Datasheet for Atticus Open Contract Dataset (AOK)(beta)

I. MOTIVATION

A. Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g. company, institution, organization)?

The Atticus Project is a non-profit organization whose mission is to harness the power of AI to accelerate accurate and efficient contract review. The Atticus Project started as a grassroots movement by experienced lawyers in public companies and leading law firms aiming to achieve high-quality, low-cost, accurate and timely contract review using AI. It was officially incorporated as a California nonprofit public benefit corporation in January 2020.

B. Did they fund it themselves? If there is an associated grant, please provide the name of the grantor and the grant name and number.

The Atticus Project relies 100% on unpaid volunteers who are organized around the single mission of changing the legal industry by leveraging AI.

C. For what purpose was the data set created? Was there a specific task in mind? If so, please specify the result type (e.g. unit) to be expected.

The original specific task and the intended use case of the dataset is to facilitate the development of accurate AI algorithms for contract review in corporate transactions. Contract review refers to the process where human lawyers physically search through hundreds or thousands of contracts to find a few types of legal provisions. This process is time consuming, costly and prone to human error. The dataset aims to accelerate this process by leveraging AI. However, as development of this dataset progressed, there has not been any datasets identifying the different components of a legal contract. We hope that this dataset can be a catalyst to spur development of AI-models focused on legal contracts that can be an aid for legal practitioners. For example, one use could be an AI-model that can identify expiration dates to remind a party when a contract needs to be renewed.

However, the dataset should not be solely used for contract drafting, contract management, dispute resolution (incl. litigation) or provision of legal advice. In those situations, qualified human lawyers should still analyze and review any output. Instead, there are still a number of possibilities that AI can be developed as tools for legal practitioners and this dataset is intended to spur the development of such tools.

II. COMPOSITION

C. What do the instances (of each type, if appropriate) that comprise the data set represent? (e.g. documents, photos, people, countries).

Each instance is a commercial contract sourced from the EDGAR (Electronic Data Gathering, Analysis, and Retrieval) system used by the U.S. Securities and Exchange Commission (SEC). Publicly traded companies in the United States are required to file certain material contracts under SEC rules. Access to these contracts is available to the public for free at https://www.sec.gov/edgar.

B. How many instances (of each type, if appropriate) are there in total?

There are approximately 3,000 labels in 200 contracts in the beta version released in October 2020. Future releases will include additional labels and contracts.

C. Does the dataset contain all possible instances or is it just a sample of a larger set? i.e. Is the dataset different than an original one due to the preprocessing process? In case this dataset is a subset of another one, is the original dataset available?

The dataset (beta) includes commercial contracts selected from 25 different types of contracts based on the contract names as shown below. Within each type, we randomly selected contracts based on the names of the filing companies across the alphabet.

Type of Contracts	**# of Docs**
Affiliate Agreement	9
Co-Branding Agreement	22
Development Agreement	23
Distributor Agreement	13
Endorsement Agreement	13
Franchise Agreement	8
Hosting agreement	4
IP Agreement	4
Joint Venture Agreement	10
License Agreement	33
Maintenance Agreement	9
Manufacturing Agreement	8
Marketing Agreement	6
Non-Competition Agreement	3
Outsourcing Agreement	4
Reseller Agreement	5
Service Agreement	4
Sponsorship Agreement	4
Supply Agreement	4
Transportation Agreement	4
Strategic Alliance Agreement	6
Promotion Agreement	4

D. Is there a label or a target associated with each of the instances? If so, please provide a description.

The labels correspond to 40 categories of legal clauses in commercial contracts that are considered important by experienced attorneys in contract review in connection with a corporate transaction. Such transactions include mergers & acquisitions, investments, initial public offering, etc.

Each category supports a contract review task which is to extract from an underlying contract (1) text context (sometimes referred to as clauses) and (2) human-input answers that correspond to each of the 40 categories in these contracts. For example, in response to the “Governing Law” category, the clause states “This Agreement is accepted by Company in the State of Nevada and shall be governed by and construed in accordance with the

laws thereof, which laws shall prevail in the event of any conflict.". The answer derived from the text context is Nevada.

Each category (including context and answer) is independent of another except as otherwise indicated in "Category List" "Group" below.

33 out of the 40 categories have a derived answer of "Yes" or "No." If there is a segment of text corresponding to such a category, the answer should be yes. If there is no text corresponding to such a category, it means that no string was found. As a result, the answer should be "No."

7 out of the 40 categories ask for answers that are entity or individual names, dates, combination of numbers and dates and names of states and countries. See descriptions in the "Category List" below. While the format of the context varies based on the text in the contract (string, date, or combination thereof), we represent answers in consistent formats. For example, if the Agreement Date in a contract is "May 8, 2014" or "8th day of May 2014", the Agreement Date Answer is "5/8/2014".

The "Expiration Date" and the "Effective Date" categories may ask for answers that are based on a combination of (1) the answer to "Agreement Date" or "Effective Date" and/or (2) the string corresponding to "Expiration Date" or "Effective Date".

For example, the "Effective Date" clause in a contract is "This agreement shall begin upon the date of its execution". The answer will depend on the date of the execution, which was labeled as "Agreement Date", the answer to which is "5/8/2014". As a result, the answer to the "Effective Date" should be "5/8/2014".

An example of the "Expiration Date" clause is "This agreement shall begin upon the date of its execution by MA and acceptance in writing by Company and shall remain in effect until the end of the current calendar year and shall be automatically renewed for successive one (1) year periods unless otherwise terminated according to the cancellation or termination clauses contained in paragraph 18 of this Agreement. (Page 2)." The relevant string in this clause is "in effect until the end of the current calendar year". As a result, the answer to "Expiration Date" is 12/31/2014.

A second example of the "Expiration Date" string is "The initial term of this Agreement commences as of the Effective Date and, unless terminated earlier pursuant to any express clause of this Agreement, shall continue until five (5) years following the Effective Date (the "Initial Term"). (Page 9). The answer here is 2/10/2019, representing five (5) years following the "Effective Date" answer of 2/10/2014.

Each category (incl. context and answer) is independent of another except otherwise indicated under the "Group" column below. For example, the "Effective Date", "Agreement Date" and "Expiration Date" clauses in a contract can overlap or build upon each other and therefore belong to the same Group 1. Another example would be "Expiration Date", "Renewal Term" and "Notice to Terminate Renewal", where the text context may be the same for two or more categories.

For example, the clause states that "This Agreement shall expire two years after the Effective Date, but then will be automatically renewed for three years following the expiration of the initial term, unless a party provides notice not to renew 60 days prior the expiration of the initial term." Consequently the answer to Effective Date is 2/14/2019, the answer to Expiration Date should be 2/14/2021, and the answer to "Renewal Term" is 3 years, the answer to "Notice to Terminate Renewal" is 60 days.

Similarly, one sentence may be responsive to both "Non-Compete" and "Exclusivity". Certain "License Grant" clauses may also correspond to "Exclusive License", "Non-Transferable License" and "Affiliate License-Licensee".

E. What is the format of the data? e.g. .json, .xml, .csv.

The files in the dataset (beta) include 201 CSV files and 200 PDF files.

- 1 master clauses CSV: a 81-column 201-row file. The first column is the names of the contracts corresponding to the PDF files in the “full_contracts” file. The remaining columns contain (1) text context (clause) and (2) human-input answers that correspond to each of the 40 categories in these contracts. See a list of the 40 categories in “Category List” below. The first row represents the file name and a list of the 40 categories. The remaining 200 rows each represent a contract in the dataset and include the text context and human-input answers corresponding to the 40 categories. The human-input answers are derived from the text context and are formatted to a unified form.

- 200 individual contract clauses CSVs: these files contain the same information as in the master clauses CSV file, but are split into 200 separate files. Each file is named as “[document name].csv” based on the corresponding PDF file in the “full_contracts” PDFs file. Each [document name].csv file contains labels for just one contract. The first column of each file contains the clauses from a contract that correspond to the categories. The second column contains the name of the category under the header “Label 1”, and the third column contains the human-input answers derived from the corresponding clauses under the header “Label 1-Answer”. The number of rows in each file varies depending on the number of relevant clauses (3-88 rows). A file may have more than 3 columns if a clause falls into multiple categories (3-7 columns).

- 200 full contract PDFs: a collection of the underlying contracts that we used to extract the labels. Each file is named as “[document name].pdf”. These contracts are in a PDF format and are not labeled. The full contract PDFs contain raw data and are provided for context and reference.

We recommend using the individual contract clause CSVs as a starting point.

F. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g. because it was unavailable). This does not include intentionally removed information, but might include, e.g. redacted text.

The dataset included all information in the contracts as extracted from EDGAR. Some clauses in the files are redacted because the party submitting these contracts redacted them to protect confidentiality. Such redaction may show up as *** or ___ or blank space. The dataset and the answers reflect such redactions. For example, the answer for “January __ 2020” would be “1/[]/2020”). Atticus did not redact or omit any information from these contracts.

G. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description. Do not include missing information here.

For any categories that require an answer of Yes/No, annotators include full sentences as text context in a contract. To maintain consistency and minimize inter-annotator disagreement, annotators select text from from period to period.

Some sentences in the files include confidential legends that are not part of the contracts. An example of such confidential legend is as follows: THIS EXHIBIT HAS BEEN REDACTED AND IS THE SUBJECT OF A CONFIDENTIAL TREATMENT REQUEST. REDACTED MATERIAL IS MARKED WITH [* * *] AND HAS BEEN FILED SEPARATELY WITH THE SECURITIES AND EXCHANGE COMMISSION. Some sentences in the files contain irrelevant information such as footers or page numbers. Some sentences may not be relevant to the corresponding category. Some sentences may correspond to a different category. Because many legal clauses are very long and contain various sub-parts, sometimes only a sub-part of a sentence is responsive to a category.

To address the foregoing limitations, annotators labeled such sentence twice. In the first instance, the annotators kept the sentence as is. In the second instance, the annotators manually deleted the portion that is not responsive. For example, if a “Termination for Convenience” clause starts with “Each Party may terminate this

Agreement if" followed by three subparts "(a), (b) and (c)," and they are all part of the same sentence from period to period, but only (c) is responsive to this category, we include all three subparts as the first label. We then include the same sentence a second time but deleted (a) and (b).

The text context in the master clauses CSV file contains references to page numbers in the form of "(Page x)", indicating the location of such text context in the underlying contract. These references are not part of the original contracts nor responsive to the categories.

H. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

See II-D above.

I. Are there recommended data splits, e.g. training, development/validation, testing? If so, please provide a description of these splits explaining the rationale behind them.

We do not split the dataset into training v. development v. testing.

J. Is the dataset self-contained, or does it link to or otherwise rely on external resources? e.g., websites, tweets, other datasets. If it links to or relies on external resources, a) Are there any guarantees that they will exist, and remain constant over time? b) Are there official archival versions of the complete dataset? i.e. including the external resources as they existed at the time the dataset was created. c) Are there any restrictions (e.g. licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, if appropriate.

The dataset is self-contained and does not link to or rely on external resources. The CSV files have a list of strings contained in the contracts in the PDF files.

K. Does the dataset contain data that might be considered confidential? e.g. data that is protected by legal privilege or by doctor patient confidentiality, data that includes the content of individuals non-public communications. If so, please provide a description.

No. The contracts in the dataset are filed with the SEC and publicly available with any confidential information already removed by the party filing the contracts.

L. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

M. Does the dataset relate to people? If so, please specify

a) Whether the dataset identifies subpopulations or not. b) Whether the dataset identifies individual people or not. c) Whether it contains information that could vulnerate any individuals or their rights. d) Any other verified information on the topic that can be provided.

The dataset includes names, addresses of individuals, their former or current employers and their relationship with other parties in the contracts. Any personal information included in the dataset is already publicly available information because these contracts were filed with the SEC in the EDGAR system.

N. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

The dataset does not contain sensitive information since the contracts were obtained from the publicly available EDGAR system maintained by the SEC and the labels are only derived from those contracts.

III. COLLECTION PROCESS

A. How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The dataset contains contracts obtained from the public EDGAR system. The CSV files contain strings from the publicly filed contracts manually labeled by volunteers of The Atticus Project, which is composed of lawyers and students. The strings are directly observable.

B. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The contracts in the PDF files were exported from EDGAR. The strings in the CSV files reflect collective know-how of experienced lawyers. Each contract in the dataset is human-labeled with multiple levels of quality-control and approved by highly sophisticated and experienced lawyers. The dataset only labels those clauses that on a standalone basis are responsive to a label.

The contracts were loaded into a software called eBrevia, which allowed the volunteers to review and label the applicable clauses. These labels represent a list of provisions in commercial contracts that are considered important by experienced attorneys in due diligence review in connection with a corporate transaction. Such transactions include mergers & acquisitions, investments, initial public offering, etc.

Our labeling process included multiple steps to ensure accuracy:

1. Law Student training: law students attended training sessions on each of the categories that included a summary, video instructions by experienced attorneys, a FAQ and a quiz. Students were then required to label sample contracts in eBrevia, an online contract review tool. The initial training took approximately 70-100 hours.
2. Law Student Label: law students conducted two levels of manual contract review and labeling in eBrevia. This step was conducted initially in sequence for the beta dataset, but will change to parallel review for future versions.
3. Key Word Search: law students conducted keyword search in eBrevia to capture additional categories that have been missed during the “Student Label” step. This step was conducted initially after the “Law Student Label” for the beta dataset, but will change to parallel review for future versions.
4. Group Review of Report: law students exported the labeled clauses into reports on a periodic basis and reviewed the reports in groups of three. If one or more students disagree with a labeled clause, such clause was highlighted for Attorney Review.
5. Initial Attorney Review: one experienced attorney reviewed the highlighted clauses and provided comments. Law students reviewed the comments and made changes in eBrevia accordingly.
6. Category-by-category Review: law students exported single-category reports and reviewed them in a group of three. If one or more students disagreed with a labeled clause, such clause was highlighted for Attorney Review.

7. Second-Level Attorney Review: two or three experienced attorneys reviewed the highlighted clauses, provided comments and addressed student questions. Attorneys then discussed such results with the students and reached consensus. Students made changes in eBrevia accordingly.
8. Third-Level Attorney Review: the full report was exported. Experienced attorneys reviewed each clause to correct mistakes and assign a confidence score (H, M & L) with different colors (white, green and yellow). Clauses that are not responsive are removed.
9. eBrevia Extras Review. Attorneys used eBrevia to generate a list of "extras", which are clauses that eBrevia AI tool identified as responsive to a category but not labeled by human volunteers. Attorneys reviewed all of the "extras" and added the correct ones. The process is repeated until all or substantially all of the "extras" are incorrect labels.
10. Final Report: Final report was exported into csv. Volunteers manually added the "Yes/No" answer column to categories that do not contain an answer.

C. If the dataset is a sample from a larger set, what was the sampling strategy? i.e. deterministic, probabilistic with specific sampling probabilities.

Contracts were selected from 22 different types of commercial contracts based on the contract names. Within each type, we randomly selected contracts based on the names of the filing companies across the alphabet.

All of the contracts are filed by publicly traded or reporting companies through EDGAR as exhibits to their periodic reports or registration statements. See here for a list of the reports and the exhibits under the SEC rules (https://www.law.cornell.edu/cfr/text/17/229.601). EDGAR is not the most ideal dataset of legal contracts because it is not representative of the overall population of legal contracts. Because only material contracts are required to be filed with the SEC, the EDGAR contracts are more complicated and heavily negotiated than the general population of all legal contracts. However, EDGAR contracts have the advantage of containing a large sample of clauses that are hard to find in the general population. One company may have only one or two contracts that contain exclusivity clauses. EDGAR contracts may have hundreds of them. It also has the benefit of being publicly available. We are aware of the inherent bias in the EDGAR contracts. However, balancing the pros and cons, we believe Edgar contracts are the best dataset for the dataset.

D. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Our volunteer annotators are experienced lawyers and law students. None of the volunteers of The Atticus Project were compensated monetarily for their time.

E. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The contracts were downloaded from EDGAR between March and July 2020, and manually labeled between March and October 2020. The dates when the contracts were signed are as set forth in the underlying contracts.

F. Did the collection process involve the participation of individual people? If so, please report any information available regarding the following questions: Was the data collected from people directly? Did all the involved parts give their explicit consent? Is there any mechanism available to revoke this consent in the future, if desired?

No. None of the data subjects are individual people. The dataset contains some contracts where individuals are parties, but all information about such individual parties is publicly available.

G. Has an analysis of the potential impact of the dataset and its use on data subjects been conducted? i.e. a data protection impact analysis. If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Not applicable. The dataset is based on publicly available contracts. None of the data subjects are individual people.

H. Were any ethical review processes conducted?

No. The dataset is based on publicly available contracts. None of the data subjects are individual people.

H. Does the data come from a single source or is it the result of a combination of data coming from different sources? In any case, please provide references.

The contracts are sourced from the EDGAR system maintained by the SEC.

IV. PREPROCESSING/CLEANING/LABELLING

A. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

See III-B for details on the labeling process.

B. Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.

Yes, the original contracts sourced from the EDGAR system are also included in the dataset.

C. Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the software used to label the instance is provided by eBrevia (https://ebrevia.com).

V. USES

A. Has the dataset been used already? If so, please provide a description.

The labeled dataset has not been used by a third party unaffiliated with The Atticus Project.

B. Is there a repository that links to any or all papers or systems that use this dataset? If so, please provide a link or any other access point.

The Atticus Project used OpenAI API to perform tests on a few legal clauses using the dataset. Preliminary results of such tests are published on the website at https://www.atticusprojectai.org/bytesized-ai.

C. What (other) tasks could the dataset be used for? Please include your own intentions, if any.

We are not aware of other tasks at this time, but we hope the dataset will spur further AI development in natural language processing in the legal field.

D. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

No. However, the dataset should not be the sole input for contract drafting, contract management, dispute resolution (incl. litigation) or provision of legal advice. The results generated from AI trained on this dataset should still be analyzed and reviewed by qualified human lawyers. This dataset is meant to facilitate development of tools to aid legal practitioners, but should not be used as a replacement for proper legal advice for non-lawyers.

Note: we make no representations or warranties regarding the underlying contracts, which are publicly available and downloadable from EDGAR.

In addition, the use of the dataset is subject to our privacy policy https://www.atticusprojectai.org/privacy-policy and disclaimer https://www.atticusprojectai.org/disclaimer.

E. Are there tasks for which the dataset should not be used? If so, please provide a description.

See V-D above.

VI. DISTRIBUTION

A. Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes. The dataset is open sourced and publicly available.

B. How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset can be downloaded from our website at https://www.atticusprojectai.org/. The contracts can be downloaded as a zip file that contains both CSV and PDF files.

C. When will the dataset be distributed?

The first release was in October 2020.

D. Is the dataset distributed under a copyright or other intellectual property (IP) license? And/or under applicable terms of use (ToU)? Please cite a verified source.

The dataset is licensed under the Creative Commons Attribution 4.0 (CC BY 4.0) license and free to the public for commercial and non-commercial use. https://creativecommons.org/licenses/by/4.0/

Note: we make no representations or warranties regarding the license status of the underlying contracts, which are publicly available and downloadable from EDGAR.

In addition, the use of dataset is subject to our privacy policy https://www.atticusprojectai.org/privacy-policy and disclaimer https://www.atticusprojectai.org/disclaimer.

E. Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

Not applicable. Note: we make no representations or warranties regarding the underlying contracts, which are publicly available for free and downloadable from EDGAR.

F. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

Not applicable.

VII. MAINTENANCE

A. Who will be supporting/hosting/maintaining the dataset?

The dataset is hosted and maintained by The Atticus Project, a non-profit organization.

B. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Contact us via email at info@atticusprojectai.org.

D. Is there an erratum? If so, please provide a link or other access point.

None at this time. All changes to the dataset will be published on The Atticus Project website (https://www.atticusprojectai.org).

E. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

All changes or updates to the dataset will be published on The Atticus Project website (https://www.atticusprojectai.org).

F. If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

Not applicable.

G. Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Yes. Older versions of the data will continue to be publicly available on the Atticus Project website (https://www.atticusprojectai.org) unless otherwise communicated on the Atticus Project website.

H. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We welcome your contribution to the dataset. Email us at info@atticusprojectai.org