BioMAISx: A Corpus for Aspect-Based Sentiment Analysis of Media Representations of Agricultural Biotechnologies in Africa

Patricia Chiril*
University of Chicago
Chicago, IL, USA
pchiril@uchicago.edu

Trevor Spreadbury*
University of Chicago
Chicago, IL, USA
tspread@uchicago.edu

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

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

David Uminsky
University of Chicago
Chicago, IL, USA
uminsky@uchicago.edu

Abstract

News articles constitute a valuable resource for opinion mining, as they contain important perspectives related to the subject matter they cover. In this paper, we explore how aspect-based sentiment analysis might help in understanding the public discourse surrounding agricultural biotechnologies in Africa. We introduce BioMAISx, the first English language dataset composed of direct quotes pertaining to agricultural biotechnologies extracted from a curated list of Africa-based news sources. We have identified and labelled entities related to key aspects of agricultural biotechnologies, providing valuable insights into public discourse. This dataset can aid in identifying challenges, improving public discourse, and monitoring the perception of agricultural biotechnologies, thus contributing to informed decision-making.

CCS Concepts

• **Computing methodologies** → **Language resources**; • **Applied computing** → **Annotation**; **Anthropology**; **Agriculture**; • **Information systems** → **Sentiment analysis**.

Keywords

aspect-based sentiment analysis, corpus, genetically modified crops, new breeding technologies, media analysis, Africa

ACM Reference Format:
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 (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3627673.3679152

1 Introduction

Genetically modified (GM) crops are one of the most consequential and controversial agricultural biotechnologies of the last century. Since their emergence in the 1980s, people worldwide have debated their efficacy and desirability. One debate in particular has gained significant attention: *whether GM crops can address issues of hunger and poverty in Africa.*

A large group of social scientific studies have examined the impacts of GM crops, and the political debates surrounding their research and development in Africa [18, 20]. More recently, researchers have begun to focus on the important role that media play as a key site for sharing information, and also vying for influence. Here, as elsewhere in the world, the media has become an important arena: advocates have invested significant funds into shaping narratives around GM crops [5], while opponents have also pursued media campaigns to express concerns [17].

The new focus on the role of media has brought a wider diversity of researchers to examine the interplay between the media and understandings of GM crops. A small group of studies have begun to shed light on key questions, including how networks and stakeholders spread and amplify particular messages [1, 4], how media narratives frame GM crops in terms of favorability and misinformation [7, 8, 11, 12], as well as measuring public opinion towards specific gene editing technologies (e.g., CRISPR) [13, 22].

A few social scientific studies of GM crops and media narratives have used large datasets [11, 12], but few, if any, have employed Natural Language Processing approaches. Zooming out to the data sciences, there have been some studies on GM crops, including whether hedge detection can be used to understand the scientific framing in debates on GM organisms [3].

Given the documented political and material interests in using the media to influence GM crop debates, and given that social scientific studies of GM crops and media narratives have yet to harness the analytical power of data science tools, there is a need and opportunity to further investigate media coverage of GM crops from an interdisciplinary, social and data science lens.

The growing abundance of accessible textual data available from social media and online news sites has made it possible for researchers to tackle more in-depth questions such as understanding, measuring, and monitoring the sentiment of users towards certain topics or events [23].

While the current studies in sentiment analysis primarily focus on identifying polarities (positive, negative, neutral), comments and opinions often pertain to a specific target or aspect of interest, and as such, finer-grained tasks can be envisioned. For example, the media

*The first two authors contributed equally to this work.

CIKM '24, October 21–25, 2024, Boise, ID, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679152

coverage of agricultural biotechnologies rarely addresses the field at large and instead is focused on specific issues such as: the role of GM crops in addressing food security; economic implications of growing GM crops (e.g., cost of production, market price, profitability, trade policies); processes that organizations undertake to develop new genetically modified crop varieties; ethical issues surrounding GM crops (e.g., issues of biodiversity, biopiracy, the rights of farmers and indigenous people); etc. Even when limited to a specific target entity, the expressed sentiments can be inconsistent between separate aspects of said entity. For example, consider (1), where the sentiment towards *local rice* is mixed. While it's acknowledged as being more expensive (which could be perceived negatively), the statement emphasizes the importance of buying and consuming local rice for its nutritional value and safety.

(1) *"Local rice is more expensive, but we say even if it means buying half bag, do it. It is better for us to eat a smaller quantity of nutritious rice than for us to take poisonous shiploads of rice."*

In particular, it has become indispensable to analyse such fine-grained polarities concerning different aspects to better understand *what* is being discussed and *how.* This is crucial given that the opinions presented in news media articles concerning certain events/scandals, such as those related to food safety, may have a significant, enduring influence on public opinion and subsequently shape policy decisions [14]. Aspect-Based Sentiment Analysis answers these important questions by identifying the aspects of given target entities and the sentiment expressed for each aspect [16].

As far as we are aware, there is no existing dataset that comprehensively covers the topic of agricultural biotechnologies at a fine-grained level.[1] Additionally, previous research has shown that certain organizations engage in research misconduct in order to influence scientific publications, and thus promote their products seeking regulatory approval[2] [9]. We believe that it is important not only to be able to identify the sentiment associated with different entities related to biotechnologies, but also to establish a link between a statement's source and the publishing news outlet. Our main contributions include:

(1) **BioMAISx (Biotechnology: Media, Agriculture, Investment, (and) Sentiment Excerpts), the first English language dataset** comprised of 1,553 direct quotes **annotated for ABSA on GM crops that is freely available to the research community**.
(2) The data presented in this study is part of a larger 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 [10].

2 Data and Annotation

2.1 Data Collection

Our corpus is new and contains news articles collected between January 1, 1997 and March 13, 2023 from the Dow Jones premium publication archive using the Factiva Snapshots API.[3] Factiva aggregates non- and for-profit media outlets, as well as government media. In order to collect only news articles pertaining to agricultural biotechnologies, we implemented a filtering mechanism that required either the article's title or body to contain at least one keyword from a predefined set of representative keywords (*gmos*, *genetically modified organism*, *agriculture*, *gm crop*, etc.). Thus, we collected around 2M news articles, which were subsequently filtered to include only articles obtained from a curated list of Africa-based publishers so that we could focus specifically on discourse emanating from, and circulating within, the continent. This resulted in a corpus comprising over 804,000 news articles that were subsequently segmented into paragraphs.[4]

Rather than randomly selecting sentences for annotation, we chose to focus on quotations, as they are frequently used in news articles to substantiate claims (thus making them a core element for persuasive communication). Moreover, attributing quoted statements to their sources not only enhances the credibility, authority, or nuance of a statement, but also promotes transparency and accountability for readers by enabling them to identify the cited sources [6]. There are three types of quotations: *direct* (enclosed within quotation marks), *indirect* (paraphrased) and *mixed.* In this paper, we focus on direct quotations, as they are the most traceable and informative type among the three. For extracting the quotes for labelling we fine-tune a distilBERT model[5] [19] on the dataset proposed by Zhang and Liu [26].[6] To focus on quotes related to GM crops, we enforced a requirement that they must include at least one keyword from a manually built lexicon related to agricultural biotechnologies (e.g., *crop names*, *organizations involved in the development of GM crops*). In this manner, a total of 4,932 quotations were extracted from 3,862 articles for labelling. The pipeline used for creating our corpus is presented in Figure 1.

2.2 Annotation Guidelines

2.2.1 *Assessment of Extracted Quotes' Quality.*

Each instance is evaluated and assigned one of three labels based on the performance of the quotation extraction model (cf. Section 2.1): *perfect* (the entire quote was accurately extracted from the provided text), *good* (at most, the extracted text span is missing two words), and *poor* (the model failed to identify more than two words, or completely missed the quote). For instances where the labels good or poor were assigned, the annotators were additionally tasked with manually selecting the quotation span.

2.2.2 *Aspect-Based Sentiment.*

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. In the following, we present the complete

[1] Note that a key limitation of existing ABSA corpora lies in the domain that is being covered (i.e., primarily restaurants and e-commerce reviews). See Chebolu et al. [2] for a survey on publicly available corpora for ABSA. For an in-depth analysis of different subtasks and models used for solving the task, see Zhang et al. [25].

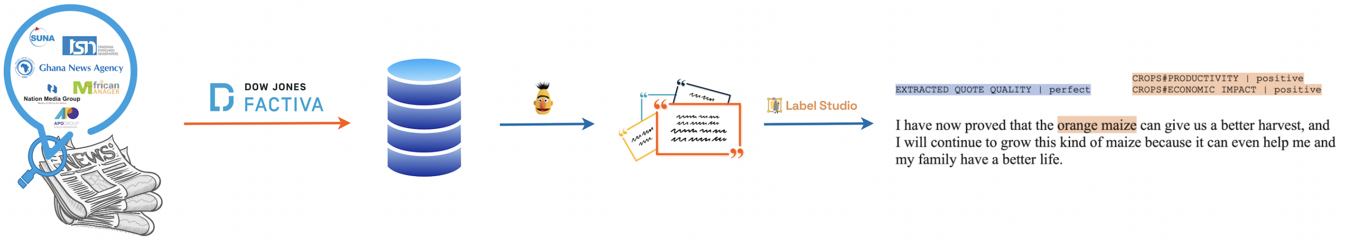
[2] https://tinyurl.com/gmo-lobbying-war

[3] https://tinyurl.com/FactivaSnapshotsAPI

[4] This was achieved by detecting the presence of two consecutive new line characters. Note that due to the format of the data extracted through the Factiva API, which includes data from a variety of sources that are not consistently standardized (i.e., it occasionally contains newline characters that do not correspond to logical line breaks), some paragraphs may appear unusual.

[5] To train the model, we used its HuggingFace PyTorch implementations [24], with default parameters.

[6] The corpus comprises more than 10,000 direct quotations extracted from news articles that have been manually annotated for quotation extraction and identification of the corresponding speakers.


SUNA
TSN TANZANIA STANDARD NEWSPAPERS
Ghana News Agency
Nation Media Group
African MANAGER
APO GROUP
NEWS
DOW JONES FACTIVA
Label Studio
EXTRACTED QUOTE QUALITY | perfect
CROPS#PRODUCTIVITY | positive
CROPS#ECONOMIC IMPACT | positive

I have now proved that the orange maize can give us a better harvest, and I will continue to grow this kind of maize because it can even help me and my family have a better life.

Figure 1: Pipeline for the creation of the BioMAISx corpus.[7]

Table 1: Examples of annotated quotations in the BioMAISx corpus that reference the entity CROPS.

Aspect Category (Entity # Attribute)	Polarity	Example
CROPS # FOOD SECURITY	positive	*"This is to ensure food security and fight malnutrition. Sweet potato vines rich in Vitamin A will be grown on 108 hectares while vegetables will be planted on the remaining part."*
CROPS # PRODUCTIVITY	positive	*"I tried the seed on my farm and got 1350kg (13 bags) on one hectare alone. That's good yield for cowpea."*
CROPS # SAFETY	negative	*"In South Africa, tobacco kills approximately 44 000 people a year."*
CROPS # ENVIRONMENTAL & ETHICAL CONCERNS	negative	*"People often immediately think of the health impact that tobacco has, but there is not enough awareness of how tremendously destructive it is for the environment too, on land, underwater and in the air."*
CROPS # RESISTANCE	positive	*"In terms of insect resistance, it is highly resistant to to the pink bollworm complex that is always ravaging and ripping it and stopping cotton from having high yield. BT cotton is a saviour to farmers."*

inventory of entities we have considered for the task at hand, along with their respective definitions:

- CROPS: any specific mention of a crop variety, or GM crops in general;
- ORGANIZATIONS: stakeholders involved in the development, approval, distribution, or regulation of GM crops;
- AGRICULTURAL PRACTICES: methods, techniques, or processes associated with GM crops;
- TECHNOLOGY: specific technologies or techniques used in the genetic modification of crops;
- GEOGRAPHIC LOCATIONS: countries or regions where GM crops are being grown, used, or regulated;
- ENVIRONMENTAL CONDITIONS references to weather, climate, or other environmental conditions;
- LEGAL ASPECTS & POLITICS: laws, policies, regulations, government statements, etc.;
- ECONOMIC FACTORS: inflation, import/export, unemployment, supply and demand, etc.;
- OTHER: entities that cannot be described by using our annotation schema.

The attributes can be assigned one of the following nine labels: RESISTANCE (ability to resist pests, diseases, and adverse weather conditions); CONSUMER PERCEPTION; SAFETY (as it relates to human and animal health); FOOD SECURITY; PRODUCTIVITY (yield or efficiency); ECONOMIC IMPACT (e.g., operational expenses, market price, profitability); RESEARCH DEVELOPMENT; ENVIRONMENTAL & ETHICAL CONCERNS; and MISCELLANEOUS (for attributes that do not fit into any of the previously mentioned categories).

2.3 Manual Annotation

The annotation process was carried out in multiple stages. In the initial pilot phase, we used 150 quotes to establish our annotation scheme and guidelines. Following this phase, five annotators[8] attempted to carry out the labelling process on small sets of quotes and subsequently discussed the results. This process was repeated over three successive rounds, and it contributed to the refinement of the annotation guidelines, thereby enhancing task clarity and promoting consistency among annotators. The inter-annotator agreement for a set of 350 quotes in terms of F1-score (a common alternative to Cohen/Fleiss kappa for NER/spans) is 83.7% for the identification of entity type, 83.3% for the attribute, and 78.2% for tuples of the form (entity-attribute pair, sentiment).

Following this, we started the primary annotation phase, ultimately resulting in the creation of the final corpus.[9] For the task at hand, we fully annotate the corpus using the open-source platform LabelStudio [21]. Given the amount of data to label, the three students were assigned distinct subsets of the corpus for annotation. Following the completion of this phase, the other two annotators reviewed and corrected the annotated instances. The final corpus

[7] Source: newspaper clip art via iStock.com/zager.

[8] Three students and two of the authors of this paper.

[9] The corpus containing all the annotated instances, as well as the annotation guidelines, are made available to the research community at: https://github.com/uchicago-dsi/BioMAISx/tree/main.

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

	CROPS	ORGANIZATIONS	AGRICULTURAL PRACTICES	TECHNOLOGY	GEOGRAPHIC LOCATIONS	ENVIRONMENTAL CONDITIONS	LEGAL ASPECTS & POLITICS	ECONOMIC FACTORS	OTHER
RESISTANCE	190	-	19	25	3	5	-	-	3
CONSUMER PERCEPTION	172	12	5	8	-	-	7	-	9
SAFETY	86	13	9	8	12	10	12	-	3
FOOD SECURITY	152	28	14	4	27	22	28	30	10
PRODUCTIVITY	492	72	68	87	104	188	55	31	43
ECONOMIC IMPACT	543	163	60	54	40	55	120	331	71
RESEARCH DEVELOPMENT	11	22	-	8	-	-	3	-	9
ENVIRONMENTAL & ETHICAL CONCERNS	57	86	20	9	14	24	20	15	28
MISCELLANEOUS	89	28	13	9	5	1	13	13	20

exclusively comprises instances where consensus was reached between two annotators, totalling 1,553 quotes (including 3,796 sentences) with 4,020 aspect categories. Table 1 presents examples of the annotation of various aspects within quotations that reference the entity CROPS.

2.4 Quantitative Results

For evaluating the quotation extraction model, we rely on two different metrics: *exact match* and an overlap metric [15]. The annotation procedure revealed that the quotation extraction model perfectly identified quotes in 11.2% of the cases (181 instances) and exhibited good performance in 82.6 % of the cases (1,333 instances).[10]

Table 2 presents the number of annotated instances per aspect category in the BioMAISx corpus. The table highlights the scarcity of data for certain entities, such as AGRICULTURAL PRACTICES and TECHNOLOGY. In contrast, CROPS frequently appear in quoted statements, though less often in the context of RESEARCH DEVELOPMENT.

Regarding the polarity distribution, 41% of the aspect categories were labelled as positive, 37.6% as negative, 19.6% as neutral, and only 1.6% were annotated as conflict (i.e., expressing both positive and negative sentiments towards the same aspect categories).[11] A closer examination of these results revealed that the entity TECHNOLOGY was discussed positively in 58.6% of the cases, while ENVIRONMENTAL CONDITIONS exhibited an overwhelmingly negative sentiment (80.8% of the cases). ORGANIZATIONS had slightly more positive instances than negative ones (47.3% and 30.9%, respectively), however, when focusing on the ENVIRONMENTAL & ETHICAL CONCERNS aspect, the sentiment was predominantly negative (60.7% of the cases). These initial results provide valuable insights into public discourse surrounding agricultural biotechnologies.

3 Conclusion and Perspectives

In this paper, we have presented BioMAISx, the first English language corpus pertaining to agricultural biotechnologies annotated for ABSA. The corpus comprises 1,553 direct quotes extracted from a curated list of Africa-based news media publications. A model trained on this dataset, in conjunction with quotation attribution, could help to establish and analyse a comprehensive network of media representations of agricultural biotechnologies in Africa. Such an analysis would allow one to explore connections between particular news outlets and quoted sources, as well as examine the sentiment conveyed within these quoted statements. When collecting the data, in addition to the English subset used in this study, we also acquired news articles written in French and Arabic. As a result, our future work involves expanding the corpus to incorporate annotations in these languages.

4 Ethics Statement

This dataset consists of annotations of media articles collected and delivered by Factiva, a for-profit global news search engine hosted by Dow Jones. Factiva aggregates non- and for- profit, as well as government media outlets. Our team of topic experts compiled a list of keywords related to agricultural biotechnologies, which we then 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 allowed us to use an expansive definition of "biotechnology", rather than rely on definitions provided by Factiva. We analysed roughly 2 million articles that matched our search criteria and timeframe. Our annotations and analysis are within the guidelines of our agreement with Factiva.
The research did not contain human subjects and therefore was not subject to an institutional review board.
Hiring policy: in addition to the authors of this article, other researchers were involved in the project. We recruited three English-speaking annotators from a pool of applicants to a summer data science program for students at the University of Chicago. All hired personnel received financial compensation.

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 research was funded by grants from 11th Hour Project. The authors have no conflicts of interest.

References

[1] 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 (2019), 222–242.

[2] 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.

[3] Eunsol Choi, Chenhao Tan, Lillian Lee, Cristian Danescu-Niculescu-Mizil, and Jennifer Spindel. 2012. Hedge detection as a lens on framing in the GMO debates: A position paper. *arXiv preprint arXiv:1206.1066* (2012).

[4] 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 (2023), e2248363.

[5] Brian Dowd-Uribe, Joeva Rock, Trevor Spreadbury, Patricia Chiril, and David Uminsky. 2023. Bridging the Gap? Public-Private Partnerships and Genetically

[10] 60 instances were annotated only for the quotation task making the denominator for these percentages 1,613.

[11] Detailed polarity statistics with respect to each aspect category are provided in the GitHub repository.

Modified Crop Development for Smallholder Farmers in Africa. (2023).

[6] 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 (2014), 229–249.

[7] Sarah Evanega, Joan Conrow, Jordan Adams, and Mark Lynas. 2022. The state of the 'GMO'debate-toward an increasingly favorable and less polarized media conversation on ag-biotech? *GM Crops & Food* 13, 1 (2022), 38–49.

[8] Eleni A Galata. 2017. The cultivation of opinions. How did the press cover the last 16 years of experience with GMOs in Canada? *Cogent Business & Management* 4, 1 (2017), 1297212.

[9] Leland Glenna and Analena Bruce. 2021. Suborning science for profit: Monsanto, glyphosate, and private science research misconduct. *Research Policy* 50, 7 (2021), 104290.

[10] Daniel Grzenda, Trevor Spreadbury, Joeva Rock, Brian Dowd-Uribe, and David Uminsky. 2022. OMGMO: Original multi-modal dataset of genetically modified organisms in african agriculture. In *International Conference on Social Informatics*. Springer, 414–425.

[11] Mark Lynas, Jordan Adams, and Joan Conrow. 2022. Misinformation in the media: global coverage of GMOs 2019-2021. *GM Crops & Food* (2022), 1–10.

[12] Mark Lynas, Selene Adams, and Karen Stockert. 2023. Gene editing achieves consistently higher favorability in social and traditional media than GMOs. *GM Crops & Food* (2023), 1–8.

[13] Martin Müller, Manuel Schneider, Marcel Salathé, and Effy Vayena. 2020. Assessing public opinion on CRISPR-Cas9: combining crowdsourcing and deep learning. , e17830 pages.

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

[15] 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*. 989–999.

[16] 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)*. Association for Computational Linguistics, Dublin, Ireland, 27–35. https://doi.org/10.3115/v1/S14-2004

[17] Joeva Rock. 2018. Complex mediascapes, complex realities: critically engaging with biotechnology debates in Ghana. *Global Bioethics* 29, 1 (2018), 55–64.

[18] Joeva Sean Rock. 2022. *We are not starving: The struggle for food sovereignty in Ghana*. MSU Press.

[19] 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* (2019).

[20] Matthew A Schnurr. 2019. *Africa's gene revolution: Genetically modified crops and the future of African agriculture*. McGill-Queen's Press-MQUP.

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

[22] Brittany Walker and Jennifer Malson. 2020. Science, god, and nature: A textual and frequency analysis of facebook comments on news articles about agricultural and environmental gene editing. , 1004–1016 pages.

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

[24] 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 (2019).

[25] Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[26] Yuanchi Zhang and Yang Liu. 2021. DirectQuote: A Dataset for Direct Quotation Extraction and Attribution in News Articles.