Data in Brief 55 (2024) 110545

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Dataset: Annotated soybean market news articles

Ivan José dos Reis Filho[a,*], Jamille de Campos Coleti[a], Ricardo Marcondes Marcacini[b], Solange Oliveira Rezende[b]

[a] State University of Minas Gerais (UEMG) at Frutal, Brazil
[b] Institute of Mathematical and Computer Sciences (ICMC) at University of São Paulo (USP) São Carlos Brazil, Brazil

ARTICLE INFO

Article history:
Received 26 February 2024
Revised 14 May 2024
Accepted 15 May 2024
Available online 22 May 2024

Dataset link: Soybean Market News Dataset (Original data)

Keywords:
Text mining
Annotation
News soybean

ABSTRACT

This dataset involves a collection of soybean market news through web scraping from a Brazilian website. The news articles gathered span from January 2015 to June 2023 and have undergone a labeling process to categorize them as relevant or non-relevant. The news labeling process was conducted under the guidance of an agricultural economics expert, who collaborated with a group of nine individuals. Ten parameters were considered to assist participants in the labeling process. The dataset comprises approximately 11,000 news articles and serves as a valuable resource for researchers interested in exploring trends in the soybean market. Importantly, this dataset can be utilized for tasks such as classification and natural language processing. It provides insights into labeled soybean market news and supports open science initiatives, facilitating further analysis within the research community.

* Corresponding author.
E-mail address: ivan.filho@uemg.br (I.J.d. Reis Filho).

https://doi.org/10.1016/j.dib.2024.110545

Specifications Table

Subject	**Computer Science: Artificial Intelligence.**
Specific subject area	Text Classification, Data Mining, Natural Language Processing.
Data format	Tables (*.csv file) and (*.pkl file)
Type of data	Text (Headline and Content) and Textual Representations from BERT models.
Data collection	Data collection involved the web scraping technique of the Brazilian website that gathers public news about the soybean market.
Data source location	website: https://www.noticiasagricolas.com.br
Data accessibility	Repository name: Soybean Market News Dataset Data identification number: 10.17632/f8fdmpp6yh.2 Direct URL to data: https://data.mendeley.com/datasets/f8fdmpp6yh/2

1. Value of the Data

- **NLP Resource:** This dataset serves as a valuable resource for NLP applications, offering a diverse set of soybean market news articles that can be leveraged for language understanding, sentiment analysis, and other NLP tasks.
- **Relevance to Computer Scientists:** Computer scientists, especially those specializing in NLP and data mining, will find this dataset beneficial. Its content and labeling process make it conducive to development, testing, and refinement in the context of soybean market news analysis.
- **Training and Evaluation for Classification:** The dataset is well-suited for training and evaluating classification models. Researchers and practitioners in the field can use it to develop and assess algorithms designed for distinguishing between relevant and non-relevant soybean market news articles.
- **Integration with Time Series Data:** With its temporal scope from January 2015 to June 2023, the dataset can be effectively integrated with time series to evaluate forecasting performance. This integration facilitates a nuanced understanding of the evolution of the soybean market over the specified period.
- **Multi-modal Model Evaluation:** The utility of the dataset extends to the evaluating multimodal models. Researchers aiming to develop models that incorporate both textual and other modalities can utilize this dataset to assess and enhance the performance of their models.

2. Background

Agribusiness companies navigate a landscape that is continuously transformed by regional, national, and global economic conditions. Financial markets, known for their intricate and ever-evolving dynamics, are influenced by a variety of factors, including economic indicators, breaking news, and investor sentiment [1]. Within this context, the capacity to precisely predict trends in the financial market is of critical importance. Such forecasts are crucial not only for investors and financial institutions but also policymakers who search for well-informed strategies.

On the internet a great number of news articles are published daily across various information sources, offering a wide diverse range of real-time market updates. These updates are indispensable resources for professionals in the field, enabling them to stay informed about market fluctuations. However, a significant portion of the news published does not offer relevant information that aids supports the decision-making process of domain experts [2]. Irrelevant news complicates analysis and extends the time needed to assess the financial market. Consequently, in recent years, there has been a push towards the development of machine learning applications to classify news according to the interests of experts.

Machine learning models are typically trained on datasets comprising numerous samples, each representing an object or event. In this context, the performance these predictive models depends on the availability of labeled data in that is plentiful and of high quality [3]. However,

for certain domains annotated data is rare, and the usual process of acquiring labels through experts examining individual samples is often costly and time-consuming. To overcome this limitation, we collected news articles and labeled them as either relevant or not relevant to the soybean market.

3. Materials and Methods

The dataset creation process was divided into four distinct phases, as illustrated in Fig. 1: web scraping, criteria and analysis, news labeling, and post-labeling. The initial phase was conducted by a data mining professional. The subsequent phase was led by a specialist from the agribusiness domain. The third phase was a collaborative effort involving a team of 12 individuals, including an agribusiness specialist, a data mining expert, and ten undergraduate students. The final step was managed once more by the data mining professional. These sequential steps were designed to ensure the development of a comprehensive and accurately labeled dataset, leveraging a mix of professional expertise and collaborative contributions from a diverse team.

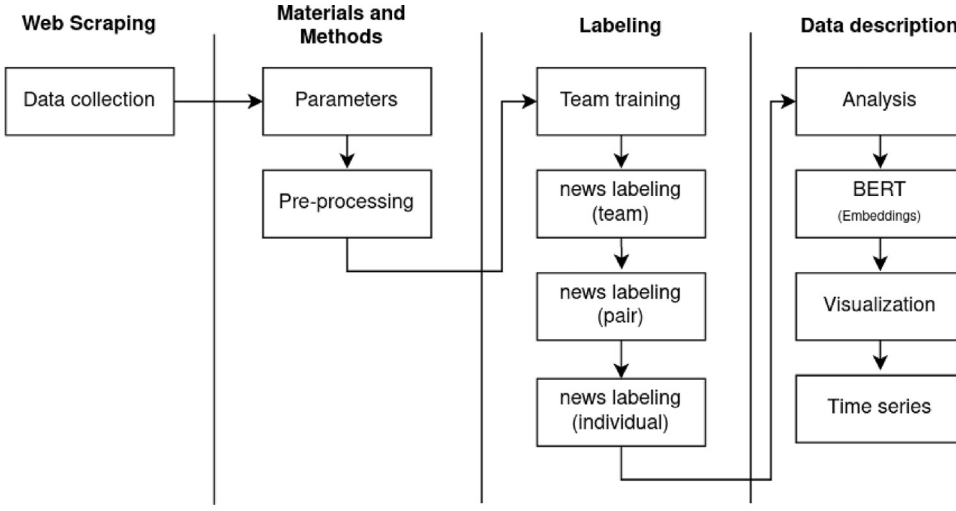

Fig. 1. Phases and processes used in constructing the dataset.

During **web scrapping** phase, a data collection algorithm was developed to gather a daily set of news articles pertinent to the soybean market. The platform selected was the agricultural news website (www.noticiasagricola.com.br), which is known for its daily publication and compilation of news from various sources related to agricultural commodities, including both original content and replicated news articles. The period of data collection extended from January 1, 2015 to June 30, 2023. Throughout this period the website daily news content regarding the soybean market was systematically gathered.

During the **criteria and analysis** phase, a series of parameters were defined to direct collaborators in the process of labeling. A domain expert identified ten key parameters: source, sources, coverage, agroterms, climate terms, research, consulting companies, technology, diseases pests, and logistics. Each parameter was defined to serve as a characteristic for extracting information from the text, with a maximum allocation of 5 points to each. A description accompanying each parameter was provided to ensure a clear understanding and application of the labeling process.

The pre-processing step involved extracting data based on the parameters and calculating the relevance level of each piece of news. For each parameter considered in extracting information

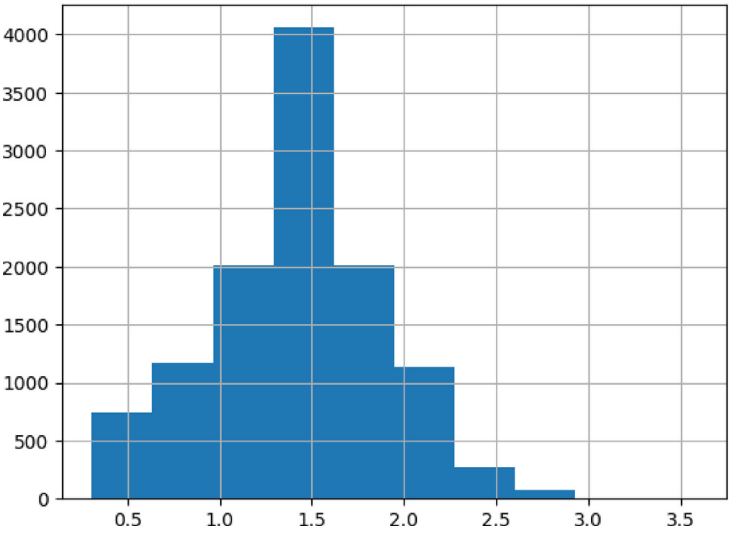

Fig. 2. Distribution of level values within the news set.

from the news content, a maximum value of 5 points was assigned. Additionally, each parameter was assigned a weight to calculate the weighted relevance level of the news. Equation 1 demonstrates how this score was assigned.

$$\begin{aligned} \text{wrl} = {} & ((\text{sr}^*5)+(\text{cv}^*5)+(\text{ct}^*5)+(\text{th}^*5)+(\text{sc}^*3)+(\text{ct}^*3)+(\text{lg}^*3) \\ & +(\text{ag}^*1)+(\text{rs}^*1)+(\text{dp}^*1))/32 \end{aligned} \tag{1}$$

An initial analysis of the distribution of values obtained during the pre-processing step was conducted and is presented in the data description section (Fig. 2).

The initial step of the third phase (**news labeling**) consisted of team training given to collaborators. A agribusiness expert led the participants through a detailed explanation of the soybean production chain, the essential elements of communication between the sender and receiver, and thorough market analyses. Practical examples of relevant and irrelevant news were presented, emphasizing the importance of distinguishing between events that are metadata, i.e., events that occurred but do not directly impact the decision-making of professionals in the field. The training also addressed the crucial distinction between news that is pertinent to the decision-making of industry experts and news that does not significantly contribute in this context.

During this labeling process, we considered that the recipients of the news are personas directly engaged in the soybean market, including farmers, grain processors, and professionals serving as traders. This strategic approach is intended guarantee that the labeling accurately reflects the usefulness of the news for various segments within the soybean market, promoting a more effective classification that is aligned with the specific requirements of these professionals.

The assignment relevance during the pre-processing step was crucial in expediting the labeling process, eliminating the need for participants to read the entire content of each news article. With the parameters set, participants gained crucial insights into the content, source, and degree of replicability of the news. Consequently, the initial phase of labeling was carried out collaboratively, with the establishment of a committee composed of all participants (news labeling - team). One collaborator was designated for reading the news, and everyone contributed with viewpoints regarding the news label. In cases of disagreement, the domain expert was called upon to provide their perspective on the label. This initial step was divided into two sessions, each addressing roughly 200 news items and conducted on a weekly basis.

During third step, labeling was carried out in pairs. Each pair received a set of news articles for collective analyses. In cases of disagreement, participants noted the indices of the news for the domain expert to assign the appropriate labels. This step was divided in two sessions, each session featuring different pairings. To further optimize the labeling process, the subsequent step was conducted on a individual basis. Each participant received a set of news articles from the same period of the year, ensuring consistency in the labeling efforts. This strategy was implemented to ensure uniformity throughout the process.

Finally, the last step involved reviewing the consistency of labels in news of the same context. For example, news with headlines such as "USDA announces soybean sale to China" were examined to confirm the uniformity of the labels assigned. This approach was adopted not only for similar news but also for different contexts, with the objective of preserving consistency in label assignment. The last phase of **post-labeling** is presented in the next section of the data description.

4. Data Description

The dataset comprises a collection of news in Portuguese and English related to the soybean market, covering international, national, and regional content from January 2015 to June 2023. The dataset totals 11,438 news articles, with daily quantities varying. The datasets are available in .csv and .pkl formats, presented sequentially in the steps outlined in Fig. 1, labeled as:

- 01_Soybean_scrapping[en/pt-br].[csv/pkl]
- 02_Soybean_10Parameters[en/pt-br].[csv/pkl]
- 03_Soybean_labeled[en/pt-br].[csv/pkl]
- 04_Soybean_Embeddings_headline[en/pt-br].[csv/pkl]
- 04_Soybean_Embeddings_headline_TimeSeries[en/pt-br].[csv/pkl]

The first step involved collecting textual data from the Brazilian website, available in the file "01_Soybean_scrapping[en/pt-br].[csv/pkl]". The dataset compiles news from various international, national, and regional sources, amounting to 544 different sources. Table 2 presents the top ten sources with the highest number of news articles.

Of the 11,438 news articles collected on the Notícias Agrícolas website, 6188 are news articles produced internally by the site. The remaining news articles are from external sources (Reuters, CEPEA, AgResource, etc.) and replicated on the website. It is important to emphasize that the collected news articles are in the public domain and can be accessed from the platform for analysis and research purposes for studies related to the agricultural market.

The "02_Soybean_10Parameters[en/pt-br].[csv/pkl]" file contains all pre-processed data for the parameters presented in Table 1. An analysis of the distribution of values obtained during the pre-processing stage was conducted. The mean value was 1.40, the standard deviation was 0.47, the first quartile was 0.93, and the third quartile was 1.87.

Out of a total of 11,438 news articles, we identified that 7738 news articles fell between the first and third quartile. Approximately 2009 news articles had values above the third quartile, while 1691 articles had values below the first quartile. To assist collaborators in making label decisions, news articles with values above the third quartile were suggested to be potentially relevant, although the final labeling decision was left to the participant. Similarly, for news articles with values below the first quartile, participants received a suggestion indicating that the news might be of lower relevance, but the labeling decision remained in the hands of the participant.

In additional analyses, the Table 3 presents the number of news that occurred at least once under the terms determined in Table 1. It is important to highlight that the source and content of the news may differ. For instance, while the news source might be national, the content could pertain to international matters. Therefore, the coverage presented in Table 3 pertains to the content coverage of the news.

Table 1
Description of the parameters used in the labeling process.

Parameters (abb): weight	Description
Source (sr): 5	Assign scores of 5, 3, and 1 for news articles originating from International, National, and Regional sources, respectively.
Coverage (cv): 5	Assign scores of 5, 3, and 1 based on the coverage of news content in international, national, and regional contexts. For instance, if the content addresses business between countries, assign an international coverage score.
Consulting (ct): 5	If the news content originates from a study or analysis conducted by a consulting company, assign scores of 5, 3, and 1 to denote international, national, and regional consultancy companies, respectively.
Technology (th): 5	The presence of technology-specific terms is evident in the news content. A total of 55 terms were identified, including Precision Agriculture, IoT, Drones, Genetic improvement and others. Assign a score of 0.5 for each occurrence, with a maximum value of 5.0.
Sources (sc): 3	Assign scores of 5, 3, and 1 for news articles that cite International, National, and Regional sources, respectively. For example, if a Brazilian website cites a source from another country, assign an international score.
Logistics (lg): 3	The news content exhibits the presence of logistic-specific terms. A total of 62 terms were identified, including freight, road, highway, and others. Assign a score of 0.5 for each occurrence, with a maximum value capped at 5.0.
Climate terms (ct): 3	The presence of climate-specific terms is evident in the news content. A total of 49 terms were identified, including Temperature, Humidity, Hail, and others. Assign a score of 0.5 for each occurrence, with a maximum value of 5.0.
Research (rs): 1	The presence of research-specific terms is evident in the news content. A total of 55 terms were identified, including research center, agricultural technology, agricultural innovation and others. Assign a score of 0.5 for each occurrence, with a maximum value of 5.0.
Agroterms (ag): 1	Calculating the sum of terms using the TF-IDF (Term Frequency-Inverse Document Frequency) method involves assessing the importance of a term within a document relative to its frequency across a collection of documents. A total of 59 terms were identified, including agriculture, harvest, planting, export and others.
Diseases pests (dp): 1	The news content exhibits the presence of disease and pest-specific terms. A total of 50 terms were identified, including Asian rust, brown stink bug, soybean caterpillar, and others. Assign a score of 0.5 for each occurrence, with a maximum value capped at 5.0.

Table 2
Top ten news sources in the dataset.

Source	News source location	Counting
Notícias Agrícola (Agricultural News)	Brazil (national)	6188
Reuters	International	1669
CEPEA (Center for advnced Studies on Applied Economics)	Brazil (national)	453
AgResource Brazil	Brazil (national)	107
La Nación	Argentina (International)	101
Gazeta do Povo (People's Gazette)	ParanáBR(regional)	99
OTCex Group Genebra	International	87
Estadão	Brazil (national)	87
Mato-Grossense Institute of Agricultural Economics	Mato Grosso - BR(Regional)	81
Só notícias (Just news)	Mato Grosso - BR(Regional)	66

The dataset named as "03_Soybean_labeled[en/pt-br].[csv/pkl]" includes date, headlines, content, and labels features. The value 0 was assigned to irrelevant news, while a value 1 was assigned to relevant news. Table 4 provides the quantity of labeled news organized by year.

Roughly 58.2 % of the news articles were labeled as irrelevant, while 41.8 % were labeled as relevant. This distribution of labels offers valuable insights into the categorization of news content within the dataset, facilitating a comprehensive understanding of the relevance proportions.

Table 3
Quantity of news that with terms specified in the parameters.

Parameters	Quantity of news
Sources	3637
Coverage	International (5410); Regional (4130); Nacional (1899)
Agroterms	11,438
Climate terms	3713
Research	3454
Consulting	6325
Technology	875
Diseases pests	573
Logistics	3852

Table 4
Quantity of labels assigned to news separated by year.

Year	Irrelevant (0)	Relevant (1)
2015	678	752
2016	888	639
2017	1198	595
2018	1044	551
2019	728	440
2020	591	592
2021	574	520
2022	633	493
2023	324	199
Total	6658	4781

Considering the applicability of textual data in natural language processing, machine learning tasks, and to encourage the use of multi-modal models, we enhanced the dataset by incorporating **BERT** embeddings and temporally aligned time series data with the news articles. The file "02_Soybean_Embeddings_Headline[en/pt-br].[csv/pkl] includes date, headline, content, label, and embeddings from four pre-trained BERT models: Paraphrase Multilingual (MiniLM-L12-v2), Distilbert Multilingual (base cased), BERTimbau (Portuguese cased model) and AgroBERT (Pre-trained model with agribusiness text). These embeddings specifically related to the news headlines and have been processed to facilitate their application in machine learning tasks and natural language processing. Fig. 3 visually represents the embeddings in a graphical format.

For some models, there is a concentration of news articles with predominant labels. For example, the Paraphrase Multilingual model displays clusters of irrelevant (0) news more concentrated at the extremes, while relevant headlines are more centralized. However, headlines do not carry much information about the content of the news, leading to indistinct graphical distribution patterns. This observation highlights the importance of conducting a more comprehensive exploration and analysis of the dataset, considering both headline and content information, to extract meaningful insights for model training and natural language processing tasks.

Furthermore, the dataset named 04_Soybean_Embeddings_headline_TimeSeries[en/pt-br].[csv/pkl] comprises date, headline, content, label, four embeddings, and soybean price **time series** data. The price series reflects the closing price and trading volume on the Chicago Board of Trade (CBOT) for the 3, 5, 7, 14, and 28 days preceding the news day. Fig. 4 presents examples representing the 14-day price series prior to the news content.

The figure illustrates examples where the green marker represents a point in the dataset (news), encompassing time series data preceding the news. The inclusion of time series data in the set aims to encourage studies that further analyze whether temporal trends have an impact on events reported in the news, and conversely, whether these events influence temporal events.

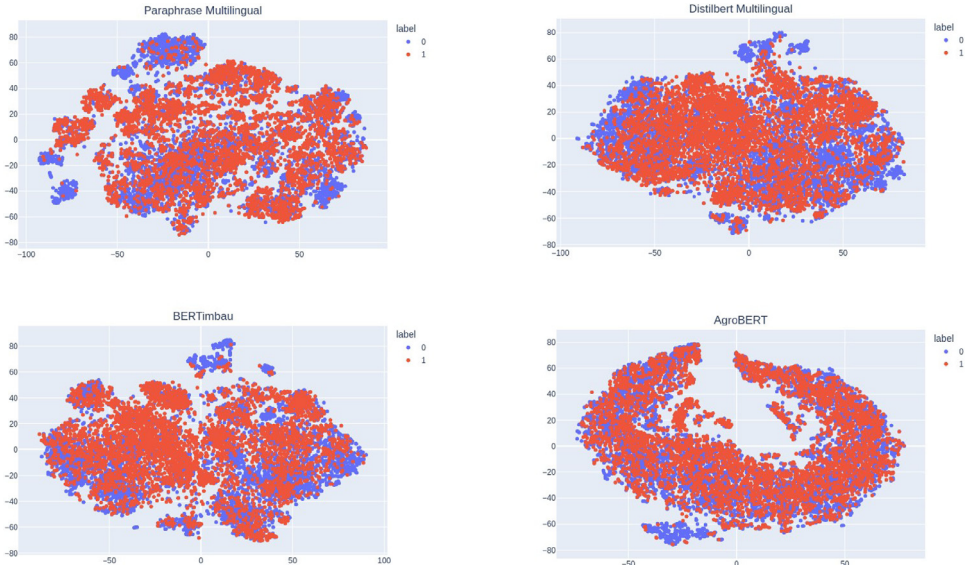

Fig. 3. Illustration of BERT embeddings for news headlines.

Fig. 4. Green marker represents a day and red line represents time series data.

Limitations

News Labeling Process: Despite employing certain strategies to reduce biases in the labeling process, determining whether content is relevant or not can be subjective and may vary among labelers. Certain specific nuances of the soybean market might not be fully encompassed by these broad categories.

Ethics Statement

This work involves annotating data collected from a public source on the web. The data collection did not involve human subjects, animal experiments, or any interaction with social media platforms. The data used in this dataset were obtained strictly by the Terms of Service of the mentioned online platform. The collected data originates from a public source and pertains to news of the public domain and access to the soybean market. The collected news does not

include personally identifiable information or sensitive individual data. Furthermore, there is no need to anonymize the source or content, as both are publicly accessible. The collected content pertains to news and opinions about the soybean market and is not altered or manipulated to distort the original context. The data collection adhered to ethical research guidelines, ensuring transparency and integrity.

Notícias agrícola: https://www.noticiasagricolas.com.br/.

CRediT Author Statement

Ivan José dos Reis Filho: Textual data scrapping, data pre-processing, Methodology, Writing, review & editing. **Ricardo Marcondes Marcacini**: Writing – review & editing, supervision. **Solange Oliveira Rezende**: Writing – review & editing, supervision **Jamile de Campos Coleti**: specialist in the field of agribusiness, conceptualization, Methodology. **Ranielly Patrícia Resende de Almeida**: labeling collaborator. **Alex Pablo Silva Pereira**: labeling collaborator. **Ernane Manoel da Silva Filho:** labeling collaborator. **Carlos Sandrioshy Ayres Oliveira:** labeling collaborator. **Ana Carolina Souza Gonçalves**: labeling collaborator. **Lucas Castro Campos:** labeling collaborator. **Igor Angelotti Marques**: labeling collaborator. **Antônio Braga do Couto Rosa Mazáro:** labeling collaborator.

Data Availability

Soybean Market News Dataset (Original data) (Mendeley Data).

Acknowledgements

This work was carried out at the Center for Artificial Intelligence (C4AI-USP) and partially supported by the São Paulo Research Foundation (FAPESP) (grant \#2019/07665-4) and the IBM Corporation. The authors of this paper thank FAPESP (Process 2019 / 25010-5) and the National Center for Scientific and Technological Development (CNPq) (process 309575/2021-4). The corresponding author thanks the Minas Gerais State Research Support Foundation (FAPEMIG) (Process PCRH BPG-00054-210).

Declaration of Competing Interest

The authors declare that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] N. Pinto, L. da Silva Figueiredo, A.C. Garcia, Automatic prediction of stock market behavior based on time series, text mining and sentiment analysis: a systematic review, in: IEEE - 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD), IEEE, 2021, pp. 1203–1208.

[2] B. Clapham, M. Siering, P. Gomber, Popular news are relevant news! How investor attention affects algorithmic decision-making and decision support in financial markets, Inf. Syst. Front. 23 (2021) 477–494.

[3] Boecking, B., Neiswanger, W., Xing, E., Dubrawski, A.: Interactive weak supervision: learning useful heuristics for data labeling. arXiv preprint arXiv:2012.06046 (2020).