SocialQuotes: Learning Contextual Roles of Social Media Quotes on the Web

John Palowitch, Hamidreza Alvari, Mehran Kazemi, Tanvir Amin, Filip Radlinski

Google Research
{palowitch,hamidrz,mehrankazemi,tanviramin,filiprad}@google.com

Abstract

Web authors frequently embed social media to support and enrich their content, creating the potential to derive *web-based*, cross-platform social media representations that can enable more effective social media retrieval systems and richer scientific analyses. As a step toward such capabilities, we introduce a novel language modeling framework that enables automatic annotation of *roles* that social media entities play in their embedded web context. Using related communication theory, we liken social media embeddings to *quotes*, formalize the page context as structured natural language signals, and identify a taxonomy of roles for quotes within the page context. We release SocialQuotes, a new data set built from the Common Crawl of over 32 million social quotes, 8.3k of them with crowdsourced quote annotations. Using SocialQuotes and the accompanying annotations, we provide a role classification case study, showing reasonable performance with modern-day LLMs, and exposing explainable aspects of our framework via page content ablations. We also classify a large batch of unannotated quotes, revealing interesting cross-domain, cross-platform role distributions on the web.

Introduction

The global adoption of online social media (SM) platforms over the past two decades has made them into *de facto* information stores of world knowledge, especially as primary sources of human stories, discussion, news commentary, and public announcements (Myers and Hamilton 2014). As a result, web authors have increasingly cited SM in online articles and on the web at-large (Gearhart and Kang 2014). Today, it is commonplace to find SM across the web in news stories, blog pieces, site info pages, and personal websites. There are standardized website tools for "embedding" SM in a page such that the post appears in-line, with platform-specific links and icons (see Figure 1). Web authors may embed SM to exemplify, explain, provide evidence, or promote, potentially using ML-backed tools for SM source-seeking (Fernandes, Moro, and Cortez 2023). Regardless, it is the web author that ultimately chooses the source appropriate for their page and frames the embedding with explanatory prose.

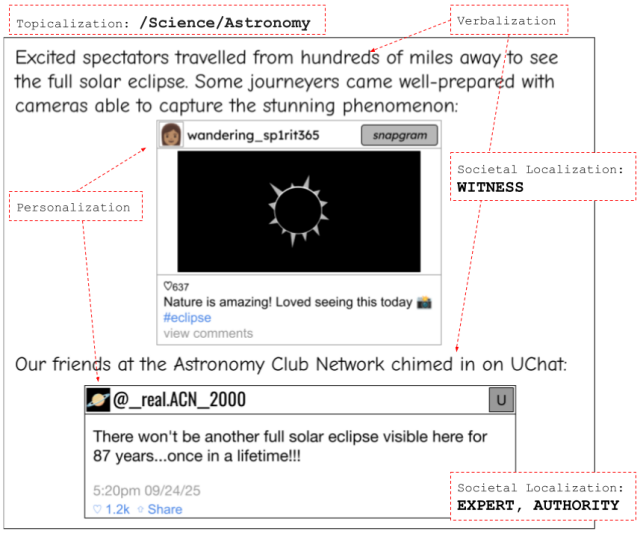

Figure 1: Fictional web article with social media quotes from fictional platforms. We model social media quotation using a 4-stage procedure from communications theory (Haapanen 2020). The "Societal localization" stage is an unobserved process wherein the web author decides which *roles* to seek for contextual support. We introduce a framework for inferring roles from the web context.

We observe that, due to the ubiquity of social media embeddings, the web contexts of such embedded SM are potential sources of raw signals for learning representations of, and annotations on, SM posts and entities. Learning such representations and annotations could be valuable for (e.g.) building SM retrieval databases, or for feeding into scientific studies of society and the web. While many SM learning paradigms exist for *on-platform* contexts (Balaji, Annavarapu, and Bablani 2021), how to effectively learn from the *web* context around SM embeddings remains an open challenge.

As a step toward addressing this challenge, we introduce a natural-language learning framework for web-embedded social media, targeting automatic understanding of the "role" of the SM within the page context. As Figure 1 illustrates, SM embeddings function as *quotes* in that they can serve to

tell a story, make a point, provide a reference, or show relevant human experiences from around the world, in some cases providing unique value beyond standard "person-on-the-street" sources (Gearhart and Kang 2014). However, such "roles" cannot be immediately known from the web context – they exist as latent, guiding categories in the mind of the web author. Our main hypothesis is that the web author's framing of SM quotes, as well as other web context around the page, can be used by a language model to infer the categorical role of the quote. If correct, modern-day language models can be profitably used to enhance SM databases with annotations about *how* particular SM entities are commonly quoted across web topics. Our contributions are:

1. Using relevant communications theory, we link the web context of social media quotes to the social media entities involved, and propose a taxonomy of roles that SM quotes play in web contexts.
2. We build and release SocialQuotes[1], a data set computed from the Common Crawl with over 32M social media quotes, accompanied by role labels for 8.3k quotes obtained from human annotators.
3. We validate our main hypothesis with SocialQuotes by showing that an LLM can predict the role of social quotes from their web context (and that performance improves with advanced reasoning techniques), and we provide a cross-domain, cross-platform case study analysis of role distributions across the web.

Two valuable aspects of our framework is that (1) it extends to all SM platforms commonly embedded on the web, and (2) it opens the door to modeling SM *without any SM platform data*, drastically cheapening the cost of data collection for researchers pursuing certain applications. We elaborate on these aspects and other future directions in our closing section.

Related Work

To the best of our knowledge, our proposed framework is the first to rigorously establish a natural language learning task for SM embedded in the web. We identified five distinct subfields that inform our work in various ways. We give a brief overview of each field to correctly frame our work's novelty and impact.

Quotation in Writing. There is a large body of work in communications and linguistics on the *form* and *function* of traditional quoting in writing and media (Zelizer 1995; Cope 2020; Harry 2014; Bublitz 2015; Zelizer 1989; Haapanen 2020). Our work hearkens to this space by focusing on quotations of social media in the open web. Specifically, our proposed taxonomy of SM quotes aims to categorize the *function* or "role" of the quote rather than its *form*, i.e. the specific language being used in the web context (though our data set release opens up this area as future work). As we describe in following sections, our approach builds on Haapanen (2020), allowing us to identify and use a novel division of SM quote instances into *societal groups* that web authors seek from SM sources.

Digital Journalism. Our work is also connected to the digital journalism space, in particular the study of SM posts appearing in online news media. Dumitrescu and Ross (2021), Gearhart and Kang (2014), and Broersma and Graham (2014) consider the differential *effect* of SM quotes from certain SM platforms and accounts across media types and audiences. Rony, Yousuf, and Hassan (2018), Kapidzic et al. (2022), and Gruppi et al. (2021) study the differential *frequency* with which SM sources are used in "unreliable"/tabloid-style outlets versus "reliable"/mainstream outlets. Mujib, Zelenkauskaite, and Williams (2022) studies the *speed* with which SM content is embedded in media outlets due to the "pressures of the 24/7 news cycle".

Many of these studies categorize social media accounts into "types", which correspond to what we call "roles" in this paper. Specifically, among these studies, there are at least three distinct 4-way taxonomies over the quoted social media accounts (Broersma and Graham 2014; Kapidzic et al. 2022; Mujib, Zelenkauskaite, and Williams 2022). In each case, the authors manually classified SM accounts into the chosen taxonomy, and used the resulting codings in their scientific analyses. Our proposed learning paradigm aims to directly impact this field by enabling automatic classification of SM entities based on their embedded web context. As such, in the next section, we adopt a role taxonomy that both covers and extends each existing taxonomy in this space.

Quotation Datasets. There are several released web-based quotation data sets. Tekir et al. (2023) propose a corpus of book quotes extracted from book reviews for the NLP task of automatic quote detection. Vaucher et al. (2021) derive a large-scale corpus of traditional (non-SM) quotes from online news sources between 2008 and 2020. In the social media domain, embedded SM have not yet been likened to quotes, yet relevant datasets still exist. Mujib et al. (2020) release NewsTweet, a data set of embedded social media found in Google News sites. Our SocialQuotes data set expands upon NewsTweet in three ways: (1) source: by using the Common Crawl, we are able to cover SM quotes from a large random crawl of the web rather than just news articles; (2) scale: our dataset covers approximately 12.7 million URLs with 32.6 million embeddings, as opposed to approximately 69K articles and 136K embeddings from NewsTweet, and three platforms instead of one (NewsTweet covers Twitter only); (3) annotations: we release topic metadata for all quotes and crowdsourced role labels for a 8.3k subset of quotes.

Influencer/Expert Detection. The task of categorizing SM quotes into functional roles is related to the tasks of influencer detection and expert finding; in fact, two roles in our chosen taxonomy are INFLUENCER and EXPERT. Influencer detection, often achieved with graph learning algorithms (Zheng et al. 2020), aims to recover structurally-central or abnormally-impactful nodes in social networks, either in general (Pei et al. 2020) or with respect to given

[1]https://www.kaggle.com/datasets/googleai/social-quotes

topics (Panchendrarajan and Saxena 2023). Expert finding is the task of retrieving members of a communication system (e.g. SM, Q&A sites, email networks) who are knowledgeable or have skills in specialized areas, and is often performed with a mixture of NLP and graph-based methods (Balog, Azzopardi, and de Rijke 2009; Balog, De Rijke et al. 2007; Lin et al. 2017). The main distinction between these paradigms and ours is that we do not attempt to classify SM users into certain roles *per se*. Instead we attempt to classify the *role* that a particular SM user's post plays in the context of a particular web page in which their post is quoted. A secondary and related distinction is that we do not use any platform data: we study how to use the surrounding web content to infer the quote's role.

Social Media in Web Data. There have been some recent works that connect social media entities and web data. For example, Wen et al. (2023) develop the task of predicting WikiData (Vrandečić and Krötzsch 2014) attributes of public figures using posts from the figures' social media accounts. Most related to our work, Hombaiah et al. (2023) develop a retrieval model for Tweets[2] given a news article. They construct (but do not release) a data set of 8M (article, Tweet) tuples, each with at least one Tweet. They evaluate various two-tower retrieval models, some of which only process BERT encodings of the article and Tweet, and others which additionally include a Tweet account encoder taking in signals from the Twitter profiles. Our work both extends and complements this work. First, we release 37M instances of social media post embeddings in websites, which cover both news and non-news sites as well as additional platforms (TikTok and Instagram). One goal in releasing this dataset is to explicitly enable future research on retrieval models similar to those introduced in (Hombaiah et al. 2023), e.g. extending them to more platforms and more types of sites. Second, instead of retrieval, we study property prediction of the embedded social media entities using the surrounding page context. The next section formalizes our model of social media web embedding, which motivates our focus on the *role* of the social post as the target of prediction.

A Model of Social Quotation

In this paper, we draw an analogy between social media embeddings in web pages and rhetorical quotation in print, and use this to motivate a new NLP task targeting the "role" of social media quotes. We base our approach on recent work in the communications literature which models quotation in news writing as a four-stage process (Haapanen 2020): (1) "topicalization": the topic of the piece is established; (2) "societal localization": societal groups representing "various roles (e.g. *people concerned*, *authorities*)" relevant to the topic are identified; (3) "personalization": representatives from the identified societal groups are chosen for inclusion in the piece; (4) "verbalization": the recorded views of the representatives are directly or indirectly woven into the piece according to the author's style and goals.

We adapt these four steps into a model of social media quotations on the web, such as those illustrated in Figure 1. Given a quotation instance i, we assume the author has arrived at a topic $t_i \in \mathcal{T} = \{T_1, \ldots, T_{n_t}\}$ (topicalization) for which they seek representative posts from social media. The author then chooses a set of roles $\rho_i = \{r_1, \ldots, r_k\} \subset \mathcal{R} = \{R_1, \ldots, R_{n_r}\}$ that indicates the "types" of social media accounts being sought (societal localization) for worthwhile quotations on t_i. Next, the author chooses a social media post p_i from a corpus $\mathcal{P} = \{P_1, \ldots, P_{n_p}\}$ created by a user u_i that the author has determined to assume one or more roles from ρ_i (personalization). Finally, the author rhetorically frames p_i within the context of the web page, producing a novel text snippet x_i.

Role Taxonomy. In the above model, $\mathcal{P}$ can be any social media corpus, and $\mathcal{T}$ can be any set of web topic categories such the Cloud NL Categories[3]. However, there is no standard set of social roles $\mathcal{R}$: indeed, our chosen modeling approach introduces a new focus on this aspect of quotation. Therefore a core aspect of our work is to formulate an appropriate $\mathcal{R}$ for social media quotes on the web. To this end, we follow two complementary desiderata:

1. $\mathcal{R}$ should cover existing social account "types" adopted by digital journalism work (see "Related Work"), so that our framework can be used to build models for similar efforts.
2. $\mathcal{R}$ should extend to SM embeddings that could plausibly be found outside of news or journalism pieces.

To achieve desiderata 1, we considered all social account "types" from Kapidzic et al. (2022), Broersma and Graham (2014), and Mujib, Zelenkauskaite, and Williams (2022), and found that they could be well-organized into four roles: EXPERT, INFLUENCER, AUTHORITY, and COMMENTER. A matching of the account types from past work to these roles is shown in Table 1.

To achieve desiderata 2, we propose to expand this initial role set in two ways, inspired by our manual review of web pages found in the Common Crawl corpus. First, we add SUBJECT and WITNESS roles to cover the many news and blog articles that (respectively) (1) focus on a certain entity and include SM quotes from that entity's account, or (2) focus on an event and include SM quotes from entities who participated or observed. Second, we add SELF-PROMOTER and MARKETER roles to cover (respectively) websites with (1) SM quotes from the same entity that owns the site (usually for promotional purposes), and (2) SM quotes that seem to be marketing a product or service.

In total, our taxonomy has eight roles, as shown in Table 1. We observe that they can be divided into Elite-type, Citizen-type, and Commercial-type roles, which we include in Table 1 to illustrate the high-level semantic coverage of the taxonomy. In the next section, we empirically validate the taxonomy by providing an "Other" option to annotators. Over a random sample of SM quotes from the web, we found

[2]The previously-named Twitter platform is now called X, though in this publication we refer to X as "Twitter" and X posts as "Tweets" for consistency both with our source data and with past publications.

[3]https://cloud.google.com/natural-language/docs/categories

Role Type	Role Name	Role Definition	Comparison Roles
Elite	EXPERT	Has professional experience, higher education, or valuable skills relevant to the page topic.	*Expert* (Broersma and Graham 2014), *Journalist* (Mujib, Zelenkauskaite, and Williams 2022)
	INFLUENCER	Is popular voice on the page topic and/or within the page's primary audience.	*Cultural Producer* (Broersma and Graham 2014), *Celebrity* (Kapidzic et al. 2022), *Personality* (Mujib, Zelenkauskaite, and Williams 2022)
	AUTHORITY	Is associated with a recognized societal position relevant to the page topic.	*Politician* (Broersma and Graham 2014), *Public Actor*, *Media* (Kapidzic et al. 2022), *Organization*, *Media Outlet* (Mujib, Zelenkauskaite, and Williams 2022)
Citizen	SUBJECT	The primary focus of the web page or article.	*(none)*
	WITNESS	Was a witness or participant in an event described on the web page.	*(none)*
	COMMENTER	Has shared thoughts or opinions about the page topic.	*Vox Populi* (Broersma and Graham 2014), *Citizen* (Kapidzic et al. 2022)
Commercial	MARKETER	Is marketing products or selling services relevant to the page.	*(none)*
	SELF-PROMOTER	Is the owner/author of the web page itself.	*(none)*

Table 1: Social media quotation taxonomy. Each role is defined with respect to the web page author's intended function for the social media post within the context of the page they are writing, via the "societal localization" step.

that annotators chose every role a non-trivial amount of times, and the "Other" option accounted for only $sim1.4\%$ of all individual annotations.

By adapting a fundamental model of traditional quoting to social media quotes on the web, we formalize the idea that the surrounding text of a social media quote can carry nontrivial information about the social media user via the "societal localization" step taken by the web author. Therefore, our primary methodological hypothesis (which we examine in our Experiments and Analysis section) is that the page context surrounding a SM quote can be used as signals to model the involved SM entities. In the next section, we describe the construction of the SocialQuotes data set, which we use in our experiments and release in full to the community.

SocialQuotes Dataset

As illustrated in Figure 1, online content frequently embeds social media content. We define an SM embedding as the rendering of a social media post directly within the web page: it must be visible to a reader with a standard web browser without being logged in to any social media services (modulo any privacy restrictions that may be selectively enabled on various websites). An embedding is considered complete if it is possible to identify a canonical URL for the social media post as well as the author of the post. Motivated by our conceptual framework outlined in the previous section, we refer to social media embeddings as "quotes" throughout the remainder of the paper.

In this section we describe the construction of the SocialQuotes data set. Starting with the widely used Common Crawl corpus[4], which consists of a frequently updated public multi-terabyte web crawl, we built a Python-based Apache Beam pipeline to identify and extract SM quotes, and used a public language model to classify the surrounding web content into one or more topics. A sample of SM quotes were also annotated for social roles. We release[5] the corpus of SM quotes and their annotations as the SocialQuotes dataset, as described fully in Appendix A Table 5. For the purposes of this paper and dataset release, we focus on Twitter, TikTok, and Instagram quotes, although our methodology is in no way limited to these platforms.

Embedding Extraction Pipeline

We determined the HTML source patterns that create SM quotes from our chosen platforms through a trial-and-error process, inspecting web pages with visible SM quotes. The source patterns that we determined are shown in Appendix A Table 6. Given the crawled HTML from a webpage, the contained social media quotes that follow patterns that we identified can be extracted using BeautifulSoup[6] functions:

```
BeautifulSoup(
    html_doc, "html.parser")
.find_all(
    tag_type, class_=tag_class)
```

[4]https://www.commoncrawl.org/

[5]https://www.kaggle.com/datasets/googleai/social-quotes

[6]https://www.crummy.com/software/BeautifulSoup/bs4/doc

Through this method, we extracted all instances of SM media quotes captured by these (`tag_type`, `tag_class`) pairs from the full Common Crawl 2022-05, 2022-40 and 2023-23 snapshots, restricting ourselves to web pages where a standard classifier indicated the content was primarily in English. Our extraction pipeline was written in Python using the Apache Beam framework, letting each worker handle SM embedding extraction from a single Common Crawl website record (identified with a unique URL). This pipeline yielded 32.6M embedded social media posts, with per-platform counts given in Table 6.

Context Parsing and Topic Annotation

In addition to extracting social media quotes following patterns in Table 6, we also extracted the page context surrounding each quote by traversing the DOM tree in either direction until a pre-specified character limit τ was met. We extracted at least $\tau = 300$ characters, and did not break in the middle of a DOM element such as a paragraph tag (thus some context snippets had more than τ characters). Note that we do not release context snippets in the SocialQuotes dataset, although these can be reconstructed from CommonCrawl and the SocialQuotes dataset. To enable the study of the intersection of SM entities and web topics, we classified each context snippet using the Google Cloud Content Classifier[7]. We release these topic annotations with SocialQuotes in the `context_topics` field, and analyze their distribution with respect to SM quotes in our Experiments and Analysis section. To avoid releasing sensitive or inappropriate content, we filter the SocialQuotes entries according to rules described in Appendix A. This filtering brings our total quote count to 32,560,806, which is the size of our final SocialQuotes release.

Role Annotation Process

From the collection of SM quotes identified, we sampled 9k SM quotes for annotation, choosing an equal number from each platform. To ensure that no URL was overly-represented (some URLs contain a huge number of SM quotes), we sampled quotes uniformly-at-random while ensuring that no URL was chosen twice. The quote instances were then each rated by 5 trained annotators as follows.

First, the annotator was presented with the URL containing a citation extract, as well as the username of a social media author. We note that while SM quotes sampled were extracted from the *static* Common Crawl collection, annotation was performed with annotators looking at *live* web pages, which may differ from the crawled versions. Thus, we asked annotators to complete an initial task (full UI displayed in Appendix A, Figure 3):

> Please look at this webpage: {`Clickable URL`}. Can you find a {`Platform`} post by {`Username`} in the main body of the page? This {`Clickable Post URL`} by {`Username`} may be embedded on the page.

[7]https://cloud.google.com/natural-language/docs/classifying-text

If the above question was answered in the negative, annotators were given a "Not Found" option to choose. If the above question was answered affirmatively[8], that the extracted citation can still be found, then a second question was shown asking the annotator to select from the social roles listed in Table 1 as follows (full UI displayed in Appendix A, Figure 4):

> The web page author likely embedded the post(s) by the {`Platform`} account {`Username`} because they thought the account _____. Please choose **all** options that apply.

For each option that was selected, the description from Table 1 was presented directly in the annotation UI to remind the annotator of the exact meaning of each label. Further, annotators were asked to review a document providing examples of annotations for specific URLs prior to commencing the annotation process, and this document was made available to them to access at any time.

We gave the annotators an "Other" option to indicate that they thought that the social post was playing a role potentially not covered by our taxonomy. In Appendix A, we provide additional statistics on the co-occurrence of the "Other" option with in-taxonomy roles (Table 7), as well as the frequency of the "Not Found" option (Figure 6). We find that the "Other" option was chosen only 289 times by any annotator, accounting for 1.4% of annotator responses in which the post could be found. Furthermore, as shown by Figure 2, annotators chose all labels a non-trivial amount of times. These results validate the coverage of our chosen taxonomy.

Ethics statement. We note that the human annotator work was carried out by participants who are paid contractors. Those contractors received a standard contracted wage, which complies with living wage laws in their country of employment. To ensure a consistent cultural understanding of social media, and to align with the fact that we only extracted social media from primarly English-language pages, the annotators were selected to be native English speakers based in the United States and in Canada.

Ground-Truth From Annotations

We filter, aggregate, and validate annotations to produce ground-truth. First, any annotation on quotes proximal to sensitive content (see "Context Parsing and Topic Annotation" and Appendix A) are neither released nor used as ground-truth, leaving 8,281 annotations. From the remaining, we define a "valid" annotation set for a given quote as one where at least two annotators were able to find the quote in the web page. Our annotation experiment produced 4,483 such sets. To check annotator agreement, we report Fleiss' κ (Fleiss 1971) across the valid annotations for each role in Table 7 in Appendix A. All roles had positive agreement,

[8]Note the reference to the content being in the main body of the page: Annotators were instructed to *not* annotate posts if they believed the post was embedded by someone other than the main page author, such as in a comment. Annotators were similarly trained to skip URLs with non-English content, or which did not load correctly.

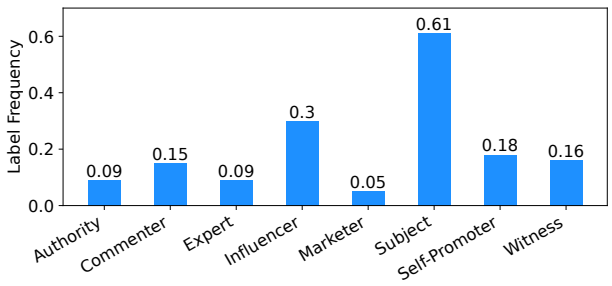

Figure 2: The frequency of each of the eight labels in the final dataset.

though with varying magnitudes. We release all completed annotations with at least one non-"Not Found" result with the SocialQuotes data set under the field `role_labels`, containing a JSON-formatted string encoding a dictionary over the roles in Table 1, giving the number of times annotators chose each role.

To derive ground-truth for our experiments in the next section, we add a role label r to the ground-truth label set for a given example i if at least two of the five annotators chose r given example i. From the 4,483 valid annotations, 4,380 had at least one role in the ground-truth label set. We use these 4,380 ground-truth annotations as test evaluation data for our experiments, which we cover in the next section. In Appendix A Figure 5, we provide a confusion matrix over the roles showing how frequently roles co-occurred in a ground-truth quote.

Experiments and Analysis

In this section we describe experiments that constitute a case-study on our main intended use-case for our framework: inferring and analyzing roles of social media entities in web contexts. We consider two primary high-level research questions:

- **RQ1**: Can the role of a social media post quoted in a web site be inferred from the surrounding web context?
- **RQ2**: Do roles, platforms, and users distribute across web domains and web topics in meaningful, intuitive, or surprising ways?

An affirmative answer to these questions would show that there is useful information about social media entities contained in the web content that quotes them, confirming our overall hypothesis laid out in the "Model of Social Quotation" section. These experiments also show the utility of SocialQuotes toward future research directions, as described further in our closing section.

Role Inference

Here we investigate our main hypothesis: whether the role of a SM quote can be inferred by a language model from the surrounding text and page context. We use a pre-trained LLM in our experiment as a representative state-of-the-art language model to expose the present-day capabilities on the role classification task. Comparisons between LLMs or between an LLM and smaller models are left for future work. Specifically, we seek to answer the following questions:

- **RQ1.1**: Can an LLM infer the role of SM quotes from the surrounding text (rhetorical framing) and the page context?
- **RQ1.2**: Does using LLM reasoning techniques such as chain-of-thought (Wei et al. 2022) and self-consistency (Wang et al. 2023) help with role classification?
- **RQ1.3**: What elements of the page context are most important for role classification?

As our LLM, we use Google's PaLM2 `text-bison` model (Google and et al. 2023), which can be accessed via their Vertex AI[9].

Task Design. As illustrated in Figure 1, in a web page with social media quotes, the web author can assign potentially multiple roles to any given social media quote. To handle this scenario, we design a task where PaLM2 is asked to make a binary decision on each role individually. In other words, for a given social media quote, we prompt PaLM2 eight times, one time for each role in our taxonomy, and then aggregate the answers into a single result. We now describe the design of the prompt and evaluation for this setting.

Prompt Structure. Our PaLM2 prompt (shown in the Appendix) includes (1) a preamble, (2) role-specific elements from the page (including the text surrounding the quote, the URL of the page, the URL of the social media post, and the handle of the social media post), and optionally (3) few-shot examples. The `<role_specific_binary_prompt>` field is filled by a question that is specific to the current role being decided upon. The question contains a description of the role resembling those in Table 1. We list all eight questions in Table 9, Appendix B:

To construct the prompt, the common crawl URL, social post URL, and extracted profile handle are each extracted from their corresponding SocialQuotes data set fields listed in Table 5. Note that the extracted profile handle is a parse of social user profile, which is a URL to the profile page. The quote context field contains the quote context as described in our previous "Context Parsing" section. The few shot examples block contains five examples pulled from our SocialQuotes corpus, containing all four fields, and a final line reading "The answer is {yes,no}". This final line instructs PaLM2 how to give a binary answer to the prompt.

Evaluation. We evaluate the binary judgements given by PaLM2 against our annotated examples using the standard F1 score over each of the eight roles in our taxonomy. As baselines, we compare against the following:

1. True-frequency: a role r is a predicted positive with probability p_r, where p_r is the ground-truth probability of role r. Note that under this baseline, the F1 score converges in probability to p_r given an asymptotic number of samples. Figure 2 shows the p_r values for each r.

[9]http://cloud/vertex-ai/docs/generative-ai/model-reference/text

2. Coin-flip: a role r is a predicted positive with probability 0.5. Under this baseline, the F1 score converges to $p_r/(0.5 + p_r)$. We list this value in the "Coin flip" column in Table 3.

Performance Analysis. The results in the "Fewshot" column of Table 3 show the F1 scores of PaLM2 provided only with the preamble, the fewshot examples, and the social media quote data fields. We find that PaLM2 outperforms the coin-flip baseline on every role, sometimes with a great margin. This provides an affirmative answer to RQ1.1: there seems to be natural language signals present in the page context that can help infer the societal localization step of quotation, which we described in our "Model of Social Quotation" section. Of note, this also provides important validation for our approach that attempts to connect traditional quotation with the process of social media embeds on the web.

Chain-of-Thought Prompting. To investigate whether providing reasoning examples helps the model, we re-ran the experiment while adding a "chain-of-thought" (CoT) paragraph (Wei et al. 2022) immediately following each few-shot example. Each CoT paragraph used the given data fields to explain in common prose why the answer was yes or no. An example of a chain-of-thought paragraph is shown in the Appendix (see "Prompt Structure" section):

From Table 3, it is clear that CoT prompting improved the classification results in every role, leading to increased macro-average F1. Even so, we notice that providing CoT leads to an over-prediction of the positive class leading to a higher recall and a lower precision compared to the Fewshot model (see Table 8 in the Appendix). We next examine some methodology to remedy this effect.

Self-Consistency and Persistence. Self-Consistency (Wang et al. 2023) is a technique used for LLM-based classification where instead of calling the model once in a greedy mode and taking its output as the final answer, one calls the model multiple times and takes the majority vote of its predictions. The model is called with temperature[10] equal to 0.5 to enable generating non-identical but highly probable samples. We tested a variant of our Fewshot and CoT models with self-consistency where besides the greedy model output, we generated another 3 sets of model samples at a temperature of 0.5. Observing that our CoT model has a high recall but low precision, we tested a special case of self-consistency, which we call **"Persistence"**, where we set the prediction for each example to be "yes" only if all model calls predict "yes". The results are reported in Table 3. We observe that both self-consistency and persistence boost the performance of the CoT model, but they are not as effective for the Fewshot model. Specifically, CoT plus persistence achieves the best results in terms of macro-average F1 across all roles. This answers **RQ1.2** in the affirmative.

Ablation Study. Finally, to address **RQ1.3**, we examine which parts of the page context contribute most to the performance of PaLM2. At a conceptual level, the URL should provide web author information, the post URL should provide platform information, the handle should provide social media author information, and the snippet should provide role-level information about the social media post. We ablate each of these signals and re-run the experiment, with results in Table 8. Looking at the macro-average scores, we find that all signals contribute some performance value to the model, however the snippet provides the most. This accords with our fundamental hypothesis laid out in our "Model of Social Quotation" section, which is that the actual language used to quote the social media author can provide key information about the role that the post is playing in the page.

[10]Larger values of the "temperature" parameter in calls to standard LLMs induce more randomness in the response, whereas zero temperature results in determinism.

Social Quotes Analysis

In this section we use the SocialQuotes data set to provide novel, cross-platform, cross-role insights into the landscape of social quotation on the web. We first use PaLM2 prompted with all page context signals and CoT to infer the roles of a large batch of social quotes (approx. 117k quotes). We use the full data set and these inferred-on quotes to investigate the following questions:

- **RQ2.1**: Do websites covering certain **topics** tend to favor quoting certain social media platforms or roles?
- **RQ2.2**: Do websites at certain **domain names** tend to favor quoting certain social media platforms or roles?

We translate these questions into the following general measurement problem: given a website attribute a and a social media attribute x, do websites with attribute a (for example, a certain domain like news.yahoo.com) disproportionately quote social media posts with attribute x (for example, a certain platform like Twitter)? We use two metrics to examine this question: the relative proportion of x found with attribute a, and the mutual information between x and a. Specifically, define N as the total number of quotes in the corpus, N_a as the number of quotes in the corpus such that the website has attribute a, N_x similarly for attribute x, and N_{ax} as the number of quotes such that the website has attribute a and the SM post has attribute x. Then we define $p(a, x) = \frac{N_{ax}}{N_a}$ as the relative proportion and $\text{MI}(a, x) = N\frac{N_{ax}}{N_a N_x}$ as the mutual information. The mutual information metric in particular represents how much more often sites with attribute a quote social media with attribute x compared with how often sites quote social media with attribute x overall.

We investigate **RQ2.1** and **RQ2.2** by computing $p(a, x)$ and $\text{MI}(a, x)$ for $a =$ domains and topics and for $x =$ platforms, handles, and roles. In Tables 10 and 4, we report results for some of the most popular domains and topics in the SocialQuotes data set, hand-picked for diversity across a variety of web topics and news markets. We rank platforms and roles by MI score, reporting all three platforms and only the top-3 roles. We note that for platforms, these statistics are computed across the entire SocialQuotes corpus, as no role annotations are needed. For roles, we use the 117k batch of inferred roles.

Role ↓, Model →	Coin flip	Fewshot	Fewshot + SC	Fewshot + P	CoT	CoT + SC	CoT + P
AUTHORITY	0.16	0.234	0.226	0.246	0.239	0.227	**0.277**
COMMENTER	0.237	0.255	0.257	0.252	0.28	0.277	**0.289**
EXPERT	0.155	0.179	0.178	0.189	0.19	0.186	**0.205**
INFLUENCER	0.371	0.454	0.457	0.45	**0.461**	0.458	**0.461**
MARKETER	0.096	0.236	0.232	0.237	0.282	0.29	**0.344**
SUBJECT	0.548	0.573	0.591	0.499	0.727	**0.732**	0.71
SELF-PROMOTER	0.27	0.562	0.627	0.37	0.629	0.634	**0.721**
WITNESS	0.246	0.286	0.287	0.282	0.302	0.301	**0.307**
Macro Average	0.26	0.347	0.357	0.316	0.389	0.388	**0.414**

Table 2: F1-Scores for PaLM2 on social roles classification with different prompt variants. We also report the expected F1 score under a random coin flip. The best results are highlighted in bold. *Fewshot* corresponds to providing fewshots with no CoT, *CoT* corresponds to providing fewshots with CoT, *SC* corresponds to self-consistency, and *P* corresponds to persistence.

Role ↓, Removed Field →	None	URL	HANDLE	POST URL	SNIPPET
AUTHORITY	0.234	0.253	0.215	0.22	0.307
COMMENTER	0.255	0.294	0.233	0.242	0.255
EXPERT	0.179	0.188	0.184	0.177	0.164
INFLUENCER	0.454	0.455	0.46	0.454	0.348
MARKETER	0.236	0.253	0.187	0.263	0.283
SUBJECT	0.573	0.546	0.532	0.521	0.496
SELF-PROMOTER	0.562	0.296	0.555	0.566	0.626
WITNESS	0.286	0.292	0.279	0.269	0.145
Macro Average	0.347	0.322	0.331	0.339	0.328

Table 3: Ablation results for social roles classification. Measuring performance by removing one component at a time.

Results. We report topic-based results in Table 4 and domain-based results in Table 10 (Appendix B). Our main observation is the following: both the topics and the domains we list can be roughly split into **culture**-oriented topics/-domains and **reporting**-oriented topics/domains, and these groups tend to favor quotes from certain platforms and certain roles. In particular:

- Culture-related topics (e.g. /Beauty, /Shopping, /Food & Drink, /Travel, /Celebrities) and domains (e.g. allure.com, upworthy.com) tend to quote preferentially from Instagram and TikTok, and preferentially from Marketer, Influencer, and Self-Promoter roles. This follows the intuition that culture-related web pages should desire social media quotes from accounts that blog about new fashion, art, and culinary trends, which tend to occur more on Instagram/TikTok and assume commercial/Influencer-type roles.
- On the other hand, reporting-related topics (e.g. /News, /Health) and domains (e.g. sportingnews.com, cnn.com, finance.yahoo.com) tend to quote preferentially from Twitter and from Authority, Commenter, Expert, and Witness roles. This aligns with research that Twitter often functions as a news site for many users (Kwak et al. 2010; Porter 2023), with the aforementioned roles microblogging about their areas of expertise/authority, or about their day-to-day experiences.

Our second observation is simpler, yet striking: almost all of these domains and topics quote from all three platforms. This supports one of our basic motivations, which is that social media quoting is now a cross-platform phenomenon. We release the SocialQuotes data set to the community so that insights such as the above (and beyond) can be studied further.

Discussion

In this paper, we introduced a new NLP paradigm for classifying social media posts that appear in web contexts. We drew an equivalence between the social media embedding process and quoting, using a formal model of quotation, and we identified a latent taxonomy of *roles* that social media posts play when they are quoted on the web. We built and released SocialQuotes, a dataset of $\sim$32M social media embeddings from $\sim$12M web pages. With carefully-designed experiments on SocialQuotes, we showed that roles can indeed be predicted from the web page context, and that the distribution of platforms, accounts, and roles over web domains can be profitably studied with our framework. We now discuss limitations of our work, future directions, and potential use-cases for SocialQuotes.

Limitations

There are three overarching limitations of our work, which we discuss below.

Coverage and Bias of SocialQuotes. Our dataset is limited by our coverage of social platforms and the reach of Common Crawl, among other factors:

Topic ID	Platforms	p	MI	Roles	p	MI
/News/Sports News	Twitter	1E+0	1.2	Authority	8E-1	5.3
	Instagram	4E-2	0.2	Commenter	9E-1	5.1
	TikTok	9E-4	0.1	Expert	8E-1	4.9
/Arts & Entertainment/Celebrities & Entertainment News	Instagram	4E-1	1.9	Influencer	8E-1	5.3
	TikTok	1E-2	1.0	Witness	2E-1	5.2
	Twitter	6E-1	0.8	Commenter	9E-1	5.1
/Beauty & Fitness/Fashion & Style	Instagram	8E-1	4.1	Marketer	4E-1	11.9
	TikTok	2E-2	1.5	Influencer	9E-1	5.9
	Twitter	2E-1	0.2	Subject	8E-1	5.3
/News/Politics	Twitter	1E+0	1.3	Authority	8E-1	5.5
	TikTok	1E-3	0.1	Commenter	9E-1	5.0
	Instagram	1E-2	0.1	Expert	9E-1	4.9
/Shopping/Apparel	Instagram	7E-1	3.4	Marketer	6E-1	17.9
	TikTok	3E-2	1.8	Influencer	9E-1	5.6
	Twitter	3E-1	0.4	Self-Prom.	5E-1	5.3
/Travel & Transportation/Tourist Destinations	Instagram	7E-1	3.7	Marketer	3E-1	9.6
	TikTok	2E-2	1.6	Witness	3E-1	6.9
	Twitter	3E-1	0.3	Self-Prom.	5E-1	5.2
/Food & Drink/Cooking & Recipes	TikTok	6E-2	3.8	Marketer	6E-1	18.9
	Instagram	7E-1	3.7	Self-Prom.	6E-1	6.0
	Twitter	2E-1	0.3	Influencer	8E-1	5.3
/Health/Public Health	Twitter	1E+0	1.2	Authority	8E-1	5.7
	TikTok	3E-3	0.2	Expert	9E-1	5.0
	Instagram	3E-2	0.1	Subject	8E-1	5.0

Table 4: Table of co-occurrence statistics for topics vs. platforms and roles. Note that the top-3 roles and the (only) 3 platforms do not correspond to each other: roles and platforms are independently ranked by MI.

1. SocialQuotes is derived only from three Common Crawl snapshots in the year range 2022-2023, out of hundreds of such snapshots spanning over a decade.
2. SocialQuotes only covers English-language websites, and only three platforms (Twitter, TikTok, and Instagram).
3. Our approach likely does not cover all social media quote HTML signatures for the three platforms.
4. Common Crawl does not visit web sites that were paywalled at crawl time, and likewise our annotators could not annotate sites that were paywalled at annotation time (nor can they rate SM posts that have become private since the crawl).

Due to the above caveats, our empirical results from SocialQuotes hold only for a subset of the web and a subset of SM quotes, in particular those parts of the web and SM that can be found without logins. We plan to release an updated version of SocialQuotes that expands on some of these dimensions; however, other dimensions such as paywalled sites and other crawl restrictions are strict limitations. Fortunately, our framework is fully-transferable to any given web corpus, including web corpi that can be built through services other than Common Crawl that may be able to access proprietary websites.

Role Taxonomy. While we consider our proposed quote taxonomy to be a well-grounded facet of our framework, covering journalistic roles used by existing analyses of SM quotes in online news, while also extending to other roles, it is by no means the only possible taxonomy, nor is it necessarily complete or useful for every given application. In particular, because our taxonomy was inspired partially by our manual review of Common Crawl websites, and Common Crawl has its own limitations (see above), it is possible we are not aware of certain roles, or better formulations of our taxonomy. Nonetheless, the we believe the taxonomy we chose was valid within our overall framework, and will be useful for practitioners and researchers studying web domain distributions similar to that found in Common Crawl. Our release of SocialQuotes – only a small percentage of which is labelled with our taxonomy – will enable further research into the many possible ways that SM authors are cited by web authors.

Source Trustworthiness. The third limitation of our data set and our overall approach is that some web authors may have minority opinions about appropriate roles for certain social media accounts or posts. In general, there is room for reasonable disagreement when characterizing a quote: whether that quote is a social media post quoted in a website, or a real quote in a news article. However, we see this as a microcosm of larger, unavoidable issues encountered when learning over large-scale web data: knowledge graph entities are rarely discussed consistently, and factuality remains a broad challenge (Augenstein et al. 2024). The SocialQuotes

dataset enables the study of these issues in the social media realm.

Long-term Impact and Use-Cases

We consider three primary potential impacts and use-cases of our framework.

Learning SM Annotations for Retrieval and Analysis. One primary impact of our work is that it unifies social media data from many platforms in a common annotation space, derived from the open web. As we showed in our case study, our data set and role taxonomy allows researchers to study cross-platform patterns over the web, generating empirical results that cover novel intersections of social media populations. By adopting our framework and LLM prompt scheme, digital journalism researchers may automatically annotate large batches of SM on the web, rather than performing the hand-coding involved in the work we discussed earlier in this paper. More broadly, using our paradigm, developers may be able to improve tools that allow analysts, journalists, and researchers to retrieve clusters of social media entities – from multiple platforms at once – that pertain to a specific topic and functional role in web contexts.

Toward Independence From SM APIs. A strong benefit of our framework is that it allows for modeling social media without depending on social platform APIs. Recent restrictions on these APIs (Paresh 2023; Porter 2023; Lawler 2022) have encouraged SM researchers to imagine a "post-API" landscape (Perriam, Birkbak, and Freeman 2020; Tromble 2021; Freelon 2018). Our work provides a foundation for modeling social media entities purely with publicly-available, off-platform data.

In some areas such as influencer detection, this also paves the way for more rigorous experimental design. In many influencer detection studies, influencer labels are derived from the same platform signals that feed into the model. For example, Zheng et al. (2020) suggest that a Twitter account should receive a positive evaluation label for a particular topic if (1) greater than 50% of their Tweets are related to the topic, and (2) their topic-specific Tweets are among the most-retweeted on that topic. However, the unsupervised influencer detection approach introduced in the paper uses both post-topic semantic similarity and the retweet graph as model inputs. Similarly, Mittal et al. (2020) evaluate using platform signals that are explicitly used as features in the detection model. This evaluation strategy (additionally observed in similar studies such as Kim, Lee, and Lee (2016) and Oro et al. (2017)) does not establish a task with meaningful headroom for models to achieve. Our framework and dataset provide a potential solution to this problem by establishing an off-platform source of labels for influencer detection models. Of course, web authors who quote a particular SM may do so because the SM entity has many followers on the platform, in which case the platform signal would be confounding. However, even accounting for this, the follower count is just one potential signal that the web author could be influenced by, making their quote a non-trivial label for an on-platform influencer model (as opposed to a label derived purely from platform signals). Moreover, our role taxonomy opens the door for modeling many other types of social accounts beyond influencers.

Toward Generative Citation. The SocialQuotes data set also exposes pointers (via Common Crawl URLs) to examples of humans writing and reasoning about social media posts in relation to their web content. This can enable the study of the *language* of web quoting and citation, and eventually lead to LLM systems that can weave relevant user-generated content into generated text with natural, purposeful, and well-reasoned exposition. We see this as a speculative yet strongly promising area for future work.

References

Augenstein, I.; Baldwin, T.; Cha, M.; Chakraborty, T.; Ciampaglia, G. L.; Corney, D.; DiResta, R.; Ferrara, E.; Hale, S.; Halevy, A.; et al. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8): 852–863.

Balaji, T.; Annavarapu, C. S. R.; and Bablani, A. 2021. Machine learning algorithms for social media analysis: A survey. *Computer Science Review*, 40.

Balog, K.; Azzopardi, L.; and de Rijke, M. 2009. A language modeling framework for expert finding. *Information Processing & Management*, 45(1): 1–19.

Balog, K.; De Rijke, M.; et al. 2007. Determining Expert Profiles (With an Application to Expert Finding). In *IJCAI*, volume 7, 2657–2662.

Broersma, M.; and Graham, T. 2014. Social media as beat: Tweets as a news source during the 2010 British and Dutch elections. In *Online Reporting of Elections*, 113–129. Routledge.

Bublitz, W. 2015. Introducing quoting as a ubiquitous metacommunicative act. In Arendholz, J.; Bublitz, W.; and Kirner-Ludwig, M., eds., *The pragmatics of quoting now and then*, 1–26. De Gruyter Mouton Berlin.

Cope, J. 2020. Quoting to persuade: A critical linguistic analysis of quoting in US, UK, and Australian newspaper opinion texts. *AILA Review*, 33(1): 136–156.

Dumitrescu, D.; and Ross, A. R. 2021. Embedding, quoting, or paraphrasing? Investigating the effects of political leaders' tweets in online news articles: The case of Donald Trump. *New Media & Society*, 23(8): 2279–2302.

Fernandes, E.; Moro, S.; and Cortez, P. 2023. Data science, machine learning and big data in digital journalism: a survey of state-of-the-art, challenges and opportunities. *Expert Systems with Applications*, 221: 119795.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5): 378.

Freelon, D. 2018. Computational research in the post-API age. *Political Communication*, 35(4): 665–668.

Gearhart, S.; and Kang, S. 2014. Social media in television news: The effects of Twitter and Facebook comments on journalism. *Electronic News*, 8(4): 243–259.

Google; and et al. 2023. PaLM 2 Technical Report. arXiv:2305.10403.

Gruppi, M.; Adalı, S.; Salemi, M.; and Horne, B. D. 2021. From Tweeting About News to Creating News Around Tweets: Characterizing Tweets Embedded in News Articles.

Haapanen, L. 2020. Modelling quoting in newswriting: A framework for studies on the production of news. *Journalism Practice*, 14(3): 374–394.

Harry, J. C. 2014. Journalistic quotation: Reported speech in newspapers from a semiotic-linguistic perspective. *Journalism*, 15(8): 1041–1058.

Hombaiah, S. A.; Chen, T.; Zhang, M.; Bendersky, M.; Najork, M.; Colen, M.; Levi, S.; Ofitserov, V.; and Amin, T. 2023. Creator Context for Tweet Recommendation. arXiv:2311.17650.

Kapidzic, S.; Neuberger, C.; Frey, F.; Stieglitz, S.; and Mirbabaie, M. 2022. How News Websites Refer to Twitter: A Content Analysis of Twitter Sources in Journalism. *Journalism Studies*, 23(10): 1247–1268.

Kim, D.; Lee, J.-G.; and Lee, B. S. 2016. Topical influence modeling via topic-level interests and interactions on social curation services. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, 13–24. IEEE.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, 591–600.

Lawler, R. 2022. Meta reportedly plans to shut down CrowdTangle, its tool that tracks popular social media posts.

Lin, S.; Hong, W.; Wang, D.; and Li, T. 2017. A survey on expert finding techniques. *Journal of Intelligent Information Systems*, 49: 255–279.

Mittal, D.; Suthar, P.; Patil, M.; Pranaya, P.; Rana, D. P.; and Tidke, B. 2020. Social network influencer rank recommender using diverse features from topical graph. *Procedia Computer Science*, 167: 1861–1871.

Mujib, M. I.; Heidenreich, H. S.; Murphy, C. J.; Santia, G. C.; Zelenkauskaite, A.; and Williams, J. R. 2020. NewsTweet: a dataset of social media embedding in online journalism. *arXiv preprint arXiv:2008.02870*.

Mujib, M. I.; Zelenkauskaite, A.; and Williams, J. R. 2022. Which tweets deserve to be included in news stories? Chronemics of tweet embedding. *arXiv preprint arXiv:2211.09185*.

Myers, C.; and Hamilton, J. F. 2014. Social Media as Primary Source: The narrativization of twenty-first-century social movements. *Media History*, 20(4): 431–444.

Oro, E.; Pizzuti, C.; Procopio, N.; and Ruffolo, M. 2017. Detecting topic authoritative social media users: a multilayer network approach. *IEEE Transactions on Multimedia*, 20(5): 1195–1208.

Panchendrarajan, R.; and Saxena, A. 2023. Topic-based influential user detection: a survey. *Applied Intelligence*, 53(5): 5998–6024.

Paresh, D. 2023. Reddit Is Already on the Rebound (https://www.wired.com/story/reddit-is-already-on-the-rebound/, accessed 2023-12-04).

Pei, S.; Wang, J.; Morone, F.; and Makse, H. A. 2020. Influencer identification in dynamical complex systems. *Journal of complex networks*, 8(2): cnz029.

Perriam, J.; Birkbak, A.; and Freeman, A. 2020. Digital methods in a post-API environment. *International Journal of Social Research Methodology*, 23(3): 277–290.

Porter, J. 2023. Twitter announces new API pricing, posing a challenge for small developers.

Rony, M. M. U.; Yousuf, M.; and Hassan, N. 2018. A large-scale study of social media sources in news articles. *arXiv preprint arXiv:1810.13078*.

Tekir, S.; Güzel, A.; Tenekeci, S.; and Haman, B. 2023. Quote Detection: A New Task and Dataset for NLP. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 21–27.

Tromble, R. 2021. Where have all the data gone? A critical reflection on academic digital research in the post-API age. *Social Media + Society*, 7(1).

Vaucher, T.; Spitz, A.; Catasta, M.; and West, R. 2021. Quotebank: a corpus of quotations from a decade of news. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 328–336.

Vrandečić, D.; and Krötzsch, M. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10): 78–85.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-consistency improves chain of thought reasoning in language models. In *2023 International Conference on Learning Representations (ICLR)*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Wen, H.; Xiao, Z.; Hovy, E.; and Hauptmann, A. G. 2023. Towards Open-Domain Twitter User Profile Inference. In *Findings of the Association for Computational Linguistics: ACL 2023*, 3172–3188.

Zelizer, B. 1989. 'Saying' as collective practice: Quoting and differential address in the news. *Text-Interdisciplinary Journal for the Study of Discourse*, 9(4): 369–388.

Zelizer, B. 1995. Text, talk, and journalistic quoting practices. *Communication Review (The)*, 1(1): 33–51.

Zheng, C.; Zhang, Q.; Young, S.; and Wang, W. 2020. On-demand influencer discovery on social media. In *Proceedings of the 29th ACM international conference on information & knowledge management*, 2337–2340.

Paper Checklist

1. For most authors...

(a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, propose an approach to social media modeling that relies only on publicly-available data from the Common Crawl.

(b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, we believe they do.

(c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, throughout the work.

(d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, we give a thorough description of our dataset.

(e) Did you describe the limitations of your work? Yes, our final discussion section includes a subsection devoted to multiple limitations.

(f) Did you discuss any potential negative societal impacts of your work? Yes, we address the societal limitations of our dataset in the discussion section.

(g) Did you discuss any potential misuse of your work? Yes, we address the societal limitations of our dataset in the discussion section.

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, we discuss mitigation strategies used in other areas of language modeling in our discussion.

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? our study does not involve hypothesis testing

(b) Have you provided justifications for all theoretical results? our study does not involve hypothesis testing

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? our study does not involve hypothesis testing

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? our study does not involve hypothesis testing

(e) Did you address potential biases or limitations in your theoretical framework? our study does not involve hypothesis testing

(f) Have you related your theoretical results to the existing literature in social science? our study does not involve hypothesis testing

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? our study does not involve hypothesis testing

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? our study does not involve theoretical proofs

(b) Did you include complete proofs of all theoretical results? our study does not involve theoretical proofs

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, we release a dataset, provide explicit model prompts, and provide necessary codepaths in the paper to reproduce the dataset.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, in our experimental section.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No, due to costs, however we will continue to run experiments and will be happy to provide them.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, we listed the cloud provider.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, in our experiment section.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? No, this question is not relevant to our study.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? We cite the creators of the Common Crawl.

(b) Did you mention the license of the assets? We refer readers to the Common Crawl repository, which contains a clear an unambiguous license.

(c) Did you include any new assets in the supplemental material or as a URL? Yes

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? n/a

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, we filtered the dataset for offensive and sensitive content, as discussed in the Appendix. Our dataset contains no PII.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes, we provide detailed information about the dataset in the paper that we believe satisifies most if not all FAIR principles. We are happy to elaborate further in the camera-ready version.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? Yes, it is attached in the supplement.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? Yes, they are in the Appendix and in dataset section.
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? No IRB required for our study.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes, they are in an Ethics statement contained in the dataset section.
   (d) Did you discuss how data is stored, shared, and deidentified? The dataset is not yet stored or shared. It does not contain PII.

Appendix A: Additional dataset information

We provide additional information about the SocialQuotes data set and the human annotation experiment.

Pipeline Details

We provide Tables 5 and 6 here to supplement our description of the pipeline in the SocialQuotes data set section.

Data Filtering

To avoid releasing URLs that contain or point to sites with inappropriate or sensitive content, we apply two rules. First, we do not release any SocialQuotes entry that has topic classifications within the following set:

- /Adult
- /Sensitive Subjects/Accidents & Disasters
- /Sensitive Subjects/Death & Tragedy
- /Sensitive Subjects/Firearms & Weapons
- /Sensitive Subjects/Recreational Drugs
- /Sensitive Subjects/Self-Harm
- /Sensitive Subjects/Violence & Abuse
- /Sensitive Subjects/War & Conflict
- /Sensitive Subjects/Other

Second, we filter any URL that contains any word from the "List of Dirty, Naughty, Obscene or Otherwise Bad Words"[12]. This list contains single words and multi-word tuples. Specifically, we filter any SocialQuotes entry with a URL that satisfies either of the following two conditions:

1. Any single word is equivalent to any word token in the URL. The URL is word-tokenized by splitting on "-", "_", and "." characters.
2. Any multi-word tuple joined by a "-" or a '-' character appears in the URL string.

[12]https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words

Human Annotation User Interface

We provide Figures 3 and 4 showing examples of the user interface that human annotators used to provide role annotations on social quotes.

Role Annotator Agreement

We provide a modified version of Fleiss' κ (Fleiss 1971) to measure role annotator agreement. The Fleiss statistic has the general form $(p_o - p_e)/(1 - p_e)$, where p_o is the average annotator agreement, and p_e is the average annotator agreement under a random null model where each annotator chooses a label with probability proportional to the ground-truth proportion of the label. The original metric assumes that each annotator provides an annotation for every example. In our case, each annotator provides an annotation for only a subset of the examples. Taking this into account, we modify the original score by taking into account the number of annotations per example:

$$p_o := \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i(n_i - 1)} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1) \qquad (1)$$

Above, N is the number of valid annotations (see Annotation Aggregation section), k is the number of classes (here $k = 2$ uniformly across the eight roles), and n_{ij} is the number of positive ($j = 1$) or negative ($j = 2$) annotations. Using this modified formula, we obtain the following Fleiss κ metrics across roles in Table 7.

Ground-Truth Role Overlap

In Figure 5 we show a confusion matrix of the ground-truth roles. Specifically, each matrix element is the proportion of time the two roles co-occurred across the ground-truth quote annotations. We find that the Influencer and Subject roles co-occur the most frequently, by a wide margin. Based on manual reviews of our SocialQuotes, we hypothesize that this is due to the wide number of blog/news articles that are written directly about a celebrity who is also a popular voice on a certain subject.

"Other" and "Not Found" Annotations

Out of all annotations on quotes that could be located at annotation time, the "Other" option was selected only 289 times. Table 7 counts, for each role r, the number of times r was selected along with "Other" (by the same annotator that selected "Other"). The "Other" annotations were not used in our experiments, however they are provided with SocialQuotes as auxiliary information.

Figure 6 shows the distribution of the number of "Not-Found" annotations per-quote, out of 8,281 annotated quotes not proximal to sensitive content. The majority of quotes were able to be located by all or none of the annotators. The quotes that were found only by some annotators may be (1) from websites that had locale-specific paywalls/functionality, or (2) located in parts of pages that were difficult to find for some annotators.

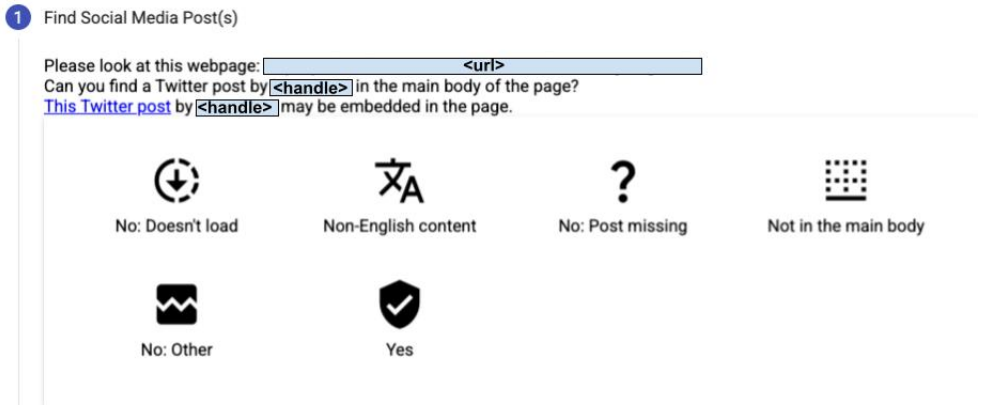

Figure 3: The first question in the human annotation user interface. Annotators were asked if they could actually find the social quote in a current-day live rendering of the web page.

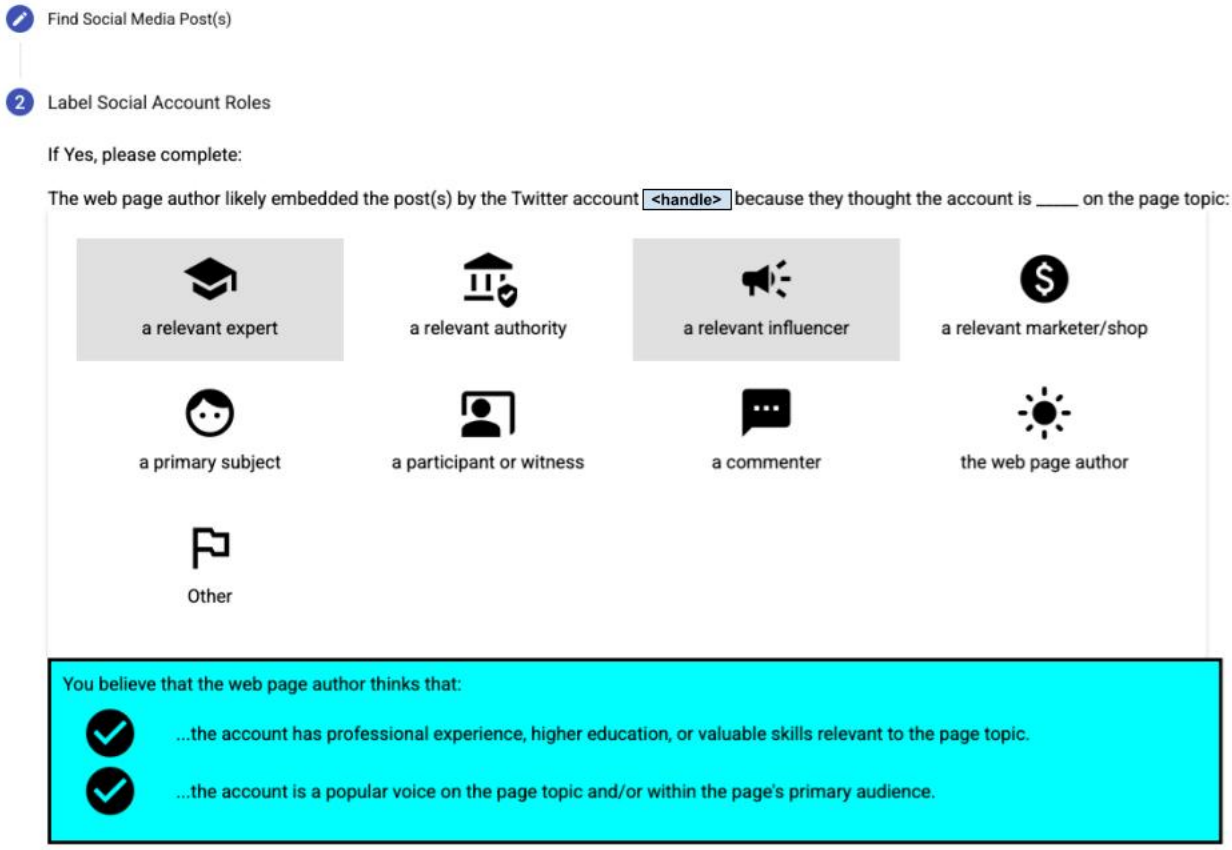

Figure 4: The second question in the human annotation user interface. Once it was determined that the annotator could find the social quote in the page, the annotator was asked to select any number of roles they thought reflected the web author's view on the social media post. When they selected the options, short snippets appeared reminding the annotator of the role definition, which are similar to our definitions laid out in Table 1.

Field	Description
`id`	Hash uniquely identifying the (embedded post, URL) pair.
`common_crawl_snapshot`	The Common Crawl crawl identifier[11] from which the SM embedding was extracted.
`common_crawl_url`	URL in Common Crawl dataset.
`social_post_url`	The URL of the embedded SM post.
`social_user_profile`	The URL of the social media account that created the embedded SM post.
`context_topics`	Up to three highest-probability topics identified in the page surrounding the SM embedding.
`role_labels`	For a subset of quotes, each social role label identified by crowd compute raters, along with the frequency with which it was selected.

Table 5: SocialQuotes dataset schema. Topics are identified via https://cloud.google.com/natural-language/docs/classifying-text.

Platform	Tag Type	Tag Class	Quotes
Instagram	`div`	`InstagramEmbedContainer`	7.6M
	`blockquote`	`instagram-media`	
TikTok	`blockquote`	`tiktok-embed`, `tiktok_lazy_shortcode`	514K
Twitter	`blockquote`	`twitter-tweet`, `twitter-video`, `tweet-blockquote`, `twittertweet`	24.4M

Table 6: Per-platform HTML patterns that we used to identify SM embeddings in Common Crawl source data, and the resulting embedding counts. "Tag Type" refers to the type of HTML tag that encloses the embedding. "Tag Class" refers to the class(es) of tags with the tag type that indicate that the tag contains a SM embedding. "Quotes" column contains embedding count.

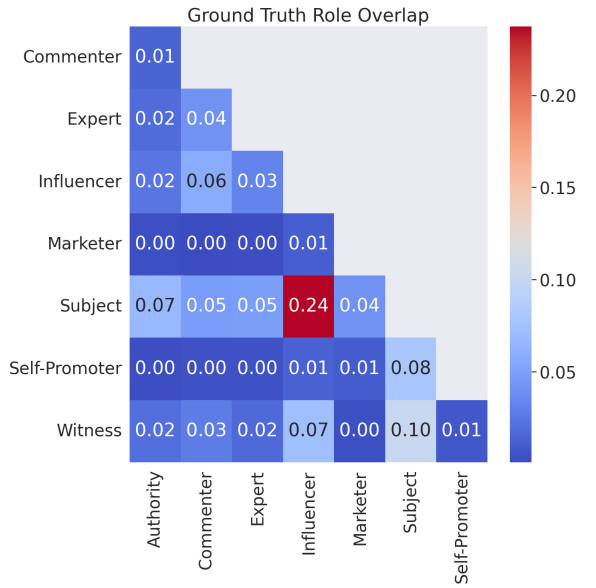

Figure 5: Proportion of times each role pair co-appeared for quote with a valid ground-truth annotation.

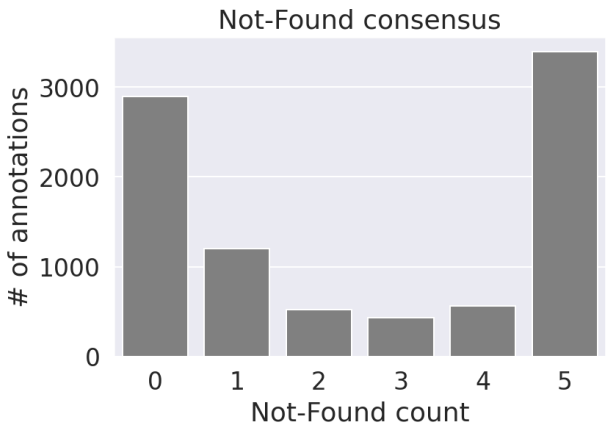

Figure 6: Number of annotations in which k number of annotators could not find the quote ($k \in \{0, \ldots, 5\}$).

Role	Fleiss' κ	"Other" overlap
AUTHORITY	0.226	0
COMMENTER	0.319	13
EXPERT	0.161	2
INFLUENCER	0.095	28
MARKETER	0.211	6
SUBJECT	0.266	15
SELF-PROMOTER	0.572	7
WITNESS	0.125	8

Table 7: Additional measurements on SocialQuotes. Fleiss' κ computes annotator agreement across all eight roles. The final column shows the number of times, for each role r, that an annotator indicated both "Other" and r for the same quote. For comparison, "Other" was chosen only 289 times in our annotation experiment, accounting for 1.4% of all individual annotator responses. This means that in 210 instances ($\sim$1%), "Other" was chosen by itself.

Appendix B: Additional experiment information

We provide additional information about our experiments in the "Experiments and analysis" section.

Prompt Design

Below we provide the main prompt structure for PaLM-2, as well as a Chain-of-Thought example for the "Commenter" role. In Table 9, we provide the complete list of prompts given to the `<role_specific_binary_prompt>` field in the PaLM2 prompt structure presented in the "Role inference" section. We note that these prompts differ slightly in wording from the definitions given in Table 1, though the core meaning is the same for each role.

Domain-Based Platform and Role Analysis

We provide Table 10 as additional insight into the SocialQuotes data set. Our analysis of the results in this table can be found in the "Social quotes analysis" section of the main text.

PaLM2 CoT example for the "Commenter" role.

```
The primary focus of the snippet
is the Pittsburgh Steelers
football team. The post is from a
fan of the team commenting on one
of the players. So the embedded
post is from someone commenting on
the topic of the webpage.
```

PaLM2 prompt structure

```
You are a social media analyst
looking at social media posts
embedded in websites. Given the
following information:

URL: The URL of a webpage with an
embedded social media post;
POST_URL: The URL of the embedded
social media post;
HANDLE: The social media username
of the author of the embedded
post;
SNIPPET: The webpage text that
appears around the embedded post;

Your job is to determine if <role_-
specific_binary_prompt>. Below are
are some examples:

<fewshot_examples>

URL: <common_crawl_url>
POST_URL: <social_post_url>
HANDLE: <extracted_profile_handle>
SNIPPET: <quote_context>
```

Model →	Fewshot		Fewshot + SC		Fewshot + P		CoT		CoT + SC		CoT + P	
Role ↓	R	P	R	P	R	P	R	P	R	P	R	P
AUTHORITY	0.9	0.13	0.9	0.13	0.86	0.14	0.98	0.14	0.98	0.13	0.95	0.16
COMMENTER	0.66	0.16	0.68	0.16	0.6	0.16	0.96	0.16	0.98	0.16	0.92	0.17
EXPERT	0.9	0.1	0.93	0.1	0.85	0.11	0.89	0.11	0.92	0.1	0.8	0.12
INFLUENCER	0.96	0.3	0.99	0.3	0.9	0.3	0.96	0.3	0.97	0.3	0.91	0.31
MARKETER	0.53	0.15	0.54	0.15	0.48	0.16	0.79	0.17	0.85	0.17	0.71	0.23
SUBJECT	0.47	0.73	0.49	0.74	0.38	0.75	0.82	0.66	0.84	0.65	0.73	0.69
SELF-PROMOTER	0.46	0.72	0.54	0.76	0.24	0.79	0.81	0.52	0.86	0.5	0.67	0.78
WITNESS	0.49	0.2	0.49	0.2	0.46	0.2	0.85	0.18	0.89	0.18	0.75	0.19
Macro Average	0.67	0.31	0.69	0.32	0.59	0.33	0.88	0.28	0.91	0.27	0.8	0.33

Table 8: Precision and recall for various models on SocialQuotes.

Role	Prompt Question
Self-Promoter	`Your job is to determine if the embedded post was created by the same entity who created the webpage (a self-promotion).`
Primary-Subject	`Your job is to determine if the embedded post is from someone who is the primary entity being discussed in the webpage.`
Expert	`Your job is to determine if the embedded post is from someone who has recognized expertise in the main topic of the webpage.`
Commenter	`Your job is to determine if the embedded post is from someone who is commenting on the main topic of the webpage.`
Influencer	`Your job is to determine if the embedded post is from someone who is a popular voice on the main topic of the webpage.`
Witness-Participant	`Your job is to determine if the embedded post is from someone who witnessed or directly participated in an event discussed in the webpage.`
Authority	`Your job is to determine if the embedded post is from someone who is a recognized public figure or institution relevant to the webpage content.`
Marketer	`Your job is to determine if the webpage is marketing or advertising a product mentioned in the embedded post.`

Table 9: Role-specific prompt questions given to PaLM2 in the preamble.

Domain Name	Platforms	p	MI	Roles	p	MI
sportingnews.com	Twitter	1E+0	1.2	Authority	9E-1	6.0
	Instagram	2E-2	0.1	Commenter	1E+0	5.4
	TikTok	9E-4	0.1	Expert	9E-1	5.3
cnet.com	Twitter	1E+0	1.2	Witness	3E-1	5.6
	TikTok	2E-2	1.1	Commenter	1E+0	5.2
	Instagram	3E-2	0.2	Influencer	7E-1	4.4
slate.com	Twitter	1E+0	1.2	Subject	9E-1	5.9
	TikTok	1E-2	0.6	Expert	9E-1	5.4
	Instagram	3E-2	0.1	Authority	8E-1	5.3
euronews.com	Twitter	8E-1	1.1	Witness	3E-1	6.0
	Instagram	1E-1	0.8	Authority	8E-1	5.4
	TikTok	4E-3	0.3	Expert	8E-1	4.8
allure.com	Instagram	1E+0	5.1	Marketer	1E+0	30.1
				Subject	1E+0	6.4
				Influencer	1E+0	6.4
vulture.com	Twitter	8E-1	1.1	Self-Prom.	7E-1	6.5
	Instagram	2E-1	0.8	Subject	1E+0	6.2
	TikTok	4E-3	0.2	Authority	8E-1	5.6
cnn.com	Twitter	9E-1	1.1	Subject	9E-1	5.6
	Instagram	1E-1	0.5	Expert	1E+0	5.5
	TikTok	8E-3	0.5	Authority	8E-1	5.4
finance.yahoo.com	Twitter	9E-1	1.1	Subject	9E-1	5.6
	Instagram	1E-1	0.7	Influencer	9E-1	5.5
				Expert	9E-1	5.4
techcrunch.com	Twitter	1E+0	1.2	Authority	9E-1	5.9
	TikTok	9E-3	0.6	Influencer	9E-1	5.5
	Instagram	1E-2	0.0	Commenter	1E+0	5.2
upworthy.com	TikTok	7E-1	46.8	Witness	3E-1	6.9
	Twitter	3E-1	0.4	Marketer	2E-1	5.0
	Instagram	2E-2	0.1	Self-Prom.	5E-1	4.9

Table 10: Table of co-occurrence statistics for domains vs. platforms and roles. Note that the top-3 roles and the (only) 3 platforms do not correspond to each other: roles and platforms are independently ranked by MI.