DIANES: A DEI Audit Toolkit for News Sources

Xiaoxiao Shang[1], Zhiyuan Peng[1], Qiming Yuan[1], Sabiq Khan[1], Lauren Xie[1], Yi Fang[1*]
Subramaniam Vincent[2*]
[1]Department of Computer Science and Engineering
[2]Markkula Center for Applied Ethics
Santa Clara University, California, USA
{xshang,zpeng,qyuan2,skhan2,lxie,yfang,svincent}@scu.edu

ABSTRACT

Professional news media organizations have always touted the importance that they give to multiple perspectives. However, in practice, the traditional approach to all-sides has favored people in the dominant culture. Hence it has come under ethical critique under the new norms of diversity, equity, and inclusion (DEI). When DEI is applied to journalism, it goes beyond conventional notions of impartiality and bias and instead democratizes the journalistic practice of sourcing – who is quoted or interviewed, who is not, how often, from which demographic group, gender, and so forth. There is currently no real-time or on-demand tool in the hands of reporters to analyze the persons they quote. In this paper, we present DIANES, a DEI Audit Toolkit for News Sources. It consists of a natural language processing pipeline on the backend to extract quotes, speakers, titles, and organizations from news articles in real time. On the frontend, DIANES offers the WordPress plugins, a Web monitor, and a DEI annotation API service, to help news media monitor their own quoting patterns and push themselves towards DEI norms.

CCS CONCEPTS

• **Information systems** → **Information extraction**; • **Social and professional topics** → **Gender**; **Race and ethnicity**.

KEYWORDS

Diversity, Equity, and Inclusion (DEI); Quote Extraction; Named Entity Recognition; Gender, Race and Ethnicity Prediction

ACM Reference Format:
Xiaoxiao Shang[1], Zhiyuan Peng[1], Qiming Yuan[1], Sabiq Khan[1], Lauren Xie[1], Yi Fang[1] and Subramaniam Vincent[2]. 2022. DIANES: A DEI Audit Toolkit for News Sources. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3477495.3531660

*Corresponding authors

SIGIR '22, July 11–15, 2022, Madrid, Spain.

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531660

1 INTRODUCTION

Diversity, equity, and inclusion (DEI) are fundamental to promoting robust journalism that supports a healthy society, by fostering well-researched, complex stories that explore different perspectives and voices. As the world is becoming more diverse, it is the news media's responsibility to reflect this. Consequently, sourcing in news is crucial since the sources that journalists choose to quote in their stories affect whose stories get told, how stories are told, whom the news is for, and what communities are served. Some studies show that the voices of women and minorities are often substantially under-quoted in media stories [3]. Thus, it is important for newsrooms to track who is quoted, how often they are quoted, what are the proportions of people quoted by gender, title, race, ethnicity, community, etc. However, for everyday reporters and editors who are part of large and small newsrooms, there is no daily system to monitor their own quoting patterns and push themselves towards the DEI norms. On the other hand, the recent advances in information retrieval and natural language processing have enabled deeper understanding of document content, which opens opportunities for (semi)-automatic DEI auditing based on computational approaches.

In this paper, we present DIANES, a DIversity Auditor for News Sources. DIANES is a DEI toolkit consisting of an NLP pipeline on the backend to extract quotes, speakers, titles, and organizations from news articles in real time. On the frontend, DIANES supports a variety of user requests on the DEI audit based on the information extracted on the backend by offering the WordPress plugins, a Web monitor, and an annotation API service. The toolkit can help reporters visualize source-diversity proportions (e.g., gender and race) for quotes in their article drafts as well as published pieces. Moreover, DIANES can inform editors of the quoting patterns of all the published articles by their newsrooms or across multiple sites. In addition, the annotation API can extract the relevant information from any news articles for downstream applications that newsrooms or other parties may wish to develop by themselves to support their specific DEI goals. Unlike one-time, manual audits, DIANES provides nearly on-demand feedback to ease barriers to assessing stories' representativeness and it may offer immediate opportunities to fix inequitable reporting. DIANES is currently in test use by several newsrooms. To the best of our knowledge, DIANES is the first toolkit for auditing DEI of news sourcing including gender and race/ethnicity.

2 RELATED WORK

Sources bring credibility and authority to news reports [11]. Traditional news sourcing practices that favor official and dominant

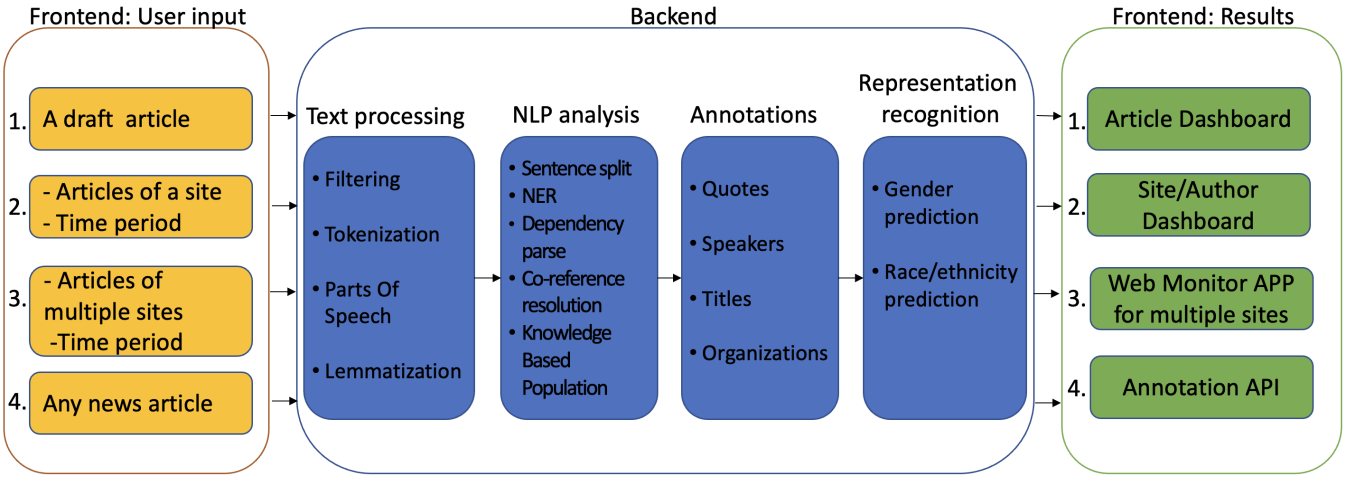

Figure 1: The diagram with the major components of DIANES.

voices have been widely documented over time and across media. The Global Media Monitoring Project in 2010 [7] studied nearly 1,300 newspapers, television and radio stations in 108 countries and found fewer than one in four news subjects were women. Results from the Project for Excellence in Journalism in 2005 [2] obtained similar results. The earlier findings were even more extreme, showing that only 10 percent of sources in newspaper front-page stories were women [4].

Some computational systems have been developed for newsrooms to track the demographics of their sources. With a tool named Dex [6], reporters and editors in National Public Radio (NPR) can submit information about their sources and later pull up reports to monitor their source diversity. The manual approaches are often not scalable. To examine gender bias in media, the Gender Gap Tracker [3] was recently created to tally the number of men and women quoted in news text using natural language processing. However, this system does not support real-time processing of news articles nor cover the important dimension of race and ethnicity in diversity.

3 SYSTEM DESIGN

DIANES supports a variety of user requests which are processed by a natural language processing pipeline on the backend and then results are returned to the frontend. The diagram with the major components of DIANES is shown in Figure 1.

3.1 Frontend

DIANES currently accepts four different types of user requests below:

(1) When writing an unpublished draft article on WordPress, a reporter can request to see all the quotes, speakers, their titles, organizations, gender and race extracted from the draft.
(2) Given all the news articles published by a newsroom in a time period, editors can request to see the gender and race distributions of the news sources (quotes) for a given author or the whole newsroom.
(3) Users can request to see the gender and race distributions across multiple newsrooms/sites.
(4) DEI application developers can access the relevant DEI information extracted from any given article to develop their own interfaces or downstream applications.

For the above requests 1 and 4, users input a single news article, which is wrapped into a JSON file along with an article ID and authentication key. For the requests 2 and 3, user inputs are news archives (often in the XML format) given by different news sites or data providers. We extract relevant information from the archives and store it in a database for subsequent processing.

Given a user request/input, DIANES will return and show the respective results as described below:

3.1.1 Article Dashboard. DIANES provides a WordPress plugin, which works in conjunction with a news article annotation server to identify quotes in the text of stories in real time. After the reporter saves the draft article, they can request the DEI data from the backend. Then we insert all the information in the JSON result returned by the backend into the corresponding tables that are created in the WordPress database. Next we show DEI tables of quotes for reporters to review. DIANES calculates the gender and titled person proportions and uses Google Charts[1] to display these proportions.

3.1.2 Site/Author Dashboard. Site/Author Dashboard is also included in the WordPress plugin. Specifically, we process archives from newsrooms in bulk for bootstrapping DEI data. When data is ready from the backend, we will insert them into the WordPress database. We can then view the relevant data (such as the proportions of genders, races, titled speakers, etc.) in the most recent month by default or in a different month that the user selects. After loading the data into variables, we use Google Charts to display the results on the frontend.

[1] https://developers.google.com/chart

3.1.3 Web Monitor for Multiple Sites. In addition to the WordPress plugin, we have prototyped a web monitor application that offers the same visualizations at the site level for thousands of news sites in the U.S. The web monitor is an application hosted and supported by React, Node.js, Express, and MySQL. We use Node.js and Express as the backend server to connect with the MySQL database, and use React.js for the user interface. This monitor application allows user to access the DEI data visualization for authorized news sites.

3.1.4 Annotation API. The Annotation API is an endpoint for news article analysis through HTTP requests. The API consists a simple Flask web server. The input is a JSON file that contains the article to be analyzed. The web API will return the results in JSON as well. All the quotes and their speakers' names, genders, titles, and community representations are included in the results.

3.2 Backend

The backend module takes news article(s) as input and annotates the quotes and speakers with their titles and organizations, and then predicts gender and race of the speakers. It consists of the following four stages and the first two stages leverage the Stanford CoreNLP library[2] [8].

3.2.1 Text Processing. We first filter some extraneous information from the news articles such as XML tags. After the filtering, the document will be tokenized. The next step is the part of speech (POS) tagging which assigns POS labels to tokens, such as whether they are verbs or nouns, and then lemmatization is applied to map a word to its dictionary form.

3.2.2 NLP Analysis. DIANES conducts sophisticated NLP analysis on every article through the CoreNLP server. It contains the following specific components.

Sentence splitting is the process of dividing text into sentences. CoreNLP splits article text into sentences via a set of rules [8]. Named Entity Recognition (NER) annotator is used to extract person names and organizations by using machine learning sequence models. To extract titles, we added the title pattern file in CoreNLP so that the NER annotator could recognize the titles based on a set of rules.

Dependency parsing analyzes grammatical relations between words in a sentence and extracts textual relations based on the dependencies which are triplets: name of the relation, governor, and dependent [10, 14]. The co-reference resolution finds mentions of the same entity in a text, such as when "Anne" and "she" refer to the same person. Knowledge Base Population (KBP) annotator [1, 16] extracts relation triples meeting the TAC-KBP[3] specifications. We used it to find titles and organizations of a person if presented in the text. Dependency parsing and co-reference resolution are required to extract KBP relations and quotes in the NLP pipeline.

3.2.3 Annotations. In this stage, we produce annotations for quotes, speakers, and their titles and organizations based on the NLP analysis. If a quote has a missing quotation mark, CoreNLP would continue searching for the closing quotation mark and include everything in-between as the quote's content, which may decrease quote resolution accuracy. We address such a situation by looking at the number of words in the quotes. Quotes are considered as long quotes if they contain more than 100 words. On the other hand, quotes with less than 5 words are considered as short quotes and will be dropped since those quotes may increase the mis-resolution rate of speakers and such cases may come from book titles or slogans containing irrelevant information.

For each person in the article, CoreNLP may detect different titles for him or her based on the contexts, where we only record the first title detected for that person. If the speaker of the quotes shares the same last name with other persons, CoreNLP would provide wrong co-reference some times. In such cases, we re-annotate the text from the beginning of the article till the paragraph where the quote is, to ensure the accuracy of speaker detection. We also find that CoreNLP has a high probability of mistaking the speaker based on their approach to processing quotes [9] if the quotes' attribute had some particular patterns. When these patterns occur, we mark the speaker as doubtful in the results.

3.2.4 Representation Recognition. In this stage, we predict the representations of the speakers based on the annotations extracted from the previous stage. DIANES currently supports two important representation attributes of DEI: gender and race/ethnicity, as they are the crucial information for understanding the representativeness of news sourcing.

Approaches to gender tagging in the text have majorly been database-reliant [12]. The key idea behind these approaches is to maintain a database of first names against which occurrences of named entities are compared [5]. We use Gender API[4] as our name-to-gender inference service since this service demonstrated competitive results in the benchmark evaluations [12]. More details can be found in [12] which also introduced other popular gender inference services.

There is much less existing work on predicting the race/ethnicity of a person given a name, and the task is more challenging than gender detection. Thus, we implemented and trained a machine learning model for this task. Our model accepts the name as input, and outputs the probabilities of the predictions which are converted to a confidence score for the predictions. A small confidence score can alert the end users about the potentially inaccurate results. Specifically, our model encodes the name as a sequence of bi-grams, passes them through Bidirectional LSTM [13], connect it with a dense layer, and the softmax layer to produce the probabilities. The model is similar with [15].

The dataset we used to train our race detector was extracted from the United States Census Bureau in 2000 and 2010[5], which contains a total of 151,670 unique names in 6 categories (White, African American, American Indian and Alaska Native, Asian, and Native Hawaiian and Other Pacific Islander). The 2-gram based vocabulary yielded 962 different 2-grams. We trained two models with one for binary classification (White vs. non-White) and another one for six-category classification.

[2] https://stanfordnlp.github.io/CoreNLP/
[3] https://tac.nist.gov/2017/KBP/
[4] https://gender-api.com
[5] https://www.census.gov/topics/population/genealogy/data/2010_surnames.html; https://www.census.gov/topics/population/genealogy/data/2000_surnames.html

Figure 2: Through the WordPress plugin, a reporter can send requests to get the source diversity information when writing a draft news article.

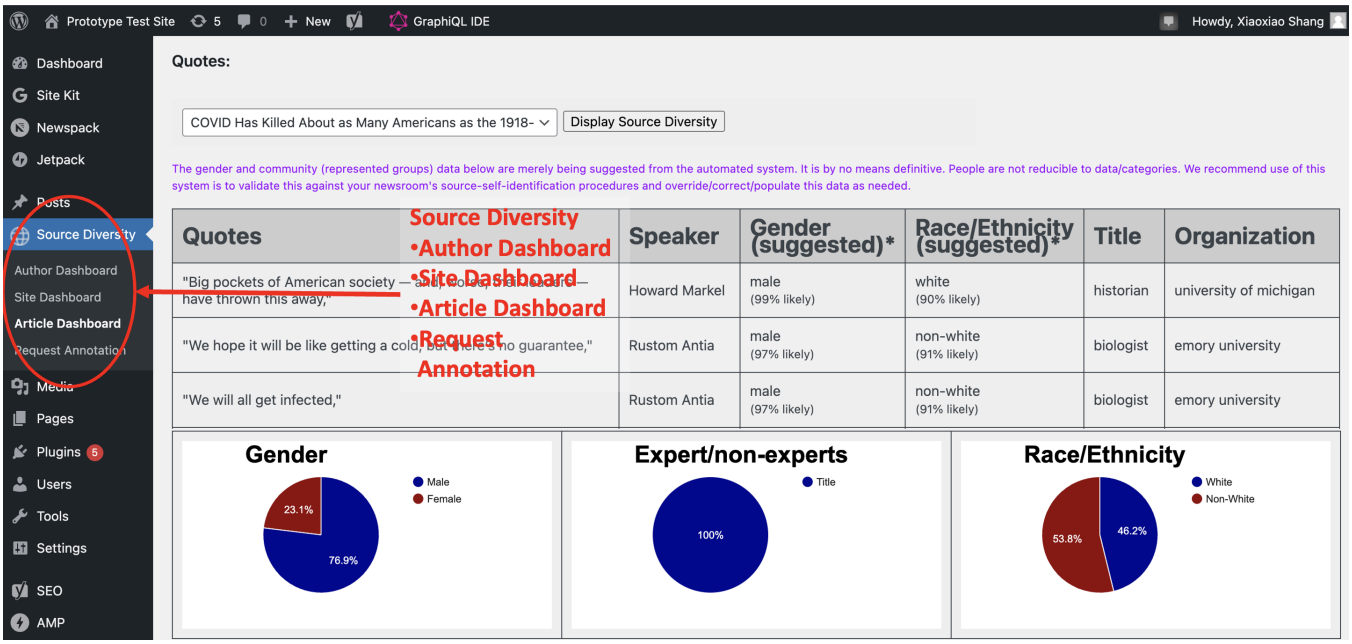

Figure 3: The Article Dashboard shows a table of quotes displayed with DEI data after a reporter or an editor requests annotation of a draft or a published article.

3.3 Evaluation

To evaluate the accuracy of the proposed race prediction model, we split the data of 2 million names collected from the US Census into 72% for training, 8% for development, and 20% for test, which resulted in 160,000 names in the test set. The accuracy for the binary classification (White vs non-White) is 82% and the accuracy for the 6-category classification is 81%.

We also created two datasets to evaluate other backend modules. One is the Mainline news sources with the topics about homelessness, which was collected at San Francisco State University. We randomly chose 20 articles from it. Another dataset is 30 articles randomly sampled from FakeNewsCorpus[6]. We manually labeled these news articles. The numbers of name occurrences in these two datasets are 375 on Mainline and 86 on FakeNewsCorpus, respetively. The numbers of quotes are 409 on Mainline and 95 on FakeNewsCorpus, respectively. The accuracy for the speaker resolution component was 92% on Mainline and 86% on the FakeNewsCorpus source. The accuracy for the title extraction component was 80% on Mainline and 67% on the FakeNewsCorpus source. The accuracy for the Gender API was 92% on the European name corpora based on the published results in the literature [12].

[6]https://github.com/several27/FakeNewsCorpus/releases/tag/v1.0

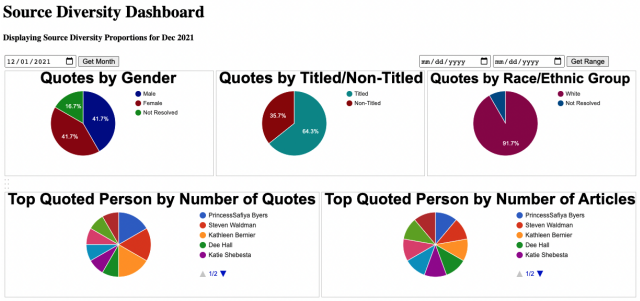

Figure 4: The Site/Author Dashboard displays the DEI annotated data for all the articles published in a specified time period from the news site or from a given author. The display also includes top-quoted persons.

4 DEMONSTRATION

In this section we briefly demonstrate the main functionalities of DIANES. A short video recording of the demonstration can be found here[7].

4.1 Article Dashboard

While a reporter is writing a draft article on WordPress, the reporter can request to get the DEI data from DIANES' backend server by clicking on "Get Source Diversity" as shown in Figure 2. This is called on-demand article annotation. As soon as the backend server completes the processing, the "View article dashboard" button will be enabled. After clicking on the button, a table of results will be shown for each captured quote and its speaker's name, title, organization, and community representation, as demonstrated in Figure 3, with pie charts generated to show the distributions of the DEI attributes of interest.

4.2 Site and Author Dashboards

If a user clicks on "Site Dashboard" in Figure 3, it will bring the user to the source diversity dashboard for the newsroom as shown in Figure 4. Users such as editors can select a month or a time range and get the visualization of the DEI data for all the articles published in that time period by the newsroom. It loads in the most recent month's data by default. Similarly, an author can retrieve the DEI data for the articles he or she wrote in a given time period by using the Author Dashboard. The result visualization is very similar to Figure 4 while it is generated based on the author instead of the whole site. This will give authors an intuitive understanding of their own quoting patterns.

4.3 Web Monitor App

If users want to go beyond a single newsroom, they can use the web monitor that DIANES offers to investigate the DEI data across multiple sites as shown in Figure 5. At full scale, the monitor system will be able to host source-diversity data for up to 7,000 U.S. sites. The resulting dashboard is similar to the one illustrated in Figure 4 with some extra information such as the statistics about the top quoted organizations.

[7]https://www.youtube.com/watch?v=RC5ielXO3Wo

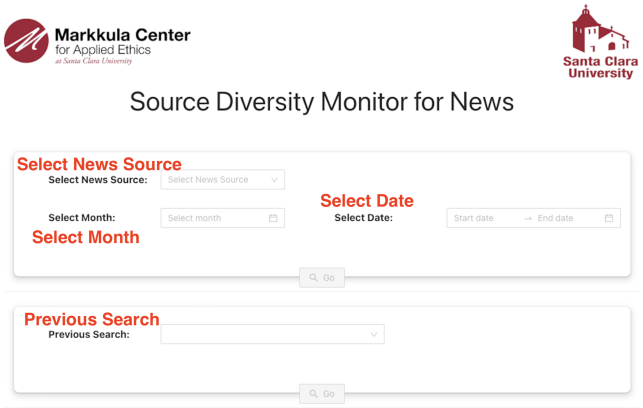

Figure 5: The user interface of the Web Monitor App.

4.4 Annotation API

With the annotation API, users can send a request for any single article, and then receive the structured information about the quotes, speakers, titles, organizations, and community representations. The results are in JSON and can help DEI application developers access the relevant DEI information extracted from any given article to develop their own downstream DEI services and interfaces.

5 CONCLUSION AND FUTURE WORK

DIANES is the first DEI audit toolkit that can analyze news articles and visualize source-diversity proportions for quotes on demand, by leveraging the recent advances of natural language processing and information retrieval. It provides easy-to-use user interfaces for reporters, editors, and DEI practitioners to monitor and track the diversity of the news sourcing.

DIANES is currently tested in the field by several newsrooms. We are improving the toolkit based on the feedback from the users. For example, we will add a feature to allow reporters to override the information extracted and predicted by DIANES when they see fit. This manual correction will not only ensure the high quality of the DEI audit but also provide valuable training data for machine learning models to further improve the performance. In addition, the results of race detection on some minority groups were noticeably worse than those on the majority group due to the limited training data available for the minority races. We will address the imbalanced classification problem by exploring new loss functions and utilizing data augmentation techniques.

ACKNOWLEDGMENTS

This project is supported by Google News Initiative and Facebook Research. Prior funding from News Quality Initiative was used to build the custom CoreNLP-based kernel routines to process news writing and build quote annotations data, which was done by Louise Li and Xuyang Wu at Santa Clara University. We also knowledge news dataset inputs and review from Laura Moorhead, Associate Professor or Journalism, San Francisco State University.

REFERENCES

[1] Gabor Angeli, Victor Zhong, Danqi Chen, Arun Tejasvi Chaganty, Jason Bolton, Melvin Johnson, Panupong Pasupat, S. Gupta, and Christopher D. Manning. 2015. Bootstrapped Self Training for Knowledge Base Population. *Theory and Applications of Categories*.

[2] Claudette G Artwick. 2014. News sourcing and gender on Twitter. *Journalism* 15, 8 (2014), 1111–1127.

[3] Fatemeh Torabi Asr, Mohammad Mazraeh, Alexandre Lopes, Vasundhara Gautam, Junette Gonzales, Prashanth Rao, and Maite Taboada. 2021. The gender gap tracker: Using natural language processing to measure gender bias in media. *PloS one* 16, 1 (2021).

[4] Jane Delano Brown, Carl R Bybee, Stanley T Wearden, and Dulcie Murdock Straughan. 1987. Invisible power: Newspaper news sources and the limits of diversity. *Journalism Quarterly* 64, 1 (1987), 45–54.

[5] Sudeshna Das and Jiaul H Paik. 2021. Context-sensitive gender inference of named entities in text. *Information Processing & Management* 58, 1 (2021).

[6] Angela Fu. 2021. *New tool allows NPR to track source diversity in real time.* https://www.poynter.org/reporting-editing/2021/new-tool-allows-npr-to-track-source-diversity-in-real-time/

[7] Sarah Macharia, Dermot O'Connor, and Lilian Ndangam. 2010. *Who makes the news?: Global media monitoring project 2010.* World Association for Christian Communication Toronto, Canada.

[8] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Baltimore, Maryland, 55–60.

[9] Grace Muzny, Michael Fang, Angel Chang, and Dan Jurafsky. 2017. A two-stage sieve approach for quote attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. 460–470.

[10] Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*. 1659–1666.

[11] Zvi Reich. 2010. Source credibility as a journalistic work tool. In *Journalists, sources, and credibility*. Routledge, 31–48.

[12] Lucía Santamaría and Helena Mihaljević. 2018. Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science* 4 (2018).

[13] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* 45, 11 (1997), 2673–2681.

[14] Sebastian Schuster and Christopher D Manning. 2016. Enhanced english universal dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*. 2371–2378.

[15] Gaurav Sood and Suriyan Laohaprapanon. 2018. Predicting race and ethnicity from the sequence of characters in a name. *arXiv preprint arXiv:1805.02109* (2018).

[16] Yuhao Zhang, Arun Tejasvi Chaganty, Ashwin Paranjape, Danqi Chen, Jason Bolton, Peng Qi, and Christopher D Manning. 2016. Stanford at TAC KBP 2016: Sealing pipeline leaks and understanding chinese. In *TAC*.