platform to collect data for machine learning, marketing research, and more. When a requester creates a new project on Toloka, they go through the following process:

1. Choose an existing project template or create a blank project. Set a name and description that performers can see. Examples of project templates that one can choose from in Toloka include image classification, object detection, and sentiment analysis of text (see Figure 3 for the requester interface).

2. Edit the task interface using either the *Template builder* or a combination of HTML, JavaScript, and CSS. Specify input and output fields for the task and write instructions for the performers.

[5]https://www.youtube.com/
[6]https://toloka.ai/
[7]Many small tasks bundled into a large project which different people over the internet complete.
[8]https://www.mturk.com/

3. Create a new *pool* – a set of paid tasks sent out for completion simultaneously. Specify the audience (e.g., the language they speak, their rank in Toloka), quality control metrics (e.g., captcha frequency, punishing fast responses), and the price paid to the performer per task.
4. Upload the data that must be labeled in a tab-separated values (TSV) file, including a special header to state input fields and types.
5. Top up the account with money that will be paid to performers. Additionally, Toloka will charge a 30% fee for every task.
6. Launch the project and wait for performers to complete the tasks. Optionally, accept or reject assignments manually.
7. Download the labeled data. Optionally, view statistics such as average time spent on a task, number of people who submitted a task, and number of suspended performers due to quality control rules.

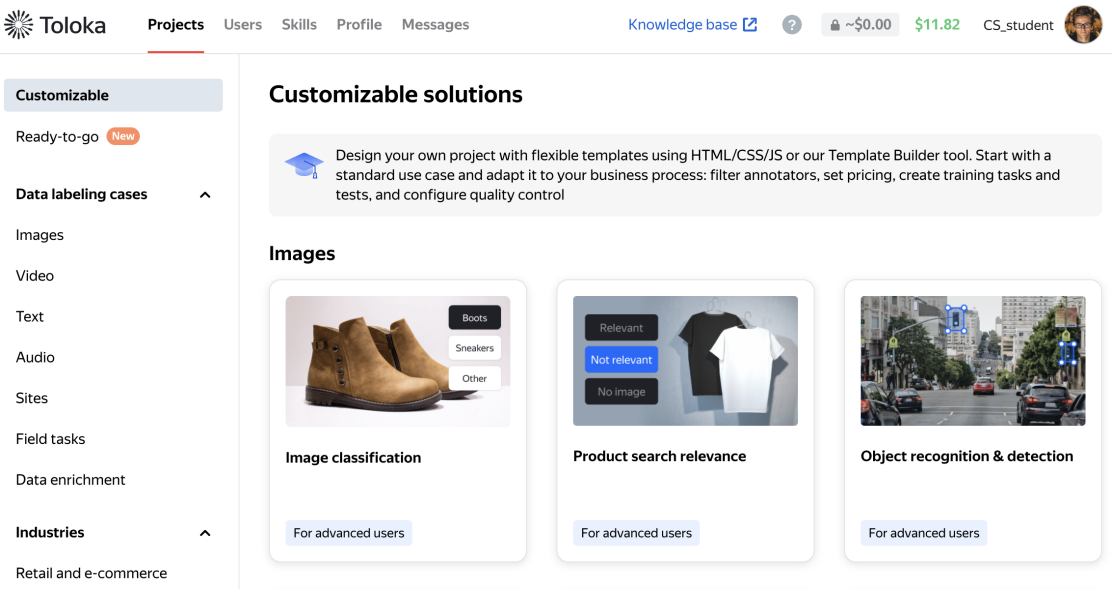

Figure 3. Toloka interface for requesters.

Anyone can sign up on Toloka to become a performer and earn money for various labeling tasks. As a performer, one can sort available tasks by price, creation date, or deadline. Performers can also filter tasks by requester or category, such as tasks requiring training and with non-automatic acceptance. After the performer has finished labeling a task, a small monetary amount is transferred to their account, and this money can be withdrawn from the site. The hourly pay that one can earn in Toloka can vary depending on the performer and project, ranging approximately from $1 to $10.

2.3 Preprocessing

Preprocessing is the manipulation of data to improve its quality, which can enhance performance in later use [6]. Said quality can be measured by different metrics, e.g., accuracy, completeness, or consistency. Preprocessing is a critical step in data mining, the process of extracting and discovering patterns in large volumes of data [7]. Preprocessing consists of the following tasks: [8]

1. Data cleaning: filling in missing values, smoothing noisy data, identifying or removing outliers and duplicates, and resolving inconsistencies. This results in data with fewer errors and discrepancies.

2. Data integration: integrating multiple databases or files.

3. Data reduction: representing the dataset with a smaller volume while preserving the same level of analytical results. This can significantly reduce the time taken for data analysis. Reduction methods include dimensionality reduction, numerosity reduction, and data compression.

4. Data transformation: mapping values of a given attribute to a set of replacement values. Transformation methods include smoothing, feature construction, aggregation, normalization, and discretization.

Data preprocessing is often used alongside exploratory data analysis (EDA), which encompasses exploring data to formulate hypotheses and help with future experiments. EDA uses visualization methods to analyze data and summarize its main characteristics. The approach has been promoted since the 1970s by John Tukey. He was of the opinion that statisticians put too much emphasis on confirmatory data analysis instead of using the data itself to suggest hypotheses [9].

Many different tools are used in exploratory data analysis, some more relevant to this study, e.g., bar charts, histograms, and boxplots. A histogram is an approximate representation of a numerical feature's distribution. Histograms are created by dividing the whole range of values into equal intervals and then counting how many values fall into each interval. Boxplots visualize the spread, locality, and skewness of data using quartiles. A line in the box shows the median value. The box represents 25% of data above and 25% below the median. In addition to the box, the plot can have lines extending from the box (also called *whiskers*), which indicate variability. Outliers are shown as dots beyond the whiskers.

2.4 Inter-annotator agreement

When working with multiple annotations per one object, it is essential to evaluate the degree of agreement among annotators. There are multiple ways to assess inter-annotator

agreement, ranging from joint probability to correlation coefficients. The joint probability of agreement is an estimated percentage of the time the raters agree.

Some joint-probability methods like kappas also correct for how often annotators might agree by chance. Examples of kappas include Cohen's kappa [10] (meant for two raters) and Fleiss' kappa [11] (for any number of raters). It is helpful or even necessary for some tasks to create a new evaluation method by integrating existing agreement measures [12]. However, all of the techniques mentioned above assume categorical or numerical data, making them unusable for text annotations.

Text similarity methods are a more straightforward option to measure agreement between two strings. These methods, such as Levenshtein, Jaccard, and Hamming distance, calculate the distance between two vectors. The smaller the distance, the closer the strings are lexically. For example, Levenshtein distance can be considered the smallest necessary number of operations to change one string into the other. Each operation is the insertion, deletion, or substitution of a character. It was named after the Soviet mathematician Vladimir Levenshtein, the first person to consider the distance [13]. The mathematical definition of Levenshtein distance between two strings *a* and *b* is shown in Formula (1), where *tail* is all but the first character of a string.

$$\operatorname{lev}(a,b) = \begin{cases} |a| & \text{if } |b| = 0, \\ |b| & \text{if } |a| = 0, \\ \operatorname{lev}\big(\operatorname{tail}(a), \operatorname{tail}(b)\big) & \text{if } a[0] = b[0] \\ 1 + \min \begin{cases} \operatorname{lev}\big(\operatorname{tail}(a), b\big) \\ \operatorname{lev}\big(a, \operatorname{tail}(b)\big) \\ \operatorname{lev}\big(\operatorname{tail}(a), \operatorname{tail}(b)\big) \end{cases} & \text{otherwise.} \end{cases} \tag{1}$$

An alternative way to estimate inter-annotator agreement is by calculating the longest common substring. This works well for two annotators but can quickly become impractical when the number of annotators rises. One workaround for that problem is to use an algorithm that determines the longest *most* common substring that does not necessarily appear in every annotation. Finding the common substring is especially beneficial when the scope of all possible combinations for the annotations is small, such as in a task where several consecutive words need to be highlighted in a short text. This method can eliminate superfluous words or characters in the described scenario, retaining only the most essential information.

Krippendorff's alpha coefficient [14] is another option for assessing agreement among observers. It is general, applying to many different types of values (e.g., nominal, ordinal, interval, polar) and any number of annotators. The method can handle missing data and correct for small sample sizes. Most relevantly for this study, Krippendorff's alpha has also been formulated for textual content analysis [15].

3 Contribution

This section is dedicated to data exploration, preprocessing, and enrichment. Section 4.1 visualizes and describes the data in detail. Section 4.2 presents how data was selected, cleaned, and integrated as part of preprocessing. Finally, section 4.3 describes the enrichment process: design, results, and analysis of the crowdsourcing survey on Toloka.

3.1 Exploration

The data used in this study was scraped in May 2021 from the website Know Your Meme, which uses a wiki system to document internet memes and other online phenomena, such as viral videos and internet celebrities. This means that anyone on the internet can upload and edit entries on the platform. For this work, only memes are considered.

Due to the nature of wiki-based websites, all memes vary in the type, amount, and quality of the information provided. As demonstrated by Figure 4, almost all entries contain an image, title, origin, tag(s), and year. Some entries also contain additional structured information, such as type, additional references, siblings, and children.

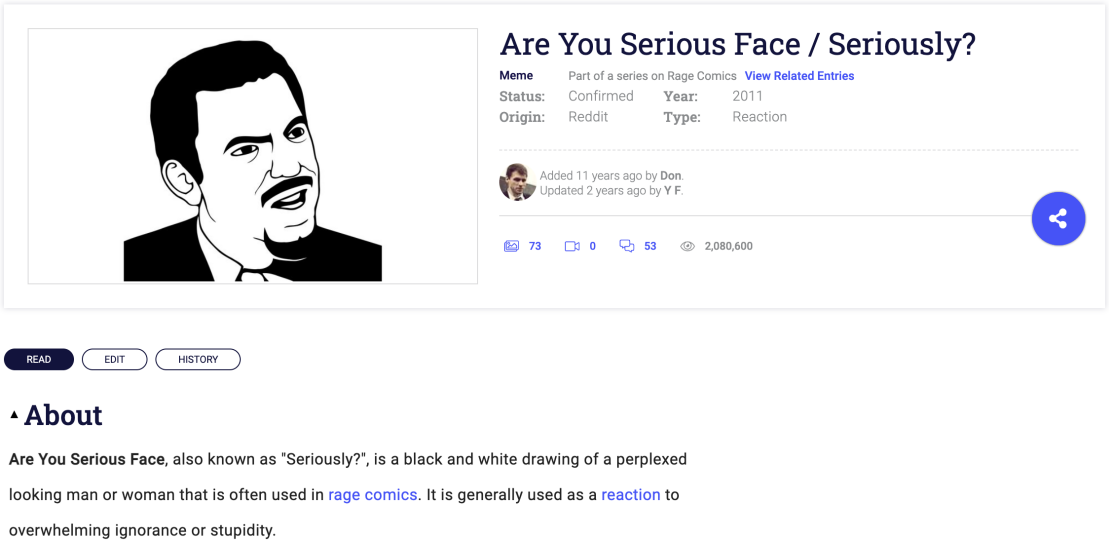

Figure 4. Header and About section of a meme on Know Your Meme.

The most interesting parts of the available metadata are year, type and tags (see Figure 5). They provide essential context and allow to investigate a more specific subset of the data, which can help to eliminate noise and draw more accurate conclusions. *Year* denotes the first appearance or source of the meme (e.g., a movie or a book). The *type* usually indicates how the meme was created, used, or what it contains. Different types can also overlap in their meaning, e.g., a snowclone is a more specific type of

exploitable. *Tags* give a meme additional context, often including information about a platform, nation, or medium. Less than 1% of memes have multiple tags or types.

Two notes regarding Figure 5: firstly, the distributions show only the most popular instances: years and types with at least 30 occurrences and tags with 200 occurrences. Secondly, it is helpful to keep in mind that approximately 91% of memes have a specified year, and 94% have at least one tag specified. Only 22% of memes have a type stated.

By observing the data, it is clear that some are incomplete. Most notably, the scraped data are missing most memes that originated after 2017. This problem has been fixed in a new iteration of scraping in March 2022[9]. As for the existing data, a large proportion of memes originate from 2010 to 2012. Most memes on Know Your Meme have been updated in the past four years, indicating that the information should be up to date.

Character, *catchphrase*, and *exploitable* are the most used types on Know Your Meme. Some types are many times more common than others, which can indicate what kinds of memes spread the most, are the easiest to create, or possibly – are the easiest to recognize and label. Even though most memes do not have a type specified, studying memes of different types separately is helpful because they might be significantly different.

The most used tags are online platforms, such as YouTube, 4chan, Twitter, Tumblr, Reddit, and Facebook (decreasing order of popularity). The tag distribution in Figure 5 may indicate that 4chan and YouTube are more popular websites for either sharing or generating new memes than Reddit and Facebook. However, this disparity could also be attributed to selection bias: it is possible that people who visit 4chan more often are also more likely to be active editors on Know Your Meme.

In addition to structured data, most memes contain three written sections called *About*, *Origin*, and *Spread*. The About section shortly describes the context in which the meme is used, Origin details the first appearance of the meme, and Spread provides information regarding the websites where the meme has spread the most. A word cloud containing the most commonly used words in a meme's About, Origin, and Spread sections can be seen in Figure 6 (more frequent words are larger).

The About section is presumably the best place to search for the meaning of a meme since no other feature seems to contain semantic information. Although it is succinct (the median length is 47 words), it may describe the emotions and ideas conveyed by the image. Most popular words in the About section, as shown in Figure 6, include types of memes (such as *video*, *image*, *song*, *character*, and *series*) but also words used to describe something (such as *refers*, *phrase*, *term*, *often*, *used*, *featuring*, and *typically*).

The median Origin section is 74 words long, measurably more than About. As expected, popular words also describe the first appearance of the meme, such as *first*, *original*, *began*, and *earliest*. The Spread section is roughly as long as About, with a median of 51 words. Frequently appearing verbs describe popularity and spread, such as *views*, *aired*, *posted*, and *released*.

[9]The new data is not considered in this thesis.

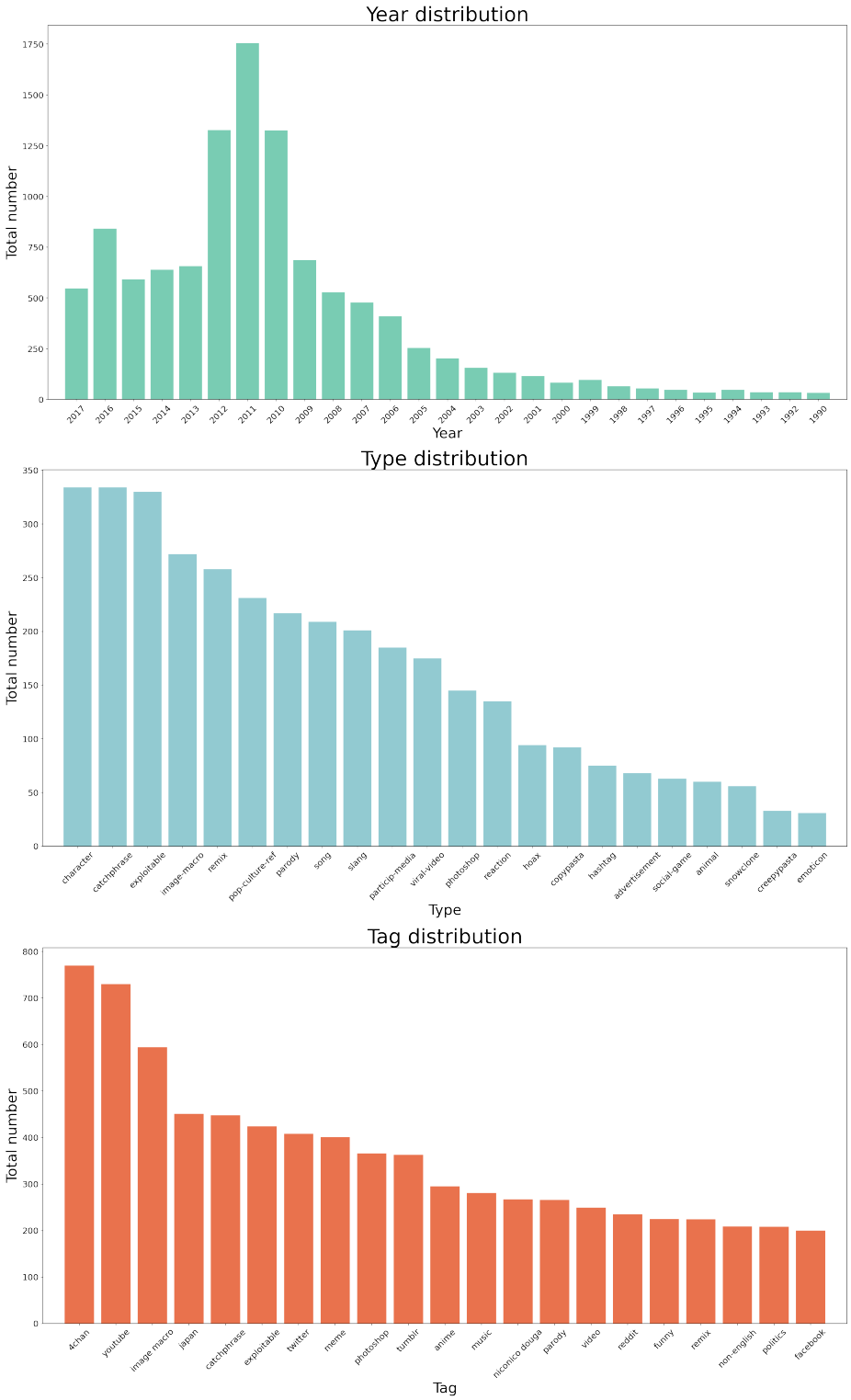

Figure 5. Distribution of parameters Year, Type, and Tag on Know Your Meme.

Figure 6. Word clouds of all About, Origin, and Spread sections.

3.2 Preprocessing

The scraped data from Know Your Meme was raw, manifesting in missing values, noise, and duplicates. Some of the more helpful information was hidden in deeper layers; some was utterly irrelevant. Therefore, a preprocessing pipeline was needed to create a cleaner and more approachable dataset. The pipeline was set up in a Jupyter Notebook, with the

essential functionalities being

1. removal of irrelevant or missing data;
2. integration with other datasets;
3. filtering based on year, type, tags, and spread; and
4. the option to export selected data in TSV or CSV format.

The dataset was read in as a Pandas[10] DataFrame with 28799 rows of data (15406 memes). After dropping duplicate entries, 12654 memes remained. The dates were converted from large integers to readable Datetime objects. The most crucial information (About section, type) was extracted from the deeper levels and added as separate columns. Automatic removal of memes with an empty About section was enabled.

Most of the preprocessing pipeline was implemented as a function with the parameters providing filtering support. To filter by time last updated, two dates can be inputted. Selecting specific combinations of types and tags is supported by logical statements. Spread on platforms is filtered by a keyword search of the Spread section. All the available options for the parameters can be displayed.

The Know Your Meme dataset was integrated with three others by using URL as a common identifier. The events dataset included dates (with statements) mentioned in the text sections on Know Your Meme. The other two datasets contained image and text entities from previous enrichment work. The integration resulted in one large dataset with information from Know Your Meme, events, and entities.

Exporting data is an essential feature of the preprocessing pipeline. For annotation, high-quality data is needed. The pipeline allows exporting memes with either an About or Origin section to ensure quality, merging the sections to provide one coherent text. It can also detect Imgflip[11] (a site containing meme instances) references in the *Additional references* section.

Figures 7 and 8 showcase the differences between raw input data and cleaned data. Note that some features contain very long information that does not fit on the screen (e.g., the length of the "content" feature for the first meme displayed is 6000 characters).

	title	url	last_update_source	category	template_image_url	meta	id	added	details	content	tags	additional_references	search_keywords	parent	siblings	children
0	This is Releva...	https://knowyo...	1547002898	Meme	https://i.kym-...	{'og:title': '...	{'@context': '...	1.229113e+09	{'status': 'co...	{'about': {'te...	[image macros,...	{}	[\"This is rel...	NaN	NaN	NaN
1	ROFLcopter	https://knowyo...	1591400337	Meme	https://i.kym-...	{'og:title': '...	{'@context': '...	1.229113e+09	{'status': 'co...	{'about': {'te...	[ascii, animat...	{'Encyclopedia...	[roflcopter]	https://knowyo...	[https://knowy...	NaN
2	Bitches Don't ...	https://knowyo...	1607986999	Meme	https://i.kym-...	{'og:title': '...	{'@context': '...	1.229113e+09	{'status': 'co...	{'about': {'te...	[myspace, 4cha...	{'Facebook,': ...	[bitches don't...	NaN	NaN	[https://knowy...
3	Leave Britney ...	https://knowyo...	1613129339	Meme	https://i.kym-...	{'og:title': '...	{'@context': '...	1.229113e+09	{'status': 'co...	{'about': {'te...	[youtube, rant...	{'Wikipedia': ...	[\"leave britn...	https://knowyo...	[https://knowy...	NaN
4	O RLY?	https://knowyo...	1615643899	Meme	https://i.kym-...	{'og:title': '...	{'@context': '...	1.229113e+09	{'status': 'co...	{'about': {'te...	[image macro, ...	{'Meme Generat...	[\"o rly\"]	https://knowyo...	[https://knowy...	NaN

Figure 7. Dataframe with raw input data.

[10] https://pandas.pydata.org/
[11] https://imgflip.com/

	url	template_image_url	about_origin	tags	year	last_update_source	other_types	imgflip	exist_about_origin
6	https://knowyourmeme.com/m...	https://i.kym-cdn.com/entr...	In Soviet Russia, also kno...	[russia, russian reversal,...	1938	2021-04-17	[]	[]	both
15	https://knowyourmeme.com/m...	https://i.kym-cdn.com/entr...	"One Does Not Simply Walk ...	[catchphrase, lotr, fantas...	2001	2021-03-18	[catchphrase, pop-culture-...	[https://imgflip.com/memeg...	both
17	https://knowyourmeme.com/m...	https://i.kym-cdn.com/entr...	"You Have My Sword, and My...	[you have my sword, and my...	2001	2021-04-06	[catchphrase]	[]	both
45	https://knowyourmeme.com/m...	https://i.kym-cdn.com/entr...	"Pretty cool guy" is a exp...	[halo, pretty cool guy, 4c...	2007	2021-01-26	[]	[]	both
432	https://knowyourmeme.com/m...	https://i.kym-cdn.com/entr...	"Bitches Love X" is an ima...	[bitches love x, tv show q...	2006	2021-01-16	[image-macro]	[]	both

Figure 8. Dataframe with cleaned data (type: snowclone).

3.3 Enrichment

In order to obtain semantic information for each meme, annotation was required. From all the data available, the About section was a good candidate. Under the assumption that the About section contains semantic information, it should be possible to highlight the relevant part in the text. Label Studio[12] was used to annotate 489 memes (see Figure 9). These memes were selected from the dataset to have either About or Origin sections and be of type reaction, exploitable, or snowclone. A large portion of those memes appeared to include the meaning. However, since meaning can be subjective, it is better to collect multiple people's opinions and check for consensus. Therefore, a questionnaire was created on the crowdsourcing platform Toloka.

3.3.1 Meme selection

Annotating all 12654 memes being financially infeasible, only 50 memes were carefully selected for annotation on Toloka. To further limit the variability of results, only three types of memes were considered: *snowclones*, *reactions*, and *exploitables*. An example of each type of meme is provided in Figure 10.

- *Snowclones* are a type of phrasal template ("fill-in-the-blank") in which certain words may be replaced with another to produce new variations with altered meanings. They can be understood as the verbal or text-based form of photoshopped exploitables.
- *Reaction* images are meant to portray a specific emotion in response to something. They are commonly used in discussion threads in a similar fashion to emoticons.
- *Exploitables* are a meme template in which a single image is manipulated through various means to achieve the intended, humorous effect. This can involve replacing words in the original image, adding words to the original image, or manipulating the positions of objects in the image to change the original's meaning.

[12] https://labelstud.io/

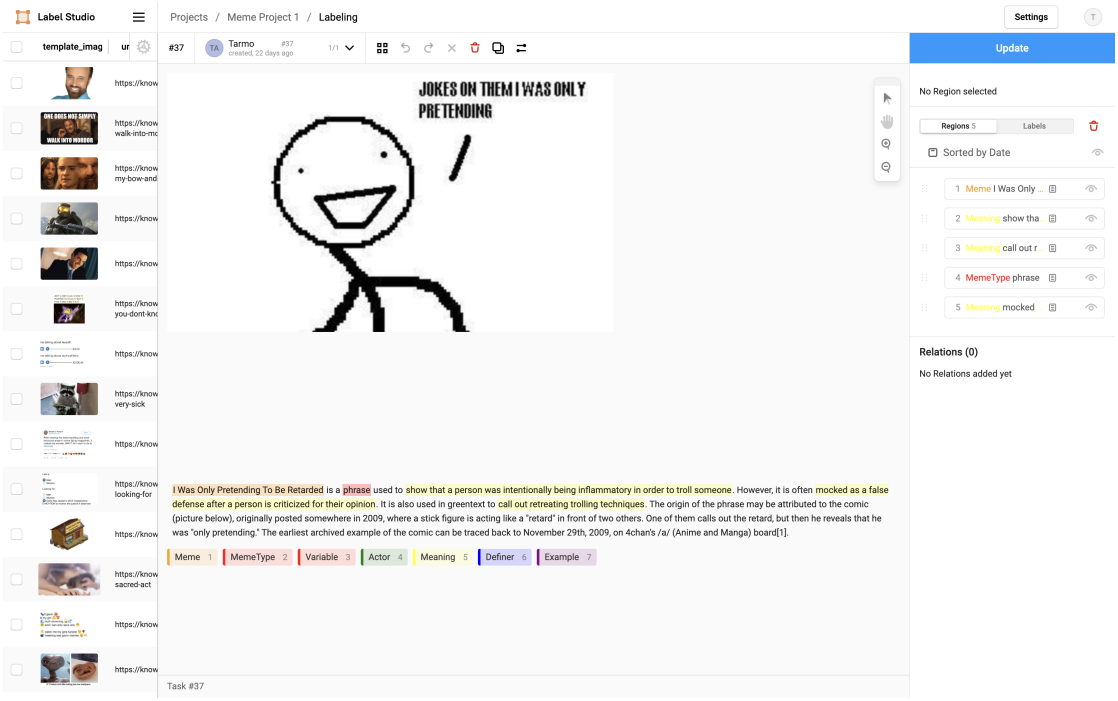

Figure 9. Label Studio interface showing a meme annotation.

Figure 10. Example of a snowclone (top left), reaction (bot left), and exploitable (right).

The reason behind selecting precisely those types was that they appeared to be more easily understandable. All memes that did not have an About section were eliminated as potential candidates. The final 50 were picked manually out of those memes that seemed

to have their meaning described in the About section. All three types were included approximately equally. The 50 memes were exported in TSV-format with a special header to indicate each input variable and its type, as per Toloka requirements. These inputs consisted of the URL of the meme's image, the About section as a string, and the URL of the meme itself.

3.3.2 Design of Toloka questionnaire

The Toloka questionnaire was written using HTML, CSS, and JavaScript. HTML and CSS were used to make up the shape and style of the survey. JavaScript enabled the creation of a custom captcha that does not allow pasting: in addition to annotation, the user had to write "I have checked the meaning" in a special box, as seen in Figure 11. The purpose of the captcha was to make sure that users spent time carefully deliberating the text and to disable bots from completing the tasks.

The captcha was only one form of quality control used in the survey. Only the top 30% of performers based on their rating on Toloka were allowed access to the survey. All participants were required to have passed an English language test on the site. Furthermore, if a user submitted fast responses (<90 seconds) too often or skipped more than two tasks in a row, they were suspended from the project. It is also possible that the slightly higher than average payment ($0.2 per task) provided extra motivation for working on one task for longer.

As shown in Figure 11, the survey consisted of a meme's image and About section. The user had to highlight the meme's title(s) and meaning(s) in the text. Collecting information about the title is helpful because (1) the dataset does not contain all alternative titles that a meme can have, and (2) the title is more objective and can estimate the quality of received annotations. Additionally, performers had to answer three multiple-choice questions regarding their personal experience with the meme and one regarding the meme type. The users also had the option to read more about the meme by clicking a link directed to Know Your Meme.

Since the task was not trivial, thorough instructions (written in HTML) were provided for the users. Some of the key instructions were to spend at least 1.5 minutes per 1 task, use the link provided to read more about the meme if necessary and highlight all alternative titles and meanings. To assess people's consensus regarding a meme later, each meme was annotated by five unique people. This means that in total, there were 250 annotations.

3.3.3 Results

The survey was fully completed with 250 annotations within 1 hour after launch. The total cost of the experiment amounted to $65: $50 went to the performers and a $15 fee to Toloka. Sixty unique performers participated, meaning that a single person submitted,

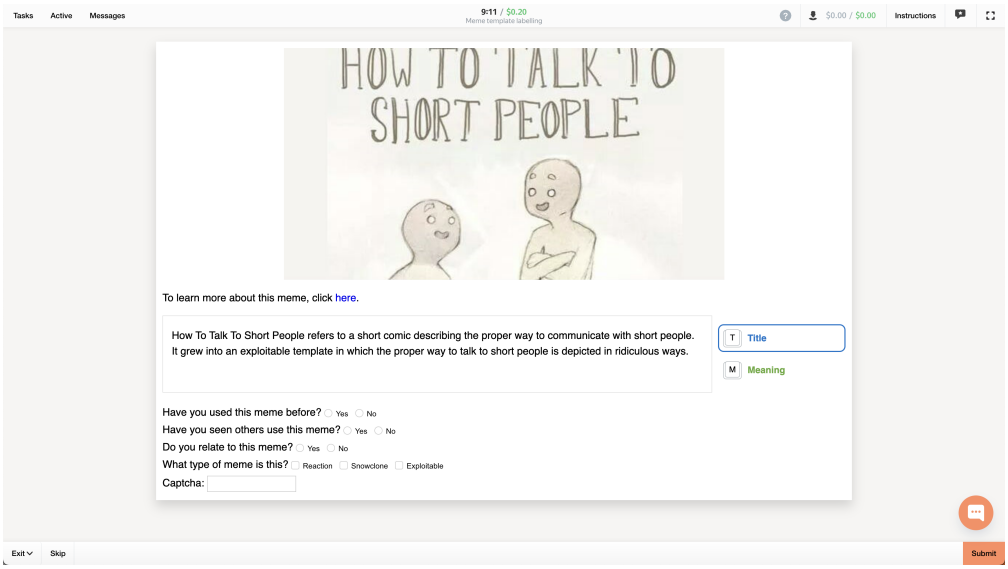

Figure 11. Toloka interface for performers.

on average, four annotations. The average time spent per annotation was 3 minutes and 35 seconds, more than was required (the maximum allowed time was 10 minutes). Twenty-three people were banned due to the set-up quality control rules: 2 for skipping assignments and 21 for submitting too many fast responses.

The outputs of the survey consisted of answers to multiple-choice questions, lists of highlighted strings, and Toloka-specific outputs, such as a unique ID, time spent per question, and less relevant parameters.

Figure 12 demonstrates that people recognized the meme under annotation 40% of the time. Annotators related to the meme roughly a third of the time and had used the meme themselves only 23% of the time. Based on these answers, the performers were somewhat familiar with the memes already, indicating the widespread usage of memes. Hopefully, this also resulted in a better understanding of the meme and, therefore, also higher annotation quality.

It is noteworthy that answers to the multiple-choice questions are moderately correlated (see Table 1). This means that the information that they capture has considerable overlap. This result was expected: if one has not seen a meme, one also cannot have used that meme before. On top of that, people tend to use memes that they relate to and vice versa.

60% of the tasks contained correct annotations of the meme type. An annotation was considered correct if at least one of the checked types matched one of the types from Know Your Meme. This figure could be inaccurate because it assumes the Know Your Meme dataset contains accurate and complete information. Other reasons to be cautious when interpreting the number include type being subjective and difficult to assess.

Table 1. Correlation matrix of multiple choice questions.

Question	Q1	Q2	Q3
Have you used this meme before? (Q1)	1	0.53	0.47
Have you seen others use this meme? (Q2)	0.53	1	0.33
Do you relate to this meme? (Q3)	0.47	0.33	1

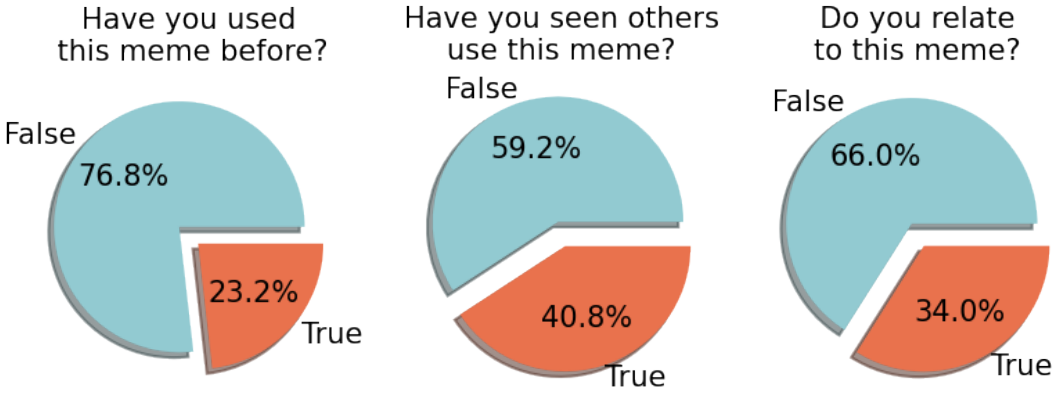

Figure 12. Proportion of answers to multiple-choice questions.

3.3.4 Evaluation of similarity

The reason behind collecting other people's annotations was to find a more objective representation of a meme's meaning. However, a representative meaning can only be achieved if there is sufficient agreement between the annotators. To evaluate inter-annotator agreement, different methods were tested. The agreement results can also be used to gauge the quality of annotations.

Number of annotations. Some performers did not annotate both Title[13] and Meaning – either they did not find a suitable label or did not attempt to. Out of the 250 total annotations, 15 missed Meaning and two missed Title. Although it was possible to highlight multiple parts of the text under one label, 194 out of 250 opted for a single continuous portion of the text (for both Title and Meaning). This figure is reasonable because the provided text is concise: only a small portion of memes have multiple titles listed in the About section.

However, the most exciting statistics are not regarding all data as a whole but using a single meme as the smallest unit of measurement. One way to assess inter-annotator agreement is the standard deviation. For 13 memes, every performer highlighted the same number of parts in the text. Any meme's highest standard deviation for Meaning

[13]Capitalized Title and Meaning refer to annotations as opposed to the general terms.

(4.8) was due to one outlier – a performer that annotated 12 different words in a single sentence. Exploring other memes with high standard deviation confirmed that a high number of annotations tend to correlate with a low annotation quality.

Length of annotations. 94% of the highlighted Titles consisted of fewer than ten words, matching our expectations. The standard deviation for the length of annotations (by characters) was 0 for 18 memes. The greatest standard deviation (101.2) was due to one performer highlighting multiple sentences as the title (perhaps mistaking the label for Meaning). Some other high standard deviations were due to some performers using one highlight for alternative titles, e.g., highlighting *"That's Where You're Wrong, Kiddo," also known as "You're Wrong Kiddo,"* with one label instead of two.

82% of the highlighted Meanings consisted of fewer than 25 words. Figure 13 shows a histogram of the individual[14] highlight lengths. Most outliers were either an entire sentence or a large part of a sentence, including superfluous information. For example, one performer's annotation for the meaning of the meme Pennywise The Clown was the following: *Online, people use the image as a reaction or to portray the things that would entice them in the face of certain death.* It could be argued that *portray the things that would entice them in the face of certain death* is the core meaning, and everything else provides context. To highlight the scope of this problem, consider the average length of Toloka annotations versus our manual annotations: 18 and 10 words, respectively.

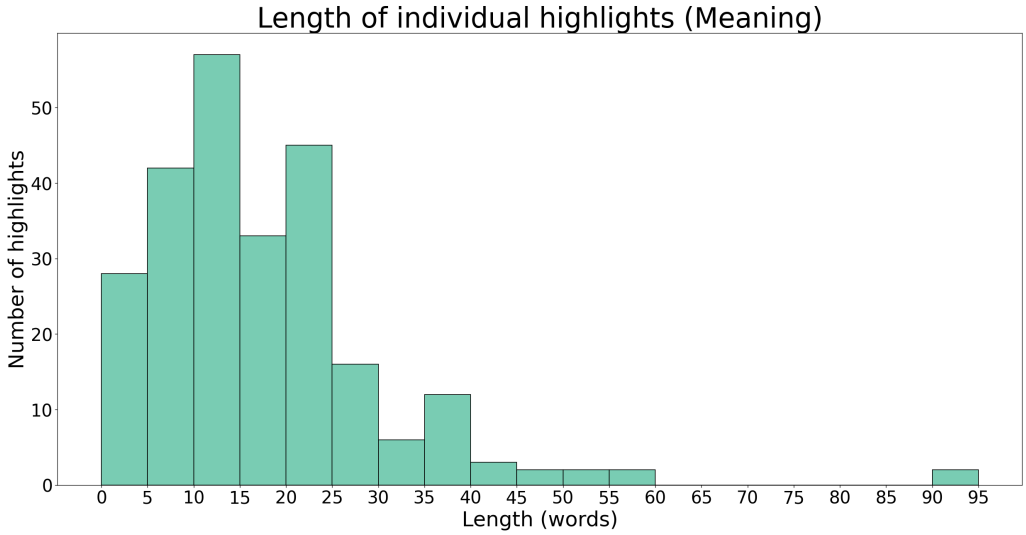

Figure 13. Distribution of the highlight lengths for Meaning annotations.

[14]For those performers who highlighted multiple sub-annotations, the average length is plotted.

Distance of annotations. The number and length of annotations do not capture either lexical or semantic similarity. To remedy this, Levenshtein distance was chosen as an easily implementable way to measure lexical similarity. The distance between two identical strings is 0 and the maximum is the length of the longest string.

We used this metric to calculate the pairwise distance between each annotation per meme. Since one annotation could include multiple strings (i.e., highlights), sub-annotations were used for the comparisons. The smallest pairwise distance of sub-annotations was chosen for each comparison between two annotators. This resulted in every meme having ten pairwise distances across its annotations.

To compare the distances, they were normalized to the range [0.0, 1.0]. At first, we tried min-max normalization, with the maximum being the greatest single calculated distance. In other words, all distances were divided by the maximum pairwise distance across the whole dataset. However, this resulted in most normalized distances being very close to 0; the maximum distance was an outlier.

Because of this problem, we settled on a different approach. For each pairwise comparison, we normalized by using their maximum possible distance, equal to the length of the longer string. A normalized distance of 0 means that the strings are identical, while 1 means that theys could not be farther apart. If an annotation was missing, it was considered to have a distance of 1 from every other annotation.

For a distribution of the normalized distances, see Figure 14. It showcases 441 pairwise distances (10 per meme minus missing data) and demonstrates that a large portion of the annotations are lexically similar while some are not. Figure 15 offers an individual breakdown of the distances, showing high variability across and within memes, indicating that only some annotators reached agreement.

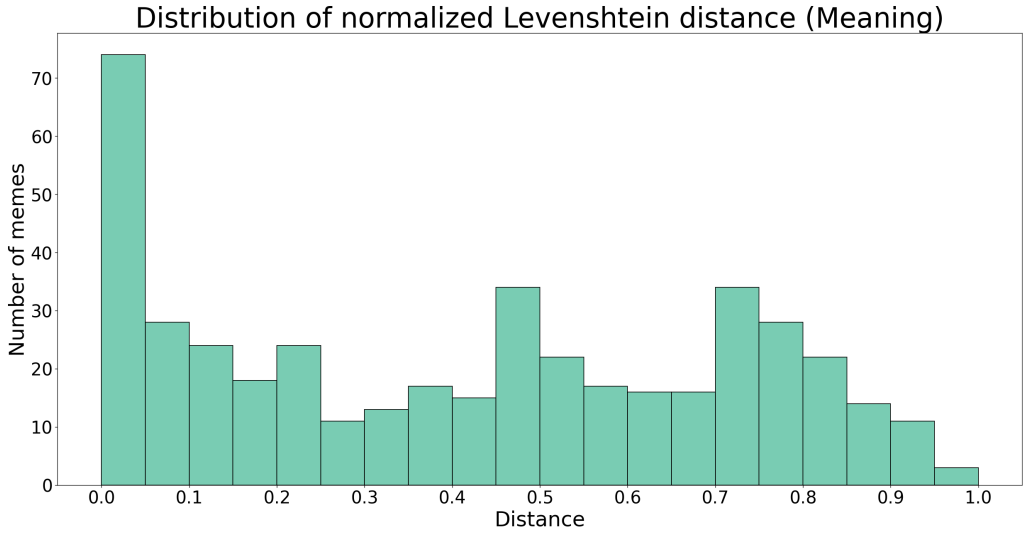

Figure 14. Distribution of normalized Levenshtein distances for Meaning annotations.

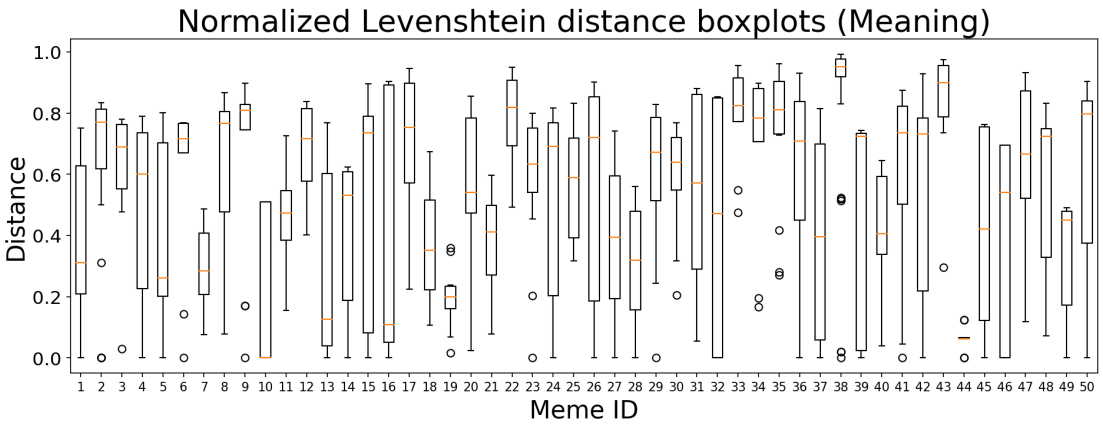

Figure 15. Boxplots of normalized Levenshtein distances for Meaning annotations.

In order to draw the line between similar and dissimilar, a *similarity threshold* was introduced. We consider any two memes with a pairwise distance less or equal to this threshold to be similar. Applying this threshold for a meme results in a symmetric similarity matrix as shown in Figure 16 where 1 represents a pairwise match and 0 represents a mismatch. A1 through A5 represent the five annotations per meme.

	A1	A2	A3	A4	A5
A1	1	1	0	1	1
A2	1	1	0	1	1
A3	0	0	1	0	0
A4	1	1	0	1	1
A5	1	1	0	1	1
Total	4	4	1	4	4

Figure 16. Meaning similarity matrix for the meme "Coincidence? I Think Not!".

Another metric is required to determine agreement: *consensus threshold*. With a consensus threshold of 0.6, the annotators are considered to agree if at least 60% of pairwise comparisons are similar. Using 0.6 as the consensus threshold, Figure 17 shows accepted memes for various similarity thresholds. The results seem unsatisfactory and may be skewed by some low-quality annotations. If one performer highlights something irrelevant or nothing at all, it eliminates four pairwise distances from being a match. The issue could be remedied by a greater number of annotations per meme. As is, the annotations might need to be manually cleaned from the noise, e.g., by removing too short or too long annotations or discarding poor annotations.

Without an ideal metric, the best way to evaluate the quality of annotations is by hand. We calculated pairwise Levenshtein distances between manual and Toloka annotations.

Based on an inspection of pairwise comparisons with large distances, the most common reasons (in decreasing order) for high distance were because the Toloka annotations:

- were very long, e.g., a whole sentence or more;
- contained a description of the image rather than the meaning; and
- contained completely irrelevant text, e.g., the Title instead of Meaning.

Although the meaning of a meme is of greater interest, Title annotations can serve as a quality metric. The title can be determined objectively and found more easily in the About section, as demonstrated by Figure 18. For the similarity threshold of 0 and consensus threshold of 0.6, 42 memes reached agreement. Future experiments may only consider Meaning annotations of those performers who correctly annotated the title.

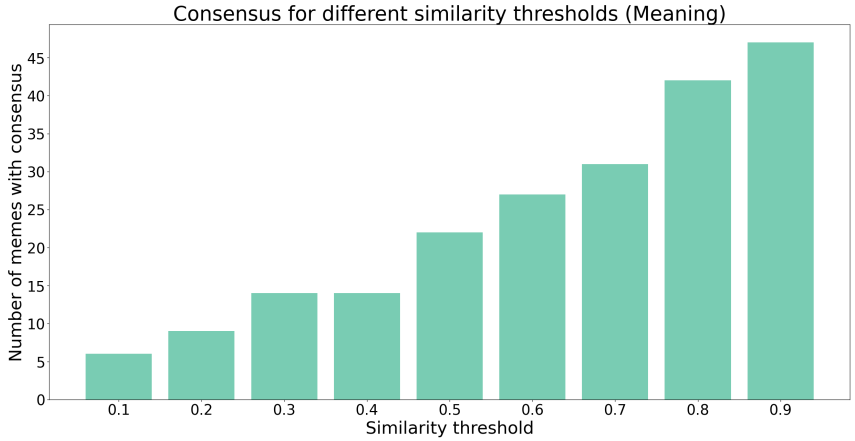

Figure 17. Number of memes with consensus for different similarity thresholds.

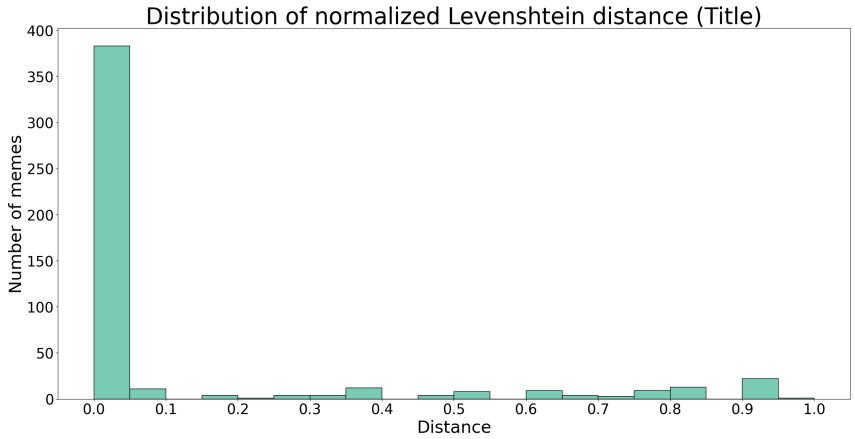

Figure 18. Distribution of normalized Levenshtein distances for Title annotations.

4 Conclusion

This thesis introduced the idea of a system capable of recommending memes based on context. To create this system, relevant data was needed. Although previous work had provided a primary dataset by scraping Know Your Meme, the goal of the thesis was to create a pipeline for preprocessing and enriching this data. The results of this pipeline can be used to determine whether it would be better to use metadata or semantic information when representing a meme.

The preprocessing pipeline was designed to explore and clean the data. Section 3.1 visualized the most critical features, such as year, type, and tags. Essential takeaways included the fact that most memes in the data originate from between 2010 and 2012, the most common types on Know Your Meme are *character*, *catchphrase*, and *exploitable*, and that tags are frequently used to specify the platform for a meme's spread.

Cleaning the data involved the removal of irrelevant or missing data and enabling filtering based on important parameters, such as year, type, tags, and spread. Another relevant feature that was implemented allowed saving the cleaned and selected data as CSV or TSV files, the latter being the format required for enrichment.

The main problem with the previously existing dataset, as diagnosed in this work, was that it was missing semantic information in a structured form. To collect this information, a crowdsourcing survey was carefully designed and conducted in Toloka on 50 memes which were selected to have high-quality existinv descriptions. Each meme was either of type *reaction*, *exploitable*, or *snowclone* and was annotated by five unique performers, resulting in 250 annotations.

The survey collected information about each meme's title, meaning, type, and the annotators' relations with the meme. 40% of performers had seen the meme in question beforehand. 60% of tasks were labeled with the correct type. However, the essential part of the survey was the semantic annotation. Different methods were tried to estimate inter-annotator agreement to evaluate the quality of received annotations. By some metrics, agreement appeared to be low. Further runs and analysis are needed to assess the agreement more accurately.

The work provided meaningful progress towards the goal of creating a meme recommending system. Potential future steps include creating a graphical and streamlined version of the pipeline and a larger rerun of the experiment. The rerun could use at least 30 annotators per 1 meme and improve the existing experiment by providing performers with more precise instructions to receive more succinct and accurate annotations. These changes could result in a higher inter-annotator agreement.

References

[1] "Meme". *Cambridge Dictionary*. Cambridge University Press. URL: https://dictionary.cambridge.org/dictionary/english/meme.

[2] Richard Dawkins. *The Selfish Gene*. Oxford University Press, 1976.

[3] Patrick Davison. "The Language of Internet Memes". In: New York University Press, Jan. 2012, pp. 120–134.

[4] Sara Cannizzaro. "Internet memes as internet signs: A semiotic view of digital culture". In: *Sign Systems Studies* 44.4 (2016), pp. 562–586.

[5] Tõnis Hendrik Hlebnikov. "Towards a Knowledge Graph of Internet Memes". Bachelor's thesis. University of Tartu Computer Science department, 2021. URL: https://comserv.cs.ut.ee/ati_thesis/datasheet.php?id=73005&year=2021.

[6] Arend Hintze. "Data Preprocessing". In: *Encyclopedia of Machine Learning and Data Mining*. 2017.

[7] Christopher Clifton. "Data mining." In: Encyclopedia Britannica. URL: https://www.britannica.com/technology/data-mining.

[8] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques (3rd Edition)*. Elsevier, 2012.

[9] J.W. Tukey. *Exploratory Data Analysis*. Addison-Wesley series in behavioral science v. 2. Addison-Wesley Publishing Company, 1977.

[10] Mary L. McHugh. "Interrater reliability: the kappa statistic". In: *Biochemia Medica* 22.3 (Oct. 2012), pp. 276–282.

[11] Joseph L. Fleiss. "Measuring nominal scale agreement among many raters." In: *Psychological Bulletin* 76.5 (1971), pp. 378–382.

[12] Marcel Bollmann, Stefanie Dipper, and Florian Petran. "Evaluating Inter-Annotator Agreement on Historical Spelling Normalization". In: Association for Computational Linguistics, Jan. 2016, pp. 89–98.

[13] Vladimir Iosifovich Levenshtein. "Binary codes capable of correcting deletions, insertions and reversals." In: *Soviet Physics Doklady* 10.8 (Feb. 1966), pp. 707–710.

[14] Klaus Krippendorff. *Content Analysis - 3rd Edition: an Introduction to Its Methodology*. Thousand Oaks: SAGE Publications, Inc, 2013.

[15] Klaus Krippendorff. "Measuring the Reliability of Qualitative Text Analysis Data". In: *Quality and Quantity* 38.6 (Dec. 2004), pp. 787–800.

Appendix

I. Accompanying files

- The source code for section 3 is available on GitHub[15].
- The datasets are available on the webpage meme4.science[16].

[15] https://github.com/tarmopungas/meme-analysis
[16] https://meme4.science/

II. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Tarmo Pungas**,
(author's name)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Towards a Semantically Rich Meme Dataset,
   (title of thesis)

   supervised by Riccardo Tommasini and Radwa El Shawi.
   (supervisor's name)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Tarmo Pungas
09/05/2022