PLOS ONE

RESEARCH ARTICLE

Representation of professions in entertainment media: Insights into frequency and sentiment trends through computational text analysis

Sabyasachee Baruah*, Krishna Somandepalli, Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory, University of Southern California, Los Angeles, California, United States of America

* sbaruah@usc.edu

OPEN ACCESS

Citation: Baruah S, Somandepalli K, Narayanan S (2022) Representation of professions in entertainment media: Insights into frequency and sentiment trends through computational text analysis. PLoS ONE 17(5): e0267812. https://doi.org/10.1371/journal.pone.0267812

Editor: Richard A Blythe, University of Edinburgh, UNITED KINGDOM

Received: August 28, 2021

Accepted: April 15, 2022

Published: May 18, 2022

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0267812

Data Availability Statement: The data can be found at https://github.com/sabyasachee/mica-profession.

Abstract

Societal ideas and trends dictate media narratives and cinematic depictions which in turn influence people's beliefs and perceptions of the real world. Media portrayal of individuals and social institutions related to culture, education, government, religion, and family affect their function and evolution over time as people perceive and incorporate the representations from portrayals into their everyday lives. It is important to study media depictions of social structures so that they do not propagate or reinforce negative stereotypes, or discriminate against a particular section of the society. In this work, we examine media representation of different professions and provide computational insights into their incidence, and sentiment expressed, in entertainment media content. We create a searchable taxonomy of professional groups, synsets, and titles to facilitate their retrieval from short-context speaker-agnostic text passages like movie and television (TV) show subtitles. We leverage this taxonomy and relevant natural language processing models to create a corpus of professional mentions in media content, spanning more than 136,000 IMDb titles over seven decades (1950-2017). We analyze the frequency and sentiment trends of different occupations, study the effect of media attributes such as genre, country of production, and title type on these trends, and investigate whether the incidence of professions in media subtitles correlate with their real-world employment statistics. We observe increased media mentions over time of STEM, arts, sports, and entertainment occupations in the analyzed subtitles, and a decreased frequency of manual labor jobs and military occupations. The sentiment expressed toward lawyers, police, and doctors showed increasing negative trends over time, whereas the mentions about astronauts, musicians, singers, and engineers appear more favorably. We found that genre is a good predictor of the type of professions mentioned in movies and TV shows. Professions that employ more people showed increased media frequency.

Funding: The study was done at Signal Analysis and Interpretation Laboratory, University of Southern California, which is supported by a research award from the U.S. Chamber of Commerce Foundation (https://www.uschamberfoundation.org/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Entertainment media, available in rich variety and diverse forms ranging from traditional feature films, television shows, and theatrical plays to contemporary digital shorts and streaming content, can profoundly impact audience perceptions, beliefs, attitudes, and behavior. Media narratives aim to inform and engage us with stories about the culture, lives, and experiences of different communities of people, including reflecting societal ideas and trends. They shed light on various social, economic and political issues, educating and creating awareness on different aspects of life. Cultivation theory suggests that prolonged exposure to the content we see on TV shapes our outlook and makes us believe that to be our reality [1]. Several social studies have explored the application of cultivation theory and confirmed its validity [2–4]. However, there are social scientists that have also questioned the validity of the cultivation theory because it ignores socioeconomic factors [5], living conditions [6], and differences in the portrayal of violence on television [7]. Still, this discussion about television and the cultivation of beliefs inspires us to examine their relationship. With the increasing number of film releases (theatrical and home/mobile entertainment market increased by 4% in 2019 [8]) and the large amount of time people spend with major media (US people watched TV for around 4 hours daily [9]), media experiences influence our views and choices in many spheres of life; this includes our professional and career choices.

Professions are paid, skilled work we perform to provide services to people and earn a livelihood. They define our role in society and allows us to contribute to a nation's economy. The distribution of the country's population in different occupations provides valuable information to business leaders, policymakers, educational institutions, students, and job seekers, to understand the labor market and make decisions. Therefore, it is critical to study and assess the changes in the occupational structure of nations. One of the primary factors that affects the professional distribution of a country is the career choices its inhabitants make regarding what educational and professional pathways to pursue. Personal interests, family expectations, contemporary culture, and media exposure influence their choices [10, 11]. Of particular interest to us in this work is the relationship between people's career choices and the media representation and portrayal of professions. Several prior works have studied this connection. Undergraduate students have indicated that the portrayal of the advertising industry in two popular TV shows—*Mad Men* and *Trust me*, prompted them to enroll in advertising courses [12]. Majoring in journalism could be predicted by the student's media and technology use [10]. US Navy recruitment went up by 500% after the release of the movie *Top Gun* because many young men were inspired by the character *Pete "Maverick" Mitchell*, a US naval aviator, played by Tom Cruise [13]. Similarly, the detective character *Dana Scully*, of the TV series *The X-Files*, played by actor Gillian Anderson, inspired many young women to pursue a career in STEM [14]. A survey of 1005 currently employed people in the US found that 58% of them attributed their career inspiration to some book, TV show, movie, podcast, or video game [15]. Therefore, the portrayal of professions in media plays a significant role in our career decisions, affecting the occupational distribution of society.

Several studies have examined the nature of the portrayal of different professions in popular media, such as lawyers [16], accountants [17], physicians [18], and police [19]. While these past studies closely examined the personality of the character portrayed in the profession of interest, their methods are not scalable because the authors manually viewed the movie or read the transcripts to infer the character's personality. Such studies can not examine more than a few hundred movies or TV shows at a time. The set of personality attributes also varied between different works. Therefore, there is a need to conduct such media studies of professions more systematically and computationally.

Our objective is to conduct a computational study of the representation of professions in media content, spanning a large set of movies and TV shows over a time period. We rely on textual data from media subtitles for this study. Specifically in this work, we use *professional mentions* as a proxy for the representation of professions in movies. Professional mentions are job titles (doctor, engineer, cop, lawyer, etc.) used to indicate a profession within an utterance. Word mentions have been previously used to study trends of different societal functions, for example, education, culture, and language use patterns [20–22]. Michel et al. introduced the Google Books corpus, which contains digitized copies of more than five million books [23]. They used n-gram frequencies to track the size of the English lexicon, word usage, grammatical structures, popularity index of individuals, etc., over time. Brysbaert et al. argued that word frequency measures of media content were better than those calculated from written sources for psycholinguistic research [24]. Inspired by these works, we use mentions of profession related words to study the representation of professions in media content. The following are the contributions of our work:

1. We develop a scalable and searchable taxonomy of professional titles, professional WordNet synsets and occupational groups of the Standard Occupational Classification (SOC) system [25].
2. We describe and share a new corpus of professional mentions, spanning 4,000 professions, 136,000 movies and TV shows, ranging over the years 1950 to 2017 (almost 7 decades) created by analyzing job title occurrences in media subtitles [26].
3. We study the relationship between real-world employment trends and mentions of different occupational groups in media content.
4. We analyze the frequency and contextual sentiment trend of professional mentions in media over time [27, 28]. We also investigate the correlation between incidence of professional mentions and the genre, title type, and country of production of the movie or TV show.

The rest of the paper is organized as follows: the Related Work section reviews some past studies about media representation of professions. It also gives the technical background of the computational models and knowledge bases we use in our work. The Data section describes the profession gazetteers and the subtitles dataset we use to search professional mentions in media content. The Methodology section is divided into two parts: Taxonomy Creation and Profession Search. The first subsection explains how we create a searchable taxonomy of job titles. The second subsection describes how we use this taxonomy to search and annotate professional mentions in media subtitles. The Analysis section analyzes the media frequency and sentiment trends of different professions, and their correlation with the real-world employment figures. We conclude with a discussion of the results of our analysis and opportunities for further research.

Related work

We summarize some past studies on media representation of professions. These studies have typically examined a small set of movies, usually manually, and their methods cannot be scaled easily to other occupations. We build upon various well-established natural language processing (NLP) methods and lexical resources to address these limitations. We leverage job title gazetteers and WordNet synsets to create a searchable taxonomy of professions. We prune non-professional mentions of job titles using word sense disambiguation and named entity

recognition, and find the targeted sentiment of the remaining mentions. We briefly explain the theoretical background of each method and highlight relevant work.

Social scientific studies

Several past works have studied the portrayal of professions in entertainment media. Asimow studied the representation of lawyers in 284 films and found most portrayals to be negative [16]. Dimnik et al. examined the representation of accountants in 121 movies and extracted six character stereotypes of the accountant personality [17]. Flores investigated physicians' image depictions in 131 films and found that they were mostly depicted as greedy and uncaring [18]. Pautz used a sample of 34 films containing more than 200 police (cop) characters and found that most cops were shown as good, hard-working, and competent law-enforcement officers [19]. Kalisch et al. analyzed 670 nurse and 466 physician characters in novels, movies and television series, and concluded that compared to physicians, nurses portrayed in the media were consistently less central to the plot, less intelligent, less rational, and less likely to exercise clinical judgement [29]. Smith et al. investigated gender representation of occupations in films, prime-time programs, and children TV shows, and found that women are grossly underrepresented compared to men in science, technology, engineering and math jobs (STEM) [30]. These works involved extensive human coding and profession-specific analysis which is not reproducible on a large scale. The present study offers a complementary view in the sense it trades off a smaller scale character-centric study for a larger scale lexical analysis, focusing on character utterances instead of personality traits. The remaining sections describe the computational methods used to achieve the same.

Named entity recognition

Named entity recognition (NER) is a classic NLP task of finding entity mentions in text and classifying their type. Traditional NER primarily targets person, organization, and location entity types. Most NER datasets only contain labels for these three types of entities [31]. The OntoNotes 5 dataset increased the number of entity types to 18 by including nationalities, products, events, and numeric values [32]. However, it does not contain any professional titles. Fine-grained NER extends traditional entity recognition by expanding the entity set to include hundreds of named categories. Ling et al. created a benchmark dataset for fine-grained NER that labels 112 different entity types but it only contains a few professional titles [33]. Sekine built an ontology for named entities and defined professional titles as vocational attributes for the person-entity type [34]. Mai et al. used this entity hierarchy to annotate English and Japanese sentences and evaluated different fine-grained NER models [35]. However, they did not include the professional attributes in their labeling set. The Text Analysis Conference Knowledge Base Population (TAC KBP) track of entity discovery and linking introduced job titles as entity types to model the *person:title* relationship [36]. The Stanford CoreNLP NER model used the KBP 2017 dataset to create regular expression-based rules for finding professional titles in text [37]. However, we observed that it missed many of the professional mentions in media text.

In the absence of labeled data, entity gazetteers are often used to find candidate spans for named entities. Gazetteers are curated lists of entities that improve NER performance when combined with supervised models. Lin et al. used gazetteers to identify text subsequences for the region-based encoder and improved the state-of-the-art NER performance on the ACE2005 benchmark [38]. Liu et al. combined dictionary lookups with semi-Markov CRF (conditional random field) architectures and achieved comparable results with more complex neural models on the CoNLL 2003 and OntoNotes 5 datasets [39]. Several gazetteers of job

titles are available (Gate job titles, fluquid). Government and international organizations use some standard gazetteers of job titles to collect occupational data. For example, the International Labor Organization uses the International Standard Classification of Occupations [40], and the US Bureau of Labor Statistics maintains the Standard Occupational Classification (SOC) system [25]. Aside from careful human coding, such gazetteers can be constructed using different automated methods, including those that leverage existing knowledge bases, such as Wikipedia and WordNet [41, 42]. In this work, we use the SOC taxonomy to get the initial list of job titles and professional groups.

WordNet

WordNet is a widely-used lexical resource in English [43]. It groups words into synonym sets called synsets. Synonymy is the semantic relationship between words of similar meaning. Polysemous words have multiple meanings and belong to more than one synset. For example, the word "conductor" is present in three synsets—*conductor.n.01* (the person who leads a musical group), *conductor.n.02* (a substance that readily conducts electricity and heat) and *conductor.n.03* (the person who collects fares on a public conveyance). The name of a synset is composed of three dot-separated literals. The first literal is the lemmatized form of the main word of the synset, the second literal is the part-of-speech, and the third literal is the sense index. Synsets are tagged by semantic classes [44]. Synset *A* is a hyponym of synset *B* if it is a more specific form of *B*. For example, *allergist.n.01* (doctor specialized in the treatment of allergies), *surgeon.n.01* (doctor specialized in surgery) and *veterinarian.n.01* (doctor practicing veterinary medicine) are all hyponyms of the more general synset, *doctor.n.01* (a licensed medical practitioner).

WordNet has been used to construct entity gazetteers. Toral et al. leveraged WordNet's noun hierarchy to build person and location gazetteers [45]. Maginini et al. used WordNet to identify trigger words and gazetteer terms for English NER [46]. Boteanu et al. expanded a shopping taxonomy for efficient product search by matching the product names to WordNet synsets [47]. In this work, we use WordNet synsets to extend the SOC taxonomy and create a searchable dictionary of professions. WordNet synsets are also the target labels for the word sense disambiguation task, an integral part of our search pipeline.

Word sense disambiguation

Word sense disambiguation (WSD) is the task of assigning words in context to their most appropriate sense. Wordnet usually serves as the sense inventory that provides the target senses. The same word can express different meanings depending upon its context. Consider the word "conductor" in the following two sentences—"Conductors communicate with the musicians through hand gestures" and "Metals are good heat conductors". The former refers to a person directing the music of an orchestra, denoted by the synset *conductor.n.01*. The latter means a heat-conducting substance, denoted by the synset *conductor.n.02* (See Sec WordNet). Many NLP tasks such as machine translation, information retrieval and question answering use WSD in their text-processing pipeline [48–51].

Knowledge-based and supervised approaches have tackled WSD, with the latter usually outperforming the former. Raganato et al. standardized the evaluation framework for WSD and used a combination of SenseEval [52, 53] and SemEval [54–56] datasets to compare the performance of different WSD models [57]. Raganato et al. treated WSD as a sequence labeling task and used an LSTM with attention layers to find the sense of all sentence words jointly [58]. Huang et al. constructed context-gloss pairs and converted WSD into a sentence pair classification task [59]. Kumar et al. produced gloss embeddings using the WordNet graph and

combined them with contextual vectors to find the word sense [60]. Bevilacqua et al. extended this method by adding hypernym and hyponym relational knowledge to construct the synset vectors and achieved state-of-the-art WSD performance [61]. In this work, we use WSD to remove non-professional mentions of job titles in media subtitles.

Sentiment analysis

Sentiment analysis or opinion mining is the task of finding the sentiments, opinions, attitudes, appraisals and emotions towards entities or their attributes expressed or implied in text [62]. Sentiment is always targeted at some entity or towards some attribute of the entity. The target entity or attribute can be a person, organization, issue, product, service, topic or event. Such target-oriented opinion mining is called aspect-based sentiment analysis (ABSA).

We use professional mentions as opinion targets for sentiment analysis to find how positively or negatively different professions are talked about in media stories. The task of profession ABSA is to find the sentiment orientation of the opinion expressed towards the person or group of persons referred to by their job title. If the job title does not refer to any profession, or people employed in the profession, the sentiment is deemed neutral. The following example sentences show the sentiment label of the job title word, marked in bold.

1. *Harry Floyd was a great* ***actor***. (POSITIVE)
   Explanation: Actor refers to Harry Floyd who is described as great.
2. *But that damn* ***vet*** *kept ordering test after test after test!* (NEGATIVE)
   Explanation: The speaker uses a swear term, *damn*, to address the veterinarian and criticizes his or her action.
3. *Fine, then we get the armor and reverse* ***engineer*** *it*. (NEUTRAL)
   Explanation: The word *engineer* refers to an action, not a profession.
4. *You're going to be a lousy* ***architect***. (NEUTRAL)
   Explanation: The person towards which the negative sentiment is expressed, is not yet an architect.

Benchmark ABSA datasets exist for several domains such as question answering forums, customer reviews and tweets [28, 63, 64]. Dong et al. proposed an adaptive recursive neural network for target-dependent twitter sentiment classification, that propagated the sentiments of words to their targets depending upon the context and syntactic relations [28]. Tang et al. used LSTMs to model the left and right context of the target entity for twitter sentiment classification [65]. Memory networks and graph convolutional neural models have also been proposed [66, 67]. Recently, several transformer networks have been introduced for the ABSA task and have achieved state-of-the-art performance [68]. Sun et al. constructed sentences containing the aspect expression and the sentiment orientation, and fine-tuned a pre-trained BERT [69] model for the ABSA task [70]. Xu et al. trained a BERT model jointly for reading comprehension and ABSA, and showed performance improvement on both tasks [71]. Zeng et al. proposed dynamically weighing or masking the attention weights of the sentence words depending upon its distance from the aspect expression [27]. In this work, we use ABSA models to find the sentiment expressed toward professions in media content.

Data

We search for mentions of job titles in entertainment media content to study the representation of professions. We use the Standard Occupational Classification (SOC) system [25] to create a searchable taxonomy of professional titles. We apply this taxonomy to find professional

mentions in media (movie) subtitles, for which we use the OpenSubtitles corpus [72]. This section describes the SOC taxonomy and the OpenSubtitles dataset.

Standard Occupational Classification taxonomy

The Standard Occupational Classification (SOC) system is a profession taxonomy, maintained by the US Bureau of Labor Statistics [25]. It arranges professions in four tiers: major, minor, broad, and detailed. The detailed tier contains a set of professions, closely related by work. Fig 1 lists all 23 major SOC groups, and shows a portion of the taxonomy's subtree rooted at *Management Occupations* major SOC group. As shown in the figure, the profession *Governor* occurs in the following SOC hierarchy: *Management Occupations* (major) → *Top Executives* (minor) → *Chief Executives* (broad) → *Chief Executives* (detailed) → *Governor* (profession). The SOC taxonomy contains 6520 unique professions.

OpenSubtitles

We used the English subset of the OpenSubtitles dataset [72], which contains 135,998 subtitle files corresponding to a variety of media content from the years 1950 to 2017. Each subtitle file is mapped to a unique IMDb title. More than 94% of the IMDb titles are movies or TV show episodes, and the rest are made up of video games, TV shorts, TV mini-series, etc. Subtitle files for IMDb titles released before 1950 are available, but we excluded them because there were very few titles for each year (less than 100) and it would not have been a representative sample for that period's media content. The subtitle files of our dataset contain around 126 million sentences and 942 million words.

Fig 2 shows some media metadata statistics of our subtitles dataset. The first panel shows the temporal distribution of movie and TV show IMDb titles by each decade. The number of media titles increases with time, with TV episodes surpassing movie releases in the later years. The number of TV episodes is more than three times the number of movies in the most recent decade (2010-2017). The second panel shows the distribution of the top ten genres of the IMDb titles in our dataset. An IMDb title can have multiple genres. Drama and Comedy are the two most common genres, covering more than 80% of the IMDb titles. The third panel shows the distribution of the top ten most common countries where the production company is based. About 68% of the time, the production company was based in the US or the UK.

Methodology

We create a corpus of professional mentions by searching job titles in the OpenSubtitles dataset. We use this corpus to study the relationship between media portrayal of professions and real-world employment trends. However, we cannot use the SOC taxonomy directly to find professional mentions in media content because of the following reasons:

1. Most of the SOC job titles are very specific multi-word phrases, for example, *Department Store Manager*, *Registered Occupational Therapist*, *Television News Video Editor*, etc. Such detailed words are rarely spoken in everyday conversations (including those captured in the subtitle transcripts of media considered here). They instead include simpler unigram professional words like *Manager*, *Therapist*, *Editor*, etc. Less than 7% of the SOC job titles are unigrams.

2. The mere occurrence of a job title in text does not mean it refers to some profession. For example, consider the sentence—*I made a peach **cobbler** for the party*. *Cobbler* is a job title, but here refers to a type of food. Both the lexical form and the context decides whether the word is a professional mention or not.

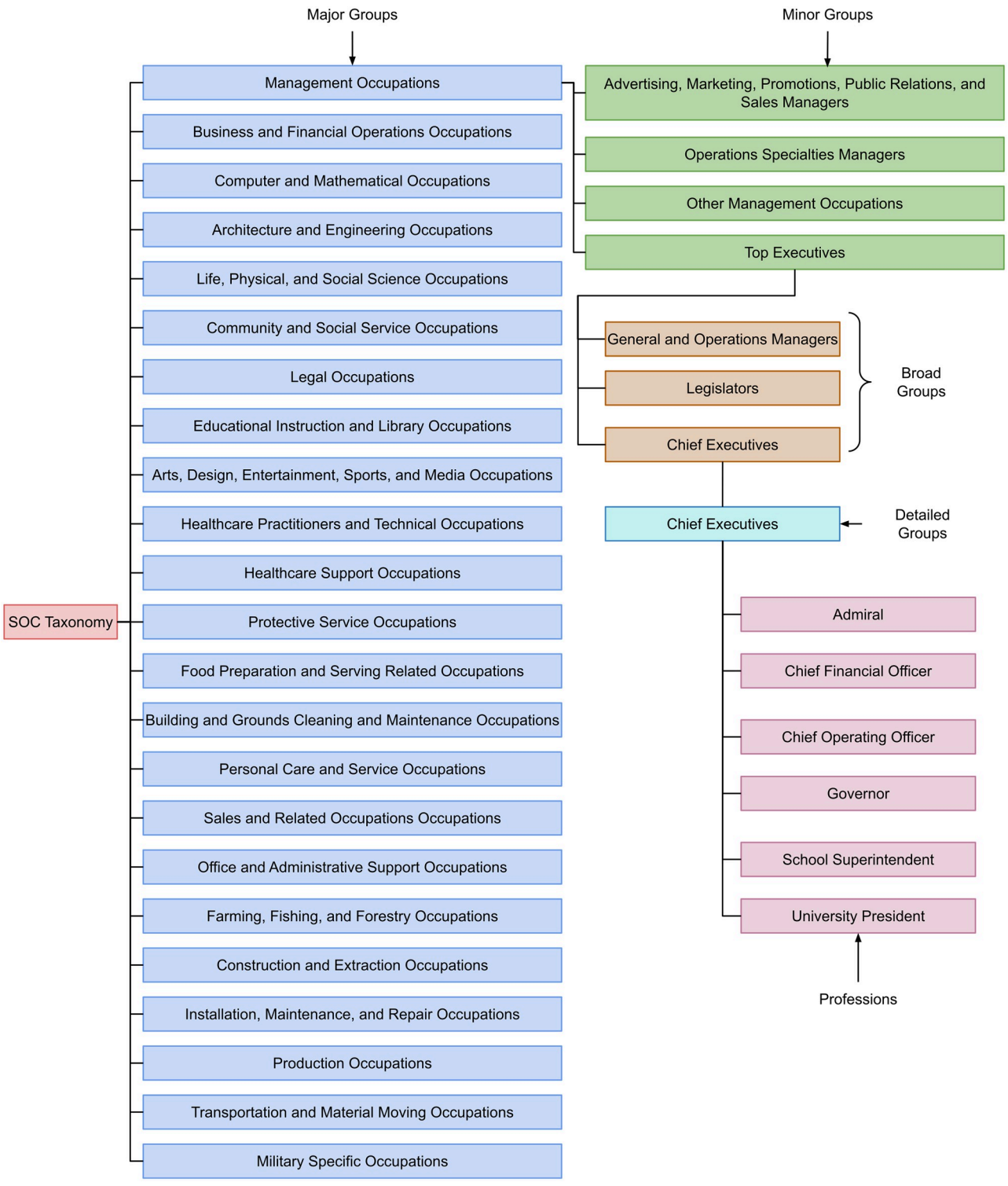

Fig 1. Standard Occupational Classification (SOC) system. The Standard Occupational Classification (SOC) is a 4-tiered profession taxonomy: major, minor, broad, and detailed groups. Detailed groups contain a set of closely-related professions. This figure shows the *Management Occupations* (major) → *Top Executives* (minor) → *Chief Executives* (broad) → *Chief Executives* (detailed) → *Admiral, Chief Financial Officer, . . ., University President* (professions) branch.

https://doi.org/10.1371/journal.pone.0267812.g001

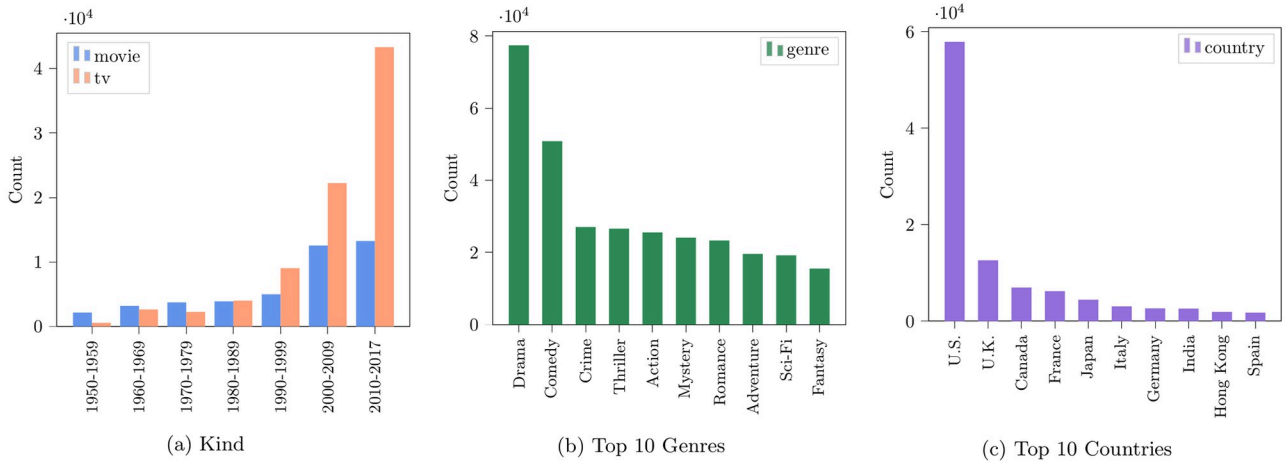

Fig 2. Descriptive statistics of the OpenSubtitles dataset. a) Distribution of IMDb title type (movie/TV) by year. b) Distribution of the top ten genres. c) Distribution of the top ten production countries. We use the OpenSubtitles dataset between the years 1950 and 2017.

https://doi.org/10.1371/journal.pone.0267812.g002

Therefore, in order to make the SOC taxonomy searchable, we need to extend its list of job titles to include simpler, more common words, and have a disambiguation model to filter non-professional usages of job titles. Fig 3 outlines the complete pipeline of expanding the SOC taxonomy, creating the corpus of professional mentions, and analyzing its frequency and sentiment trends. The figure uses the *cobbler* profession to exemplify the corpus creation method. As shown in the figure, the Taxonomy Creation section describes how we expand the SOC system, and create a searchable profession taxonomy. The Profession Search section explains the NLP techniques we apply to find the professional mentions and its targeted sentiment.

Taxonomy creation

We use WordNet synsets to create a searchable profession taxonomy. Fig 4 shows an example of finding new professions from the SOC job titles: *Orchestra Conductor* and *Train Conductor*.

We use the following method to expand the SOC taxonomy.

Find Substrings: Given a SOC job title, we split it into substrings and join them cumulatively from the end to find candidate job titles. For example, given the job title *Chief Executive Officer*, we find the new candidate titles—*Officer* and *Executive Officer*. We do not split titles that contain conjunctions, prepositions, punctuations, or those that are abbreviations (all letters are in upper case), or if they contain more than five words. In Fig 4, *Conductor* is a new job title we find from the SOC words, *Orchestra Conductor* and *Train Conductor*.

Find WordNet synsets: We find WordNet synsets of the candidate job titles. We retain only those synsets whose semantic class is *noun.person* or *noun.group*. We add the hyponym synsets of the retained synsets.

Remove Non-Professional Synsets: We manually check the list of synsets and remove those that do not refer to any profession. As shown in Fig 4, *conductor.n.02* denotes some heat or electricity conducting substance, which is not a profession and therefore, it is removed. The final curated list contains **1615** professional synsets.

Synonym Expansion: We collect all synonyms of the professional synsets. These are the new job titles, which we add to the SOC taxonomy. As shown in Fig 4, synsets *conductor.n.01* and *conductor.n.03* contribute the new job titles: *Music Director*, *Bandleader*, *Bandmaster* and

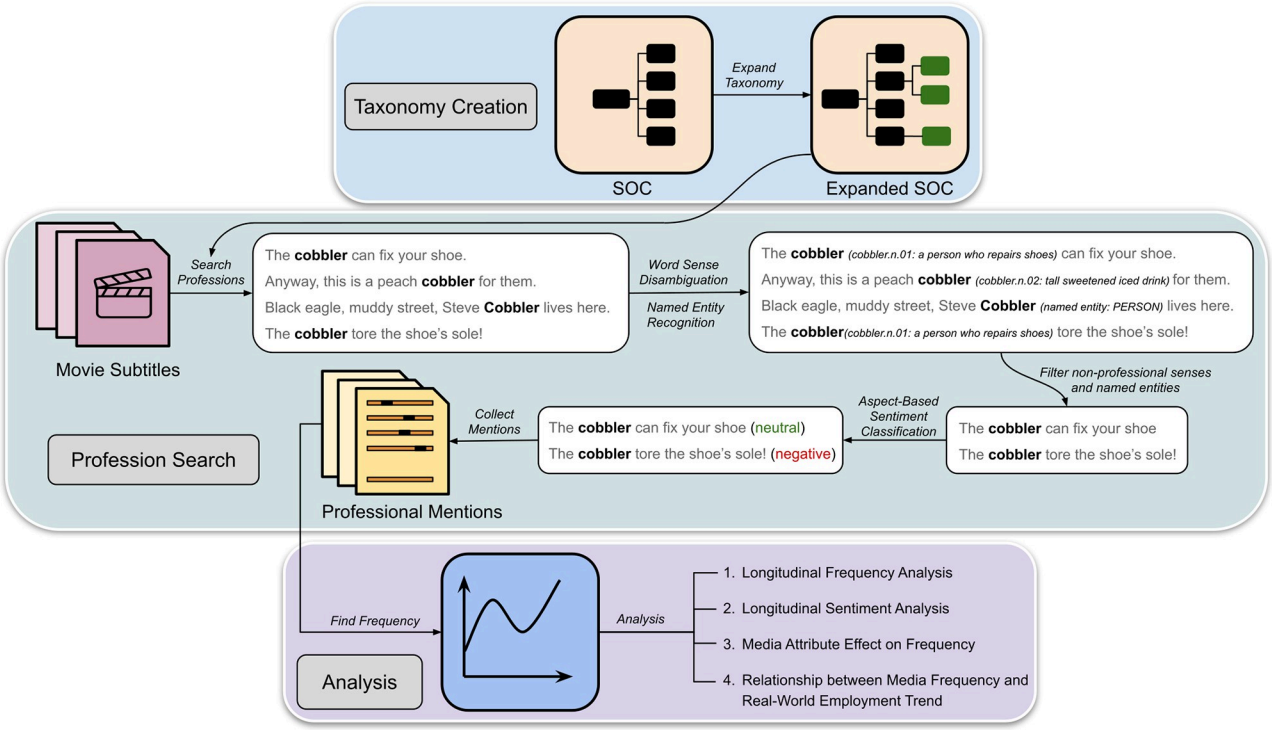

Fig 3. Profession taxonomy creation and analysis. The full pipeline for creating the profession taxonomy and the corpus of professional mentions, and analyzing the subtitle frequency and sentiment of professions.

https://doi.org/10.1371/journal.pone.0267812.g003

Conductress. The final taxonomy contains **10952** job titles. The number of unigram titles increased from 426 in SOC to 1881 in the expanded taxonomy.

SOC Mapping: We map the job titles in the expanded taxonomy to SOC major groups. This allows us to compare the employment in different SOC professional groups with their frequency in media content. We create the mapping through a semi-automatic process. Given a job title *x*, we first find the SOC major groups that contain *x* or has some job title which contains *x* as a substring. If there exists exactly one such SOC major group, we map *x* to it. Otherwise, we examine the professional synsets of *x* and find the mapping manually. Often, different synsets of a job title map to different SOC groups. For example, from Fig 4, we observe that the two synsets of *Conductor*, *conductor.n.01* and *conductor.n.03*, map to two different SOC major

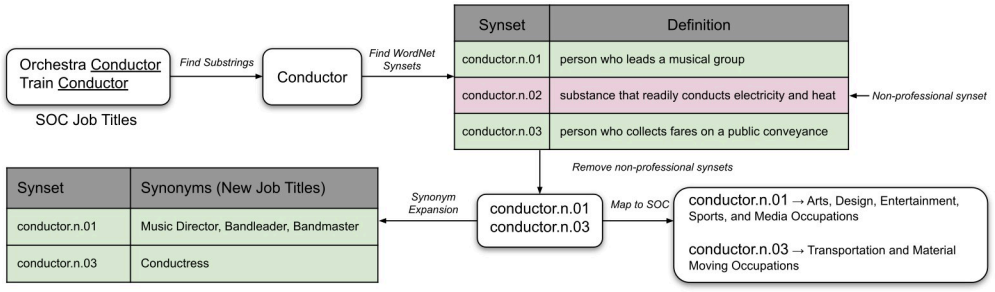

Fig 4. Profession taxonomy creation. We use WordNet synsets to expand the SOC taxonomy and map to SOC major groups. This figure shows how we find new job titles from the SOC job titles, *Orchestra Conductor* and *Train Conductor*.

https://doi.org/10.1371/journal.pone.0267812.g004

Table 1. Profession taxonomy sizes.

	Professions	Unigram Professions	Synsets
SOC Taxonomy	6520	426	-
Expanded Taxonomy	10952	1881	1615
SOC-mapped Taxonomy	500	409	562

We expand the SOC taxonomy to create the Expanded taxonomy. The SOC-mapped taxonomy is a subset of the Expanded taxonomy. It has been mapped to SOC major groups.

https://doi.org/10.1371/journal.pone.0267812.t001

groups, depending upon their respective definitions. We only perform the mapping for the top **500** most occurring job titles in media subtitles. These job titles belong to **562** professional synsets and cover more than 94% of the professional mentions.

Table 1 shows the number of job titles and synsets in the different profession taxonomies. The SOC taxonomy does not contain any synsets and therefore, is not searchable. The expanded taxonomy adds professional synsets and quadruples the number of unigram job titles, making it searchable. The SOC-mapped taxonomy is a subset of the expanded taxonomy which we use to study the relationship of media frequency of professions with their employment trend. Fig 5 shows the structure of the SOC-mapped profession taxonomy. It contains three tiers: SOC major groups, WordNet synsets and job titles. The figure only shows five SOC major groups of the complete taxonomy.

Profession search

We search mentions of the expanded taxonomy's job titles in the OpenSubtitles corpus. We apply NER and state-of-the-art WSD techniques to prune non-professional mentions. Finally, we train an ABSA model on sentiment-annotated subtitle sentences, and use it to tag professional mentions with their sentiment polarities.

Mention search. We search the subtitle sentences to find mentions of job titles. We create a word-document search index using the Whoosh Python package [73] for quick retrieval of mentions. We also search for the plural form of the job title while finding its mentions. Paranthesized mentions, speaker references (*Referee: The match will begin shortly!*) and lyrical mentions are removed.

Removing non-professional mentions. As discussed in the Methodology section, not all job title mentions refer to some profession. We remove non-professional mentions using WSD and NER methods. We apply the EWISER (Enhanced WSD Integrating Synset Embeddings and Relations) WSD model to find the mention's sense [61]. The EWISER model achieved state-of-the-art performance on the WSD benchmark dataset [57], reporting an overall F1 of 80.1. We apply the Stanford CoreNLP NER model [37] to find the named entity tags of words. We use the following rule to find professional mentions. A job title mention refers to a profession if 1) the predicted sense belongs to the set of professional WordNet synsets of the expanded taxonomy (see section Taxonomy Creation), and 2) it is not the name of an organization, or of a person who is cast in the corresponding IMDb title. We remove the non-professional mentions of job titles using the above method. The remaining mentions form our corpus of professional mentions. Fig 3 shows the steps to remove non-professional mentions for the *cobbler* job title.

To evaluate our rule-based model of finding professional mentions, we randomly sample 200 job title mentions and manually annotate their professional label. The test set contained 123 professional mentions and 77 non-professional mentions. Our model correctly predicted

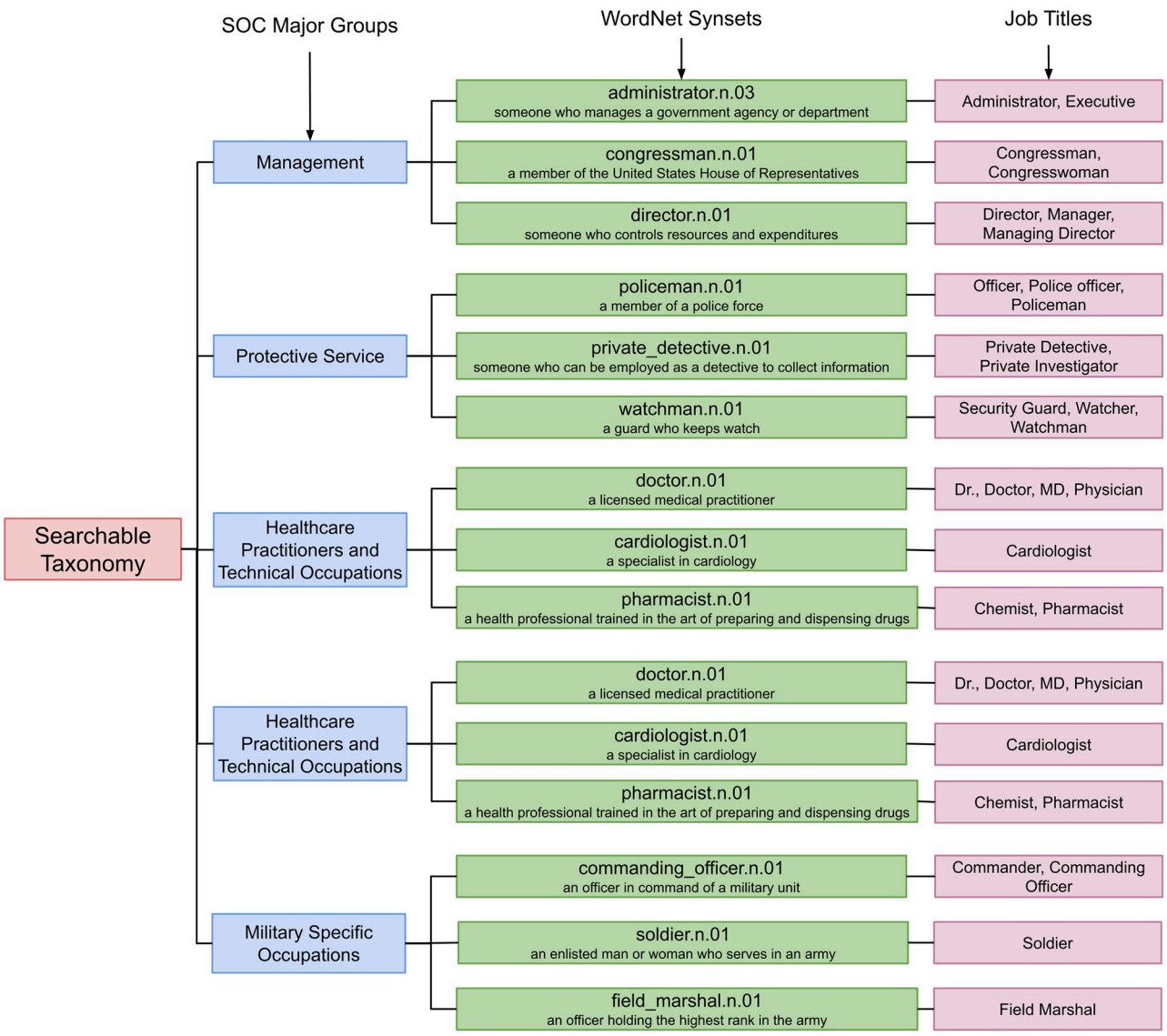

Fig 5. Searchable profession taxonomy. The SOC-mapped expanded profession taxonomy contains 3 tiers: SOC major groups, WordNet synsets, and job titles. The synsets and unigram job titles make the taxonomy searchable. This figure shows a few nodes of the SOC-mapped taxonomy, which contains 500 job titles and 562 synsets.

https://doi.org/10.1371/journal.pone.0267812.g005

the professional label for 83.5% of the mentions, with 94.12% precision and 78.05% recall. Therefore, our corpus of professional mentions has a 5.88% false-positive rate.

Determining expressed sentiment. We tag each professional mention with the sentiment (positive, negative, or neutral) expressed towards them in the subtitle sentence (see section Sentiment Analysis). We apply the LCF (Local Context Focus) BERT model to find the targeted sentiment [27]. The LCF model defines a value called semantic relative distance (SRD) for each word in the sentence. SRD is the absolute difference between the word position and the target (professional mention) position in the sentence. The model masks or weighs down the output features of words whose SRD is larger than some threshold, that is, the model reduces the effect of words that are farther away from the profession word in determining the

Table 2. Sentiment-annotated professional mentions dataset.

	Positive	Negative	Neutral	Professions
Train	2,431	1,409	3,915	85
Validation	345	167	366	11
Test	540	107	333	11
Total	3,316	1,683	4,614	107

The annotations are crowdsourced using Amazon Mechanical Turk. The train, validation and test sets do not share professions.

https://doi.org/10.1371/journal.pone.0267812.t002

sentiment. The LCF model achieved state-of-the-art accuracy on the Twitter ABSA task [28], recording 75.78 F1 score. It also obtained high scores on the customer reviews dataset [64] of laptops (79.59 F1) and restaurants (81.74 F1). We train the LCF model using sentiment-annotated professional mentions.

We crowdsourced sentiment annotations using Amazon Mechanical Turk. We trained Turkers using expert-labelled examples, and then asked them to annotate the sentiment of five sentences. We selected only those annotators who correctly annotated all five sentences. 52 annotators qualified our test. These annotators then labelled the sentiment of 15,000 professional mentions: two annotations per mention. We retained only those mentions with identical annotations. We were left with 9613 sentiment-annotated professional mentions: 3,316 positive, 1,683 negative, and 4,614 neutral. The dataset contains mentions of 107 professions. To train the LCF model, we divide the dataset into train, validation and test sets. Professions, whose mentions occur in the training set, do not appear in the validation and test set. This prevents the model from overfitting to the target professions of the training set, encouraging it to learn the sentiment from the context and handle mentions of unseen professions. Table 2 shows the distribution of the sentiment classes and the number of professions in each set.

We tune the following hyperparameters of the LCF model on the validation set: *SRD*, architecture type (*CDM* or *CDW*), L2-regularization, dropout, embedding dimension, and hidden dimension. The model achieved 87.76% accuracy and 83.22 F1 on the test set. We apply the trained model to find the targeted sentiment of each professional mention of our corpus.

In total, our corpus of professional mentions contains 3,657,827 mentions, covering 4,073 professions. The corpus contains mentions from 133,133 IMDb titles, ranging between the years 1950 to 2017. The top 500 most occurring professions, which have been mapped to SOC major groups (see section Taxonomy Creation), cover more than 94% of the mentions.

Analysis

We study profession representation in media according to the frequency of their mentions and the sentiment expressed towards them in subtitles. We also analyze the effect of media attributes like genre, location of the production company, and title type, on the incidence and sentiment of different professions. Lastly, we investigate the relationship between the trends of media frequency and the real-world employment statistics of professions.

Profession frequency

We calculate the media frequency of a profession as the total number of professional mentions (both singular and plural form, for example, *advocate* and *advocates*) divided by the total number of *n*-grams in the subtitles. Here, *n* equals the number of words in the profession phrase, for example, *doctor* is a 1-gram, *chief executive officer* is a 3-gram, etc. We calculate the frequency of SOC major groups by adding the frequencies of professions mapped to it (see

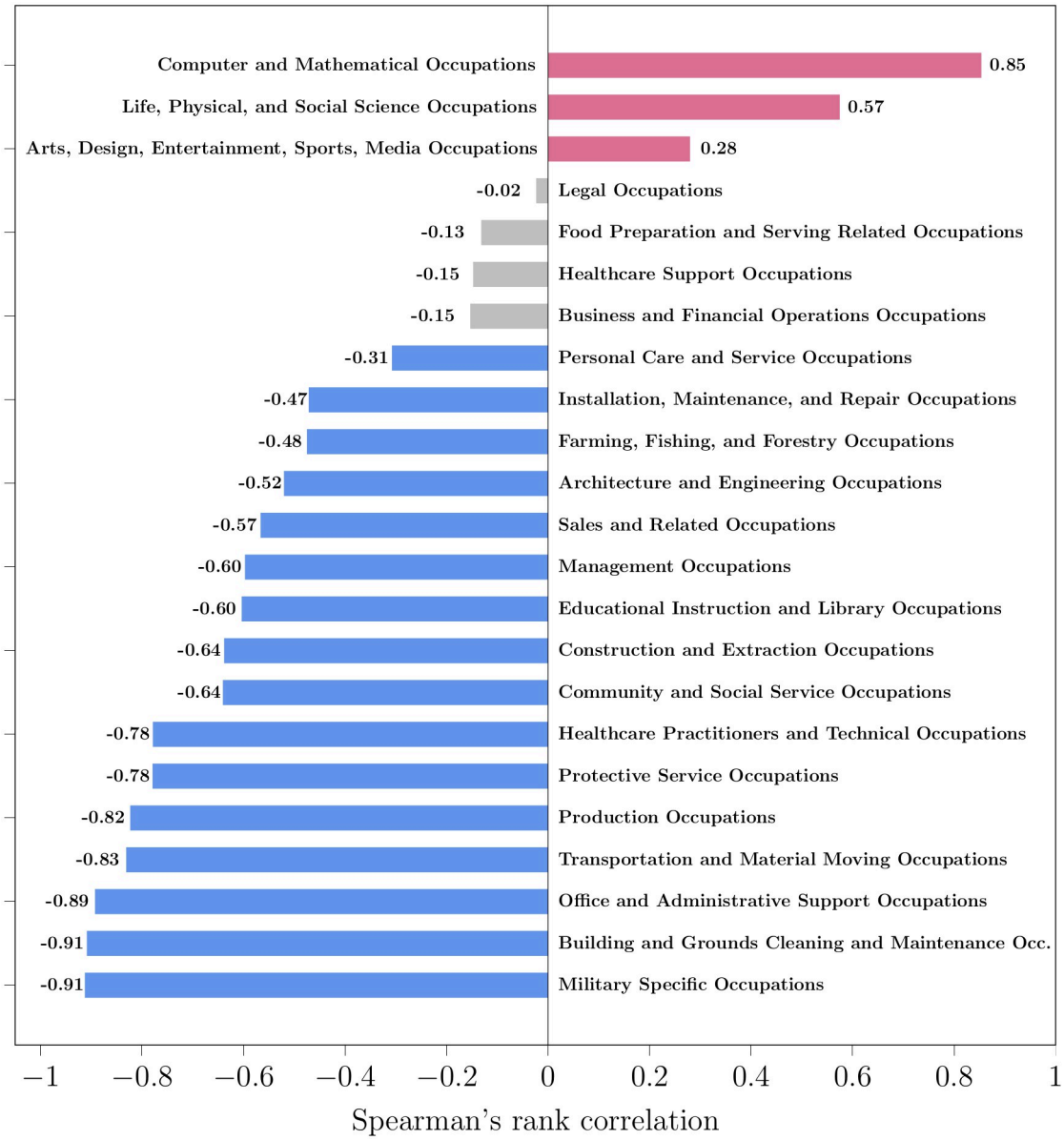

Fig 6. Spearman's rank correlation coefficient of the media frequency of SOC groups vs time. The red-colored bars have positive correlation (increasing trend), and the blue-colored bars have negative correlation (decreasing trend). The grey-colored bars mean that the correlation is not statistically significant ($\alpha = 0.05$).

https://doi.org/10.1371/journal.pone.0267812.g006

section Taxonomy Creation). This frequency measure is motivated by the Google-ngrams study [23]. We calculate the trend of a profession or SOC major group as the Spearman's rank correlation coefficient [74] of its media frequency against time. A significant positive correlation denotes an increasing trend, and a significant negative correlation implies a decreasing trend over time ($\alpha = 0.05$).

Fig 6 shows the trend of the 23 major SOC groups, from most positive to most negative. Table 3 lists these groups and their frequency trends. Only 3 SOC groups showed an increasing frequency trend in mentions over time, while 16 SOC groups decreased in frequency over time. The rank correlation was not significant for the remaining 4 SOC groups. These SOC

Table 3. SOC major groups with increasing, decreasing or no frequency trend over time.

Increasing Frequency	No Trend
Computer and Mathematical Occupations	Legal Occupations
Life, Physical, and Social Science Occupations	Food Preparation and Serving Related Occupations
Art, Design, Entertainment, Sports, Media Occupations	Healthcare Support Occupations
	Business and Financial Operations Occupations
Decreasing Frequency	
Personal Care and Service Occupations	Community and Social Service Occupations
Installation, Maintenance, and Repair Occupations	Healthcare Practitioners and Technical Occupations
Farming, Fishing, and Forestry Occupations	Protective Service Occupations
Architecture and Engineering Occupations	Production Occupations
Sales and Related Occupations	Transportation and Material Moving Occupations
Management Occupations	Office and Administrative Support Occupations
Educational Instruction and Library Occupations	Building, Grounds Cleaning and Maintenance Occupations
Construction and Extraction Occupations	Military Specific Occupations

A SOC major group has increasing or decreasing frequency trend if its Spearman's rank correlation is significantly positive or negative respectively.

https://doi.org/10.1371/journal.pone.0267812.t003

groups contain 500 professions (see section Taxonomy Creation). We analyze the trend of some professions belonging to these SOC groups.

Computer and Mathematical Occupations: This SOC group includes mathematicians and computer-related professions. Fig 7a) shows the frequency trend of two of its professions: hacker and programmer. Both occupations showed a positive frequency trend, but mentions of hackers increased more than programmers.

Life, Physical, and Social Science Occupations: This SOC group includes archaeologists, astronauts, biologists, chemists, geologists, etc. Almost all its professions showed an increasing frequency trend over time. Fig 7b) shows the trend of four occupations: geologist, biologist, anthropologist, and economist. Mentions of geologists and biologists are consistently more frequent than economists and anthropologists.

Arts, Design, Entertainment, Sports, and Media Occupations: This SOC group includes entertainment, sports, music, and media-related professions. Fig 7c), 7d), 7e) and 7f) show the frequency trends of some of its professions. Mentions of actor dominate actress mentions in media content. The word actor can be used as a gender-neutral term, explaining part of this trend. The frequency of sports-related professions dipped in the 1960s but has increased since then. Pianist mentions decreased, whereas mentions of bass players, guitarists, and drummers increased over time. Mentions of journalists and reporters are more frequent than correspondents and columnists.

Legal Occupations: Legal occupations include lawyers, judges, attorneys, prosecutors, etc. Fig 7g) shows the frequency trend of defense attorneys and prosecutors. Mentions of prosecutors (a lawyer who conducts a case against a defendant) are more frequent than defense attorneys (a lawyer who defends the client against criminal charges).

Food Preparation and Serving Related Occupations: This SOC group includes professions related to food serving and food preparation. Fig 7h), 7i) and 7j) show the frequency trends of some of its occupations. Mentions of waiters are more frequent than waitresses overall. Similar to actors, waiters can refer to either gender, so this trend is not surprising. However, mentions of waiters have decreased over time, whereas mentions of waitresses have

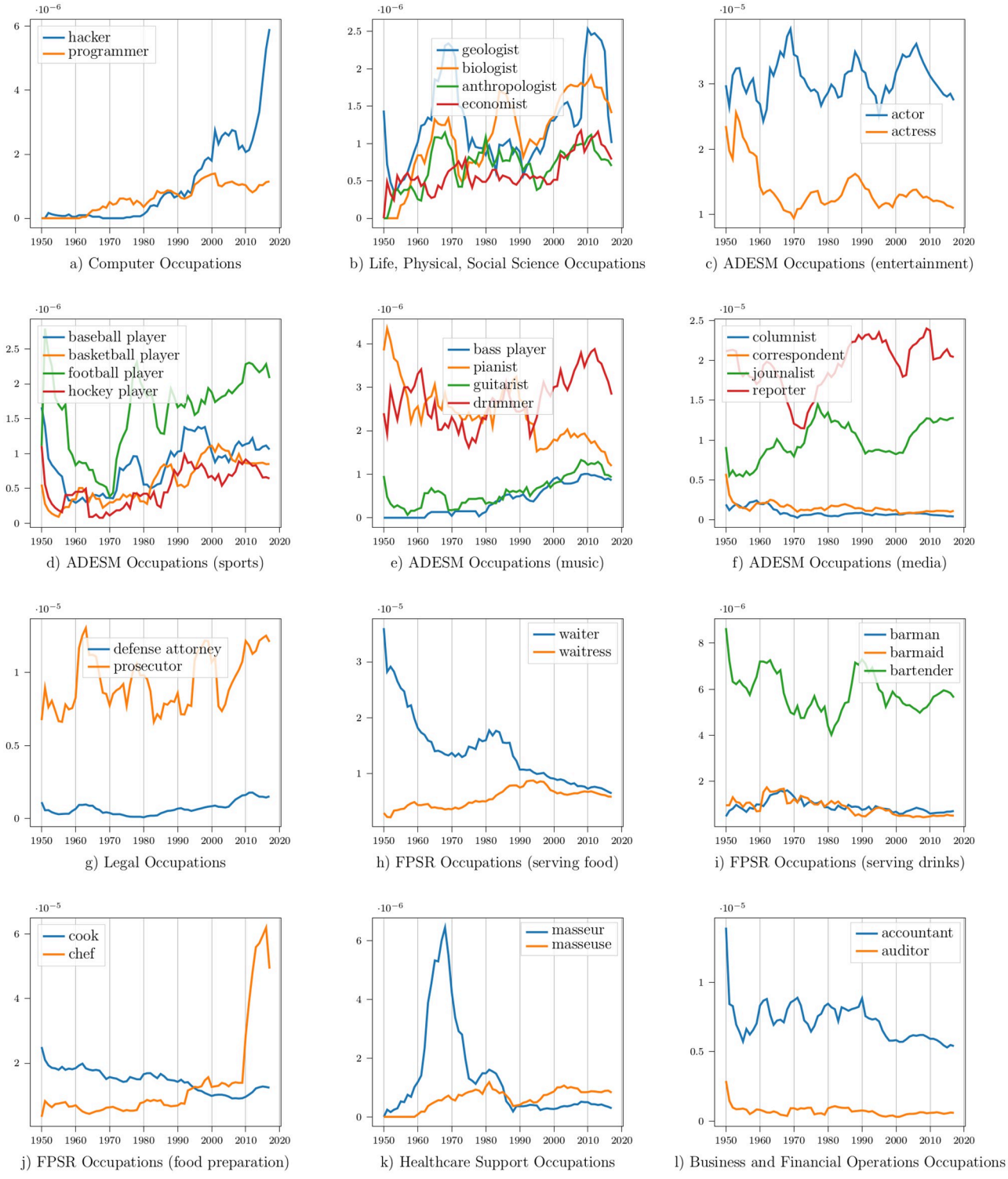

Fig 7. Frequency trends of different professions over time in media subtitles. ADESM = Arts, Design, Entertainment, Sports, and Media. FPSR = Food Preparation and Serving Related.

https://doi.org/10.1371/journal.pone.0267812.g007

increased. Gendered professional mentions like barman and barmaid are less frequent than the gender-neutral term: bartender. Mentions of cooks were more common than chefs, but the latter became more frequent in media content in the 2010s.

Healthcare Support Occupations: This SOC group includes healthcare assistants, nursing aides, massage therapists, etc. Fig 7k) shows the frequency trend of two gendered professions: masseur (male) and masseuse (female). The frequency of masseurs has significantly decreased over time after it peaked around 1970. Mentions of masseuses have become more common than masseurs. The frequency of the gender-neutral term, massage therapist, has increased over time.

Business and Financial Operations Occupations: This SOC group includes accountants, contractors, auditors, etc. Fig 7l) shows the frequency trend of accountants and auditors. Their frequencies have mostly remained steady over time. Accountants are more frequent than auditors.

Personal Care and Service Occupations: This SOC group includes barbers, valets, nannies, ushers, etc. Fig 8a) shows the frequency trend of three child-care-related professions: nanny, babysitter, and governess. The mention frequency of nannies and babysitters has increased over time. The word governess is an archaic term, and its usage has declined.

Installation, Maintenance, and Repair Occupations: This SOC group includes mechanics, electricians, locksmiths, etc. Fig 8b) shows the frequency trend of electricians and mechanics. The mentions of both these professions have decreased in media subtitles.

Farming, Fishing, and Forestry Occupations: This SOC group includes farmers, shepherds, herders, fishermen, etc. Fig 8c) shows the frequency trend of farmers, fishermen, and hunters. Mentions of farmers and fishermen have decreased, whereas mentions of hunters have increased over time.

Architecture and Engineering Occupations: This SOC group includes architects, designers, engineers, surveyors, etc. Fig 8d) shows the frequency trend of engineers and architects. The frequency of architect mentions has remained steady over time, whereas mentions of engineers have diminished.

Sales and Related Occupations: This SOC group includes sales and real-estate-related professions. Fig 8e) and 8f) show the frequency trends of some of its occupations. Mentions of retailers, distributors, vendors, and brokers have increased over time. The increase in frequency is even more prevalent in real-estate jobs: estate agent, realtor, and real estate agent.

Management Occupations: This SOC group includes administrative professions and occupations related to the management of educational institutions. Fig 9g) and 9h) show the frequency trend of some management occupations. Mentions of both congressman and congresswoman have increased over time. Although congressmen are mentioned more than congresswomen, the frequency of congresswomen has increased at a higher rate (not discernible from the graph). Mentions of senators, governors, and mayors have decreased over time. The frequency of headmistress mentions has increased, whereas headmaster mentions have decreased in media content. The frequency of the gender-neutral term, principal, has increased.

Educational Instruction and Library Occupations: This SOC group includes teachers, professors, librarians, etc. Fig 8i) shows their frequency trends. Professor mentions have decreased greatly over time, whereas the frequency of teacher and librarian mentions have increased slightly.

Construction and Extraction Occupations: This SOC group includes builders, carpenters, masons, miners, etc. Fig 8j) shows their frequency trends. The frequency of masons has remained steady, but miner and carpenter mentions have reduced over time.

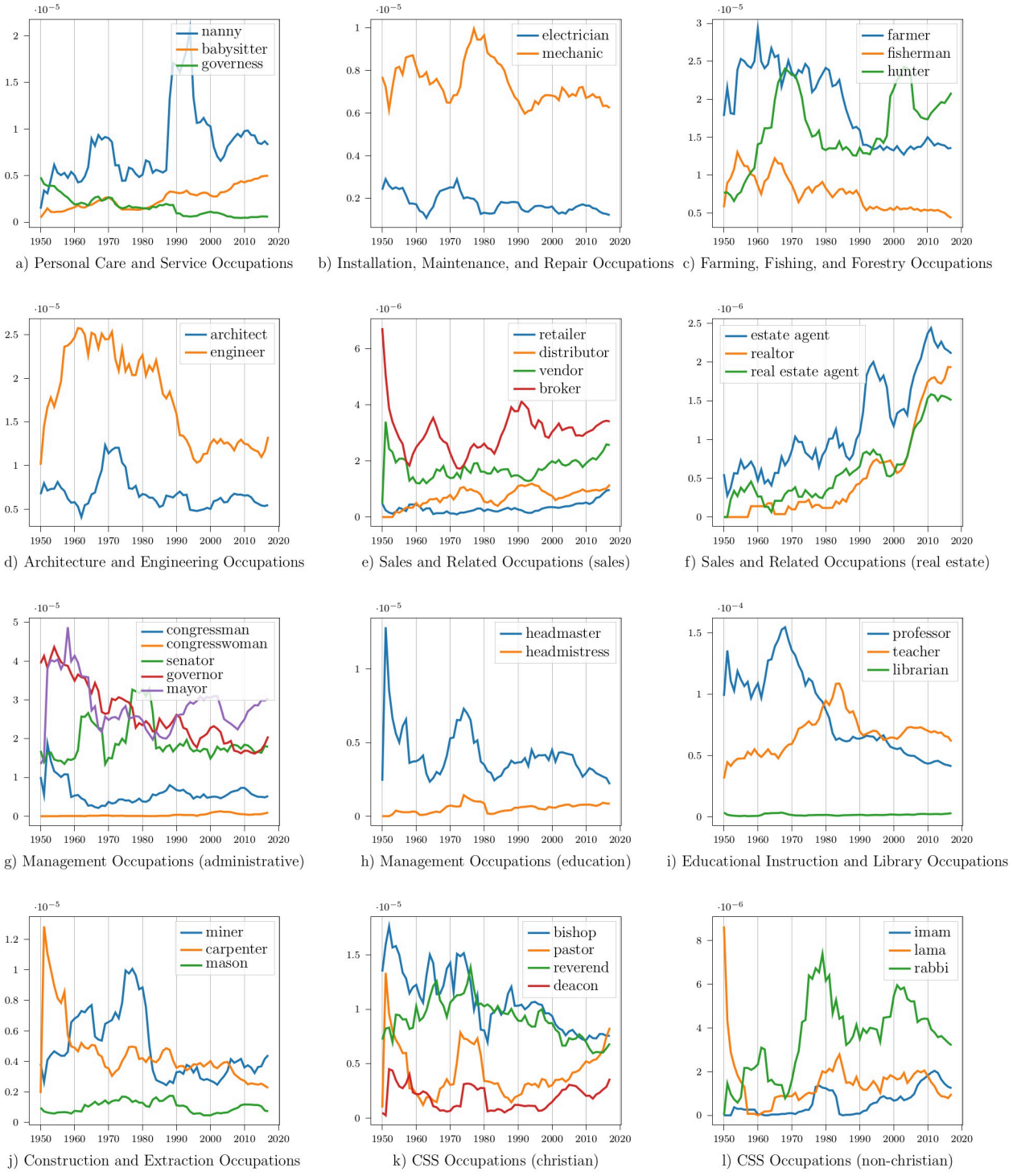

Fig 8. Frequency trends of different professions over time in media subtitles. CSS = Community and Social Service.

https://doi.org/10.1371/journal.pone.0267812.g008

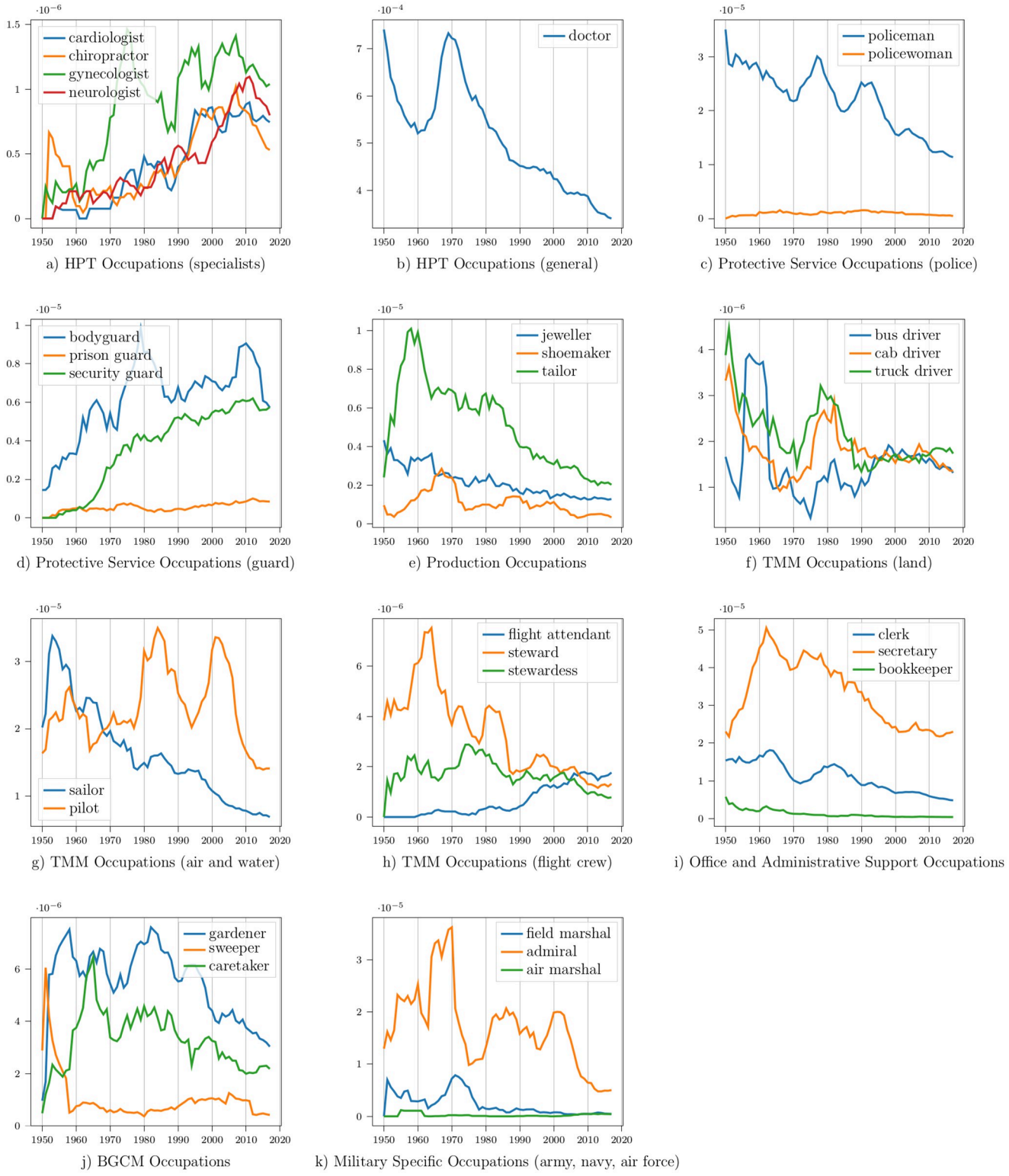

Fig 9. Frequency trends of different professions over time in media subtitles. HPT = Healthcare Practitioners and Technical. TMM = Transportation and Material Moving. BGCM = Building and Grounds Cleaning and Maintenance.

https://doi.org/10.1371/journal.pone.0267812.g009

Community and Social Service Occupations: This SOC group includes religious and social workers. Fig 8k) and 8l) shows the frequency trend of some religious occupations. The mention frequency of bishops and reverends has decreased over time, whereas the frequency of pastors and deacons has increased. Mentions of imams, lamas, and rabbis have also increased over time, but the frequency of priests has decreased.

Healthcare Practitioners and Technical Occupations: This SOC group includes healthcare-related professions. Fig 9a) and 9b) show the frequency trend of some of its professions. Mentions of healthcare specialists like cardiologists, chiropractors, gynecologists, and neurologists have increased over time, but the frequency of the generic term, doctor, has decreased.

Protective Service Occupations: This SOC group includes law enforcement and protective service workers. Fig 9c) and 9d) show the frequency trend of some of its professions. Mentions of policemen have decreased over time but are still much higher than mentions of policewomen, which have remained steady. Guard occupations like bodyguards, prison guards, and security guards have become more frequent in media content.

Production Occupations: Production occupations include artisans, cobblers, jewelers, millers, etc. Fig 9e) shows the frequency trend of some of its professions: jeweler, shoemaker, and tailor. The mention frequency of all three professions has decreased over time.

Transportation and Material Moving Occupations: This SOC group includes professions related to the transportation of goods and people. Fig 9f), 9g) and 9h) show frequency trends of different transportation-related professions. Mentions of bus drivers increased, cab drivers remained steady, and truck drivers decreased over time. Mentions of sailors decreased more rapidly than pilots. The frequency of gendered professional terms like steward and stewardess decreased over time. The frequency of the gender-neutral term, flight attendant, increased.

Office and Administrative Support Occupations: This SOC group includes clerks, receptionists, tellers, notaries, etc. Fig 9i) shows the frequency trend of some of its professions: clerk, secretary, and bookkeeper. The frequency of all three professions decreased over time.

Building, Grounds Cleaning and Maintenance Occupations: This SOC group includes construction workers, maids, chamberlains, janitors, etc. Fig 9j) shows the frequency trend of some of its professions: gardener, sweeper, and caretaker. Mentions of all three professions decreased over time in media content.

Military Specific occupations: Military occupations include army, naval and air-force-related professions. Fig 9k) shows the frequency trend of some of its professions. Mentions of field marshal (army) and admiral (navy) decreased over time. The frequency of air marshal mentions increased (not discernible from the graph).

The frequency trend of the SOC group does not reflect the frequency trend of all its professions. For example, Fig 8e) and 8f) show increasing trends for many sales-related professions, but the overall frequency of the Sales Occupations SOC group decreased. A large proportion of Sales Occupations mentions are comprised of bankers and cashiers, whose frequencies have decreased.

Profession sentiment

We find the sentiment expressed toward each professional mention in our dataset using the LCF model (see section Profession Search). The computed sentiment can take the following values: positive, negative, or neutral. We call all mentions tagged with non-neutral sentiment as opinionated mentions. We represent the sentiment expressed towards a profession or a major SOC group as the number of positive sentiment mentions divided by the total number of opinionated mentions. We find the sentiment trend of professions by calculating the Spearman's rank correlation [74] between the proportion of positive sentiment mentions and time.

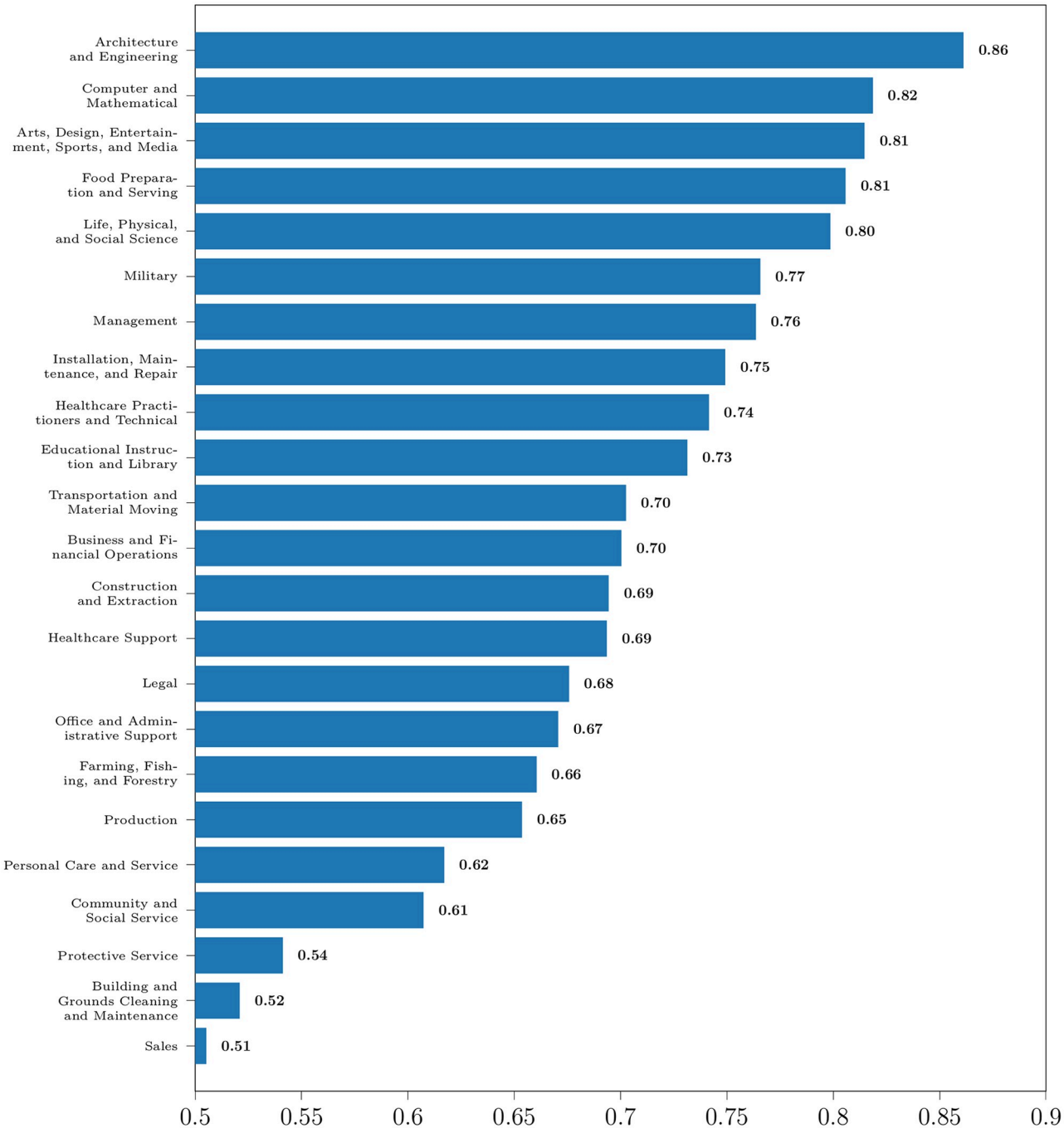

Fig 10. Proportion of positive sentiment mentions in opinionated mentions of the SOC groups.

https://doi.org/10.1371/journal.pone.0267812.g010

Fig 10 shows the proportion of positive sentiment mentions of the 23 major SOC groups. From the figure, we observe that the proportion is always greater than 0.5. Therefore, the number of positive sentiment mentions is greater than the number of negative sentiment mentions for all the SOC groups. This might be because of an inherent positive bias in the sentiment of media narratives. Architects and engineers are talked about most positively, and sales-related

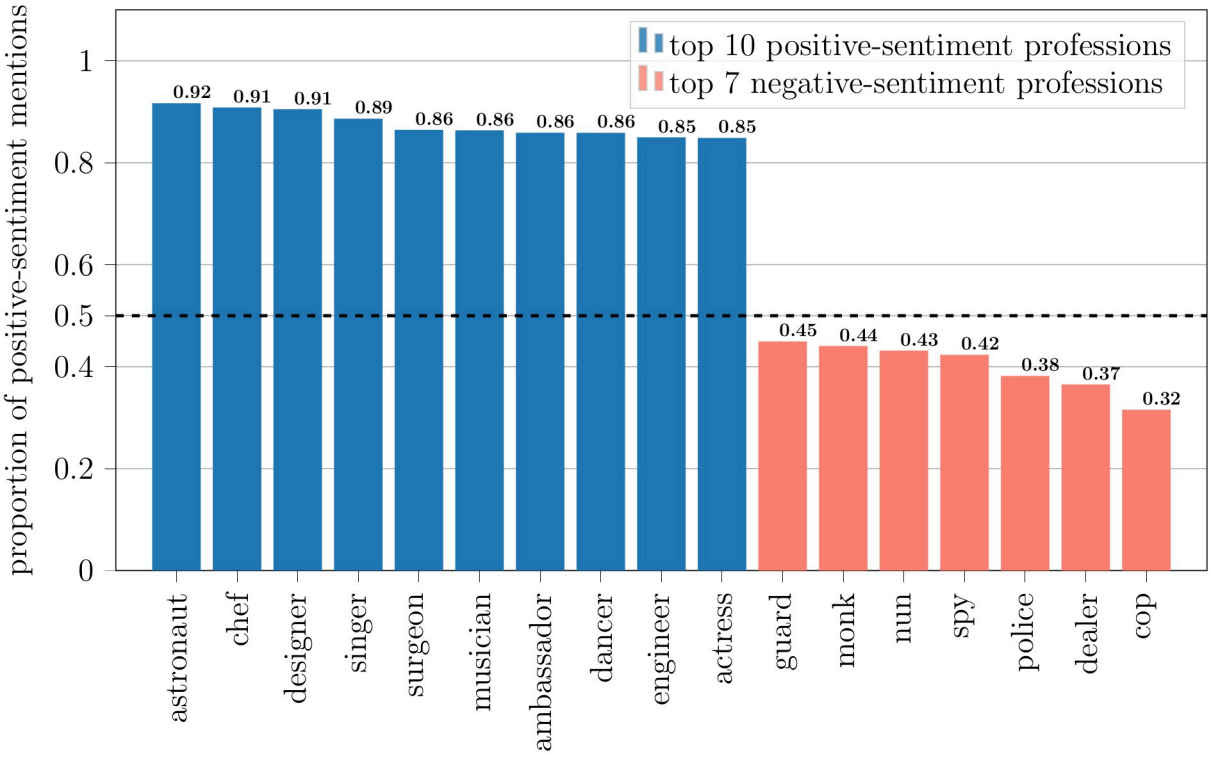

Fig 11. Top 10 positive-sentiment and top 7 negative-sentiment professions. The blue-colored professions have the top 10 highest proportion of positive sentiment mentions. The red-colored professions have the top 10 highest proportion of negative sentiment mentions.

https://doi.org/10.1371/journal.pone.0267812.g011

occupations are talked about most negatively. STEM professions are generally expressed in positive sentiment. Four out of the top ten SOC groups ordered by the proportion of positive sentiment contain STEM professions: Architecture and Engineering; Computer and Mathematical; Life, Physical, and Social Science; and Healthcare Practitioners and Technical Occupations. Professions involving manual labor are generally expressed with negative sentiment. Four out of the bottom ten SOC groups contain blue-collar jobs: Building, Grounds Cleaning and Maintenance; Production; Farming, Fishing, and Forestry; and Construction and Extraction Occupations.

We analyze the proportion of positive sentiment mentions for individual professions. Fig 11 shows the top ten professions with the highest proportion of positive sentiment mentions and the bottom seven professions with the lowest proportion of positive sentiment mentions from the list of most occurring top 100 professions in our subtitle corpus. More than half of the opinionated mentions of police and cops are negative. Religious workers like monks and nuns are talked about more negatively than positively. More than four-fifths of the mentions of singers, musicians, and dancers are positive. STEM professions like astronauts and engineers are also mentioned positively in media content.

We also study the sentiment trend of professions in the analyzed media content. Fig 12 shows the Spearman's rank correlation coefficient of the proportion of positive-sentiment mentions over time for some of the professions. The figure lists the top six professions with the highest correlation and the bottom ten professions with the lowest correlation from the list of most occurring top 100 professions in our subtitle corpus. The sentiment expressed towards therapists, spies, rangers, and detectives in media content trends toward becoming more

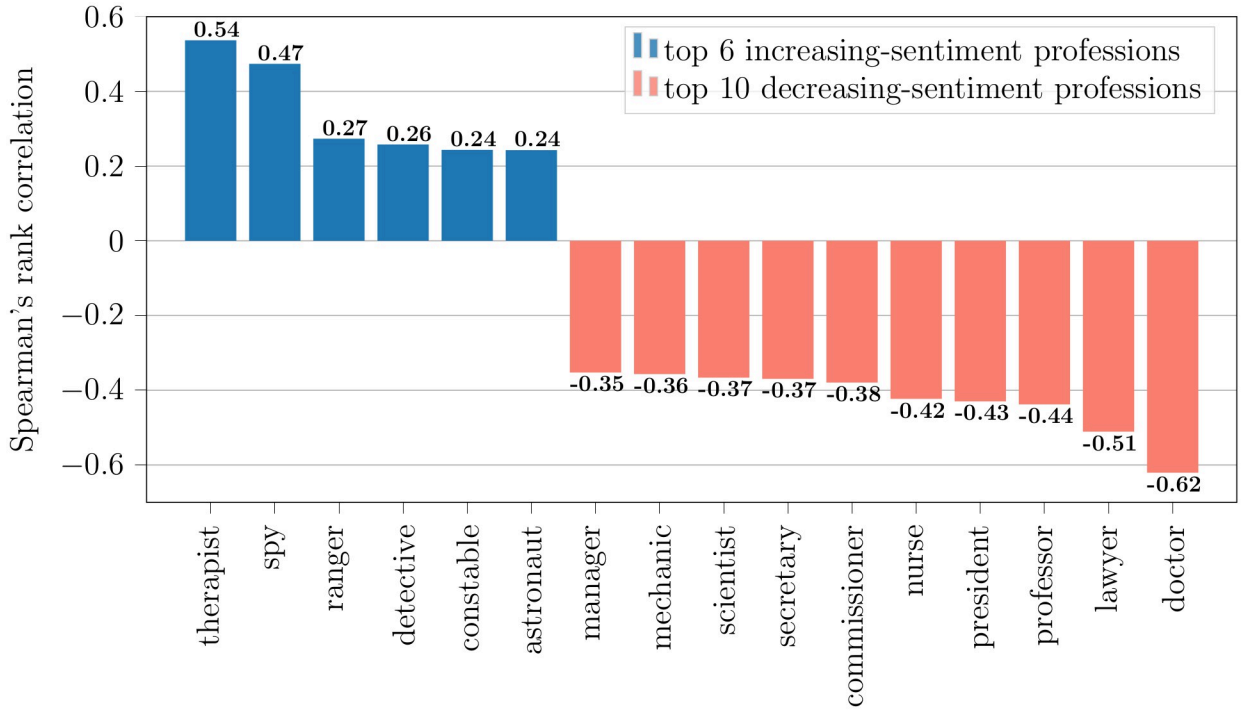

Fig 12. Professions with the top increasing and top decreasing sentiment trends. The blue-colored professions have the top 6 most positive Spearman's rank correlation of proportion of positive mentions vs time. Therefore, the proportion of positive sentiment mentions shows increasing trend over time for these professions. The red-colored professions have the top 10 most negative correlation and their proportion of negative sentiment mentions is increasing over time.

https://doi.org/10.1371/journal.pone.0267812.g012

positive over time, whereas the sentiment expressed toward doctors, lawyers, professors, and scientists trends more negative. Lawyer mentions have the second-lowest sentiment correlation over time, behind doctors. Astronauts not only have the highest proportion of positive sentiment mentions overall (see Fig 11), but also one of the most positive sentiment trends.

Media attributes

We have observed that both the frequency of professional mentions and the sentiment expressed towards them change over time. In this section, we study the relationship of the following media attributes: year, genre, title type, and country of production (see section OpenSubtitles) with the observed media frequency and sentiment trends of professions and SOC major groups.

We perform a regression analysis to analyze the effect of media attributes. The response is the frequency or sentiment of the profession or the SOC major group. The predictors are year, genre, title type, and country of production. All predictors are categorical variables except for year, which is numeric. Genre and country are multi-valued. An IMDb title can belong to multiple genres, and its production can take place in multiple countries. Therefore, we create a categorical variable for each possible genre and country, taking values 0 or 1. We set it to 1 if the IMDb title belongs to the corresponding genre or its production takes place in the corresponding country. We ignore media attribute configurations for which the number of IMDb titles is less than 30. We use logistic regression because both frequency and sentiment are proportions, bounded between 0 and 1 (defined in sections Profession Frequency and Profession

Sentiment). We use a generalized linear model with the binomial family and logit link. We provide the total number of ngrams and opinionated mentions (defined in section Profession Sentiment) as prior weights to the model to specify the number of trials when the response is frequency or sentiment, respectively.

Figs 13 and 14 show the significant predictors ($\alpha = 0.05$) when the response is the frequency of professions and SOC groups, respectively, in the form of a heatmap. The color of the cells indicates the sign of the coefficient (blue = +, red = −) and the intensity of the color denotes its magnitude. The white cells mean that the predictor is not significant. We observe some interesting relationships between the frequency of professional mentions and media attributes. The frequency of actors increases, but the frequency of actresses decreases when the genre is adventure, documentary, or thriller. The reverse is true when the genre is romance. Mentions of lawyers increase in crime, drama, and mystery genre media content. The frequency of lawyers and attorneys increase, and the frequency of prosecutors decreases when the country of production is the United States. United States-produced movies and TV shows mention cops and sheriffs more than inspectors and police. The opposite is true for United Kingdom-produced titles. The frequency of detectives and spies increases in mystery genre titles. Comedy, reality-TV, and music genres increase the frequency of dancers, singers, and artists. The frequency of doctors, nurses, and surgeons in movies is higher than in TV shows. In science-fiction and family media titles, the frequency of doctors increases, and the frequency of nurses and surgeons decreases. Documentary titles increase the frequency of reporters and journalists. Mentions of engineers and scientists increase when the genre is science-fiction or documentary and decreases in comedy and fantasy genres. The frequency of teachers and professors decreases in action and adventure genres. Movies mention teachers more than TV shows, but the opposite holds for professors. Mentions of senators, mayors, and presidents increases in news and thriller genres. The frequency of lieutenants and soldiers increases in action and war media titles and decreases when the genre is fantasy and romance.

Fig 14 shows the coefficient heatmap of media attributes when the response is the frequency of SOC groups. We highlight some relationships between SOC frequency and media attributes. The frequency of Management, Business, and Financial Operations occupations increases in biography and news genre media titles. Documentaries and science-fiction movies and TV shows frequently mention STEM professions like Computer, Mathematical, Architecture, Engineering, Life, Physical, and Social Science occupations. Mentions of Community and Social Service occupations decrease in reality TV shows, sports, and family movies. The frequency of Legal occupations increases in news genre titles. The frequency of Arts, Design, Entertainment, Sports and Media occupations increases in music, sports, game show, and biography genre titles. Mentions of Healthcare Practitioners increase in news and drama genres and decrease in musicals. Food Preparation and Serving related professions occur highly in reality TV shows and decrease in music and adventure movies. The frequency of manual labor jobs like Construction, Extraction, Production, Building, Grounds Cleaning, and Maintenance occupations decreases when the country of production is the United States. Mentions of Military occupations increase in war and action movies and decrease in comedy and family shows.

We study the effect of media attributes on the proportion of positive sentiment mentions of professions and SOC Groups. Unlike frequency, the number of significant predictors are fewer. Therefore, instead of a heatmap, we tabulate the significant media attributes with the most positive and negative coefficients ($\alpha = 0.05$) in Table 4. The proportion of positive sentiments expressed towards marshalls, mayors, professors, and prosecutors increases when the genre is mystery. The proportion of negative sentiment increases for congressmen, priests, and prosecutors in crime genre media titles. The average sentiment of SOC groups containing STEM occupations like Computer, Mathematical, Life, Physical, and Social Science

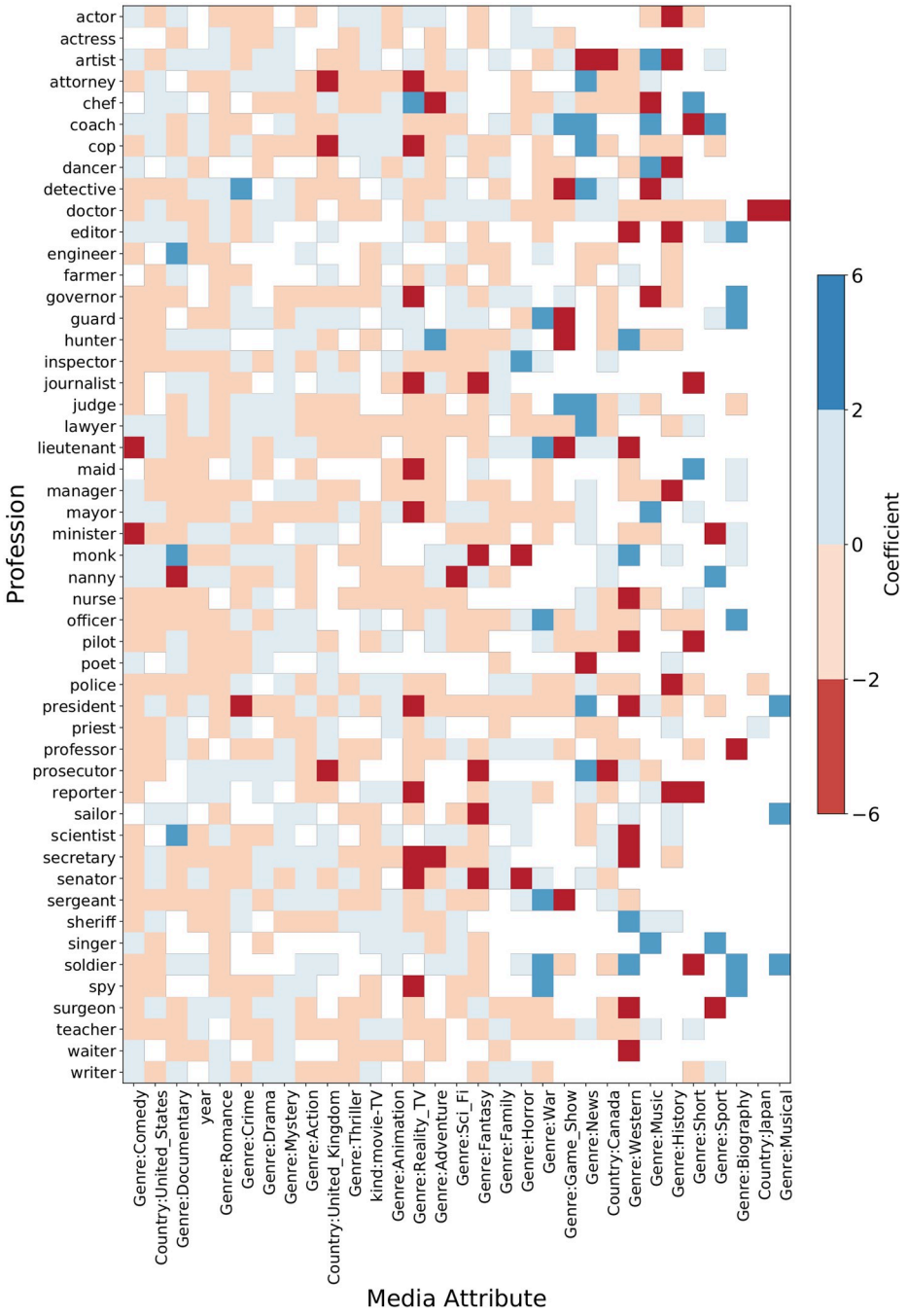

Fig 13. Heatmap of coefficient values of media attributes in predicting profession frequency. The color denotes the sign of the coefficient (blue = +, red = −) and the intensity of the color is proportional to the magnitude of the coefficient. The blank cells indicate that the media attribute is not a significant predictor for the frequency of the corresponding profession.

https://doi.org/10.1371/journal.pone.0267812.g013

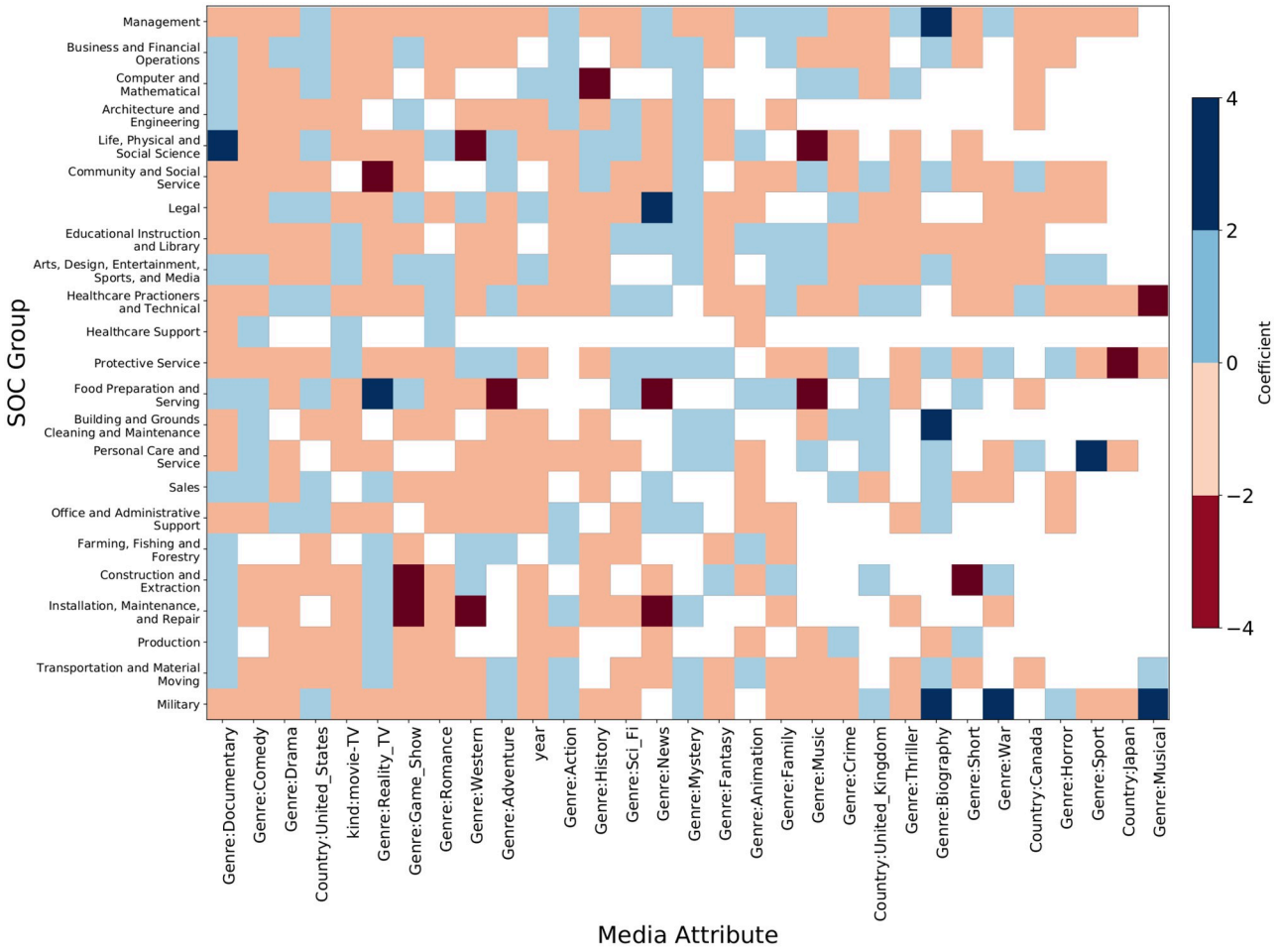

Fig 14. Heatmap of coefficient values of media attributes in predicting SOC group frequency. The color denotes the sign of the coefficient (blue = +, red = −) and the intensity of the color is proportional to the magnitude of the coefficient. The blank cells indicate that the media attribute is not a significant predictor for the frequency of the corresponding SOC group.

https://doi.org/10.1371/journal.pone.0267812.g014

occupations becomes more positive in documentaries. Movies express more negative sentiment than TV shows towards Legal occupations.

Employment

We analyzed the frequency and sentiment trends of different professions and SOC groups in media content. We also studied the effect of media attributes on these representations. However, our analysis has been limited largely to the entertainment media domain, by using NLP on the subtitles of media content. In this section, we study the correlation between media frequency and real-world events, namely employment figures. We obtained the employment data of SOC major groups from the Occupational Employment Statistics survey (OES). The survey does not provide employment numbers for individual professions. Therefore, we only conduct our analysis on SOC groups. We calculate the Spearman's rank correlation [74] between the media frequency of the SOC group and the proportion of the working population employed in any of the professions of the SOC group. We compute the correlation for the period 1999-2017. The employment data for the earlier years is not available.

Table 4. Effect of media attributes on the sentiment of professions and SOC groups.

Profession / SOC Group	Media Attributes	
	Positive Coefficients	**Negative Coefficients**
actor		Genre:Drama
actress		Country:United Kingdom
architect	Genre:Documentary	
artist	Genre:Game Show	Genre:Fantasy
banker		kind:movie-TV
chef	Genre:Game Show	Country:United Kingdom
congressman		Genre:Crime
cop	Genre:Adventure	Genre:Family
dealer	Country:United States	kind:movie-TV
detective	Genre:Crime	
district attorney		kind:movie-TV
doctor	Genre:Action	Genre:Horror
engineer	Genre:Documentary	kind:movie-TV
judge		kind:movie-TV
lawyer	Genre:Romance	Genre:Thriller
manager	Genre:Family	
marshall	Genre:Mystery	
mayor	Genre:Mystery	
musician		Genre:Comedy
nurse		Genre:Documentary
officer	Country:United States	Genre:Fantasy
police	Genre:Animation	Genre:News
priest	Genre:Thriller	Genre:Crime
professor	Genre:Mystery	Genre:Adventure
prosecutor	Genre:Mystery	Genre:Crime
scientist		Genre:Reality TV
secretary	Genre:Action	
sheriff		Genre:Action
social worker	Country:United Kingdom	
soldier	Genre:Western	
spy	Genre:Animation	Country:United Kingdom
Architecture and Engineering	Genre:Documentary	kind:movie-TV
Arts, Design, Entertainment, Sports, and Media	Genre:Game Show	Genre:Crime
Building and Grounds Cleaning and Maintenance	Genre:Documentary	Genre:Crime
Business and Financial Operations	Genre:Adventure	Genre:News
Community and Social Service	Genre:Mystery	Genre:Family
Computer and Mathematical	Genre:Documentary	
Construction and Extraction		Genre:Sci Fi
Educational Instruction and Library	Genre:Documentary	Genre:Sport
Farming, Fishing and Forestry		Genre:Western
Food Preparation and Serving	Genre:Documentary	Genre:Crime
Healthcare Practitioners and Technical	Genre:Action	Genre:Horror
Legal	Genre:Romance	kind:movie-TV
Life, Physical and Social Science	Genre:Documentary	Genre:Reality TV
Management	Genre:War	Genre:Crime
Military	Genre:Western	Genre:Comedy

(*Continued*)

Table 4. (Continued)

Profession / SOC Group	Media Attributes	
	Positive Coefficients	**Negative Coefficients**
Personal Care and Service	Genre:Documentary	Genre:Animation
Production	Genre:Action	Genre:Fantasy
Protective Service	Genre:Reality TV	Genre:News
Sales	Genre:Horror	Genre:Crime
Transportation and Material Moving		Genre:History

This table lists the media attributes with the highest significant positive and negative coefficients in predicting the sentiment of professions and SOC major groups.

https://doi.org/10.1371/journal.pone.0267812.t004

Fig 15 shows the correlation between media frequency and employment for the SOC groups, from most positive to most negative. The correlation is positive for 14 out of the 22 SOC groups (64%) and negative for the rest. Therefore, the trend of media frequency correlates with the employment trend for most SOC groups. The number of media mentions of the SOC group increases as more people are employed in its professions, and decreases as people move away to other occupations. The correlation is significantly negative for only two SOC groups: Healthcare Practitioners and Social Service occupations.

Discussion

The frequency and sentiment expressed towards professionals vary in media content depending on the character archetype, the story being told, year of production, and genre, among other variables. We developed a visualization tool [75] to examine the frequency, sentiment, media attribute, and employment trends in our subtitle corpus. While examining the wealth of results using this tool, a few general findings and specific trends stood out to us. In this section, we highlight these trends, discuss some reasons that may be suggestive for explaining the observed results, and try to reconcile these findings with those in relevant studies, both from the real world and the media domain. We underscore that our interpretation is purely speculative and requires careful, controlled experiments and surveys to further validate the claims. Also, this discussion is by no means an effort to explain all our findings, and we hope future research in this space may further explain some of the results we found in this large-scale analysis.

Findings from frequency analysis

We observed that gender-neutral terms like massage therapists and flight-attendants are becoming more frequent than their gendered counterparts. We suggest that this trend is due to the increased awareness stemming from conversations about gender-neutral terms among the youth, parents, and the LGBTQ+ community [76, 77]. Analyzing explicit gendered profession terms, we observed that the frequency of some female job titles such as waitresses, congresswomen, and policewomen has either increased or remained steady relative to the corresponding male job titles (i.e., waiters, congressmen, and policemen). Equal opportunity legislation, women's rights movements [78], and constitutional amendments for women's suffrage [79] have improved the representation of women in police and government. This may have driven a similar trend of increasing female representation in media content. While there is a change in trend between the explicitly gendered job titles over the years, the overall frequency of most male job titles exceeds their female counterparts.

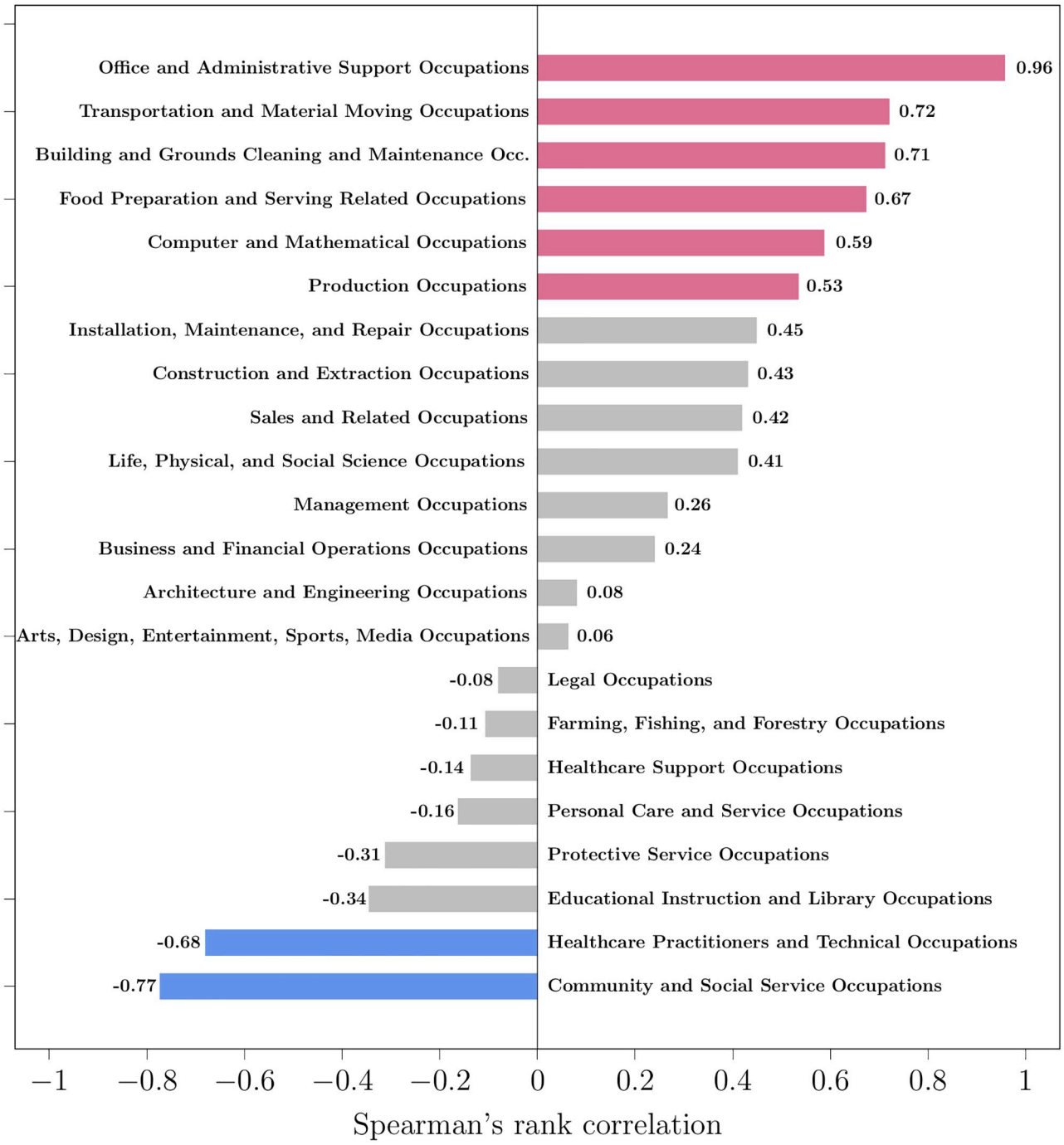

Fig 15. Spearman's rank correlation coefficient of the media frequency of SOC groups with employment. The red-colored bars have positive correlation: media frequency and employment have the same trend. The blue-colored bars have negative correlation: media frequency and employment have opposite trends. The grey-colored bars mean that the correlation is not statistically significant ($\alpha = 0.05$).

https://doi.org/10.1371/journal.pone.0267812.g015

The frequency of STEM, sports, arts, and design occupations has increased, whereas the frequency of construction, farming, and manual labor jobs has decreased. Improvements in cinematography and visual effects technology, space travel, the emergence of the world wide web, and sports documentaries have contributed to the increased fascination of science and sports

in media [80, 81]. These advances and interests potentially explain the increase in the frequency of STEM-related and sports portrayals on the big screen. The frequency of specialized professions like cardiologists, gynecologists, and neurologists has increased, but the frequency of generic terms like doctors and nurses has decreased. The evolution of medicine and research, students' preference for higher-paying specialties [82], and the emergence of novel diseases have increased the demand for specialized professions over general ones in the medical field [83]. We posit that as new and specialized professions become popular career choices and essential in catering to the needs of the continuously evolving society, they also attract the interest of media creators.

Thus far, we have discussed the frequency trends of individual professions. It is important to note that this does not always reflect the overall trend of the subsuming SOC group. Sales-related occupations like retailers, vendors, brokers, and realtors show an increase in frequency, but the aggregate frequency of the associated SOC group shows a decreasing trend. Technology improvements in computerized financial transactions (barcode scanning, automatic teller machines, and mobile banking) and customers' preference for self-service checkouts probably contributed to the decrease in employment of bankers and cashiers [84, 85], or at least their explicit mentions in media narratives.

Findings from sentiment analysis

STEM occupations are favorably (i.e., positive sentiment) mentioned in the subtitles corpus we analyzed. However, the sentiment expressed toward professions that predominantly involve manual labor—the so-called "blue-collar" jobs—is largely negative. Overall, police, monks, and nuns are mentioned negatively, whereas musicians and engineers are mentioned positively.

The sentiment expressed toward therapists, astronauts, and detectives was observed to trend more positive over time, while the sentiment toward doctors, lawyers, professors, and presidents, was more negative. The decreasing trend of positive sentiment toward lawyers is consistent with Asimow's work [16] that showed the decline in public opinion of lawyers. This may be one of the explanatory factors behind the increasingly negative portrayal of lawyers in movies. In contrast, the negative sentiment expressed towards doctors, scientists, and police that we found in our media corpus does not appear to agree with the findings of public polls. Several studies over the years [86–88] show that people usually trust these professions. We note that public opinion also depends on several social and demographic factors such as political leanings and the level of educational attainment [87, 88]. Factors like the popular stereotype of the "mad scientist" (someone who is more concerned about their scientific findings than human welfare) [18] in movies, grievances of the health care system [89], and general distrust towards people and institutions in positions of authority, might have contributed towards these negative portrayals in media.

We have made several hypotheses to explain the observed frequency and sentiment trends and require controlled experiments to verify these claims. It is challenging to design such studies because we are trying to relate media narratives to real-world events—two very different domains. As of now, we are unsure regarding what types of experiments are required, but we suggest some ideas. We can survey scriptwriters and movie directors to find what societal events inspired their stories and creative decisions. We can also try to map change points in media trends to real-world events to explain their possible cascading effects on media narratives and genres. We leave this task for future research.

Findings from media attribute analysis

Media attributes like genre, country of production, and content type affect the frequency and sentiment trends of professions in our subtitle corpus. Genre, in particular, seems to be a good predictor of the type of profession mentioned in the subtitles. For example, science-fiction movies often mention engineers and scientists, action and war genres mention lieutenants and soldiers, and mystery titles contain detectives and spies. Adventure/thrillers contain more references to male actors than female actors. However, the opposite is true for romantic movies. Consistent with findings in previous work [90], this suggests that gender bias is prevalent in these media genres.

In examining the country of production, we found that sheriffs are mentioned more than inspectors in titles produced in the US, and the opposite holds for movies produced in the UK, reflecting the differences in English usage, and law enforcement structures in these countries. Regarding the type of content, we did not find significant differences in the mentions of job titles between movies and TV shows in our subtitle corpus.

Findings from employment analysis

Interestingly frequency of job titles in media correlates with real-world employment statistics of corresponding professions. We observed a significant positive correlation between media frequency and employment trends of most SOC groups. Professions that typically employ more people were also more frequently mentioned in media content. We note that these correlation findings do not imply causality; a question that requires further systematic study.

Limitations

There are several limitations of the present work. First, subtitles provide only a limited view of the complete movie or TV show. Features, such as the character's behavior on screen, their interactions with other people, and their character arc can only be studied by carefully viewing the entire movie or TV episode. Therefore, our representation measures of frequency and sentiment of job titles in subtitles do not capture all aspects of their media portrayal. Second, our frequency and sentiment analyses do not control for the frequency or sentiment expressed toward these professions in everyday language. For example, the frequency of policewomen might have been less than the frequency of policemen in media subtitles because that is how their usage evolved in literature. A possible solution is to use the corresponding Google ngram frequencies [23] of the professional words as covariates in our analysis. Third, our study has primarily focused on the representation of professions in the US and the UK (see the distribution of production country in section OpenSubtitles). One needs to adapt our proposed taxonomy and models for different cultures and languages. Our profession taxonomy does not contain all job titles in other countries. Depending on the application of this taxonomy, one may have to add professions specific to a local region for a comprehensive study.

Conclusion

In this work, we have created a searchable taxonomy of professions to facilitate job title search in short context documents like media subtitles. We used WordNet synsets and word sense disambiguation methods to retrieve professional mentions in movie and TV show subtitles. We automatically classified the sentiment (positive, negative, or neutral) expressed towards these professional mentions in the subtitle sentence. We analyzed the frequency and sentiment trends of professions and SOC groups, the effect of media attributes on these trends, and showed that media frequency of professions correlates with their employment statistics. Future

work entails extending our analysis to include industries and businesses, and to explore other media domains like news and social media. Importantly, future work should also consider investigating causal relations, beyond correlations, between media representations and employment trends.

The profession taxonomy and sentiment-annotated subtitle corpus is publicly available [26]. We have also released a visualization tool to explore and view the trends in our subtitle corpus [75]. We hope to add more features and better infographics to the tool in the future.

Author Contributions

Conceptualization: Sabyasachee Baruah, Krishna Somandepalli.

Data curation: Sabyasachee Baruah.

Formal analysis: Sabyasachee Baruah.

Funding acquisition: Shrikanth Narayanan.

Investigation: Sabyasachee Baruah.

Methodology: Sabyasachee Baruah, Krishna Somandepalli.

Project administration: Sabyasachee Baruah.

Software: Sabyasachee Baruah.

Supervision: Shrikanth Narayanan.

Validation: Sabyasachee Baruah, Krishna Somandepalli.

Visualization: Sabyasachee Baruah.

Writing – original draft: Sabyasachee Baruah.

Writing – review & editing: Sabyasachee Baruah, Krishna Somandepalli, Shrikanth Narayanan.

References

1. Gerbner G, Gross L. Living with television: The violence profile. Journal of communication. 1976; 26 (2):172–199. https://doi.org/10.1111/j.1460-2466.1976.tb01397.x PMID: 932235
2. Hawkins RP, Pingree S. Television's influence on social reality. Television and behavior: Ten years of scientific progress and implications for the eighties. 1982; 2:224–247.
3. Potter WJ. Cultivation theory and research: A conceptual critique. Human Communication Research. 1993; 19(4):564–601. https://doi.org/10.1111/j.1468-2958.1993.tb00313.x
4. Morgan M, Signorielli N. Cultivation analysis: Conceptualization and methodology. Cultivation analysis: New directions in media effects research. 1990; p. 13–34.
5. Hughes M. The Fruits of Cultivation Analysis: A Reexamination of Some Effects of Television Watching. The Public Opinion Quarterly. 1980; 44(3):287–302. https://doi.org/10.1086/268597
6. Chandler D. Cultivation Theory; 1995. http://visual-memory.co.uk/daniel//Documents/short/cultiv.html.
7. Newcomb H. Assessing the Violence Profile Studies of Gerbner and Gross: A Humanistic Critique and Suggestion. Communication Research. 1978; 5(3):264–282. https://doi.org/10.1177/009365027800500303
8. Motion Picture Association of America. Theme Report: A comprehensive analysis and survey of the theatrical and home/mobile entertainment market environment for 2019; 2019.
9. Statista. Average time spent with major media per day in the United States as of April 2020, by format; 2020. www.statista.com/statistics/276683/media-use-in-the-us/.
10. Hoag A, Grant AE, Carpenter S. Impact of media on major choice: Survey of communication undergraduates. Nacada Journal. 2017; 37(1):5–14. https://doi.org/10.12930/NACADA-15-040

11. Akosah-Twumasi P, Emeto TI, Lindsay D, Tsey K, Malau-Aduli BS. A Systematic Review of Factors That Influence Youths Career Choices—the Role of Culture. In: Frontiers in Education. vol. 3. Frontiers; 2018. p. 58.
12. Tucciarone K. Influence of Popular Television Programming on Students' Perception about Course Selection, Major, and Career. The Popular Culture Studies Journal. 2014; 2(1):172–193.
13. Robb DL. Operation Hollywood: How the Pentagon shapes and censors the movies. Prometheus Books; 2011.
14. The Geena Davis Institute on Gender in Media, & J Walter Thompson Intelligence. The "Scully effect": I want to believe. . . in STEM; 2018.
15. ZenBusiness. Influence of Media on Careers; 2020. https://www.zenbusiness.com/blog/influence-of-media-on-careers/.
16. Asimow M. Bad lawyers in the movies. Nova L Rev. 1999; 24:533.
17. Dimnik T, Felton S. Accountant stereotypes in movies distributed in North America in the twentieth century. Accounting, organizations and society. 2006; 31(2):129–155. https://doi.org/10.1016/j.aos.2004.10.001
18. Flores G. Mad scientists, compassionate healers, and greedy egotists: the portrayal of physicians in the movies. Journal of the National Medical Association. 2002; 94(7):635. PMID: 12126293
19. Pautz MC. Cops on Film: Hollywood's Depiction of Law Enforcement in Popular Films, 1984-2014. PS, Political Science & Politics. 2016; 49(2):250. https://doi.org/10.1017/S1049096516000159
20. Moskovkin V, Saprykina T, Pupynina E, Belenko V, Shumakova I. Examination of trends in education with the Google Books Ngram Viewer. International Journal of Engineering and Advanced Technology. 2019; 8:1083–1090. https://doi.org/10.35940/ijeat.F1323.0886S219
21. Younes N, Reips UD. The changing psychology of culture in German-speaking countries: A Google Ngram study. International Journal of Psychology. 2018; 53:53–62. https://doi.org/10.1002/ijop.12428 PMID: 28474338
22. Basile P, Caputo A, Luisi R, Semeraro G. Diachronic analysis of the italian language exploiting google ngram. CLiC it. 2016; p. 56.
23. Michel JB, Shen YK, Aiden AP, Veres A, Gray MK, Pickett JP, et al. Quantitative analysis of culture using millions of digitized books. science. 2011; 331(6014):176–182. https://doi.org/10.1126/science.1199644 PMID: 21163965
24. Brysbaert M, New B. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior research methods. 2009; 41(4):977–990. https://doi.org/10.3758/BRM.41.4.977 PMID: 19897807
25. Bureau of Labor Statistics, U S Department of Labor. Standard Occupational Classification (SOC) System; 2018. https://www.bls.gov/soc/2018/home.htm.
26. Baruah S, Somandepalli K, Narayanan S. GitHub repository: Representation of professions in entertainment media; 2022. Available from: https://github.com/sabyasachee/mica-profession.
27. Zeng B, Yang H, Xu R, Zhou W, Han X. LCF: A local context focus mechanism for aspect-based sentiment classification. Applied Sciences. 2019; 9(16):3389. https://doi.org/10.3390/app9163389
28. Dong L, Wei F, Tan C, Tang D, Zhou M, Xu K. Adaptive Recursive Neural Network for Target-dependent Twitter Sentiment Classification. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Baltimore, Maryland: Association for Computational Linguistics; 2014. p. 49–54. Available from: https://www.aclweb.org/anthology/P14-2009.
29. Kalisch PA, Kalisch BJ. A comparative analysis of nurse and physician characters in the entertainment media. Journal of Advanced Nursing. 1986; 11(2):179–195. https://doi.org/10.1111/j.1365-2648.1986.tb01236.x PMID: 3635547
30. Smith SL, Choueiti M, Prescott A, Pieper K. Gender roles & occupations: A look at character attributes and job-related aspirations in film and television. Geena Davis Institute on Gender in Media. 2012; p. 1–46.
31. Tjong Kim Sang EF, De Meulder F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003; 2003. p. 142–147. Available from: https://www.aclweb.org/anthology/W03-0419.
32. Pradhan S, Moschitti A, Xue N, Ng HT, Björkelund A, Uryupina O, et al. Towards Robust Linguistic Analysis using OntoNotes. In: Proceedings of the Seventeenth Conference on Computational Natural Language Learning. Sofia, Bulgaria: Association for Computational Linguistics; 2013. p. 143–152. Available from: https://www.aclweb.org/anthology/W13-3516.

33. Ling X, Weld DS. Fine-Grained Entity Recognition. In: Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence. AAAI'12. AAAI Press; 2012. p. 94–100.

34. Sekine S. Extended Named Entity Ontology with Attribute Information. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). Marrakech, Morocco: European Language Resources Association (ELRA); 2008.Available from: http://www.lrec-conf.org/proceedings/lrec2008/pdf/21_paper.pdf.

35. Mai K, Pham TH, Nguyen MT, Nguyen TD, Bollegala D, Sasano R, et al. An Empirical Study on Fine-Grained Named Entity Recognition. In: Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA: Association for Computational Linguistics; 2018. p. 711–722. Available from: https://www.aclweb.org/anthology/C18-1060.

36. Ellis J, Getman J, Fore D, Kuster N, Song Z, Bies A, et al. Overview of Linguistic Resources for the TAC KBP 2015 Evaluations: Methodologies and Results. In: TAC KBP Workshop 2015: National Institute of Standards and Technology, Gaithersburg, MD.; 2015.

37. Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D. The Stanford CoreNLP Natural Language Processing Toolkit. In: Association for Computational Linguistics (ACL) System Demonstrations; 2014. p. 55–60. Available from: http://www.aclweb.org/anthology/P/P14/P14-5010.

38. Lin H, Lu Y, Han X, Sun L, Dong B, Jiang S. Gazetteer-Enhanced Attentive Neural Networks for Named Entity Recognition. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics; 2019. p. 6232–6237. Available from: https://www.aclweb.org/anthology/D19-1646.

39. Liu T, Yao JG, Lin CY. Towards Improving Neural Named Entity Recognition with Gazetteers. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics; 2019. p. 5301–5307. Available from: https://www.aclweb.org/anthology/P19-1524.

40. Ganzeboom HB. A new International Socio-Economic Index (ISEI) of occupational status for the International Standard Classification of Occupation 2008 (ISCO-08) constructed with data from the ISSP 2002-2007. In: Annual Conference of International Social Survey Programme, Lisbon. vol. 1; 2010.

41. Kazama J, Torisawa K. Inducing Gazetteers for Named Entity Recognition by Large-Scale Clustering of Dependency Relations. In: Proceedings of ACL-08: HLT. Columbus, Ohio: Association for Computational Linguistics; 2008. p. 407–415. Available from: https://www.aclweb.org/anthology/P08-1047.

42. Nadeau D, Turney PD, Matwin S. Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. In: Conference of the Canadian society for computational studies of intelligence. Springer; 2006. p. 266–277.

43. Miller GA. WordNet: a lexical database for English. Communications of the ACM. 1995; 38(11):39–41. https://doi.org/10.1145/219717.219748

44. Princeton University. WordNet lexicographer file names and numbers; 2010. https://wordnet.princeton.edu/documentation/lexnames5wn.

45. Toral A, Muñoz R. A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia. In: Proceedings of the Workshop on NEW TEXT Wikis and blogs and other dynamic text sources; 2006.Available from: https://www.aclweb.org/anthology/W06-2809.

46. Magnini B, Negri M, Prevete R, Tanev H. A WordNet-Based Approach to Named Entites Recognition. In: COLING-02: SEMANET: Building and Using Semantic Networks; 2002.Available from: https://www.aclweb.org/anthology/W02-1109.

47. Boteanu A, Kiezun A, Artzi S. Synonym expansion for large shopping taxonomies. In: Automated Knowledge Base Construction (AKBC); 2018.

48. Neale S, Gomes L, Agirre E, de Lacalle OL, Branco A. Word Sense-Aware Machine Translation: Including Senses as Contextual Features for Improved Translation Models. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). Portorož, Slovenia: European Language Resources Association (ELRA); 2016. p. 2777–2783. Available from: https://www.aclweb.org/anthology/L16-1441.

49. Pu X, Pappas N, Henderson J, Popescu-Belis A. Integrating Weakly Supervised Word Sense Disambiguation into Neural Machine Translation. Transactions of the Association for Computational Linguistics. 2018; 6:635–649. https://doi.org/10.1162/tacl_a_00242

50. Zhong Z, Ng HT. Word Sense Disambiguation Improves Information Retrieval. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Jeju Island, Korea: Association for Computational Linguistics; 2012. p. 273–282. Available from: https://www.aclweb.org/anthology/P12-1029.

51. Ramakrishnan G, Jadhav A, Joshi A, Chakrabarti S, Bhattacharyya P. Question Answering via Bayesian Inference on Lexical Relations. In: Proceedings of the ACL 2003 Workshop on Multilingual

Summarization and Question Answering. Sapporo, Japan: Association for Computational Linguistics; 2003. p. 1–10. Available from: https://www.aclweb.org/anthology/W03-1201.

52. Edmonds P, Cotton S. SENSEVAL-2: Overview. In: Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems. Toulouse, France: Association for Computational Linguistics; 2001. p. 1–5. Available from: https://www.aclweb.org/anthology/S01-1001.

53. Snyder B, Palmer M. The English all-words task. In: Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text. Barcelona, Spain: Association for Computational Linguistics; 2004. p. 41–43. Available from: https://www.aclweb.org/anthology/W04-0811.

54. Pradhan S, Loper E, Dligach D, Palmer M. SemEval-2007 Task-17: English Lexical Sample, SRL and All Words. In: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007). Prague, Czech Republic: Association for Computational Linguistics; 2007. p. 87–92. Available from: https://www.aclweb.org/anthology/S07-1016.

55. Navigli R, Jurgens D, Vannella D. SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). Atlanta, Georgia, USA: Association for Computational Linguistics; 2013. p. 222–231. Available from: https://www.aclweb.org/anthology/S13-2040.

56. Moro A, Navigli R. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Denver, Colorado: Association for Computational Linguistics; 2015. p. 288–297. Available from: https://www.aclweb.org/anthology/S15-2049.

57. Raganato A, Camacho-Collados J, Navigli R. Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. Valencia, Spain: Association for Computational Linguistics; 2017. p. 99–110. Available from: https://www.aclweb.org/anthology/E17-1010.

58. Raganato A, Delli Bovi C, Navigli R. Neural Sequence Learning Models for Word Sense Disambiguation. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics; 2017. p. 1156–1167. Available from: https://www.aclweb.org/anthology/D17-1120.

59. Huang L, Sun C, Qiu X, Huang X. GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics; 2019. p. 3509–3514. Available from: https://www.aclweb.org/anthology/D19-1355.

60. Kumar S, Jat S, Saxena K, Talukdar P. Zero-shot Word Sense Disambiguation using Sense Definition Embeddings. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics; 2019. p. 5670–5681. Available from: https://www.aclweb.org/anthology/P19-1568.

61. Bevilacqua M, Navigli R. Breaking Through the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. p. 2854–2864. Available from: https://www.aclweb.org/anthology/2020.acl-main.255.

62. Liu B. Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge university press; 2015.

63. Saeidi M, Bouchard G, Liakata M, Riedel S. SentiHood: Targeted Aspect Based Sentiment Analysis Dataset for Urban Neighbourhoods. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan: The COLING 2016 Organizing Committee; 2016. p. 1546–1556. Available from: https://www.aclweb.org/anthology/C16-1146.

64. Pontiki M, Galanis D, Pavlopoulos J, Papageorgiou H, Androutsopoulos I, Manandhar S. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland: Association for Computational Linguistics; 2014. p. 27–35. Available from: https://www.aclweb.org/anthology/S14-2004.

65. Tang D, Qin B, Feng X, Liu T. Effective LSTMs for Target-Dependent Sentiment Classification. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan: The COLING 2016 Organizing Committee; 2016. p. 3298–3307. Available from: https://www.aclweb.org/anthology/C16-1311.

66. Wang S, Mazumder S, Liu B, Zhou M, Chang Y. Target-Sensitive Memory Networks for Aspect Sentiment Classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational

Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics; 2018. p. 957–967. Available from: https://www.aclweb.org/anthology/P18-1088.

67. Zhang C, Li Q, Song D. Aspect-based Sentiment Classification with Aspect-specific Graph Convolutional Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics; 2019. p. 4568–4578. Available from: https://www.aclweb.org/anthology/D19-1464.

68. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in neural information processing systems; 2017. p. 5998–6008.

69. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171–4186. Available from: https://www.aclweb.org/anthology/N19-1423.

70. Sun C, Huang L, Qiu X. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 380–385. Available from: https://www.aclweb.org/anthology/N19-1035.

71. Xu H, Liu B, Shu L, Yu P. BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 2324–2335. Available from: https://www.aclweb.org/anthology/N19-1242.

72. Lison P, Tiedemann J, Kouylekov M. OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA); 2018.Available from: https://www.aclweb.org/anthology/L18-1275.

73. Chaput M. GitHub repository: Whoosh; 2018. Available from: https://github.com/mchaput/whoosh.

74. Spearman C. The Proof and Measurement of Association between Two Things. The American Journal of Psychology. 1904; 15(1):72–101. https://doi.org/10.2307/1412159

75. Baruah S, Somandepalli K, Narayanan S. *Trend Visualizer*, Representation of Professions in Entertainment Media, Center for Computational Media Intelligence; 2021. https://sail.usc.edu/~ccmi/representation-of-professions/#visualization.

76. Berger M. A guide to how gender-neutral language is developing around the world; 2019. https://www.washingtonpost.com/world/2019/12/15/guide-how-gender-neutral-language-is-developing-around-world/.

77. Williams D. The Rise of Gender Neutrality and its Impact on Language; 2018. https://www.translatemedia.com/us/blog-usa/rise-gender-neutrality-impact-language/.

78. Nicholson P. The second wave: A reader in feminist theory. Psychology Press; 1997.

79. Brown JK. The Nineteenth Amendment and women's equality. The Yale Law Journal. 1993; 102 (8):2175–2204. https://doi.org/10.2307/796863

80. Geraghty L. American science fiction film and television. Berg; 2009.

81. Jenns N. Why Suddenly Real-Life Sports Documentaries Trend in Film Industry?; 2021. https://www.tetongravity.com/story/news/why-suddenly-real-life-sports-documentaries-trend-in-film-industry.

82. Hauer KE, Durning SJ, Kernan WN, Fagan MJ, Mintz M, O'Sullivan PS, et al. Factors Associated With Medical Students' Career Choices Regarding Internal Medicine. JAMA. 2008; 300(10):1154–1164. https://doi.org/10.1001/jama.300.10.1154 PMID: 18780844

83. Dalen JE, Ryan KJ, Alpert JS. Where have the generalists gone? They became specialists, then subspecialists. The American journal of medicine. 2017; 130(7):766–768. https://doi.org/10.1016/j.amjmed.2017.01.026 PMID: 28216448

84. Reimink T. Are Bank Tellers and Retail Cashiers Experiencing Similar Transformations?; 2017. https://www.linkedin.com/pulse/bank-tellers-retail-cashiers-experiencing-similar-timothy-reimink/.

85. Wong JC. End of the checkout line: the looming crisis for American cashiers; 2017. https://www.theguardian.com/technology/2017/aug/16/retail-industry-cashier-jobs-technology-unemployment.

86. Flanagan TJ, Vaughn MS. Public opinion about police abuse of force. Police violence: Understanding and controlling police abuse of force. 1996; 30(5):397–408.

87. Funk C, Hefferon M, Kennedy B, Johnson C. Trust and mistrust in Americans' views of scientific experts. Pew Research Center. 2019; 2:1–96.

88. Doherty C, Kiley J, Daniller A, Jones B, Hartig H, Dunn A, et al. Majority of public favors giving civilians the power to sue police officers for misconduct; 2020.

89. Kennedy BR, Mathis CC, Woods AK. African Americans and their distrust of the health care system: healthcare for diverse populations. Journal of cultural diversity. 2007; 14(2). PMID: 19175244

90. Ramakrishna A, Martinez VR, Malandrakis N, Singla K, Narayanan S. Linguistic analysis of differences in portrayal of movie characters. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada: Association for Computational Linguistics; 2017. p. 1669–1678. Available from: https://aclanthology.org/P17-1153.