Comparison of Unicef's Annual Reports by Topic Extraction

Emily Bünker
emily.buenker@hhu.de

Abstract

In the age of Big Data it is very time consuming to be informed about important subjects. Extracting the main topics of a text can help to get a brief understanding what the text is about. Further, it is possible to generate a summary out of these topics in order to get a deeper but still brief understanding of the underlying document. This can be done by extracting important sentences from the original document and concatenate them into a summary. This is called ”extractive summarization”. Another option is the ”abstractive summarization” which uses Deep Learning techniques in order to generate a summary with its own unique sentences. (Raja et al., 2019)

1 Introduction

The project described in this article aims to identify the most important topics of Unicef's annual reports from 1982 to 2021 respectively as well as to determine overall similarities and differences between the years. In order not to get some topics out of context and generally to get a better understanding of the work, summaries are created for every report. This is done by using the extractive summarization method.

Ideally, the topics extracted in this project match with the topics Unicef officially deals with. These are according to their website (Unicef, 2022):

1. Child protection and inclusion
2. Child survival
3. Education
4. Social policy
5. UNICEF in emergencies
6. Gender
7. Innovation for children
8. Supply and logistics
9. Research, evidence and analysis

2 Data and Resources

As described in the previous section the annual reports from 1982 to 2021 of Unicef are used. This selection was made because earlier reports, i.e. before 1982 , are published as scans of printed reports. Therefore, it was not possible to load those reports as text. The report of 2021 is the last one published. They can be downloaded on the website. (Unicef, 2022)

For the preprocessing and analysis steps, the following libraries were used: pandas, numpy, nltk, fitz, os and pathlib.

The pandas DataFrame was used as the data structure in the project.

The python version 3.9 was used.

3 Method

Firstly, the pdf files needed to be loaded into a pandas dataframe. The fitz library converted the pdf text to strings. Than, the dataframe was created with two columns: The year of the annual report and the content text of the report.

Secondly, the text needed to be preprocessed. Using nltk the text was tokenized into words. Also, POS tagging is applied to every word. That was necessary to extract all the nouns which are needed to extract the topic. All part of speech tags for nouns, meaning NN (singular noun), NNS (plural noun), NNP (proper noun, singular) and NNPS (proper noun, plural) were extracted. After that, the list of words is lemmatized using the nltk lemmatizer. In order to get a “clean” string, stopwords were deleted. The stopwords used, are part of the nltk corpus. Also, some additional words were appended to the stopword list. These are words that aren't purposeful (i.e., Unicef or report). Which additional words to delete was determined later on in the process by looking at the most frequent words. The dataframe now consists of 5 columns:

The year of the annual report, the whole text of the pdf file, only the nouns of the text as type list, the lemmatized list of nouns and the nouns without the stopwords as type list.

Having done the preprocessing, the main feature can be implemented, which is, getting the most important topics by looking at the most frequent nouns. Pandas function ”value_counts” is used to get these words for every annual report. It counts how often a word appears in the list of words and sorts these words in descending order. Additionally, the most frequent nouns overall (i.e., of all reports combined) are determined using the same function. As a preliminary step the list-of-nouns-column had to be combined to one list. After these two steps 60 columns were added to the original dataframe: 30 columns with the most frequent words and 30 columns with the quantities of these words. Also, an additional dataframe was created for the most frequent words over all reports. This dataframe has 60 columns matching the columns previously introduced. The number of the most frequent words (30 words) is arbitrarily chosen.

As an additional feature, the sentences in which these most frequent words per report occur are filtered out of the “text” column and are added to built a summary - resulting in 40 summaries respective to the 40 reports. Because it was still a relatively big summary a briefer summary was created (see the Results and Discussion section). This time, only the first 10 words were used minus the word ”child” because it is obvious that Unicef deals with children since it is a children’s fund. To do so, the text of each annual report is tokenized into sentences using the nltk tokenizer. After that, the sentences in which the words occured were appended to a list. To keep the summary even shorter, only the sentences that contain less than 50 words were considered.

4 Results and Discussion

Annotation: The entire results for the extracted topics and the summary can be found in the repository. This paper only shows some examples for reasons of clarity and space.

4.1 Extracted Topics

4.1.1 Comparison of the Reports

Looking at the dataframe with the most frequent words overall, it can be investigated what the main topics for Unicef are. The word “Child” occurs about 12925 times in the reports. Followed by the word ”Country”. Table 1 shows the 10 words with the highest quantity as well as the number of occurences, i.e. in how many reports the specific word is part of the 30-most-frequent-words.

Table 1: Most frequent words over all annual reports, their quantity and the number of occurences

Word	Quantity	Occurences
Child	12925	40
Country	5823	40
Programme	4097	40
Health	3745	40
Education	2744	40
Development	2673	39
World	2626	39
Woman	2502	32
Resource	2222	37
Committee	2220	32

It can be seen that the occurence column is not ordered perfectly in descending order. But, the 10 words with the most occurences are nearly all listed in that column. Only ”Government” and ”Service” are not listed and instead ”Women” and ”Committee” are in that column. That shows that these reports are important in nearly all reports.

Table 2 shows a selection of words that occur in only one or two reports in order to determine topics that depend on the events of that year. Overall there are 50 words to which that applies (in total there are 120 unique words over all 40 reports).

Table 2: Year-specific words with the respective year

Word	Year
Rwanda	1995
Climate	2021
Japan	2002
Refugee	2016
Drug	1989
Polio	2001, 2004
China	2008, 2011
Change	2019, 2021
Covid-19	2020, 2021
Conflict	2000, 2021

It can be seen that some year-specific topics appear in the reports. For example ”Rwanda” in the year 1995 which refers to the Rwandan genocide in 1994. Or Covid-19 in the years of 2020 and

2021 for obvious reasons. Also, ”Climate” in 2021 which refers to the climate change. But it must also to be said that there are topics of different years that don’t occur in the words list but are a relevant topic in the report. These are for example “Ebola” in the years 2014 and 2015 or Tsunami in the year 2005, which refers to the Indian Ocean earthquake and tsunami of 2004.(Unicef, 2005)

4.1.2 Comparison to Unicefs Goals

The topics extracted from the reports match Unicef’s goals listed in the introduction.

1, 2 Child protection and inclusion, Child survival: The following words which are part of the 30-most-frequent-words of several reports are related to these goals: ”child”, ”health”, ”support”, ”protection”, ”nutrition”, ”water”, ”access”, ”care” and ”partnership”

3 Education: Words like ”education”, ”school”, ”development” and ”training” can be assigned to this.

4 Social policy: This means that Unicef wants to achieve equal chances for every child by reducing child poverty and therefore its consequences (Unicef, 2022). The words ”rights” and ”development” stress this goal.

5 Unicef in emergencies: This can be found in many reports. For example the topics ”Rwanda”, ”covid-19” and ”climate” as described in the previous section.

6 Gender: Unicef aims to reach that every girl fulfills their potential (Unicef, 2022). The topics ”woman” and ”girl” in the list of most frequent words show this.

7 Innovation for children: This is described as follows on the website (Unicef, 2022): ”UNICEF works with partners in every sector to co-create innovative solutions that accelerate progress for children and young people.” Therefore, words such as ”partner”, ”partnership”, ”resources”, ”project”, ”programme”, ”foundation”, ”sector” and ”funding” reflect this goal.

8 Supply and logistics: This is represented by: ”sector”, ”resources”, ”access”, ”area” in the extracted topic list.

9 Research, evidence and analysis: Unicef wants to use data and evidence to drive their results (Unicef, 2022). This goal is not represented in the list. Possibly, because this is the way they want to work and not what they thrive to achieve.

4.2 Generated Summary

As a quantifiable result for the summary, the proportions each summary has in respect to the original text are taken into account. Without any cuttings, the mean proportion and the standard deviation are 61.002 % ± 9.860 %. Having done the preliminary steps to shorten the summary as described in the method section, these values are 36.152 % ± 5.313 %. Regarding that the reports have an average length of ca. 58 pages there are still about 20 pages per report left to read.

As a closer look to a summary the one of 2021 is used as an example, since this is the shortest one. Reading this summary where ”change” is one of the extracted nouns it becomes clear that this word is mostly used in the context of climate change (7 out of 11 times). By just looking at the word itself in the list of extracted topics, it could also be interpreted more positively. In 4 out of 11 sentences this is the case. This is one of these sentences: ’Through these systemic changes transformative progress can be achieved on vaccine equity education mental health and addressing the climate crisis reaching those who have been left behind.’ This shows that an extracted topic list alone can lead to misinterpretations due to a lack of context.

Moreover, the summary can also give context for the year-specific topics. For example the word ”China” itself is not very meaningful but by looking in the summary of that year (2008) it becomes clear, that the earthquake in China which killed thousands of children is meant.

Furthermore, the main topics Unicef dealt with in a specific year can be identified and the reading flow is mostly good. It is interrupted a few times by what has been headlines, headers or footers. These pop up a few times at any point in the sentences. Also, the results cannot be compared to a human-written summary or maybe even an automatic summary which is done using the abstractive summarization method because there is no structure, i.e., introduction and cocnlusion are missing. Also, one of the most important things for writing a summary is to use your own words and not copy sentences - since copying sentences is the main idea of the extractive summarization method, it is obvious that this could not be achieved.

In terms of improving the understanding of a text, the summarization method as used in this project succeeded, but there are still things to work on - especially regarding the length of the summary.

5 Challenges and Open Issues

5.1 Extracted Topics

The topic extraction worked well but there were some issues. For example there were still some verbs and punctuations left after extracting the nouns. ”Reimagine”, ”Responding”, ”-” and ”*” had to be added to the stopword list in order to get a list that only consists of nouns.

The topic extraction done in this project is focused on finding the most frequent words. There is also the possibility to exctract words from the title, bold written words or italic written words (see Edmundson, 1969 or Raja et al., 2019). When only looking at one text and not at multiple ones as done in this project, a subdivision of the various paragraphs of this text is also possible. This could be done to determine and compare the main ideas of these paragraphs. Since this project aimed to identify similarities and differences between the reports and not within one single report, this did not take place.

As described above some year-specific topics do not occur in the 30 most-frequent-word-list. To avoid that, the number of extracted words could be increased. As another way to determine the year-specific topics, it is possible to calculate the tf-idf value of the words, which means ”Term Frequency–Inverse Document Frequency”. The importance of a word in a collection of documents is evaluated. To use this, it is important to know the number of documents, which is the case in this project. (Erra et al., 2014)

5.2 Generated Summary

Regarding the automated summarization, there are a few more things to improve. Allthough the automated summary is shortend by some preliminary steps, it is still pretty long as described in the results and discussion section. A possibility to reduce the summary without missing out on too much information could be to strike out similar sentences. Farouk, 2019 presented 3 different methods to determine sentence similarity: Word based, structure based and vector based. Word based means that the sentence similarity is measured by word-to-word similarity. Whereas structure based similarities are defined by syntactic and semantic information like word order or Part-of-Speech. The vector based approach converts the sentences into vectors where the columns are the features like dependency pairs extracted through parsing the sentence.

Another solution could be to not just consider the nouns but give every word a weight based on their tf-idf value. This weight can be adjusted if its written in bold or is part of title as explained at the beginning of this section. Raja et al. (2019) did so by summing up these word-specific weights per sentence. The sentences with the highest sums are extracted and included in the summary. Furthermore, it is possible to adjust the overall weight of a sentence, if cue words like ”significant,” ”impossible” or ”hardly” are included as Edmundson (1969) did. For this approach it is much more important to delete the stopwords, because most of them aren’t nouns (i.e., by filtering out the nouns as done in this project most of the stopwords are already eliminated). Using this way, it is also possible to better control the size of the summary.

Nevertheless, it must be said that an extracted summary cannot be as good as an abstractive summary in points of reading flow and structure.

6 Summary and Conclusion

All in all, the topic extraction was quite successful. It was possible to determine topics that are relevant in every year and represent general goals for Unicef as well as to determine differences between the years. These differences were impacted by the events that happened in that year.

Regarding the automated summary, several improvements have to be made. Most importantly the summary needs to be shorter.

References

[1] Edmundson, H.P., (1969). New Methods in Automatic Extracting *Journal of the .Association for Computing Machinery, Vol. 16, No.2*

[2] Erra, U., Senatore, S., Minnealla, F., Caggianese, G., (2014) Approximate TF–IDF based on topic extraction from massive message stream using the GPU. *Information Sciences*

[3] Farouk, M., (2019). Measuring Sentences Similarity: A Survey. *Indian Journal of Science and Technology*

[4] Raja, N.K., Bakala, N., Suresh, S., (2019). NLP: Text Summarization By Frequency And Sentence Position Methods. *International Journal of Recent Technology and Engineering (IJRTE)*

[5] Unicef. (2022, August 29) *What we do*. https://www.unicef.org/what-we-do

[6] Unicef. (2005). *ANNUAL REPORT 2004*. New York: Unicef.

[7] Unicef. (2022, August 10) *UNICEF Annual Report*. https://www.unicef.org/unicef-annual-report