Seeds of Discourse: A Multilingual Corpus of Direct Quotations from African Media on Agricultural Biotechnologies

Patricia Chiril* University of Chicago pchiril@uchicago.edu

Trevor Spreadbury* University of Chicago tspread@uchicago.edu

Joeva Sean Rock Stony Brook University joeva.rock@stonybrook.edu

Brian Dowd-Uribe University of San Francisco bdowduribe@usfca.edu

David Uminsky University of Chicago uminsky@uchicago.edu

Abstract

Direct quotations play a crucial role in journalism by substantiating claims and enhancing persuasive communication. This makes news articles a rich resource for opinion mining, providing valuable insights into the topics they cover. This paper presents the first multilingual corpora (English and French) featuring both manually annotated (1,657) and automatically extracted (102,483) direct quotations related to agricultural biotechnologies from a curated list of Africa-based news sources. In addition, we provide 665 instances annotated for Aspect-Based Sentiment Analysis, enabling a fine-grained examination of sentiment toward key aspects of agricultural biotechnologies. These corpora are freely available to the research community for future work on media discourse surrounding agricultural biotechnologies.

1 Introduction

Climate change presents novel challenges and has spurred significant debate regarding how to ensure global food security while maintaining environmental sustainability. One technological tool that animates these debates is agricultural biotechnology. Often referred to as genetically modified organisms (GMO), agricultural biotechnologies are varieties of crops that have been bred using the tools of modern molecular biology to alter a plant's genetic composition. For some, these technologies are a powerful means to address severe events like drought and pest infestation. Others, however, question their connections to large agribusiness conglomerates, arguing that agricultural biotechnologies are good for business, but not necessarily the environment or food sovereignty.

Debates regarding the potential benefits of biotechnology unfold in media outlets. Agribusiness companies have placed ads for their products in newspapers (Nestle, 2019; Glover, 2010), activists have penned op-eds, and investigative journalists have uncovered vast networks of influence amongst agribusiness and non-profits.[1] More recently, initiatives seeking to bring GM crops to African farmers have explicitly developed and funded media organizations as a core part of their work (Rock and Schurman, 2020). This indicates that proponents of GM crops see the media as an essential component of crop development and promotion.

Given the importance of media for both civic debates and the development of agricultural biotechnologies, academics have examined how networks amplify messages around GM crops (Calabrese et al., 2019), asked if and how misinformation spreads within the media (Lynas et al., 2022), and analysed social media networks (Crossland-Marr et al., 2023).

Understanding the use of media in a complex space like agricultural biotechnology is a task well-suited for combining tools of data and social science. Data science allows for the retrieval and analysis of millions of articles, while social science offers the contextual knowledge required to conduct deep analysis. Our main contributions include: (1) the first multilingual corpora (in English and French), featuring both manually annotated (1,657) and automatically extracted (102,483) direct quotations from news media on agricultural biotechnologies, (2) along with 665 instances annotated for Aspect-Based Sentiment Analysis (ABSA) that is freely available to the research community.[2]

2 Related Work

Quotation extraction. Quotations play a key role in news articles by supporting claims and enhanc-

*The first two authors contributed equally to this work.

[1] https://www.theguardian.com/us-news/2024/sep/26/government-funded-social-network-attacking-pesticide-critics

[2] https://github.com/uchicago-dsi/BioMAISx/tree/main.

Findings of the Association for Computational Linguistics: NAACL 2025, pages 8509–8515
April 29 - May 4, 2025

ing persuasive communication. While many works focus on building resources for direct quotations (Pouliquen et al., 2007; O'Keefe et al., 2012; Zhang and Liu, 2022), some also explore indirect and mixed quotations (Pareti et al., 2013; Zhang et al., 2023; Petersen-Frey and Biemann, 2024). In this paper we are focusing on direct quotations due to their traceability and informativeness, which enhance credibility, authority, and transparency (Esser and Umbricht, 2014).

Aspect-Based Sentiment Analysis. The widespread availability of textual data from social media and online news platforms has opened up new opportunities to analyse public sentiment towards various topics (Wankhade et al., 2022). ABSA, in particular, offers a more fine-grained approach by examining sentiment towards different aspects of an entity in a text (Pontiki et al., 2014). However, existing ABSA corpora are mostly limited to domains such as restaurant and e-commerce reviews, highlighting a need for broader coverage (Chebolu et al., 2023). In this work, we build on the corpus proposed by Chiril et al. (2024), which includes 1,553 English language instances pertaining to agricultural biotechnologies.

3 Data and Annotation

3.1 Data Collection

To build our dataset, we relied on a subset of 1.2M articles from the available corpus collected by Chiril et al. (2024). This dataset comprises nearly 2M news articles published over a 26-year period,[3] and were sourced from the Dow Jones premium publication archive using the Factiva Snapshots API.[4] The material was collected using a set of representative keywords (e.g., *crop names*, *organizations involved in the development of GM crops*) and it originates from a diverse range of sources, including non- and for-profit media outlets, as well as government media. We further refine this collection by selecting only articles from a curated list of Africa-based publishers, allowing us to focus specifically on discourse originating from, and disseminating within, the continent. This resulted in a corpus comprising 804,000 English and 300,000 French news articles.

[3]From January 1, 1997, to March 13, 2023.

[4]`https://tinyurl.com/FactivaSnapshotsAPI`.

3.2 Methodology

For the task at hand (i.e., quotation extraction), we experimented with several models that have shown remarkable performance on various NLP tasks (Qiu et al., 2020). To this end, we fine-tune the following models: `distilBERT` (Sanh et al., 2019), `BERT-base-multilingual` (Devlin et al., 2019), and `XLM-RoBERTa` (Conneau, 2019)[5] on the dataset proposed by Zhang and Liu (2022).[6] Given the sequence length limitation of these models, we segment each article into individual paragraphs (based on the presence of two consecutive newline characters), and use these smaller text segments as our test set. While it is possible for a quotation to span multiple paragraphs (e.g., in literary works, academic writing), this is less frequent in news articles, where journalists often break up lengthy quotations by paraphrasing or summarizing a speaker's statements. In this manner, over 5,000 quotations were extracted from articles for labelling.

The main drawback of these models, is their need for large amounts of training data. Inspired by recent advancements in NLP, where Large Language Models (LLMs) have evolved to manage increasingly complex language generation tasks,[7] we include a `LLaMA` model (Dubey et al., 2024)[8] for comparison.

In addition to quotation extraction, our interest extends to attributing these statements to their corresponding speakers and identifying how the speakers are introduced to the reader. Consider the following direct quotation:

> *"You can't build a peaceful world on empty stomachs and human misery."*

This quote takes on additional significance when associated with the speaker, Nobel Peace Prize laureate Dr. Norman Ernest Borlaug (i.e., the identity of the speaker can amplify the impact and urgency of the message conveyed (Gagich et al., 2014)). In simple cases, attribution follows clear syntactic patterns, where the speaker is explicitly mentioned (e.g., SPEAKER *said* QUOTATION). However, more complex scenarios can occur, such as when the

[5]To train the models, we used their HuggingFace PyTorch implementations (Wolf et al., 2019), with default parameters.

[6]This corpus consists of 19,706 text segments extracted from news articles, with manually annotated quotation spans and corresponding speakers.

[7]For a comprehensive overview of LLM capabilities see Guo et al. (2023) and Chang et al. (2024).

[8]`https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`

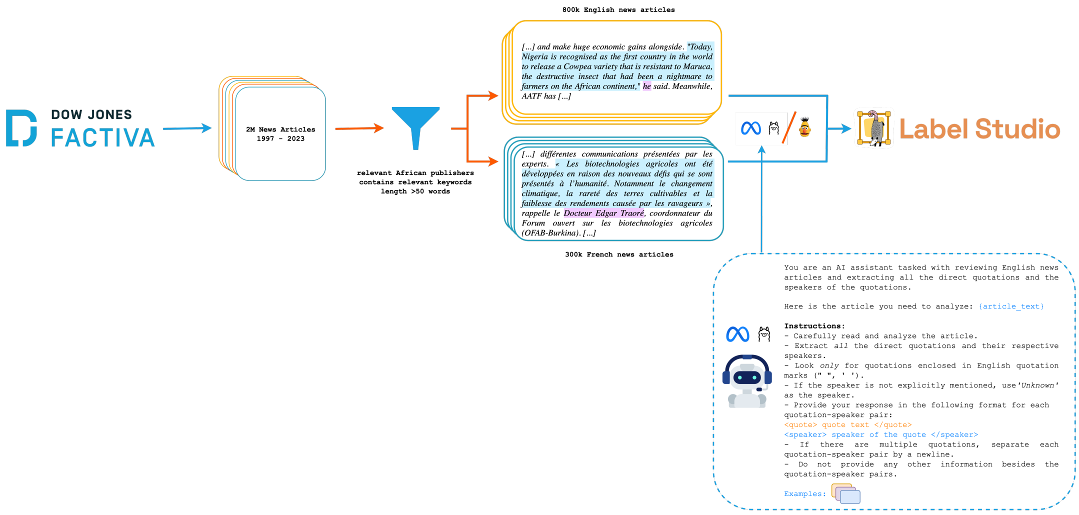

Figure 1: Pipeline for the creation of the corpus. The dashed-line box illustrates the three-shot prompt used for extracting the quotation-speaker pairs from the English articles. A corresponding three-shot prompt, adapted for French, was used for the French-language articles.

speaker is indirectly referenced through pronominal anaphora (e.g., *"he said"*, cf. Figure 1) or nominal anaphora (e.g., *"the farmer said"*). To handle these challenges, we leverage the capabilities of the `LLaMA` model: instead of processing only a single paragraph, we provide the entire article and prompt the model to extract all the direct quotations along with their corresponding speakers, accounting for cases where the reference is indirect or distant. The pipeline used for creating our corpus is presented in Figure 1.

3.3 Annotation Protocol

Assessment of Extracted Quotes' Quality. The primary goal of this annotation task is to evaluate the performance of a suite of models on the task of quotation extraction and attribution. Each instance is evaluated and assigned one of three labels based on the performance of the quotation extraction model: *perfect*, *good* (the extracted span is missing at most two words), and *poor* (the extracted span is missing more than two words, or the model completely failed to capture the quotation). For instances labelled as good or poor, annotators were further required to manually select the correct span. In addition, for assessing the quality of the `LLaMA` model performance on quotation attribution, annotators reviewed 100 (full) articles and evaluated the accuracy of the extracted speakers.[9]

Aspect-Based Sentiment Analysis. To better understand *what* is being discussed and *how*, the main purpose of this annotation task is assigning a polarity (*positive*, *neutral*, *negative*, *conflict*) to each of the aspect categories (entity-attribute pairs) identified within the quotation. To this end, we extend the English corpus (along with the annotation guidelines) proposed by Chiril et al. (2024) to also include a small set of French data. The entity types considered for the task at hand include: *crops, organizations, agricultural practices, natural resources, geographic locations, technology, legal aspects & politics, environmental conditions, economic factors, programs & initiatives*, and *other*. The attributes can be assigned one of the following nine labels: *resistance, consumer perception, safety, food security, productivity, economic impact, research development, environmental & ethical concerns*, and *miscellaneous*.

3.4 Annotation Results

The annotation process was conducted in multiple stages. During the initial pilot phase, we aimed to refine the annotation scheme, and ensure task understanding and annotation consistency. In this phase, we used a total of 150 quotations . Following this stage, the eight annotators,[10] worked on small sets of quotations, with any arising issues be-

[9]Many of the manually labelled paragraphs included speakers referenced indirectly. While transformer models are capable of correctly identifying these speakers, an additional step (i.e., coreference resolution) would have been necessary to resolve the mentions.

[10]Six students (three proficient in English, three in French) and two of the paper's authors/four females and four males.

Aspect Category Entity # Attribute	Polarity	Example
CROPS # CONSUMER PERCEPTION CROPS # PRODUCTIVITY	positive positive	*"the beauty of the bt variety is the quality of the yarn produced from it. It is better and the yield is higher so why would we stick to the old variety that doesn't give the quality and quantity the bt variety gives?"*
LEGAL ASPECTS & POLITICS # SAFETY	positive	*"Text and pictures advertising health warnings shall appear together and shall occupy no less than 75% of the packet display. All tobacco products should conform to the regulations"*
ENVIRONMENTAL CONDITIONS # PRODUCTIVITY ENVIRONMENTAL CONDITIONS # ECONOMIC IMPACT	negative negative	*"Exacerbées par les impacts du conflit, la sécheresse de 2018 et les inondations survenues en 2019 ont été catastrophiques pour les petits exploitants agricoles, les empêchant de cultiver de vastes parcelles de terres ou de constituer des surplus de stocks de semences qui pourraient ensuite être utilisés lors des prochaines plantations. Ce qui a également perturbé les marchés locaux."* *("Exacerbated by the impacts of the conflict, the 2018 drought and the floods of 2019 were catastrophic for smallholder farmers, preventing them from cultivating large areas of land or building surplus seed stocks that could later be used for future planting. This also disrupted local markets.")*
ENVIRONMENTAL CONDITIONS # SAFETY	negative	*"Des pluies torrentielles ont depuis la mi-septembre causé des dégâts considérables dans l'ensemble du pays. Ces inondations sont les pires au Bénin depuis plus d'un siècle."* *("Since mid-September, torrential rains have caused considerable damage throughout the country. These floods are the worst in Benin in over a century.")*
ENVIRONMENTAL CONDITIONS # ECONOMIC IMPACT	negative	*"D'ici 2050, le changement climatique entraînera une augmentation de 25 % du prix des céréales, en comparaison avec un scénario sans le facteur de changement climatique."* *("By 2050, climate change will lead to a 25% increase in cereal prices, compared to a scenario without the climate change factor.")*

Table 1: Examples of annotated quotations in the corpus that reference various aspect categories.

	CROPS	ORGANIZATIONS	AGRICULTURAL PRACTICES	TECHNOLOGY	GEOGRAPHIC LOCATIONS	ENVIRONMENTAL CONDITIONS	LEGAL ASPECTS & POLITICS	ECONOMIC FACTORS	NATURAL RESOURCES	PROGRAMS & INITIATIVES	OTHER
RESISTANCE	33	8	5	10	0	42	1	0	0	1	1
CONSUMER PERCEPTION	42	8	4	1	0	3	0	1	2	0	3
SAFETY	27	12	12	19	2	48	10	2	0	3	6
FOOD SECURITY	53	11	20	2	12	15	11	23	1	5	6
PRODUCTIVITY	101	24	51	38	27	100	9	18	5	12	3
ECONOMIC IMPACT	87	130	99	20	45	57	103	407	28	64	44
RESEARCH DEVELOPMENT	4	9	14	2	1	2	6	2	0	16	2
ENVIRONMENTAL & ETHICAL CONCERNS	11	100	29	15	12	101	72	27	4	18	17
MISCELLANEOUS	17	9	5	4	1	10	15	4	3	8	8

Table 2: Statistics for the aspect categories present in the corpus.

ing discussed on a weekly basis. The labelling tasks varied widely in difficulty. For some labelling tasks, we observed moderate inter-annotator agreement, highlighting the complexity of the task. For others, near perfect agreement is reflective of the straightforward nature of the task. The inter-annotator agreement for a set of 200 quotations, measured in terms of F1-score (a common alternative to Cohen/Fleiss kappa for NER/spans), is 100% and 99.5% for quotation extraction, 83.7% and 70.8% for the identification of entity type, 83.3% and 76.7% for the attribute, and 78.2% and 61.6% for tuples of the form (entity-attribute pair, sentiment), in English and French, respectively. Given the amount of data to label, the students were each assigned distinct subsets of the corpus for annotation, using the open-source platform LabelStudio (Tkachenko et al., 2020-2022). After this phase was completed, the two senior annotators reviewed and corrected the annotated instances. The final corpus consists only of instances where minimum two annotators reached consensus, totalling 1,657 direct quotations (1,299 in English and 358 in French). Of these, 665 (357 in French and 308 in English) quotes were labelled for ABSA, containing 1,719 relevant entities. Table 1 presents examples of the annotation of various aspects within quotations in both English and French, while Table 2 presents the number of annotated instances per aspect category in the corpus. These initial results provide valuable insights into public discourse surrounding agricultural biotechnologies.

3.5 Quantitative Results

For evaluating the quotation extraction model, we rely on two different metrics: *exact match* and an *overlap metric* (Pareti et al., 2013). The annotation procedure revealed that the quotation extraction model perfectly identified quotes in 11.9% of the cases and exhibited good performance in 82.1% of the cases. Table 3 shows the performance for the task (cf. Section 3.2). Surprised by the relatively poor performance of LLaMA, we performed a manual error analysis, and identified several issues: the model occasionally splits quotations, extracts a quotation followed by a substring of said quotation, or, in rare cases, includes mentions (i.e., a scientific term placed in quotes). Based on this analysis, we implemented filters to exclude quotations that are substrings of another quotation, quotations without an identified speaker, and quotations shorter than 4 words.

We then repeat the same experiments on a larger set of articles, including only the quotations on which all models reached consensus in the final dataset, resulting in a total of 102,483 instances (59,501 and 42,982 quotations in French and English, respectively).

Model	Language (# of instances)	
	English (1,302)	French (314)
distilBERT	96.5	–
BERT-base-multilingual	96.6	98.1
XLM-RoBERTa	96.3	98.6
llama-3.1-8B	88.6	90.3

Table 3: Quotation extraction results.

4 Conclusion and Perspectives

In this paper, we have presented the first multilingual corpora which includes direct quotations and ABSA annotations pertaining to agricultural biotechnologies. A model trained on this data, combined with quotation attribution, could enable a comprehensive analysis of media representations of agricultural biotechnologies in Africa, uncovering connections between news outlets and quoted sources, while examining the sentiment expressed in these quotations. Although the broader applications of such analysis lie beyond the scope of this article, we believe that our corpora will be a valuable resource for future work in analysing media representations of GM crops, and exploring the discourse structure surrounding agricultural biotechnologies.

Ethics Statement

The corpora presented in this study is part of a broader initiative in which we aim to harness the tools of social and data science to deliver unique insights into GM crop development and use on the African continent. This work consists of three unique data sets on GM crops related to development, financial support, and now, media coverage. The data is composed of direct quotes and annotations of media articles that were collected and delivered by Factiva, a for-profit global news search engine hosted by Dow Jones. Factiva aggregates content from both non-profit and for-profit organizations, as well as government media outlets. Our team of topic experts assembled a lexicon containing keywords associated with agricultural biotechnologies, which we used to query the Factiva database. We also queried their database based on keywords appearing in the articles or the articles being tagged with relevant industry codes. This strategy allowed us to adopt a broad definition of "biotechnology" rather than relying on Factiva's definitions. We analysed approximately 2 million articles that matched our search criteria and time frame. All our annotations and analyses were conducted in compliance with our agreement with Factiva.

The research did not involve human subjects and was therefore exempt from an institutional review board.

Hiring policy: apart from the authors of this article, other researchers contributed to the project. The annotators were recruited from pools of applicants to a summer data science program and a research assistant role for students at our University. All hired personnel were financially compensated for their work.

Limitations

The corpora presented has several limitations. Direct quotations from published news articles do not capture the full spectrum of news dissemination. This dataset lacks audio/radio news, social media posts, and television news. The time frame is cut off in 2023, leaving out more current events. Furthermore, due to the labour-intensive nature of annotating for Aspect-Based Sentiment Analysis, the dataset is limited in size. It is important to keep these limitations in mind when using the dataset.

Acknowledgments

The authors would like to thank Daniel Grzenda and research assistants at the University of Chicago and San Francisco for their work on the project over the years. The mBio project is supported by grants from the 11th Hour Project (2020–2022), the University of San Francisco, and University of Chicago.

References

Christopher Calabrese, Brittany N Anderton, and George A Barnett. 2019. Online representations of "genome editing" uncover opportunities for encouraging engagement: a semantic network analysis. *Science Communication*, 41(2):222–242.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and Thamar Solorio. 2023. Survey of aspect-based sentiment analysis datasets. In *Proceedings of the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 13th International Joint Conference* *on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics.

Patricia Chiril, Trevor Spreadbury, Joeva Sean Rock, Brian Dowd-Uribe, and David Uminsky. 2024. Biomaisx: A corpus for aspect-based sentiment analysis of media representations of agricultural biotechnologies in africa. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5338–5342.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Lauren Crossland-Marr, Alexandru Giurca, Maya Tsingos, Matthew A Schnurr, Adrian Ely, Dominic Glover, Glenn Davis Stone, and Klara Fischer. 2023. Siloed discourses: a year-long study of twitter engagement on the use of crispr in food and agriculture. *New Genetics and Society*, 42(1):e2248363.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Frank Esser and Andrea Umbricht. 2014. The evolution of objective and interpretative journalism in the western press: Comparing six news systems since the 1960s. *Journalism & Mass Communication Quarterly*, 91(2):229–249.

Melanie Gagich, Emilie Zickel, and Terri Pantuso. 2014. Rhetorical appeals: Logos, pathos, and ethos defined. *WRITING ARGUMENTS IN STEM*, page 34.

Dominic Glover. 2010. The corporate shaping of gm crops as a technology for the poor. *The Journal of Peasant Studies*, 37(1):67–90.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. 2023. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*.

Mark Lynas, Jordan Adams, and Joan Conrow. 2022. Misinformation in the media: global coverage of gmos 2019-2021. *GM Crops & Food*, pages 1–10.

Marion Nestle. 2019. *Food politics: How the food industry influences nutrition and health*. University of California Press.

Tim O'Keefe, Silvia Pareti, James R Curran, Irena Koprinska, and Matthew Honnibal. 2012. A sequence labelling approach to quote attribution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 790–799.

Silvia Pareti, Tim O'keefe, Ioannis Konstas, James R Curran, and Irena Koprinska. 2013. Automatically detecting and attributing indirect quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 989–999.

Fynn Petersen-Frey and Chris Biemann. 2024. Dataset of quotation attribution in german news articles. *arXiv preprint arXiv:2404.16764*.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Bruno Pouliquen, Ralf Steinberger, Clive Best, et al. 2007. Automatic detection of quotations in multilingual news. In *Proceedings of Recent Advances in Natural Language Processing*, volume 2007, pages 487–492. Borovets, Bulgaria.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China technological sciences*, 63(10):1872–1897.

Joeva Rock and Rachel Schurman. 2020. The complex choreography of agricultural biotechnology in africa. *African Affairs*, 119(477):499–525.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771.

Wenjia Zhang, Lin Gui, Rob Procter, and Yulan He. 2023. Newsquote: A dataset built on quote extraction and attribution for expert recommendation in fact-checking. *arXiv preprint arXiv:2305.04825*.

Yuanchi Zhang and Yang Liu. 2022. Directquote: A dataset for direct quotation extraction and attribution in news articles. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6959–6966.