Generative system for categorizing and relating photos

Tim Frohlich

Layout: typeset by the author using LaTeX.
Cover illustration: Unknown artist

Generative system for categorizing and relating photos

Tim Frohlich
11233982

Bachelor thesis
Credits: 18 EC

Bachelor *Kunstmatige Intelligentie*

University of Amsterdam, Faculty of Science
Science Park 904, 1098 XH Amsterdam

Supervisor
dr. B. Bredeweg

Informatics Institute, Faculty of Science
University of Amsterdam, Science Park 907, 1098 XG Amsterdam

Supervisor
MSc. M. ten Brink PhD

Civic Interaction Design, Amsterdam University of Applied Sciences
Systemic Chance, Eindhoven University of Technology

June, 2021

Abstract

In this paper, a generative system is proposed to support the process of categorizing and arranging abstract self-made photos. It aims to do so by suggesting category labels for photo collections as well as axis labels that can be used to arrange photo collections. Photo tags generated by image recognition software are transformed to a vector representation by the Word2Vec model in combination with a class based tf-idf weighting scheme for comparison. The system then clusters photos with k-means based on similarity and uses WordNet to retrieve ontological relations displaying hierarchical structure as well as semantic patterns in a photo collection. The system appears unable to capture truly symbolic relations and WordNet seems unable to capture ontological relations between tags correctly. However, evaluation shows promising results by using class-based tf-idf for topic extraction and Word2Vec for finding relevant semantic patterns. The proposed system can potentionally contribute to various subfields in the photography domain, particularly educational photovoice studies.

Contents

Chapter 1

Introduction

Computer vision software has, due to the rapid developments of computational infrastructures, machine learning techniques and big data, made impressive progress in recent years. Image recognition software in particular has proven to be extremely effective in describing and recognizing photos individually (Markoff 2014). A next and more complex step for artificial intelligence (AI) in general however, is to categorize and arrange photos based on semantic relations. Cognitive processes such as categorizing photos and arranging them in meaningful ways, are abilities that come naturally for humans. These cognitive processes display underlying patterns within a group of photos and can be captured by a) identifying clusters of semantically similar photos and b) constructing relevant concept pairs that describe semantical relations between groups of photos (Figure 1.1).

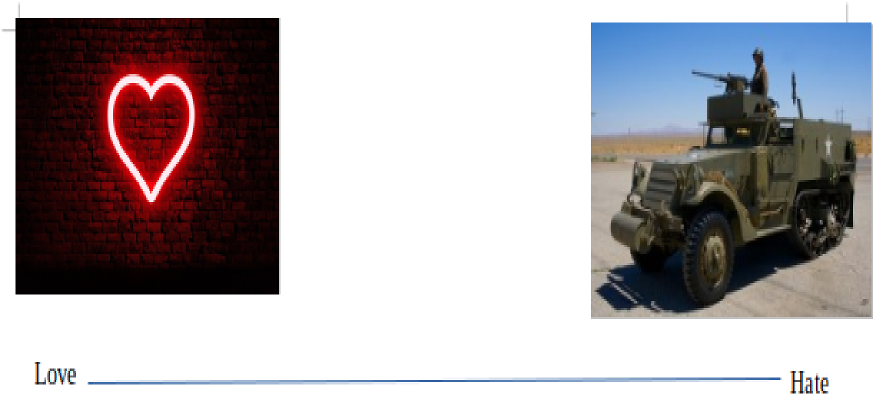


Figure 1.1: *Example of a semantic relation between photos*

Although research has been done on the categorization of photos in general (Barnard, Duygulu, and Forsyth 2001) and identifying semantical relations between keywords in general (Rattenbury and Naaman 2009; Derrac and Schockaert 2015). There has been a lack of research on categorizing and relating photos with the purpose of displaying implicit relations in photo collections.

This paper aims to bridge the gap between concrete image recognition and abstract knowledge representation based on images by proposing a generative system that can categorize photos and conceptualize semantic patterns in a photo collection. This will be achieved by answering four sub-questions:

1. How can accurate representations of photos be obtained?

2. How can clusters of photos be identifed?

3. How can a generative system identify higher hierarchical concepts of a cluster?

4. How can a generative system construct dimension labels for an axis on which photos can be arranged?

The system proposed in this paper contributes to various applications of photography. One of such applications is an educational photovoice study, where students complete categorization and relational exercises on self-made photo collections to stimulate critical reflection (Brink, Nack, and Schouten in press). The collected data in this paper is from an actual educational photovoice study and the proposed system is evaluated on this data as well.

The paper is structured as follows: Chapter two describes the theoretical background of the series of techniques used in this paper. In chapter three, context of this study is given as well as answering sub-question one by presenting the used image recognition software. Chapter four introduces the word embedding method that represent tags as weighted vectors in a multidimensional space and functions as a basis for the clustering method. Chapter five answers sub-question two by describing how the word embeddings of tags are used in clustering photos. In chapter six sub-questions three and four are answered by describing the process of identifying semantical relations between clusters. Finally, in chapter seven the generative system is evaluated based on human expertise by a) match on data and b) expert review.

Chapter 2

Theoretical background

The first step in the proposed system is to accurately represent a photo with image recognition software. There are numerous of renowned artificial intelligence companies that offer an image recognition service (*Clarifai* 2021; *Microsoft Azure Computer Vision* 2021; *ibm Watson Visual Recognition* 2021; *Amazon Rekognition* 2021; *Google Cloud Vision* 2021). Performance of these services are evaluated in Korot et al. (2020) for image classification tasks and demonstrated similar performance.

These services typically describe a photo with natural language in the form of a list of keywords (or tags). Due to the unstructured nature of tags, simple statistical metrics such as term frequency - inverse document frequency (tf-idf) can be highly efficient for ranking tags of a photo collection (Rattenbury and Naaman 2009). In Rattenbury and Naaman (2009) an adaptation of the original term frequency - inverse document frequency (tf-idf) is used so that tags that only occur in a specific class are ranked higher than tags that occur widely spread over multiple classes. This adaptation of tf-idf suggests a better retrieval of relevant keywords than the original tf-idf formula when documents (or photos) are organized in classes. Although tf-idf seems relevant for ranking photo tags, it cannot capture semantic relations between tags.

To capture semantic relations between tags, a word embedding technique must be used. A well known word embedding technique is to represent words as vectors. Vector representations of words improve learning algorithms by providing a means to store information and identify relations between words (Mikolov et al. 2013a). Vector representations are based on semantical and syntactical relations to other words, thus similar words are grouped together. The word vectors can be used in combination with a tf-idf weighting scheme to represent weighted word vectors (White et al. 2015; Huang, Yin, and Hou 2011; Liu et al. 2018).

To identify clusters of semantic groups, a commonly used clustering method in natural language processing is the K-means algorithm (Zhang et al. 2018; Ma and Zhang 2015; Ramage et al. 2009). K-means tends to decrease in performance with increasing dimensions because of its distance function (Tanioka and Yadohisa 2012). The algorithm does not work in high dimensional data because distances in high dimensions between data points cannot be measured in euclidean distance. Another shortcoming of k-means, is that it increases in computational power with increasing dimensions. To improve the performance of K-means and reduce its' computational power, dimensionality reduction techniques must be applied to the data. A well-known dimensionality reduction technique is principal component analysis (PCA), Ding and He (2004) have proven that PCA in combination with K-means is an effective clustering method.

There are various ways to extract topic labels from data. This paper focuses on ontology based topic modelling only, because photos in the collected data are already categorized in an abstract way (Chapter 3). Logically the photos in these categories share a common higher hierarchical term (hypernym). In Barnard, Duygulu, and Forsyth (2001) photos are organized hierarchically by combining photo captions with hypernymy relations from an ontological database. In Allahyari and Kochut (2015) a weighted graph based topic model is proposed. Combining hypernymy relations with semantical relations have proven to be efficient topic models (Barnard, Duygulu, and Forsyth 2001; Allahyari and Kochut 2015).

Chapter 3

Application context and data

One of the fields this paper contributes to is photovoice as an educational tool. The data in this paper and the specific categorizing and relating tasks originate from an actual photovoice study. In this chapter the photovoice method of this study as well as the structure of the collected data will be explained.

3.1 Application context

Photovoice is a visual participatory based research method where participants of the study capture and reflect upon their needs by sharing a photographic documentation of their lives. Its' main goals are to inspire others, to stimulate critical dialogue and to reach policy makers (Wang and Burris 1994). Originally photovoice has been used mainly to empower people by giving them a voice and consequently influence policy. Alternatively it can also be applied as an educational tool, to stimulate critical thinking and encourage discussion among students (Andina-Díaz 2020; Brink, Nack, and Schouten in press; Schell et al. 2009; Cooper, Sorensen, and Yarbrough 2017). A photovoice research typically follows a series of specific steps that can vary depending on the desired learning outcomes. The photovoice process of the study that the collected data originates from is described briefly:

1. Initial concepts for taking photos are identified.

2. The photovoice method is introduced to students.

3. Students take photos.

4. Students reflect upon the photos and engage in discussions about them.

The photovoice study that this paper collected data on is set up as follows: 21 Bachelor design students participated in the photovoice study and completed weekly assignments. Every student was assigned three (out of six) concepts with the instruction to make a minimum of nine photos per concept, ultimately resulting in a total of 574 photos. The six possible concepts are identified by researchers and are defined as a general concept, 'Anger' for example. Additional assignments for students include: ranking, associating and classifying the photos of a concept within a thinking frame. Thinking frames in the context of this research are defined as a set of graphic frames that help organize the photos in appropiate categories or timelines that lets students see the concepts from different perspectives and consequently encourage discussion and stimulate critical reflection. The axis labels in the graphic frames are formulated by the students. Although this is an unfinished study, one of the bottlenecks that researchers found, was that a self-guided process leading to critical reflection is not self-evident. Triggers to support students in this process might be needed.

3.2 Photo dataset

The initial dataset consists of 574 photos distributed over six concepts as defined by the researchers. The concepts are defined as general themes:

- Anger
- Impuls
- Network
- Climate change
- Student well being
- Garbage is non existent

The photos are more or less evenly distributed because students were asked to come up with at least nine photos per theme. It is important to note that because of the nature of a photovoice study, photos can be extremely abstract and multi-interpretable. Even for humans it can be challenging to accurately describe some of the photos in this dataset. An abstract example from this dataset is the photo shown in Figure 3.1 assigned to the theme 'Anger'.

Figure 3.1: *Example of an abstract photo in the collected data*

3.3 Preprocessing the data

Directory structures as well as file naming of collected data was inconsistent. To properly process the photos, they had to be structured and named in a logical way. The following pre-processing steps were applied on the original data:

- Recognizable faces were blacked out for privacy reasons.
- Unrecognizable formats were converted to .jpg file extensions.
- Photos were renamed to a number in the range [1, 573].
- Student names were converted to a number in the range [1,21].

3.4 Constructing a dataset of photo tags

Image recognition software can now be used to create an accurate representations of photos. Clarifai is an artifical intelligence company that uses advanced machine learning techniques to identify text, video and photos (*Clarifai* 2021). Clarifai offers pre-trained deep learning networks for small and broadly ranged datasets like the set of photos in our data. The photos are sent to the Clarifai API for image classification, the API then returns a list of 20 concepts describing the photo. Clarifai offers a variety of models designed for different categories, however the photos in our dataset are extremely diverse, Clarifai's general model is therefore the most suitable.

The collected data is then combined in a logically structured excel file that functions as the new dataset where further data exploration can be performed upon (Figure 3.2)

Photo-id	Photo source path	student-id	theme	tag1	tag2	...	tag20

Figure 3.2: *Dataset structure*

Chapter 4

Word embedding

4.1 Tf-idf

Despite its' simplicity, term frequency - inverse document frequency (tf-idf) is still a widely used statistical measure in natural language processing (Aizawa 2003). It considers the amount of times a word appears in a document as well as in how many documents the word is in, to disregard words that are common in all documents. From now on, in the context of tf-idf, a document will refer to a set of tags that represent a single photo. The most important downsides of this metric are that it cannot capture sequence information within a sentence or capture semantics. In this study however most of these disadvantages are not relevant because the text in our dataset consists of a list of limited concepts generated by Clarifai. Unlike a book that can contain a large vocabulary with a wide variety of synonyms in its' text, the tags in our dataset originate from Clarifai's general image recognition model which contains 11.000 unique concepts (*Clarifai* 2021). The tf-idf metric is used as a means to assign weight to all unique concepts to measure its' relevance within a defined concept of the photovoice study e.g. 'Anger'.

4.2 Class based tf-idf

The class based tf-idf formula is an adaptation of the original tf-idf formula that generates scores for a term based on the class that they are in (Grootendorst 2020). Rather than the original formula it will not base its score on individual photos but find the most relevant terms within a class while ignoring common terms in the vocabulary set of all classes. The original tf-idf formula is adapted to the form shown in equation 4.1 where t is the frequency of a tag within each class i, divided by the total number of tags w. The total number of photos is then divided by the total frequency of a tag t within all n classes.

$$\frac{t_i}{w_i} \text{ x } \log(\frac{m}{\sum_j^n t_j}) \tag{4.1}$$

4.3 Word vectors

Representations of words in a vector space is one of the most promising technologies in natural language processing. Unlike primitive approaches such as tf-idf, word vectors can capture semantic and syntactic relationships between words.

4.3.1 Word2Vec

Word2vec is a neural network model that learns from text input using either continous bag of words(CBOW) or skip-gram (Mikolov et al. 2013b; Mikolov et al. 2013a). Word2vec represents words in a vector space where semantically similar words are in close proximity. Similarity between words can be measured as the cosine distance between two word vectors. Besides clustering similar words, these vectors also have another interesting property: they can be used in basic mathematical operations. A common word2vec example of this property is: vec('king') + vec('woman') = vec('queen').

Depending on the data and goals, the size of a training dataset for this neural network should be significant therefore also requiring a considerable amount of computational resources. As mentioned before, the dataset in this study is widely ranged and computational resources are scarcely available, therefore it would be infeasible to succesfully train this algorithm on a training dataset large enough to capture all of the generated unique terms by Clarifai. As part of the Word2Vec project, researchers have released a pre-trained word vector dataset that has been trained on roughly 100 billion words which will be used in this project. The data contains 300-dimensional pre-trained vectors for 3 million words and phrases.

4.3.2 Weighted word vectors

A generated Clarifai tag will be represented as a Word2Vec vector multiplied by its respective class based tf-idf weight. This is demonstrated in Equation 4.2, where w is the tf-idf score of this tag within its respective class and concept[0-300] are the floating point numbers of this vector.

$$w \ * \begin{bmatrix} concept_0 \\ concept_1 \\ \vdots \\ concept_{300} \end{bmatrix} \tag{4.2}$$

A single generated Clarifai tag is represented as a weighted 300-dimensional vector, a single photo however contains 20 tags. We can use basic mathematical operations on Word2Vec vectors to combine these 20 vector representations to represent a photo (Equation 4.3).

$$\begin{bmatrix} concept1_0 * w \\ concept1_1 * w \\ \vdots \\ concept1_{300} * w \end{bmatrix} + \begin{bmatrix} concept2_0 * w \\ concept2_1 * w \\ \vdots \\ concept2_{300} * w \end{bmatrix} \tag{4.3}$$

This method is known as Sum of Word Embeddings (SOWE) and has been proven to be an effective way to embed the semantics of a sentence (White et al. 2015). In White et al. (2015) this method was evaluated on sentences, whereas in this study the 20 concepts are not syntactically connected the same way as in a sentence. Performance should therefore, be even better in this study than in White et al. (2015).

Chapter 5

Clustering photos

Every photo now has a single vector representation that can be used to cluster similar photos. The amount of clusters will determine the amount of suggested categories and the clusters in itself will determine the topics of these clusters. To find these clusters the k-means clustering algorithm from sk.learn (Buitinck et al. 2013) is used. K-means is a simple but effective clustering method, it does come with some downsides however. The algorithm tends to decrease in performance with increasing dimensions because of its distance function (Tanioka and Yadohisa 2012). The algorithm does not perform well in high dimensional data because distances between data points in high dimensions cannot be measured in euclidean distance. Another disadvantage is that k-means increases in computational power with increasing dimensions. To improve the performance of the algorithm and reduce its' computational power, dimensionality reduction techniques must be applied to the data.

5.1 Principal component analysis

Principal component analysis (PCA) is a dimensionality reduction technique that reduces the number of variables while preserving most information. The remaining variables are called the principle components and share a distributed amount of variance that each of them capture. By reducing the dimensions of the data while preservering its information, computational speed increases as well as clustering performance. A useful application of PCA when transforming high dimensional data, is that it allows visualization of the data. Transforming the dataset of this project with PCA however, showed that roughly 10-15% of the variance was preserved after reducing 300-dimensional data to two-dimensional data. Common practice is to create a scree plot and choose the number of principal components based upon maximizing variance. To increase clustering performance and speed up

K-means as well as preserving information, a minimum of 75% captured variance has been set for PCA. The reasoning behind this threshold, is that scree plots show that dimensions remain very high when maximizing variance on the collected data.

5.2 Finding the optimal number of clusters

To automatically generate clusters without human interaction, the k-means algorithm selects the optimal number of clusters by using an empirical method known as the elbow method. Traditionally, this method uses the distortion score (Sum of Squared Errors) as a metric and selects a value for K where marginal gain starts to decrease (Bengfort et al. 2018). Alternatively, when data is not clearly clustered the silhouette score can be used as a metric to determine the quality of clusters and select a value for K based on local maxima (Bengfort et al. 2018). The elbow method is applied for K in the range of [2-9] because of a lack of computational resources as well as the fact that a theme only contains about 90 photos. In both metrics a strong inflection point in the plot indicates an optimal value for K. An example of this method is shown in Figure 5.1 where no clear inflection point is visible for the distortion score. The system in this paper therefore makes use of the silhouette metric to determine the optimal number of clusters.

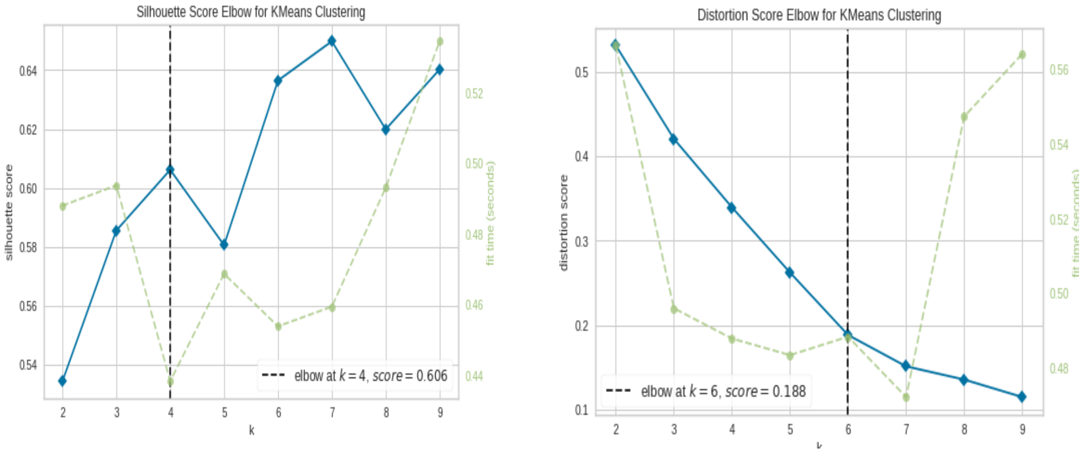


Figure 5.1: *An example of the elbow method on data in the theme 'Network' using the silhouette metric (left) and the distortion metric (right)*

Chapter 6

Semantic relations

As part of an educational photovoice assignment, students will share and discuss photos. In this stage of a photovoice process, learning and self-reflection takes place. The first and easiest step in this process is categorization, where students distribute photos over a number of categories that they find suitable. A hypernym is the equivalent of the term 'category' in linquistic jargon and is the term that will be used from now on in this paper.

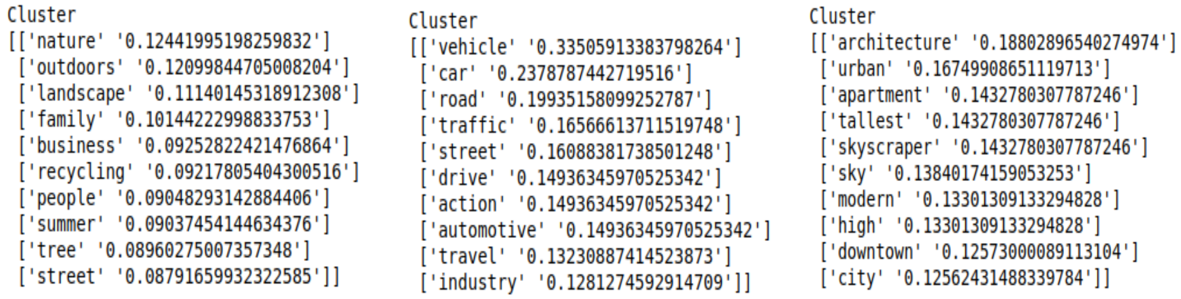
Next, students increased complexity by creating an axis in which pairs of opposite words (For example: "Hot", "Cold") are introduced as axis labels to arrange photos. An antonym is the term used in linquistic jargon to describe a pair of opposite words and is the term that will be used from now on in this paper. Although other relations between photos were introduced in Brink, Nack, and Schouten (in press), the key semantic relations were hypernyms and antonyms.

6.1 Keywords of a cluster

After performing K-means clustering within a theme, photos within a cluster should be relatively similar. To generate the most relevant keywords within a cluster we apply class-based tf-idf (Equation 4.1) once again where clusters in a theme are considered classes. The result is a ranked list of keywords for every cluster in every theme. An example of this is shown in Figure 6.1 that shows the keywords of three clusters in the theme 'Climate change'.

6.2 WordNet

WordNet is a large lexical database of the english language. It contains nouns, verbs adjectives and adverbs grouped into sets of synonyms described as synsets.



```
Cluster
[['nature' '0.12441995198259832']
 ['outdoors' '0.12099844705008204']
 ['landscape' '0.11140145318912308']
 ['family' '0.10144222998833753']
 ['business' '0.09252822421476864']
 ['recycling' '0.09217805404300516']
 ['people' '0.09048293142884406']
 ['summer' '0.09037454144634376']
 ['tree' '0.08960275007357348']
 ['street' '0.08791659932322585']]
```

```
Cluster
[['vehicle' '0.33505913383798264']
 ['car' '0.2378787442719516']
 ['road' '0.19935158099252787']
 ['traffic' '0.16566613711519748']
 ['street' '0.16088381738501248']
 ['drive' '0.14936345970525342']
 ['action' '0.14936345970525342']
 ['automotive' '0.14936345970525342']
 ['travel' '0.13230887414523873']
 ['industry' '0.1281274592914709']]
```

```
Cluster
[['architecture' '0.18802896540274974']
 ['urban' '0.16749908651119713']
 ['apartment' '0.1432780307787246']
 ['tallest' '0.1432780307787246']
 ['skyscraper' '0.1432780307787246']
 ['sky' '0.13840174159053253']
 ['modern' '0.13301309133294828']
 ['high' '0.13301309133294828']
 ['downtown' '0.12573000089113104']
 ['city' '0.12562431488339784']]
```

Figure 6.1: *An example of the ranked lists of keywords corresponding to clusters in a theme*

All synsets are linked by semantical and lexical relations. These relations include but are not limited to: hyponymy (hypernyms), antonymy (antonyms) and synonymy (synonyms).

6.3 Hypernyms

6.3.1 Inital approach

Creating a weighted hypernymy graph

An important property of the hypernymy structure in WordNet is that all possible noun hypernyms ultimately go up to its' root node: 'Entity'. Logically the top keywords of a cluster should be closely related, therefore creating a weighted graph with the keywords as leaf nodes creates a graph connecting every node eventually through the common root node. Combining this with the class-based tf-idf score as weight for every keyword creates a weighted graph for every cluster within a theme. To find the best hypernym in this graph, the approach is to maximize weight and find the hypernym that is at most 1-3 nodes from a leaf node.

Observations suggest that hypernyms for various clusters are very general and/or identical. There are various explanations for this behavior: Firstly the WordNet hypernym database is not flawless and seems limited in its coverage. Secondly, keywords within a cluster (generated by Clarifai) can already be described as general concept, (e.g. 'nature'). When finding the hypernym of these keywords, results are very abstract and not relevant to the photovoice process. Lastly, keywords within a cluster are not always as similar as one would hope. Once again this results in very abstract and general concepts.

A simplified example that summarizes these shortcomings is shown in Figure 6.2, where a hypernym graph is created with three keywords: 'Nature', 'Flower' and 'tree'. Logically, these three keywords should have a common hypernym and in the very least be connected. The graph however, shows that the common connected node in this case is 'Entity', the root node in the WordNet database.

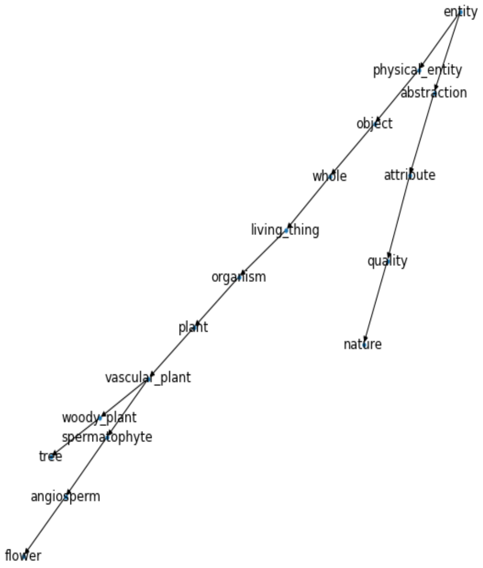


Figure 6.2: *An example of a hypernym graph based on WordNet*

6.3.2 Final approach

The weighted graph approach suggests that keywords might be too abstract already to go up a level hierarchically. Consequently, the keywords themselves may be suitable topic suggestions for the categorization process. In this new approach, the top 10 keywords of a cluster are extracted as possible topic labels. Although some of these keywords are relatively similar, they can be quite diverse. To capture every unique semantic concept in this list, the top 10 keywords are clustered in K optimal clusters where the keyword with the highest weight is extracted from each cluster. As a result there are K topic suggestions for every cluster in a theme. Once again the keyword with the highest tf-idf score is extracted as a possible topic label for a cluster.

Alternatively to clustering, unique semantic concepts could be extracted from the keyword list based on their hierarchical level in WordNet. However, this approach does not work because observations suggest that the length of a hypernym's path to the root node does not neccesarily reflect its level of abstractness.

6.4 Antonyms

In the WordNet database, an antonym can be an adjective, verb or noun. A list of all possible antonyms is retrieved from the WordNet database, ultimately containing 3243 antonyms. The top 10 keywords of a theme are then represented as a single cluster vector by using the SOWE method (Equation. 5.3). The cosine similarity is then calculated between every antonym and a cluster vector and stored in a lists of similarity scores. The 10 antonyms with the highest similarity scores for a theme are then suggested as relevant axis labels.

Chapter 7

Evaluation

7.1 Set up

Although the photovoice study that this paper collected data from includes hypernym and antonym results from students, results can be incomplete and insufficient overall to truly evaluate the performance of the proposed system. For that reason and to evaluate the generated concepts from Clarifai, three human experts were recruited. All human experts are students and followed the same process as in the photovoice study described in chapter three (Application context), apart from taking photos themselves. Each human expert evaluated the generated Clarifai concepts of 10 photos as well as the generated results (hypernyms and antonyms) of one theme. Both the Clarifai concepts as well as the hypernyms and antonyms were evaluated two-ways. Firstly by matching on data, where the formulated concept of a human expert must match the generated concept exactly to be regarded as correct. Secondly by expert review, where human experts classify a generated concept as relevant or irrelevant. The evaluation process for each human expert can be partioned in three phases: tags, hypernyms and antonyms.

7.1.1 Tags

Each human expert was shown a total of 10 photos and asked to formulate 10 concepts that describe a shown photo. The human expert also reviewed generated Clarifai concepts by classifying a photo's concepts as relevant if 10 or more concepts were relevant to this photo and irrelevant otherwise. Although there are twenty generated concepts per photo, the system represents a photo as a single vector. The reason behind a threshold of 10 or more relevant concepts, is that if at least half of the generated concepts is relevant, a vector representation of this photo will have a suffient representation of a photo's meaning. Another reason for this threshold is that generated Clarifai concepts can be synonyms and/or words that are not included in the Word2Vec dataset, therefore not included in the vector representation and thus disregarded.

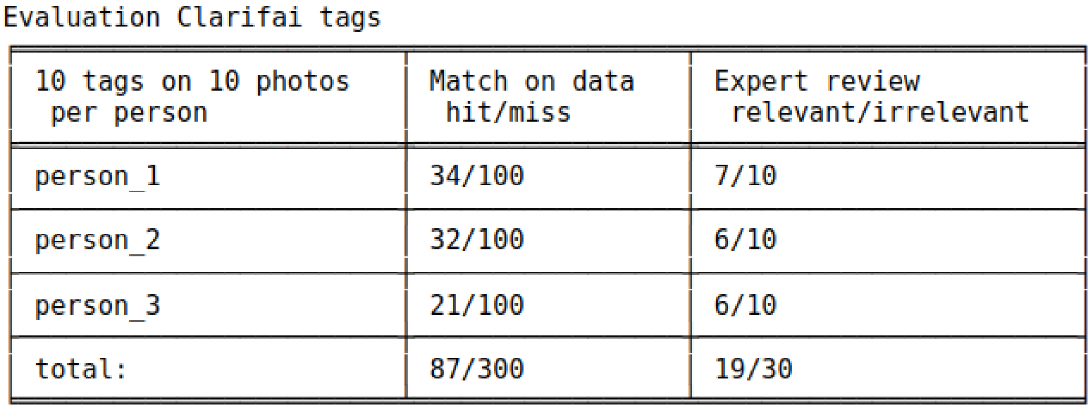
7.1.2 Hypernyms and antonyms

The human expert was shown all photos in its' assigned theme which contained roughly 90 photos. Next, the human expert formulated five hypernyms and five antonyms. The formulated hypernyms and antonyms were then compared with the generated hypernyms and antonyms by matching on data and expert review.

7.2 Results

7.2.1 Tags

On average 29/100 concepts formulated by the human expert matched exactly with the generated Clarifai concepts of a photo. The human experts classified the Clarifai concepts as relevant 6/10 times on average. The summarized results per person are shown in Figure 7.1.



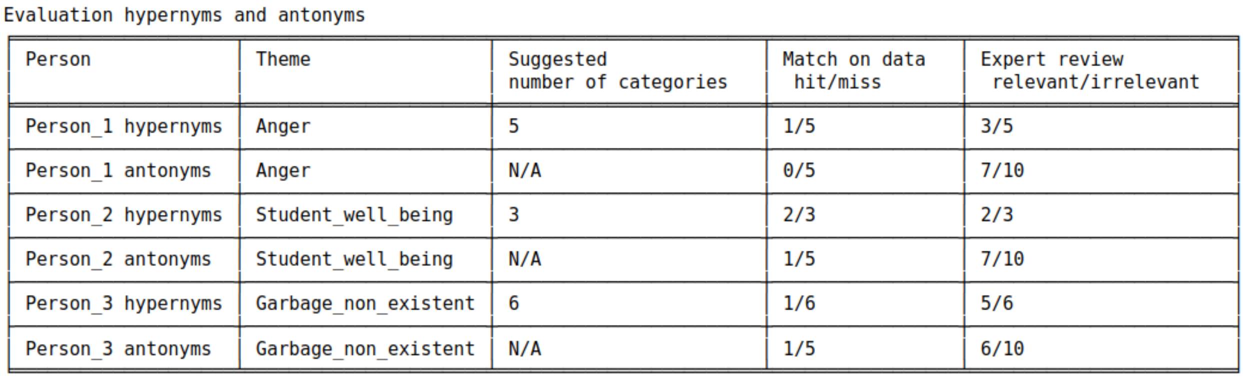
Evaluation Clarifai tags

10 tags on 10 photos per person	Match on data hit/miss	Expert review relevant/irrelevant
person_1	34/100	7/10
person_2	32/100	6/10
person_3	21/100	6/10
total:	87/300	19/30

Figure 7.1: *Tag evaluation table*

7.2.2 Hypernyms and antonyms

The formulated hypernyms by human experts matched the generated hypernyms exactly 4/14 times. The formulated antonyms matched the generated antonyms exactly 2/15 times.
Human experts classified generated hypernyms and generated antonyms as relevant 10/14 and 20/30 times respectively. Evaluation results per person are shown in Figure 7.2. A more detailed view of the hypernym and antonym evaluation is shown in Appendix A.



Evaluation hypernyms and antonyms

Person	Theme	Suggested number of categories	Match on data hit/miss	Expert review relevant/irrelevant
Person_1 hypernyms	Anger	5	1/5	3/5
Person_1 antonyms	Anger	N/A	0/5	7/10
Person_2 hypernyms	Student_well_being	3	2/3	2/3
Person_2 antonyms	Student_well_being	N/A	1/5	7/10
Person_3 hypernyms	Garbage_non_existent	6	1/6	5/6
Person_3 antonyms	Garbage_non_existent	N/A	1/5	6/10

Figure 7.2: *Hypernym and antonym evaluation table*

Chapter 8

Conclusion

The proposed system in this paper has shown that it is capable of displaying basic levels of human intelligence in a photovoice process. By using a combination of computer vision and natural language processing techniques it can transform concrete image recognition results into an abstract knowledge representation.

The foundation of this system is the set of Clarifai tags that represent a photo. Even though evaluation shows that the image recognition software's performance is not perfect on collected data, results can be considered sufficient for the purpose of identifying larger patterns in a photo collection.

To cluster similar photos, photos can be represented as the sum of their weighted multi-dimensional tag vectors. Dimensionality reduction is then applied on vectors to improve computational performance as well as k-means clustering.

Furthermore did an ontology based topic modelling approach, produce concepts far too abstract and thus appear unsuitable for topic extraction. Topic extraction based on the tf-idf scores of keywords in a cluster while filtering out similar words shows promising results. Although exact match ratios on hypernyms are quite low, human experts have classified the majority of the hypernyms as relevant.

Lastly, the WordNet database seems suitable for antonym extraction due to its vocabulary size. By measuring the cosine similarity between the vector representation of an antonym and the vector representation of a cluster of photos, relevant axis labels for the purpose of arranging photos can be generated. Similarly to the generated hypernyms, match on data ratios were low for generated antonyms. Expert reviews however, were generally positive on the relevance of generated antonyms.

The system in this paper can potentially contribute to the learning curve of students by suggesting hypernyms in a categorization exercise as well as antonyms in a relational exercise that reflects deeper relations between photos. Research has shown that a self-guided photovoice process resulting in critical reflection is not self-evident and can be challenging for some students (Brink, Nack, and Schouten in press). By providing new perspectives to students, especially students that find a self-guided photovoice process challenging, intented learning outcomes such as initiating dialogue and ultimately critical reflection can be stimulated.

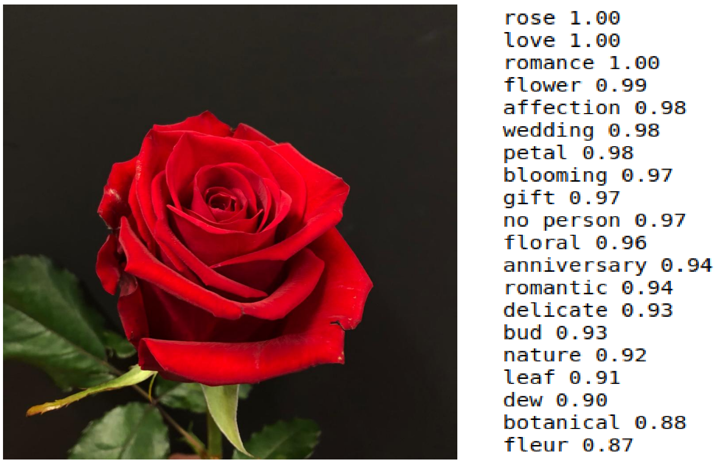
Chapter 9

Discussion

9.1 Findings

The design of a generative system, designed specifically for categorization and relational exercises on photos, has been presented in this paper. The system essentially makes use of the results of two generative processes.

Firstly, the concepts generated by the image recognition software. Even though evaluation indicates that some information is lost in the process, the quality of the image recognition software seems sufficient for the purpose of this system: finding patterns in a set of photos. But do the generated concepts describe a photo literally or do they also capture more symbolic meanings?
Evaluation of this software has been carried out globally without acknowledging the different semantic relations a photo can represent. A photo can describe three relationships as defined by semiotic terminology: iconic, indexical and symbolic. The system in this paper depends greatly on the performance of the available image recognition software and its ability to capture the various relationships. An iconic relation means the object in a photo represents the meaning of a photo literally. An indexical relation means the object in a photo has a causal relation that is implicitly expressed, for example: a football may represent recreation. A symbolic relation means the object has an arbitrary relation with the meaning of a photo. A symbolic relation must be learned, for example: a red rose may represent romance. Photos in a photovoice study tend to represent larger abstract concepts and therefore often contain indexical and symbolic relations. Although the various relations have not been evaluated separately, observations show that Clarifai does try to capture indexical and symbolic relations. The red rose example is shown in Figure 10.1 and shows that the symbolic meaning of a red rose such as romance and love can be captured by Clarifai.



```
rose 1.00
love 1.00
romance 1.00
flower 0.99
affection 0.98
wedding 0.98
petal 0.98
blooming 0.97
gift 0.97
no person 0.97
floral 0.96
anniversary 0.94
romantic 0.94
delicate 0.93
bud 0.93
nature 0.92
leaf 0.91
dew 0.90
botanical 0.88
fleur 0.87
```

Figure 9.1: *Clarifai concepts on a red rose*

Secondly, the suggested hypernyms and antonyms for a photo collection. Due to the subjective nature of photos in a photovoice study, evaluating the hypernyms and antonyms can be challenging because results are extremely subjective. One student could for example classify a generated hypernym or antonym as relevant while another student would classify it as irrelevant. Nonetheless, hypernyms scored relatively well in expert review and less so when matching on data. A logical explanation for this contrast is the wide range of possible correct results. Relevant hypernyms might be hierarchically a level higher and thus not match but still be considered relevant. Evaluated hypernyms might also differ slightly from one another but still be relevant (e.g. business and industry). The majority of the found antonyms is classified as relevant by human experts, however antonyms can be multi-interpretable and are not always perceived by the human expert as the system intented. An example of this is the antonym: reflective - nonreflective. Photos often contained reflective surfaces such as windows or solar panels, consequently the system quite literally found this antonym to be the most relevant. Human experts however may interpret this antonym differently: as an idea or consideration.

9.2 Limitations

First and foremost, the performance of the proposed system depends greatly on the quality of the photos it receives. As stated before, photos can be extremely abstract and multi-interpretable in a photovoice study: e.g. they almost never represent just an iconic relation. As the system is only evaluated on one dataset, performance can differ tremendously depending on the quality and nature of a photo collection.

Furthermore, to represent the semantics of a word, the system relies on the ability of the Word2Vec dataset to contain a diverse vocabulary. Due to a lack of computational resources, only half of this dataset (1.5 million words and phrases) has been used. As a result, several keywords have been disregarded which could potentionally have been significant to the data.

Finally, evaluation has been carried out by a relatively small group of human experts whom did not participate in the original photovoice study. To truly evaluate performance, the system would have to be evaluated by the students of the original photovoice study for they also understand the underlying meaning of their photos and are more familiar with the total set of photos.

9.3 Future work

Future research can focus on improving a series of steps in the proposed method. Nearly every step in this process can most definitely be improved.

As a start, vector representations of photos can be improved by removing outliers from the set of generated concepts of a photo. Photos almost never need 20 concepts to accurately describe them, on the contrary additional concepts are often not relevant to the photo and could offset an accurate vector representation. A possible method to remove outliers from a photo's concepts, is to measure the average cosine similarity score of a vector to all other vectors in a set. With these scores, a normal distribution can be created and within this distribution, data that lies more than approximately three standard deviations from the mean should be removed.

Performance can also be enhanced by using more capable computational resources, for this could double the vocabulary size of the Word2Vec model from 1.5 million to 3 million.

Furthermore can the hypernym process be improved by incorporating a level of abstractness property on a concept. Although this most likely did not work in this project due to a lack of similarity between keywords as well as a high level of abstracness of keywords, other lexical databases than WordNet containing ontological relations might perform better. For example: DBpedia (*Home - DBpedia Assocation* 2021) or Wikidata (*Wikidata* 2021).

Additionally, if the hypernym process is optimalized and thus accurately describes various clusters within a set of photos, the antonym process can make use of these clusters. Unlike the current method, clusters could represent an antonym on both sides of an axis which can potentionally improve the relevance of suggested antonyms.

Finally, after perfecting the proposed method in this paper, the proposed system could be extended with a machine learning algorithm. By collecting feedback from human experts on given suggestions, the system can potentionally improve its' performance. Alternatively, it could also make use of a large labeled training dataset, possibly various photovoice studies or similar studies containing categorization and relational photo exercises.

References

[AD20] Elena Andina-Díaz. “Using Photovoice to stimulate critical thinking: An exploratory study with Nursing students”. In: *Revista Latino-Americana de Enfermagem* 28 (2020).

[Aiz03] Akiko Aizawa. “An information-theoretic perspective of tf–idf measures”. In: *Information Processing & Management* 39.1 (2003), pp. 45–65.

[AK15] Mehdi Allahyari and Krys Kochut. “Automatic topic labeling using ontology-based topic models”. In: *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. IEEE. 2015, pp. 259–264.

[Ama] *Amazon Rekognition*. 2021. URL: `http://www.aws.amazon.com` (visited on 06/24/2021).

[BDF01] Kobus Barnard, Pinar Duygulu, and David Forsyth. “Clustering art”. In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*. Vol. 2. IEEE. 2001, pp. II–II.

[Ben+18] Benjamin Bengfort et al. *Yellowbrick*. Version 0.9.1. Nov. 14, 2018. DOI: `10.5281/zenodo.1206264`. URL: `http://www.scikit-yb.org/en/latest/`.

[BNS p] Marije T Brink, Frank Nack, and Ben Schouten. “Framing students’ reflective interactions based on photos”. In: (in press).

[Bui+13] Lars Buitinck et al. “API design for machine learning software: experiences from the scikit-learn project”. In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. 2013, pp. 108–122.

[Cla] *Clarifai*. 2021. URL: `http://www.clarifai.com` (visited on 06/24/2021).

[CSY17] Cheryl Cooper, William Sorensen, and Susan Yarbrough. “Visualising the health of communities: Using Photovoice as a pedagogical tool in the college classroom”. In: *Health Education Journal* 76.4 (2017), pp. 454–466.

[Dbp] *Home - DBpedia Assocation.* 2021. URL: `http://www.dbpedia.org` (visited on 06/24/2021).

[DH04] Chris Ding and Xiaofeng He. “Principal component analysis and effective k-means clustering”. In: *Proceedings of the 2004 SIAM International Conference on Data Mining.* SIAM. 2004, pp. 497–501.

[DS15] Joaquin Derrac and Steven Schockaert. “Inducing semantic relations from conceptual spaces: a data-driven approach to plausible reasoning”. In: *Artificial Intelligence* 228 (2015), pp. 66–94.

[Goo] *Google Cloud Vision.* 2021. URL: `http://www.cloud.google.com` (visited on 06/24/2021).

[Gro20] Maarten P Grootendorst. *c-TF-IDF.* `https://github.com/MaartenGr/cTFIDF`. 2020.

[HYH11] Cheng-Hui Huang, Jian Yin, and Fang Hou. “A text similarity measurement combining word semantic information with TF-IDF method”. In: *Jisuanji Xuebao(Chinese Journal of Computers)* 34.5 (2011), pp. 856–864.

[Ibm] *ibm Watson Visual Recognition.* 2021. URL: `www.ibm.com` (visited on 06/24/2021).

[Kor+20] Edward Korot et al. “Cross Platforms Comparison between Automated Machine Learning Models for Fundus Photos and OCT Classification”. In: *Investigative Ophthalmology & Visual Science* 61.7 (2020), pp. 2040–2040.

[Liu+18] Cai-zhi Liu et al. “Research of text classification based on improved TF-IDF algorithm”. In: *2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE).* IEEE. 2018, pp. 218–222.

[Mar14] John Markoff. “Researchers announce advance in image-recognition software”. In: *New York Times* 17 (2014).

[Mic] *Microsoft Azure Computer Vision.* 2021. URL: `http://www.azure.microsoft.com` (visited on 06/24/2021).

[Mik+13a] Tomas Mikolov et al. “Distributed representations of words and phrases and their compositionality”. In: *arXiv preprint arXiv:1310.4546* (2013).

[Mik+13b] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[MZ15] Long Ma and Yanqing Zhang. "Using Word2Vec to process big text data". In: *2015 IEEE International Conference on Big Data (Big Data)*. IEEE. 2015, pp. 2895–2897.

[Ram+09] Daniel Ramage et al. "Clustering the tagged web". In: *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. 2009, pp. 54–63.

[RN09] Tye Rattenbury and Mor Naaman. "Methods for extracting place semantics from Flickr tags". In: *ACM Transactions on the Web (TWEB)* 3.1 (2009), pp. 1–30.

[Sch+09] Kara Schell et al. "Photovoice as a Teaching Tool: Learning by Doing with Visual Methods." In: *International Journal of Teaching and Learning in Higher Education* 21.3 (2009), pp. 340–352.

[TY12] Kensuke Tanioka and Hiroshi Yadohisa. "Effect of data standardization on the result of k-means clustering". In: *Challenges at the Interface of Data Analysis, Computer Science, and Optimization*. Springer, 2012, pp. 59–67.

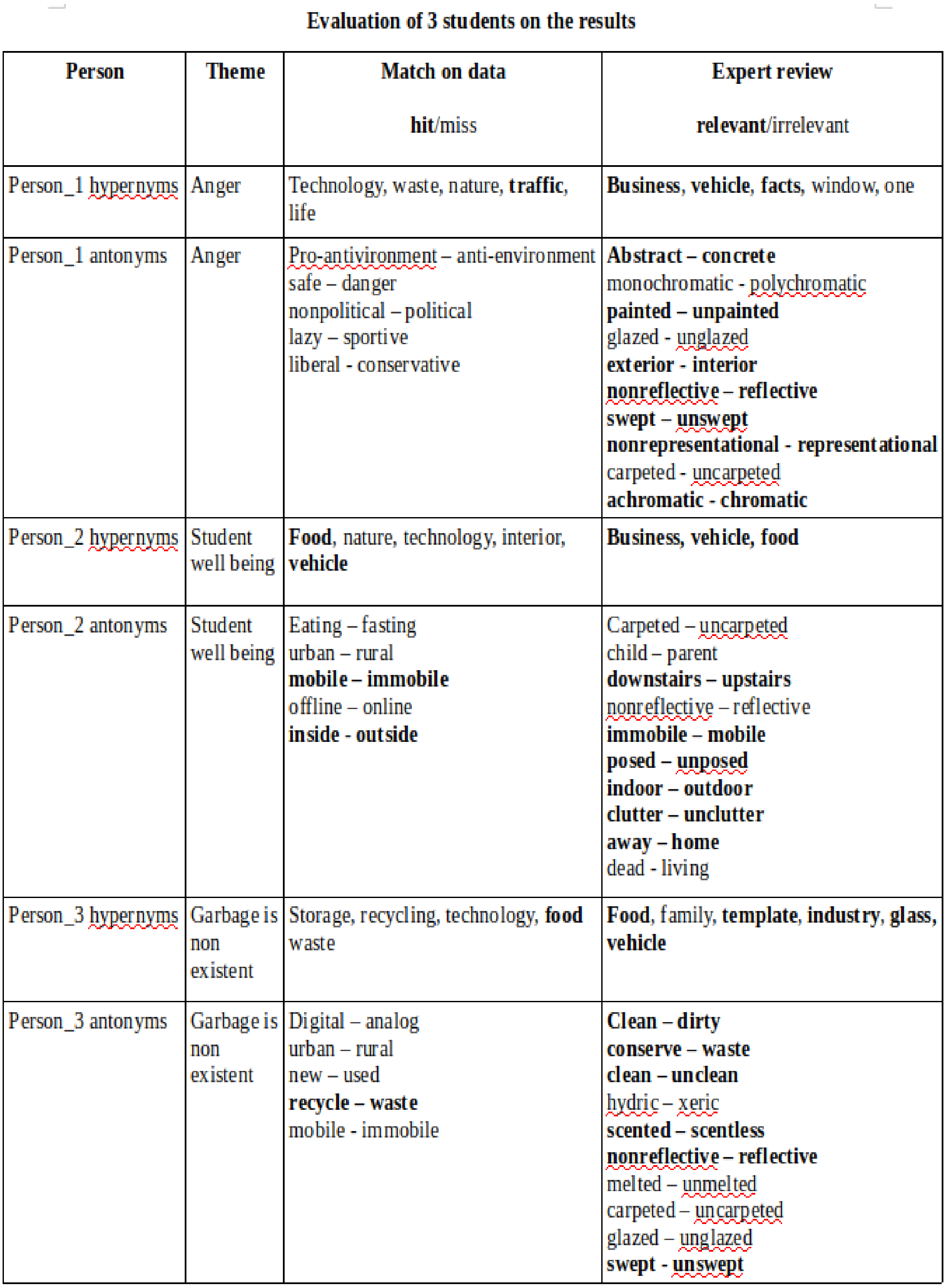
[WB94] Caroline Wang and Mary Ann Burris. "Empowerment through photo novella: Portraits of participation". In: *Health education quarterly* 21.2 (1994), pp. 171–186.

[Whi+15] Lyndon White et al. "How well sentence embeddings capture meaning". In: *Proceedings of the 20th Australasian document computing symposium*. 2015, pp. 1–8.

[Wik] *Wikidata*. 2021. URL: `http://www.wikidata.org` (visited on 06/24/2021).

[Zha+18] Yi Zhang et al. "Does deep learning help topic extraction? A kernel k-means clustering method with word embedding". In: *Journal of Informetrics* 12.4 (2018), pp. 1099–1117.

Appendix A



Evaluation of 3 students on the results

Person	**Theme**	**Match on data** **hit**/miss	**Expert review** **relevant**/irrelevant
Person_1 hypernyms	Anger	Technology, waste, nature, **traffic**, life	**Business**, **vehicle**, **facts**, window, one
Person_1 antonyms	Anger	Pro-antivironment – anti-environment safe – danger nonpolitical – political lazy – sportive liberal - conservative	**Abstract – concrete** monochromatic - polychromatic **painted – unpainted** glazed - unglazed **exterior - interior** **nonreflective – reflective** **swept – unswept** **nonrepresentational - representational** carpeted - uncarpeted **achromatic - chromatic**
Person_2 hypernyms	Student well being	**Food**, nature, technology, interior, **vehicle**	**Business, vehicle, food**
Person_2 antonyms	Student well being	Eating – fasting urban – rural **mobile – immobile** offline – online **inside - outside**	Carpeted – uncarpeted child – parent **downstairs – upstairs** nonreflective – reflective **immobile – mobile** **posed – unposed** **indoor – outdoor** **clutter – unclutter** **away – home** dead - living
Person_3 hypernyms	Garbage is non existent	Storage, recycling, technology, **food** waste	**Food**, family, **template**, **industry**, **glass**, **vehicle**
Person_3 antonyms	Garbage is non existent	Digital – analog urban – rural new – used **recycle – waste** mobile - immobile	**Clean – dirty** **conserve – waste** **clean – unclean** hydric – xeric **scented – scentless** **nonreflective – reflective** melted – unmelted carpeted – uncarpeted glazed – unglazed **swept - unswept**

Figure A.1: *Detailed evaluation results on hypernyms and antonyms*