Natural Language Processing Project Phase 1 Report

Diyar Hamedi
Department of Computer Engineering
Iran University of Science and Technology
diyar_hamedi@comp.iust.ac.ir

Spring 2023

Abstract

This report presents the methodology employed to gather the dataset for the text topic classification task, including data source details and the data collection process. Additionally, we provide insights into the statistics of the collected dataset, showcasing the label distribution, the unit of data segmentation, and the effects of data cleaning. The statistical analysis serves as a foundation for the subsequent development of a text topic classification model using machine learning techniques. The findings and insights gained from this dataset, along with the associated challenges and limitations, contribute to the broader field of natural language processing (NLP) and text classification research. The script files for collection, cleanup and analysis of data can be accessed in the following repository: https://github.com/DiyarH/nlp1402project

1 Introduction

In the era of abundant textual data, efficient organization and extraction of valuable information from vast document collections pose significant challenges. Topic classification, a fundamental task in natural language processing (NLP), addresses these challenges by categorizing texts into specific topics or domains. This project report focuses on the initial phase of developing a video subtitle topic classification model. Our primary objective is to gather and clean up data, extract useful metrics and statistics, and lay the groundwork for subsequent model development.

2 Motivation

The motivation behind developing a video subtitle topic classification model lies in the wealth of textual information embedded within video content. As videos continue to dominate various platforms, such as YouTube, educational websites, and online tutorials, the accompanying subtitles serve as a valuable resource for understanding and accessing the content within these videos.

In this phase of the project, we embark on gathering and cleaning up the video subtitle data while extracting useful metrics and statistics. The subsequent phase will focus on model development and training using the preprocessed data.

The benefits of video subtitle topic classification are as follows:

- Information Retrieval and Search: Categorizing video subtitles into specific topics or domains facilitates efficient organization and retrieval of textual data. This categorization provides a structured framework for indexing and searching relevant information within large video collections, enabling users to find specific content quickly.
- Content Recommendation and Personalization: Topic classification allows for personalized content recommendations. By understanding the topics of interest for individual users, recommendation systems can deliver tailored content aligned with their preferences. This personalization enhances user experiences, improves satisfaction, and increases the relevance of suggested content.
- Social Media Monitoring and Sentiment Analysis: Topic classification is valuable for monitoring social media discussions and analyzing sentiment trends. Categorizing social media posts or comments into different topics enables businesses and organizations to gain insights into public opinions, customer feedback, and emerging trends related to specific domains or products. This analysis helps in understanding customer sentiment, identifying key concerns, and adapting marketing strategies accordingly.
- Automated Content Tagging and Organization: Topic classification enables the automatic tagging and organization of textual content, benefiting content management systems, news portals, or document repositories. By categorizing videos' subtitles based on topics, the system can streamline content organization, retrieval, and navigation. This automation saves time and effort in manually tagging and structuring the content, enhancing overall efficiency and user experience.

By gathering and cleaning up video subtitle data and extracting useful metrics and statistics, we aim to establish a solid foundation for the subsequent phase of model development. The following sections of this report will outline the data gathering and cleaning process, metrics and statistics analysis, and pave the way for the forthcoming model development phase.

3 Data Description

The dataset used for this project is the HowTo100M dataset[1], which is a large-scale collection of narrated videos with a specific focus on instructional content. The dataset contains a vast amount of video clips accompanied by captions, providing valuable textual information that can be used for various NLP tasks, including topic classification.

The key features of the HowTo100M dataset are as follows:

- Video Clips and Captions: The dataset comprises a total of 136 million video clips sourced from 1.2 million YouTube videos. These videos span a timeline of 15 years, offering a diverse and extensive collection of instructional content. The captions associated with each video are automatically downloaded from YouTube and provide textual descriptions of the visual content on the screen.
- Activities and Domains: HowTo100M dataset covers a wide range of activities from various domains. These activities include cooking, hand crafting, personal care, gardening, fitness, and more. The dataset's focus on instructional videos ensures that the content creators explicitly intend to explain the visual elements and provide detailed narration to aid understanding.

For this project, only the subtitles extracted from the videos are utilized, as the primary objective is to perform topic classification on textual data rather than working directly with the videos themselves. By using the subtitles, we can leverage the rich instructional content present in the HowTo100M dataset while keeping the project within the scope of the course's requirements.

The data is available in two files:

- Captions File[2]: The JSON file containing (video id, caption) pairs can be accessed at here. This file provides a structured representation of the video captions, allowing easy access to the textual data.

[1] `https://www.di.ens.fr/willow/research/howto100m/`
[2] `https://www.rocq.inria.fr/cluster-willow/amiech/howto100m/asr.json`

- Labels File[3]: The labels for each video id (such as main category, subcategory and task id) are available within a zip file. These labels can be used to annotate and categorize the videos based on their topics or domains.

By utilizing the subtitles from the HowTo100M dataset, we can leverage a vast collection of instructional content for our topic classification model. The subsequent sections of this report will detail the data gathering process, including accessing the captions and labels files, preprocessing the subtitle data, and performing statistical analysis to gain insights into the dataset's characteristics.

4 Data Gathering and Cleanup

To gather the data for the project, the HowTo100M dataset serves as the primary source. The dataset already contains a comprehensive collection of narrated videos with associated captions. Therefore, the data gathering process involves downloading the required files from the HowTo100M dataset.

For the cleanup phase, a subset of the data is selected to ensure a manageable size for the initial training. In this phase, a maximum of 1000 samples per category are chosen from the available dataset. This selection is made to strike a balance between having sufficient data for training the model and avoiding potential performance issues due to an excessively large dataset. However, in the subsequent phases, additional samples may be considered if the initial model training does not yield satisfactory results.

During the cleanup process, the selected samples are subjected to several cleaning operations. This includes splitting each sample into sentences and words using the nltk Python package. The sentence splitting step helps in organizing the text into coherent units, allowing the model to capture the finer nuances of the topic. Furthermore, during the word-splitting process, punctuation marks are removed. This decision is made considering the nature of the text topic classification task, where the focus is primarily on the content and context conveyed by the words themselves.

Additionally, special characters such as parentheses, non-standard characters like emojis, and any other irrelevant textual elements are eliminated. These steps aim to ensure that the data remains focused on the essential content while removing potential noise or distractions that may hinder the model's learning process.

Currently, each sample consists of the entire video subtitle for simplicity. However, in the second phase of the project, an alternative approach will be explored. The subtitles will be split into three distinct segments: the introduction, body, and outro. This division aims to analyze the impact of such segmentation on the model's performance and its ability to capture different aspects of the topic. By comparing the results obtained from the segmented subtitles with the initial approach, valuable insights can be gained into the effectiveness of different input representations for the topic classification task.

After the cleanup, segmentation, and word-splitting processes, the samples are organized into individual categories. Each category's samples are saved into separate CSV files, with the file names corresponding to their respective categories. This organization facilitates efficient data management and subsequent model training.

By following this data gathering and cleanup methodology, the project ensures a well-defined and manageable dataset for the initial phase. This approach allows for focused model training while retaining the flexibility to adjust the dataset size, composition, and input representation in later stages if necessary.

5 Data Statistics and Metrics

In this section, we present an overview of the dataset along with various metrics and statistics that have been extracted. The raw and clean dataset sizes are provided, followed by an analysis of different linguistic features and measurements.

5.1 Overview

The HowTo100M dataset used in this project comprises more than 1.2 million narrated videos with associated captions. In this phase, a maximum of 1000 sample from each category has been extracted. The raw and cleaned-up dataset sizes are as follows:

[3] `https://www.rocq.inria.fr/cluster-willow/amiech/howto100m/HowTo100M.zip`

category	raw dataset size	clean dataset size
Family Life	1035	1000
Youth	1100	1000
Finance and Business	779	779
Relationships	39	39
Holidays and Traditions	26976	1000
Health	14963	1000
Philosophy and Religion	487	487
Hobbies and Crafts	248653	1000
Personal Care and Style	15938	1000
Food and Entertaining	493203	1000
Arts and Entertainment	9680	1000
Work World	583	583
Education and Communications	15240	1000
Pets and Animals	30501	1000
Cars and Other Vehicles	67144	1000
Sports and Fitness	15844	1000
Travel	127	127
Computers and Electronics	5253	1000
Home and Garden	267776	1000
Total	1215321	16015

Table 1: Size of the raw and clean datasets per category.

These sizes provide a perspective on the scale of the dataset and its potential impact on the subsequent analysis.

5.2 Linguistic Features and Measures

In order to gain insights into the linguistic characteristics of the dataset, several metrics and statistics have been computed. The following metrics are presented:

5.2.1 Scale Analysis

1. Count of Sentences: The number of sentences in each category, as well as the average number of sentences per sample.
2. Count of Word: The number of words in each category, as well as the average number of words per sample.
3. Count of Unique Words: The number of unique words across all documents, as well as the number of categorywise common and unique words.

5.2.2 Frequency Analysis

1. Most Frequent Words: The top 10 most frequent words unique to each category.
2. RNF Metric Top Words: The top 10 common words with highest RNF value for each category.
3. TF-IDF Metric Top Words: The top 10 common words with highest TF-IDF value for each category.
4. Histogram of Unique Word Frequencies: A visualization representing the frequency distribution of top 100 most frequent unique words in descending order.

category	number of sentences	avg. number of sentences
Holidays and Traditions	57904	57.9
Home and Garden	49556	49.56
Computers and Electronics	46822	46.82
Youth	53442	53.44
Personal Care and Style	46373	46.37
Philosophy and Religion	21649	44.45
Hobbies and Crafts	57229	57.23
Relationships	1941	49.77
Sports and Fitness	57118	57.12
Work World	34239	58.73
Family Life	47836	47.84
Health	49702	49.7
Travel	5445	42.87
Finance and Business	45760	58.74
Education and Communications	50978	50.98
Arts and Entertainment	59291	59.29
Cars and Other Vehicles	53486	53.49
Pets and Animals	55260	55.26
Food and Entertaining	57011	57.01
Total	851042	53.14

Table 2: Number of sentences per category.

category	number of words	avg. number of words
Holidays and Traditions	913128	913.13
Home and Garden	838231	838.23
Computers and Electronics	829238	829.24
Youth	882685	882.68
Personal Care and Style	809060	809.06
Philosophy and Religion	377839	775.85
Hobbies and Crafts	942512	942.51
Relationships	31772	814.67
Sports and Fitness	924885	924.88
Work World	549581	942.68
Family Life	788726	788.73
Health	888328	888.33
Travel	82183	647.11
Finance and Business	809054	1038.58
Education and Communications	846122	846.12
Arts and Entertainment	1040803	1040.8
Cars and Other Vehicles	936685	936.68
Pets and Animals	888658	888.66
Food and Entertaining	903375	903.38
Total	14282865	891.84

Table 3: Number or words per category.

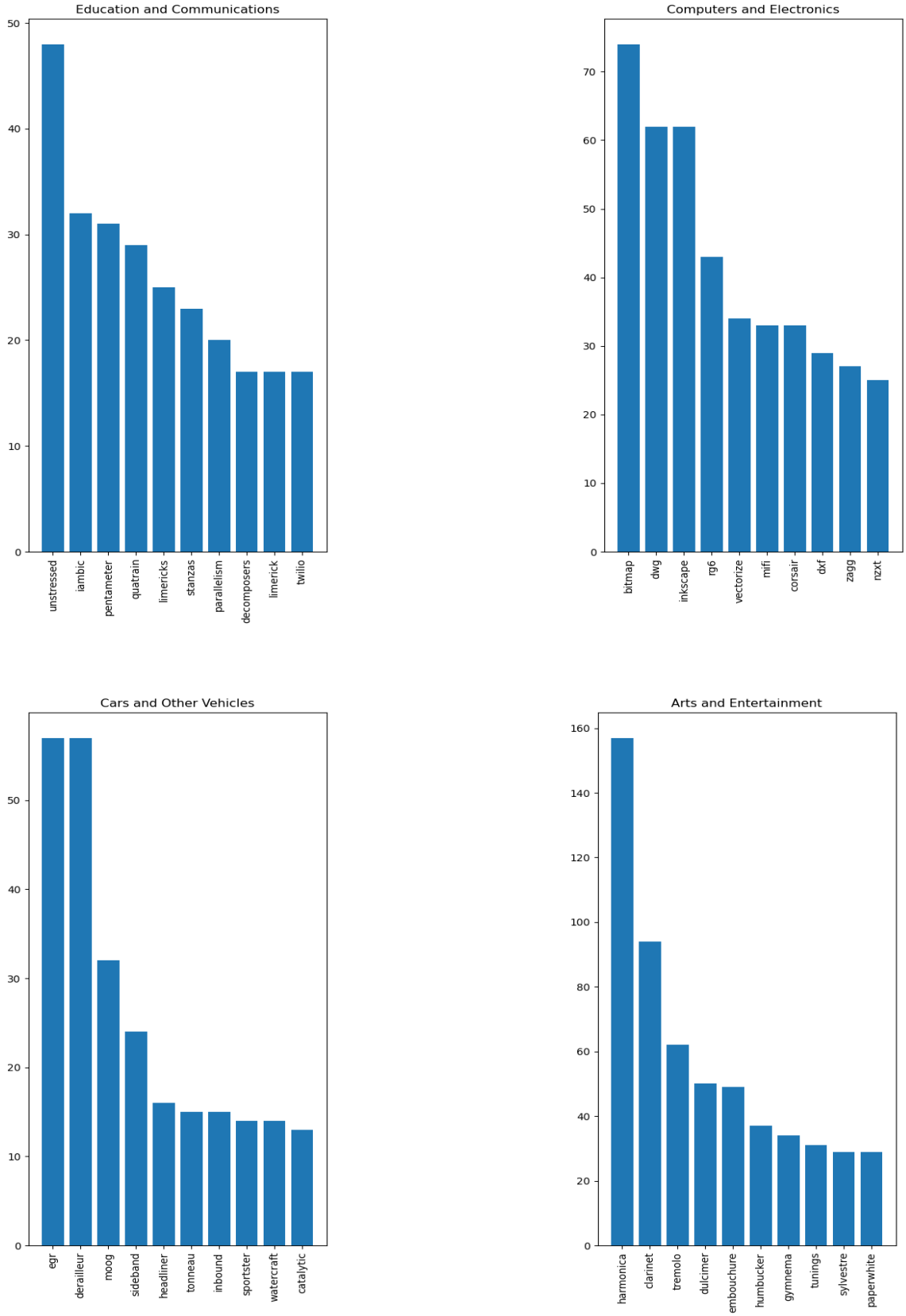

Figure 1: Most frequent words per category

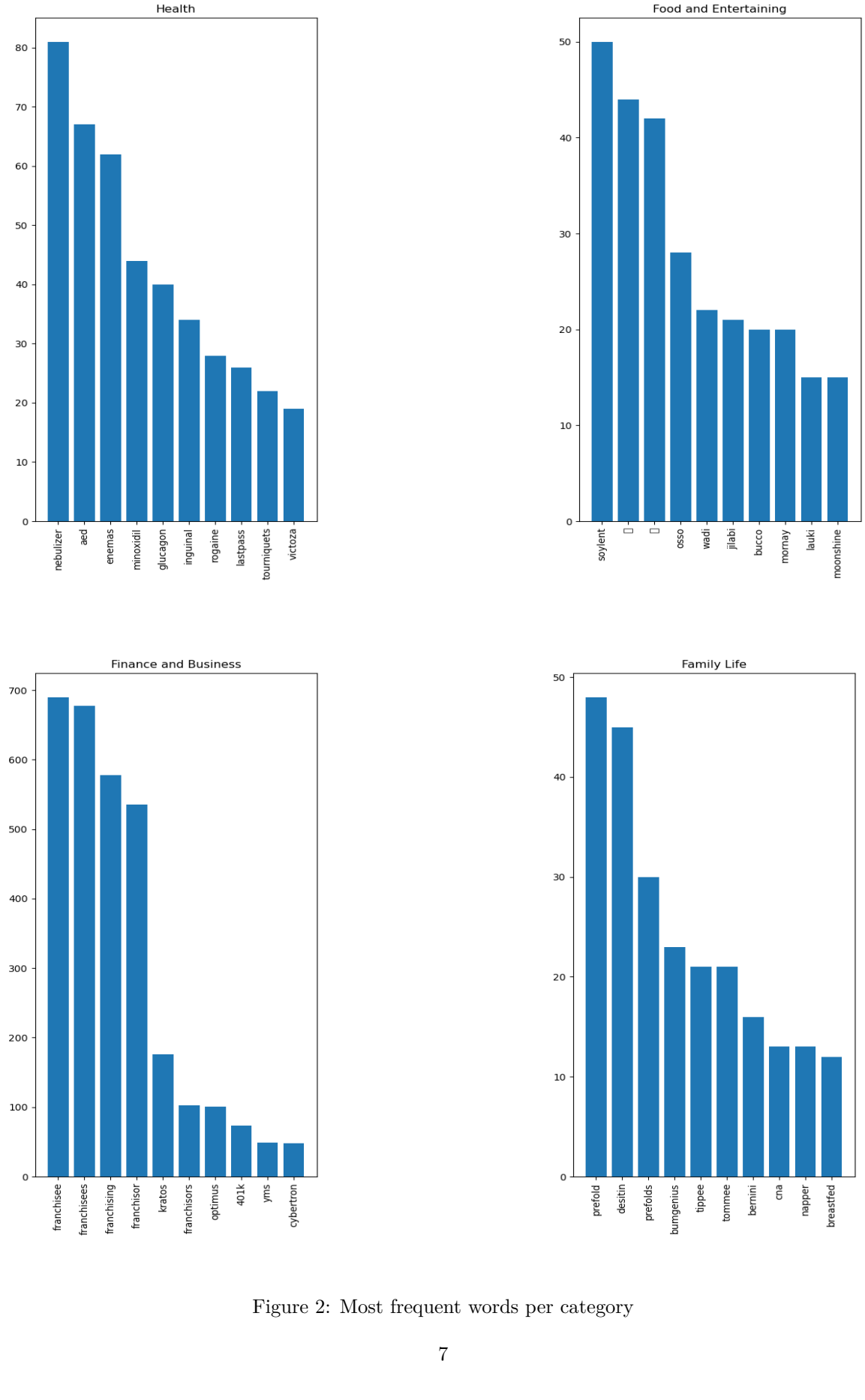

Figure 2: Most frequent words per category

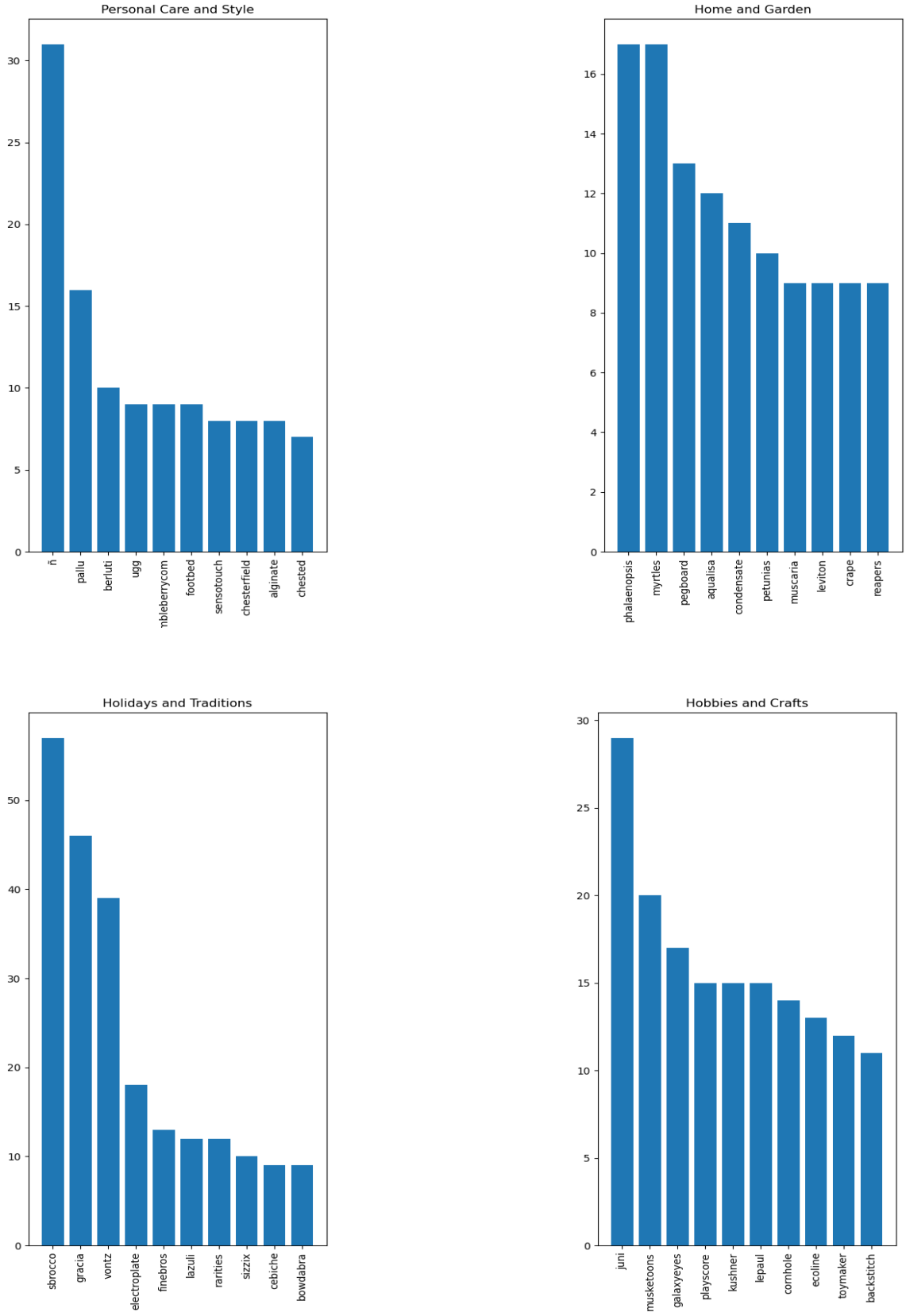

Figure 3: Most frequent words per category

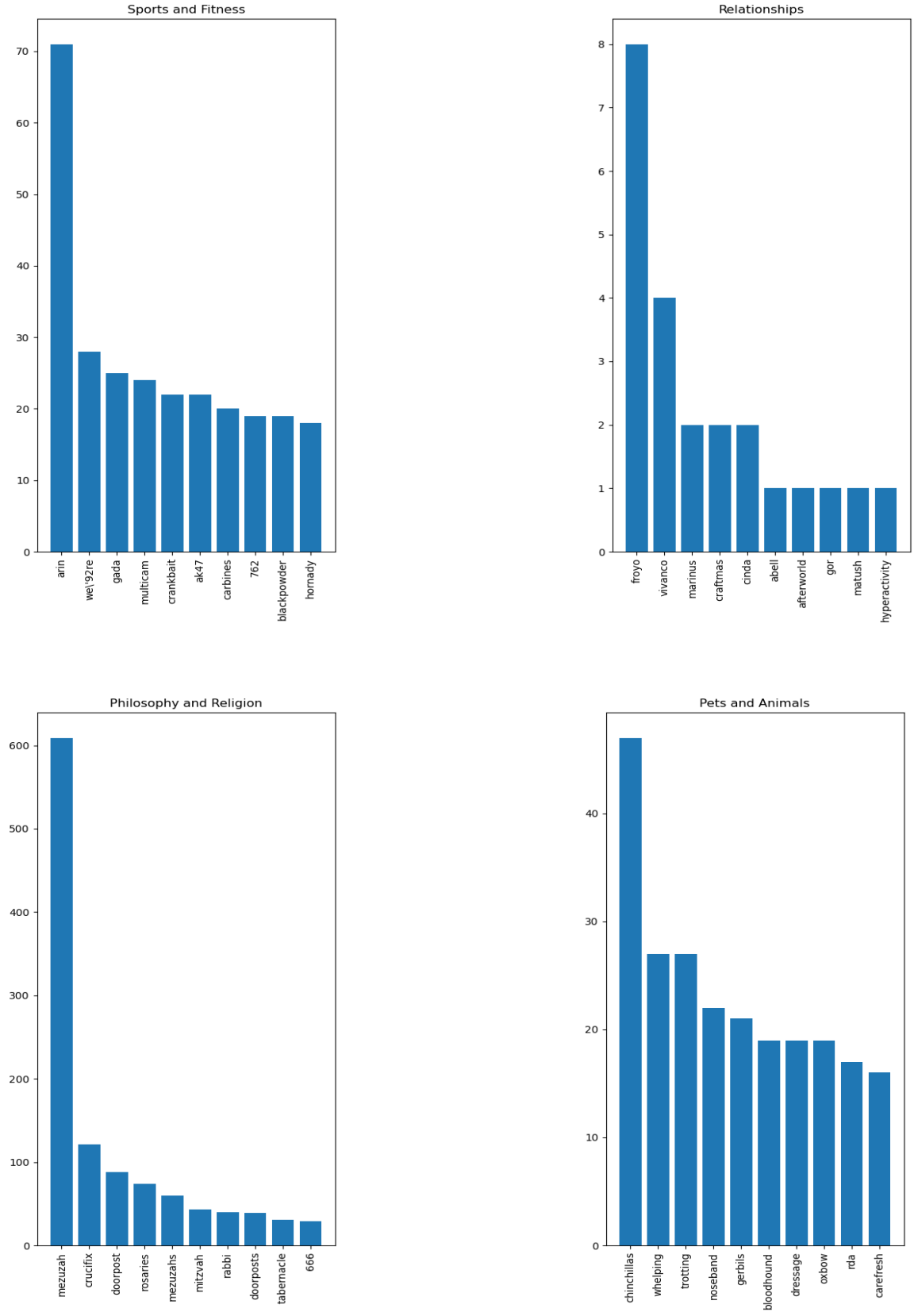

Figure 4: Most frequent words per category

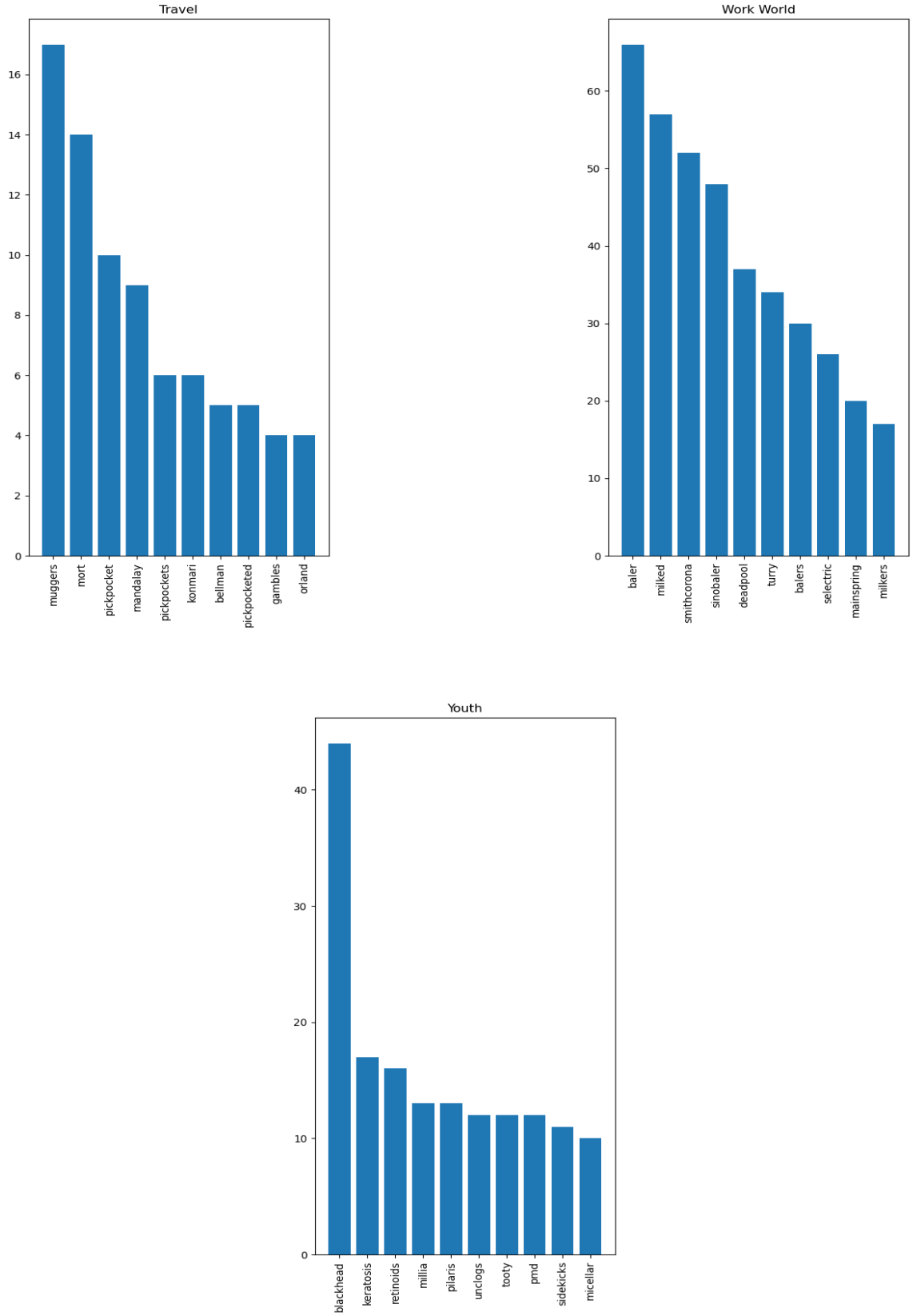

Figure 5: Most frequent words per category

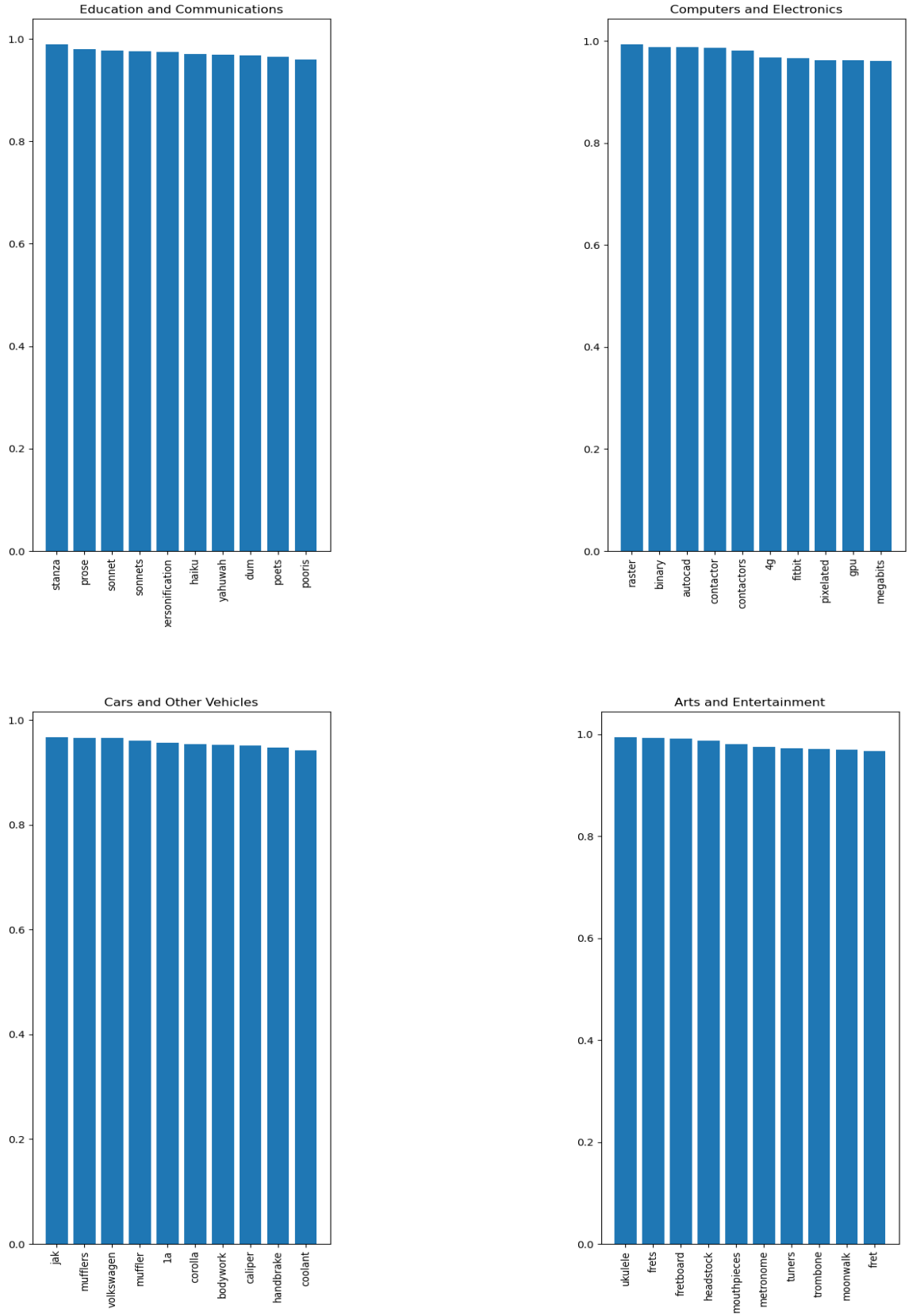

Figure 6: RNF metric words per category

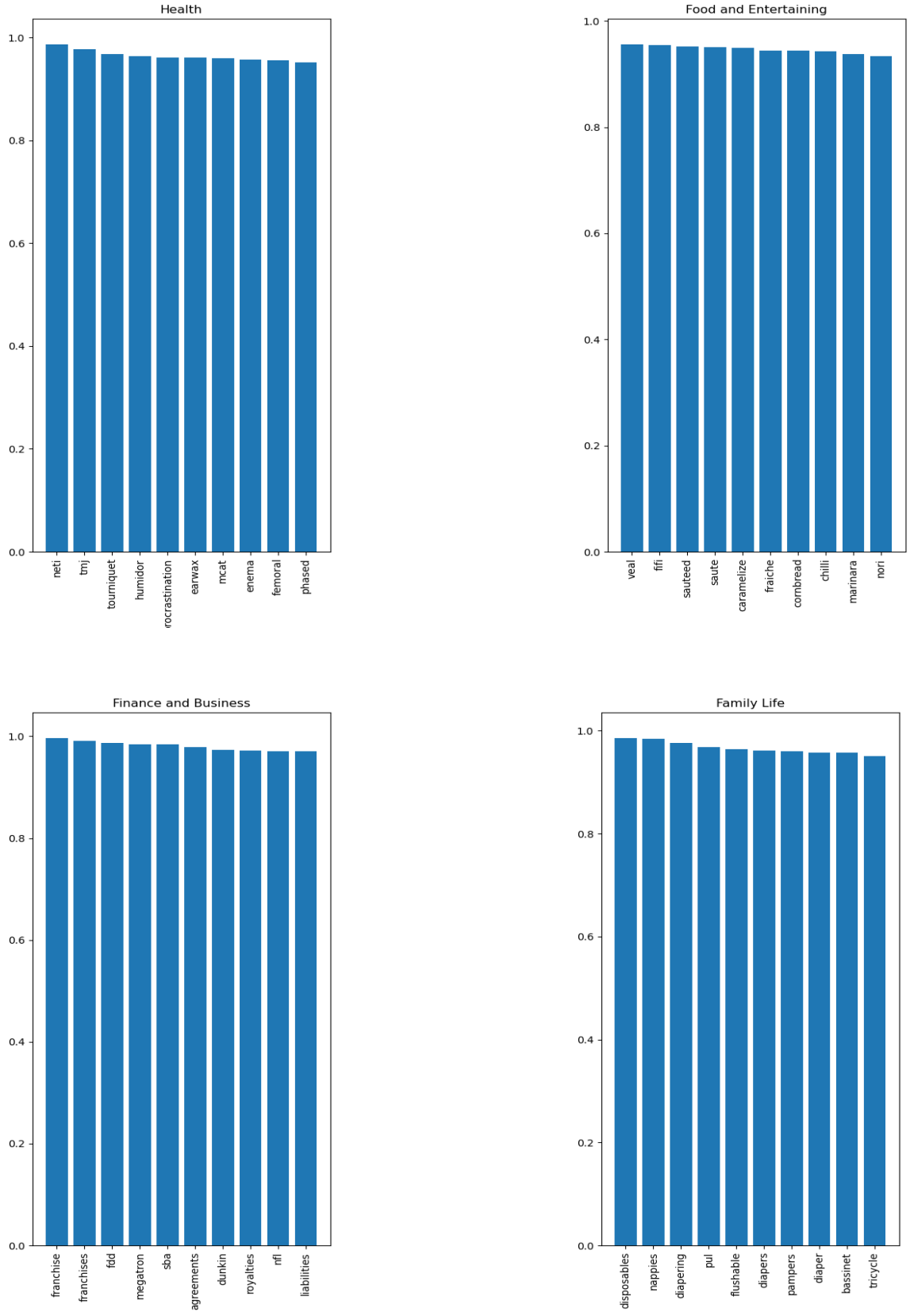

Figure 7: RNF metric words per category

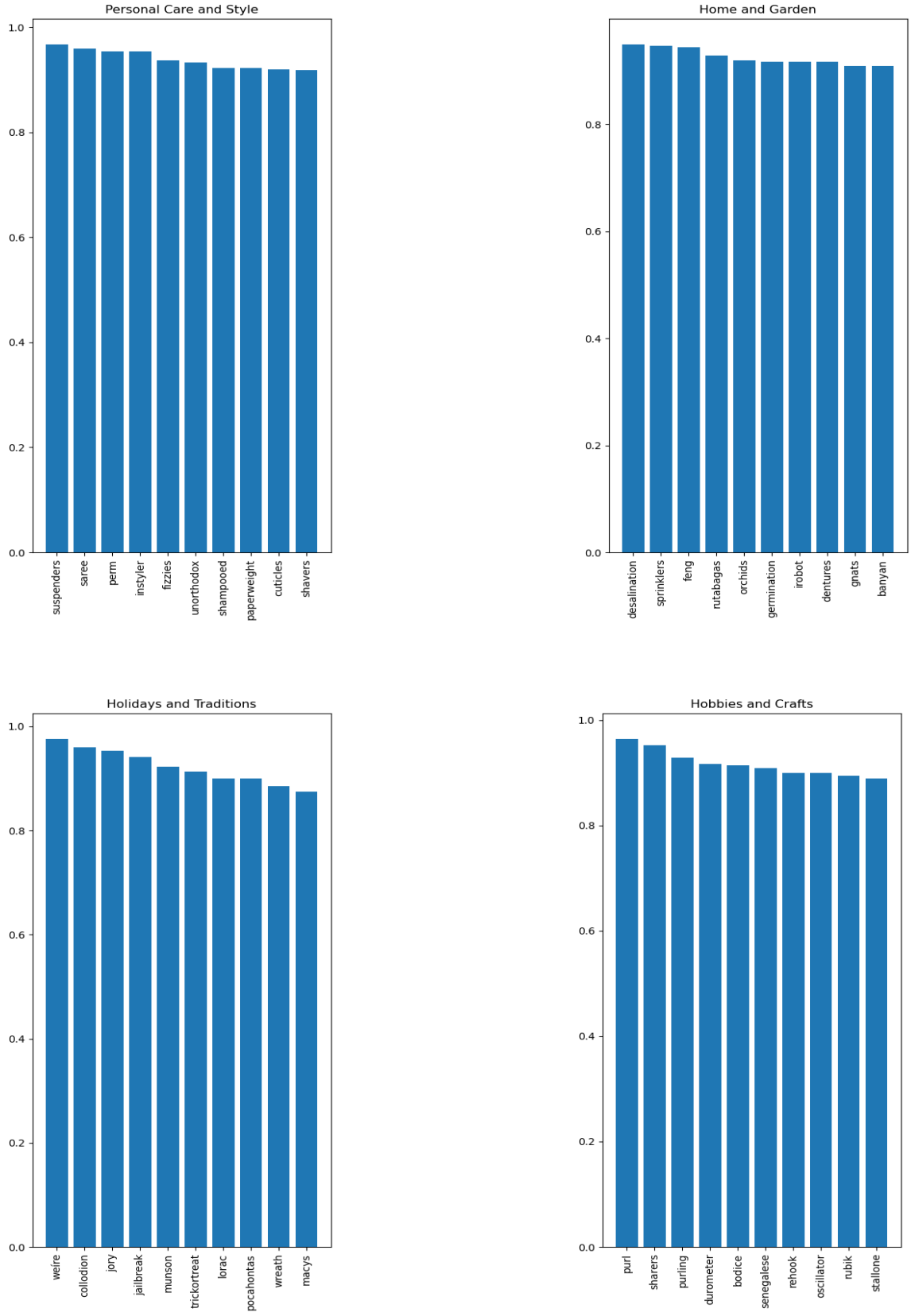

Figure 8: RNF metric words per category

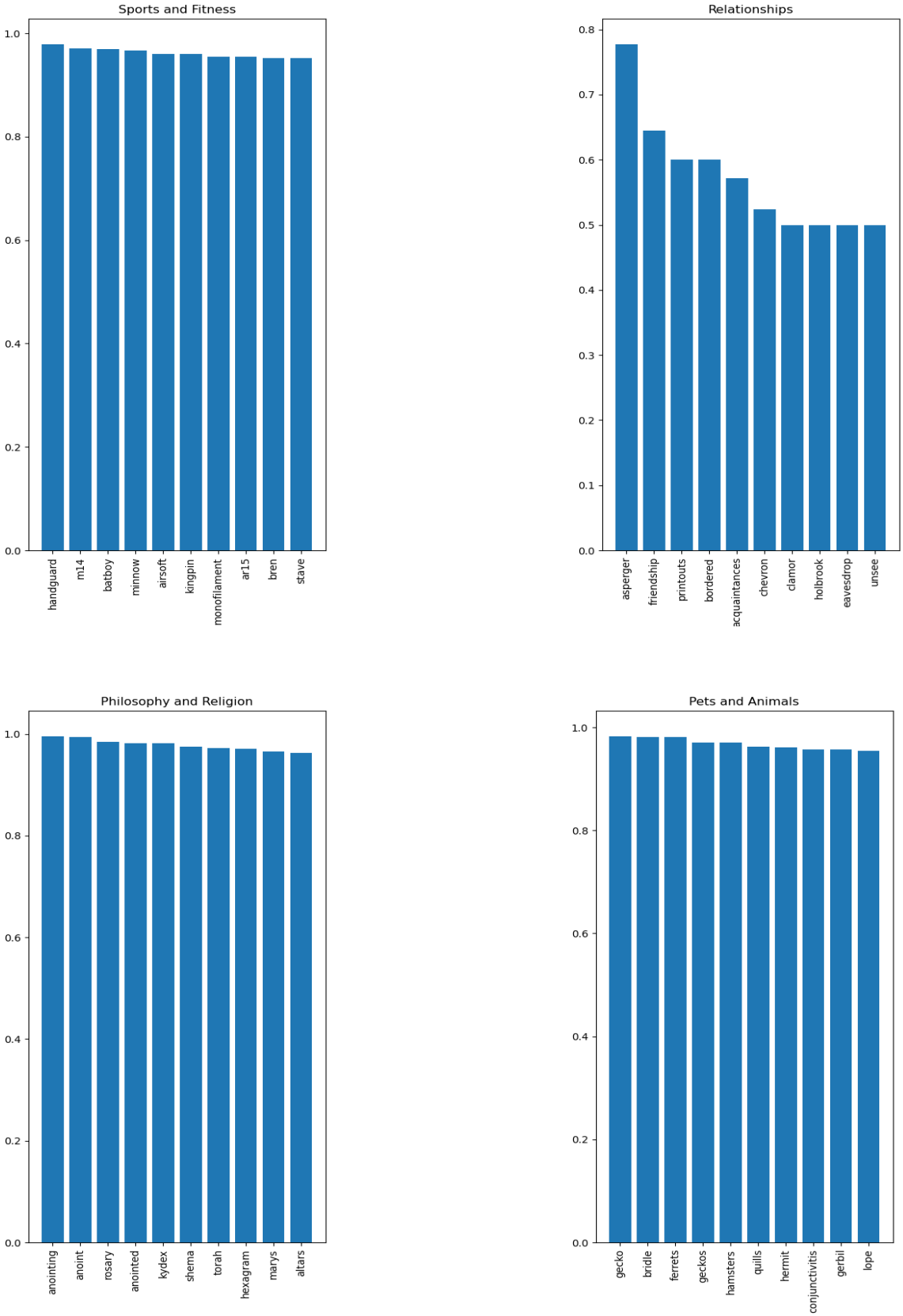

Figure 9: RNF metric words per category

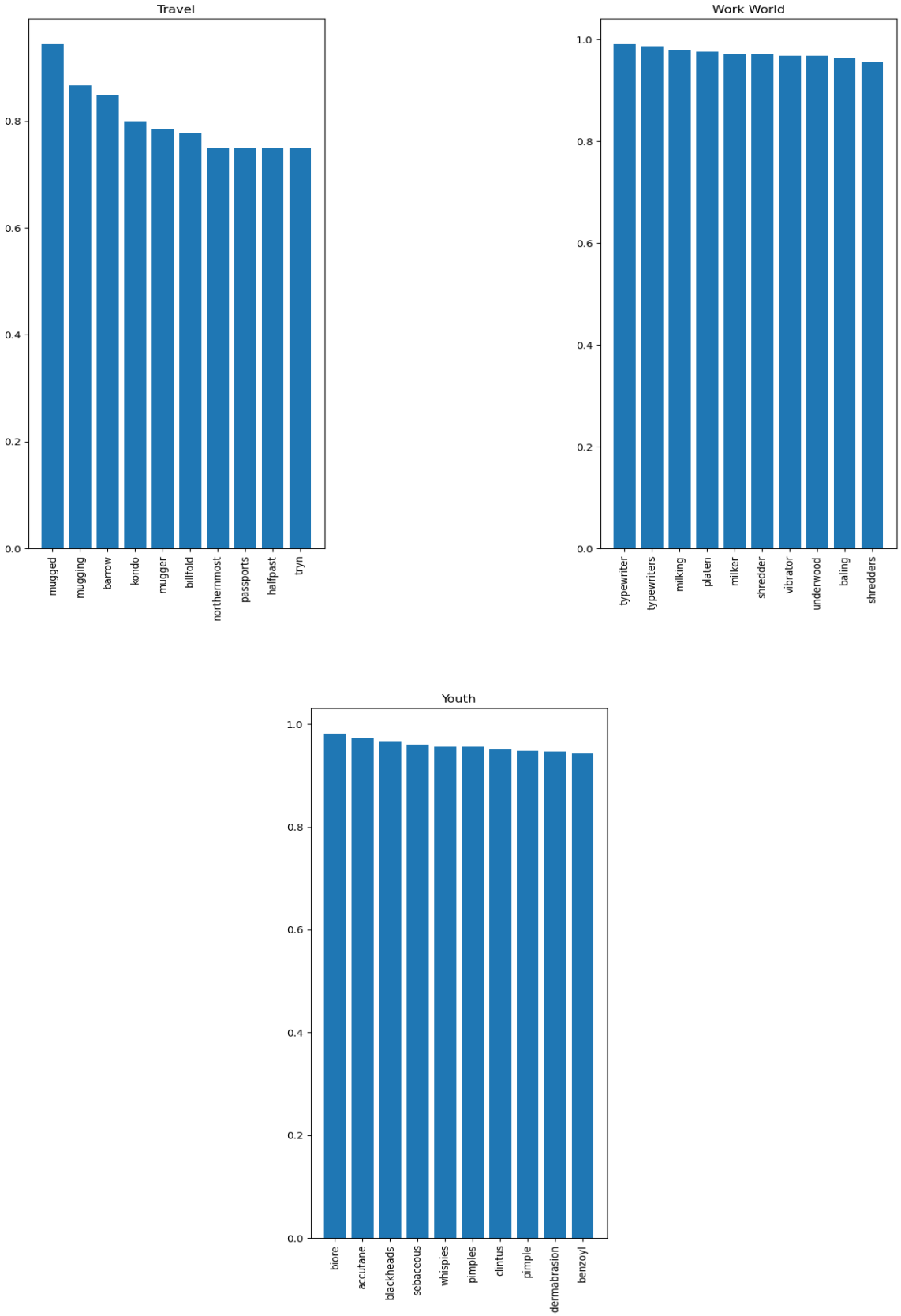

Figure 10: RNF metric words per category

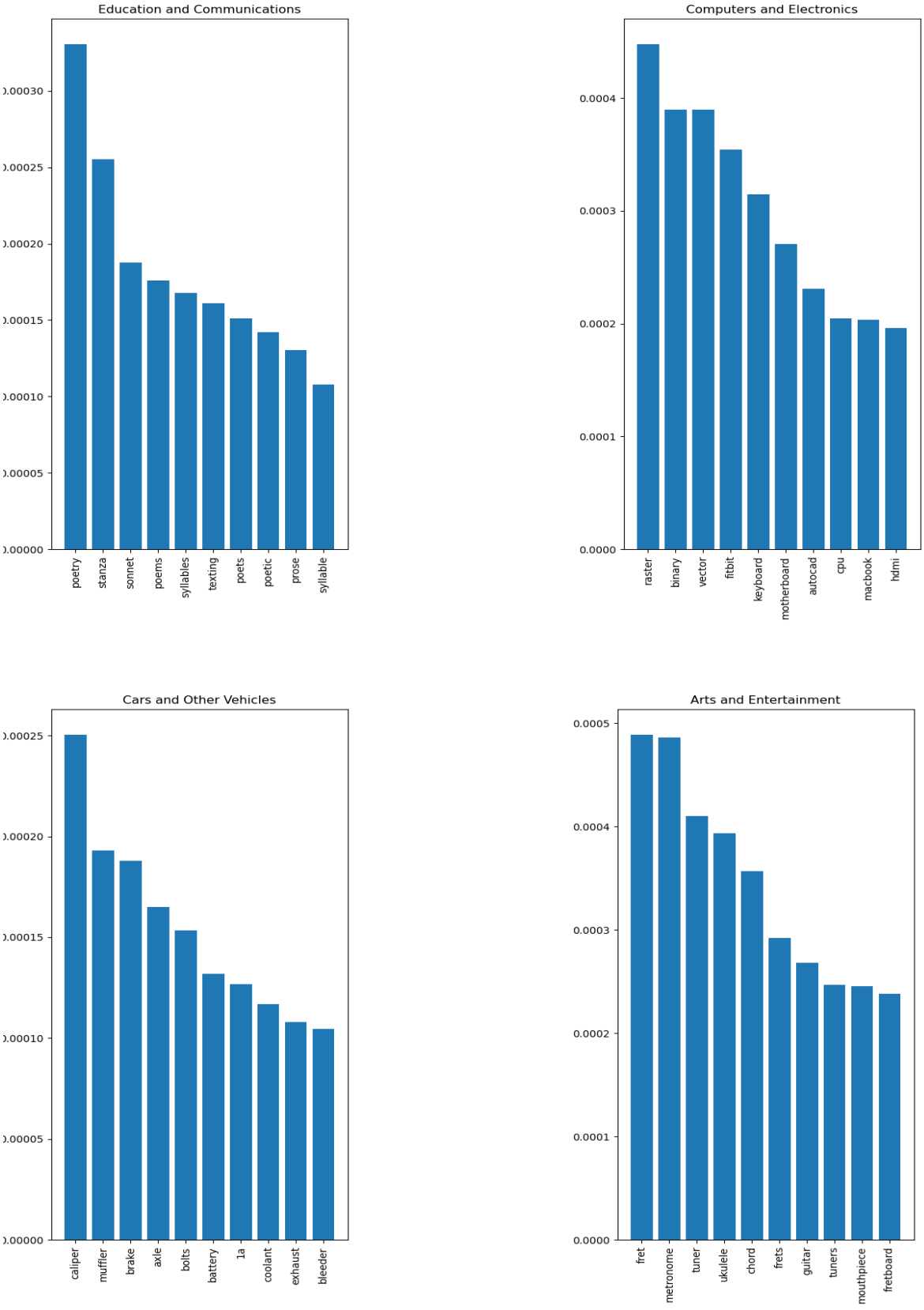

Figure 11: TF-IDF metric words per category

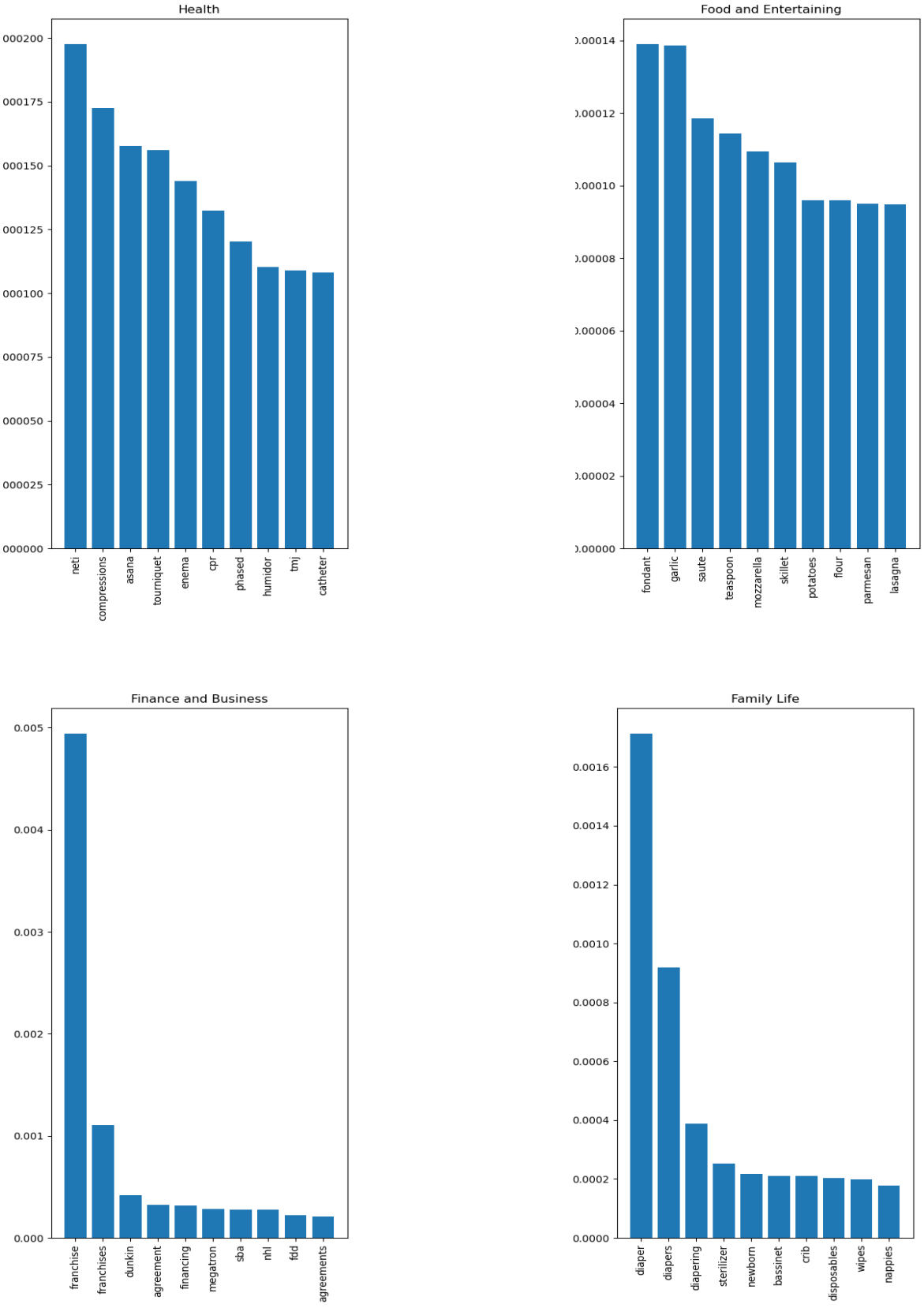

Figure 12: TF-IDF metric words per category

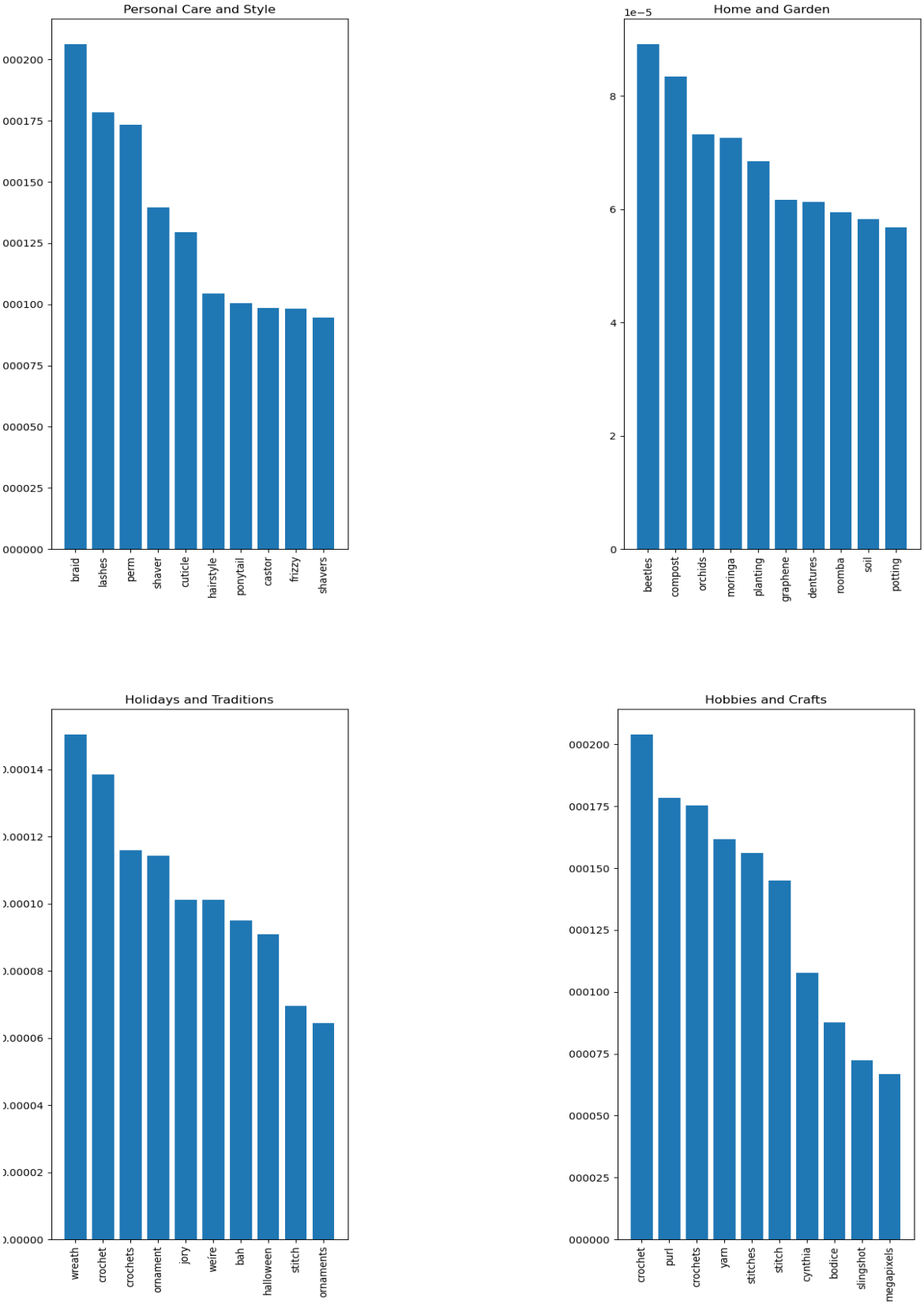

Figure 13: TF-IDF metric words per category

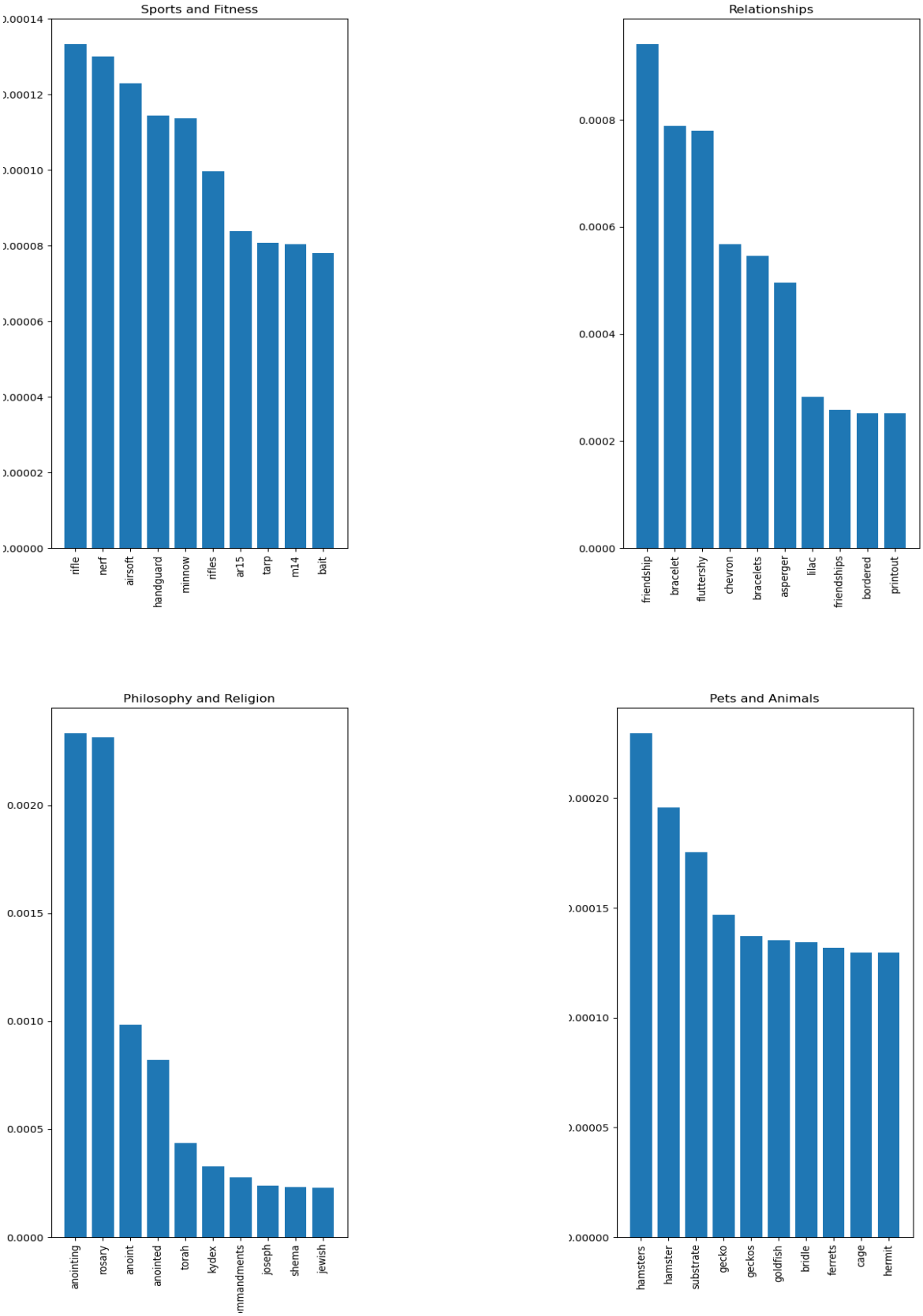

Figure 14: TF-IDF metric words per category

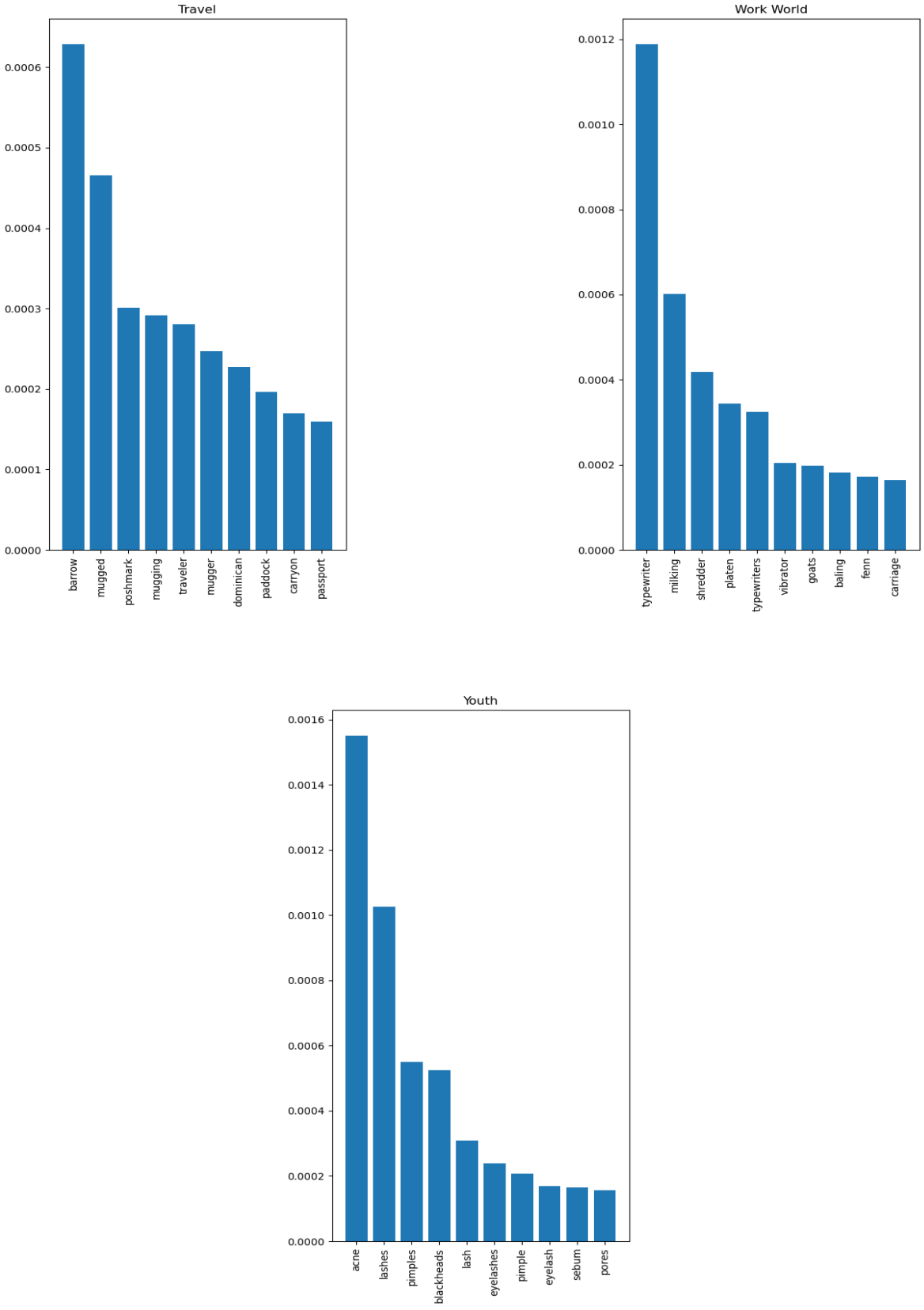

Figure 15: TF-IDF metric words per category

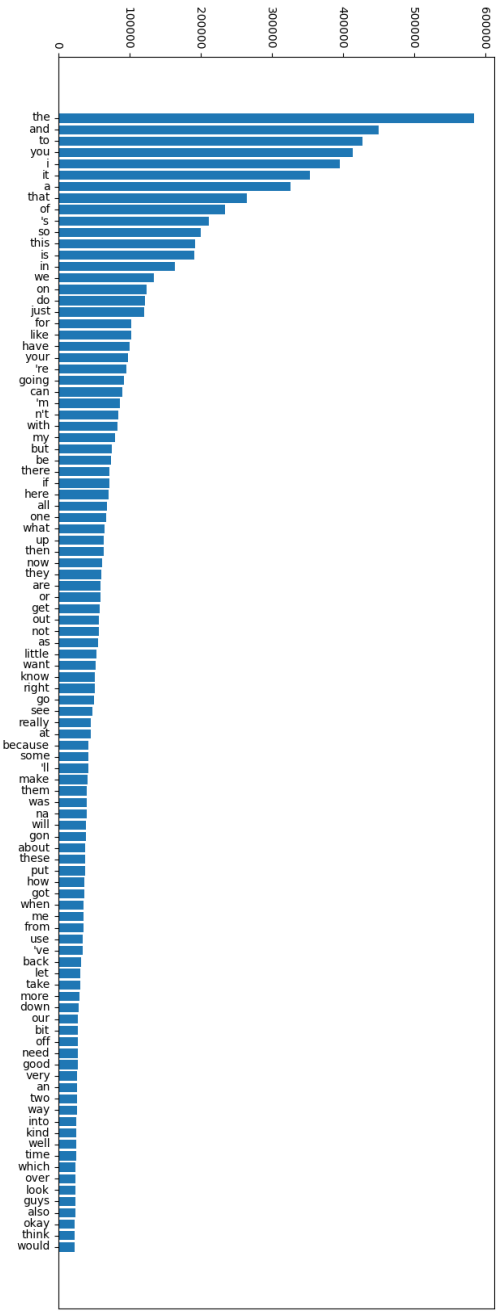

Figure 16: Frequency histogram for top 100 most frequent words

category	number of unique words
All	93835
Common	46940
Holidays and Traditions	2316
Home and Garden	2314
Computers and Electronics	2253
Youth	2401
Personal Care and Style	1764
Philosophy and Religion	2023
Hobbies and Crafts	2466
Relationships	66
Sports and Fitness	2992
Work World	2493
Family Life	1842
Health	3126
Travel	313
Finance and Business	4785
Education and Communications	4013
Arts and Entertainment	3016
Cars and Other Vehicles	2365
Pets and Animals	2654
Food and Entertaining	3693

Table 4: Number or unique words per category.

6 Conclusion

In summary, this project has successfully completed the crucial steps of data gathering, cleaning, and extracting useful statistics and metrics from the HowTo100M dataset. These preliminary findings will serve as a solid foundation for the development of the video subtitle topic classification model in the next phase. By leveraging the insights gained and incorporating advanced techniques, we aim to create a powerful model capable of accurately classifying video subtitles into various topics, thereby enabling applications in information retrieval, content recommendation, sentiment analysis, and automated content organization.