PRELIMINARY RELEASE

RegData U.S. 5.0 User's Guide

July 31, 2023

Patrick McLaughlin, Michael Gilbert, Jonathan Nelson, and Thurston Powers

Abstract

We describe RegData U.S. 5.0, the latest iteration of the United States Federal Regulation dataset in the RegData Project. The dataset was preliminarily released on July 31, 2023, and created for the purpose of facilitating third-party usage and independent research. This preliminary release does not include industry relevance or restriction estimates. This User's Guide explains the feature additions of the 5.0 release, general RegData methodology used to build data, and describes the recommended methods to download the data and to interact with the data.

Introduction

RegData is both a methodology and a database focused on the quantification of various dimensions of regulation. We use custom-made text analysis and machine-learning algorithms to create statistics designed to measure several features of regulation, including volume, restrictiveness, complexity, and relevance to different sectors and industries. The RegData Project was launched in 2012 with the express purpose of facilitating research that was previously infeasible. Regulations have been an important and widely used policy tool for decades, but empirical analysis of the actual effects of regulation has historically been hampered by a paucity of data.

RegData was designed to solve this data problem. RegData U.S. 5.0 maps United States federal regulations to the sectors and industries affected by them. This opens the way for new research about both the causes and effects of regulation and how it relates to specific sectors of the economy. RegData captures the restrictiveness of various regulations by counting words and phrases that indicate a specific prohibited or required activity; these words and phrases are called regulatory restrictions. RegData also reports the word count, complexity, agency, department, and an estimate of industry relevance for all regulations contained in the *Code of Federal Regulations*. RegData is produced with the open-source QuantGov policy analytics platform. Because of its open-source nature, QuantGov can be used to produce modified versions of the RegData dataset or datasets based on other documents, such as guidance documents or regulations from other jurisdictions. The technical details on using the QuantGov platform are available at http://docs.quantgov.org. Data from other jurisdictions (including national and subnational jurisdictions in the United States, Canada, and Australia) are also available at https://www.reghub.ai/data.

A Note on the Decline in Restrictions

In the 53 years of the Code of Federal Regulations that RegData has analyzed, the total number of restrictions in the CFR has only declined from the previous year a handful of times (under President

Reagan in 1983 and 1985, under President Clinton in 1996 and 1997, and under President Trump in 2019). For this reason, we were surprised to find that the total number of restrictions in CFR decreased by 5,606 restrictions from 1,321,041 restrictions in 2021 to 1,315,435 restrictions in 2022. After looking into the cause of the decline, we found that most of the decrease in restrictions came from two regulatory actions by the Environmental Protection Agency in which they moved regulations from the CFR into incorporation by reference (IBR) documents.[1]

This extensive use of IBR led us to conduct a further investigation into the extent to which IBR is within the CFR. Incorporation by reference documents are cited in 557 parts, about 6.4% of the CFR. We also found that many agencies use IBR to some extent. The Environmental Protection Agency, the Coast Guard, and the Department of Transportation are the top three users of IBR. Approximately 38% of all IBR citations come from EPA regulations.

We plan to conduct an in-depth analysis of IBR documents in the future, which will help researchers and other users of RegData better understand the extent of IBR in the U.S. regulatory system.

New Features Included in Version 5.0

New NAICS Classifier:

RegData U.S. 5.0 introduces a brand new NAICS classifier, which uses modern, large language models to improve the classification process. Previously, the team used an older style of model, a logistic regression, with bag-of-words preprocessing. This approach consisted of the following steps:

1. Tokenization: Splitting any text the model is trained on into separate words a.k.a. tokens.
2. Lemmatization: Transforming any token in the text from many versions of that token into one version of that token. For example, transforming the words medicinal, medication, medicated and medical into medic.
3. TF-IDF vectorization: Vectorizing the text into numerical values using the Term Frequency – Inverse Document Frequency (TF-IDF) process. TF-IDF vectorization calculates each word's frequency within each document and its frequency within all the documents and computes a score that tries to accentuate the words that appear frequently within a few documents and downplay the words that appear rarely in a particular document or appear in all documents.
4. Model fitting: Once the text is transformed into a numerical matrix, a logistic regression was fit for each code/label.

The modern LLM approach was used in RegData U.S. 5.0. LLMs are superior to the older, bag-of-words approach in the following ways:

1. LLMs consider word order and are considerate of the many different forms a word can take. The previous approach is called a bag-of-words approach because it looks at each document as simply a collection of words which could appear in any order, tense, plurality, etc. This process ignores anything other than the words that appear in the texts. LLMs have vastly more parameters and can consider word order and meaning.
2. LLMs are pretrained on a large corpus of documents and benefit from the knowledge gained from any relationships in those texts. This training is then supplemented by further training on a more

[1] 85 FR 78412, responsible for a decline of **4,933** restrictions, and 86 FR 34308, responsible for a decline of **4,745** restrictions. 4,072 restrictions were added elsewhere to bring the total decrease to 5,606 from 2021 to 2022.

specific group of texts. The previous approach only learns from the texts the QuantGov team has within their own corpus of documents.

For a more detailed explanation of the process used to train the current iteration of the model, see the section “Methodology: NAICS algorithm” below.

With the release of the new algorithm, our metric “industry-relevant restrictions” has changed significantly. Researchers who are using industry data should be aware that the new industry data are not compatible with older editions of RegData, and versions should not be mixed. For any projects going forward, researchers should use RegData U.5. 5.0 data only for U.S. Federal regulatory analysis.

New Complexity Data Series:

RegData U.S. 5.0 improves upon a variety of complexity data series. The ***conditionals*** data series was improved by adding a standardized ***conditionals per 100 sentences*** data series in order to account for the fact that total conditionals is highly correlated with total word count. This new data series more effectively estimates the complexity of a particular part. A score of 100 means that, on average, there is a conditional term in every sentence of the part. ***Conditionals*** include the terms “if,” “but,” “except,” “provided that,” “when,” “where,” “whenever,” “unless,” “notwithstanding,” “in the event,” “in no event,” and (new in version 5.0) “to the extent that.”

Two new groups of data series are also introduced: acronyms and long word score. Acronyms are defined as a string of 3 or more capital letters either explicitly defined in the part (in a section titled “acronyms” or “definitions”) or implicitly defined by being surrounded by parentheses (for example, “Office of Government Information Services (OGIS)”). There are three data series for acronyms: ***total acronyms*** counts the total number of acronyms in the part; ***acronyms per 100 sentences*** standardizes the number to account for longer or shorter parts; and ***unique acronyms*** counts how many of the acronyms are unique and must be remembered by a reader.

The ***long word score*** data series estimates the complexity of a piece of text by assigning the text a score based the average lengths of its words. More weight is given to longer words, so the score is calculated using the formula $long\ words\ score = \frac{\sum_{i=0}^{n} wordlength_i^2}{n}$ where *i* is the *i*th word of the text and *n* is the total number of words. A higher ***long word score*** infers that the text has longer words on average and thus may be more difficult to understand.

Improved Last Updated Data Series:

RegData U.S. 5.0 improves methodology for determining the last date that a given part was updated. In previous versions of RegData U.S., the date was determined by calculating the date that the given part’s word count had changed by at least 1 percent. This methodology was lacking for several reasons. First, word count is somewhat noisy and can fluctuate based on the source of the text (for example, CFR annual editions versus annualized eCFR), leading to false positives. Second, changes can be masked if, for example, a subpart of 500 words was replaced with a whole new subpart of 500 words, leading to false negatives.

The new methodology finds dates in the source text to determine the last date that the part was updated. The dates are found in either the “source” of the part in the header or the Federal Register citation in the footer of each section (see Title 40, Part 10 as an example). The most recent date between these dates is determined to be the "last updated date" for that part. This methodology is more accurate because it comes from the text itself and more precise because it gives us an exact date that the part was last updated. When a date was unable to be found for a particular year-title-part combination, the date was forward filled from previous years, giving us a smooth time series.

New Year of Data:

RegData U.S. 5.0 expands the date range of RegData by adding analysis for the 2022 version of the CFR. The years 2017 through 2022 use text made available through the eCFR (the electronic version of the CFR), which is published daily and annualized by us. Later sections in this User Guide will go into the specifics of how this text is chosen and cleaned for RegData U.S. 5.0.

Improved Agency Metadata:

With the release of RegData U.S. 5.0, we have taken some additional steps in improving the accuracy of the associated metadata. Agency metadata was improved by using the agency information available on the eCFR website. Previously, we had a single agency value for each part, which was inconsistent in its use of department or subagency labels. We now have clean, consistent department and agency metadata for every year-title-part combination from 1970 to 2022. Users can query the data to gain greater insights into the composition of regulations and at both the department and subagency level, and how these regulations have changed over time. This improved metadata will be especially useful to researchers and policy analysts doing research on specific federal departments or agencies.

In some cases, a historical part may be assigned to an agency that did not exist at the time (for example, while the Department of Homeland Security was not founded until 2002, much of title 6 is assigned to DHS for years prior to 2002 because DHS is responsible for those regulations now). This produces a trade-off between historical accuracy and a more useful, smooth time series, and for convenience and policy research purposes, we have chosen to err on the side of the smooth times series.

Improved Source Text:

The text for 2017 to 2022 was also cleaned to correct an error in which text within data tables was missing whitespaces and led to some words being erroneously concatenated. The fix for 5.0 has led to a slight increase in word counts for the years 2017 to 2021 (compared to 4.1) and some improvements in the accuracy of the complexity data series and industry classifier.

Improved Document Granularity for Tax Code Data:

The text for the tax code (Title 26, Part 1) in previous iterations of RegData was provided as one file. This file constitutes a large portion of any one year of Federal Regulations and as a unit of analysis can be unwieldy. Additionally, large files do not work well with the NAICS code classification algorithm, which truncates documents before classification. Therefore, for this release of RegData, the tax code was further split at the section level. This added, on average, about 3,472 documents per year. For years after 2002, the code was split by using the XML data provided by the U.S. Government, which provided very clean splits. For 2002 and prior, the entire tax code was not available in XML format and so the code was split proportionally according to the number of words in each of 2003's splits. Therefore, the splits pre-2003 for Title 26, Part 1 are not precise and should be used with caution. The texts can always be aggregated back up to form a complete document.

Primary Features of RegData

Unit of Analysis

The CFR is divided into fifty topical titles, each published across one or more volumes. Titles are subdivided, with varying levels of consistency, into chapters, subchapters, parts, sections, subsections, paragraphs, and subparagraphs. These subdivisions generally correspond to levels of topical specificity.

While the CFR is divided and subdivided into several different demarcated portions—such as title, chapter, part, subpart, section, and paragraph—all downloadable RegData U.S. 5.0 datasets use the CFR part as the unit of analysis. We analyze the CFR at the part level for several reasons. First, part-level division is present in every title of the CFR, and the parts in a title collectively contain all non-appendix regulatory text. Second, parts tend to focus on a set of related issues that are likely to have similar relevance to industries throughout. Third, sources of regulatory authority are cited at the part level.

Primary Data Series of Regulation

While RegData U.S. 5.0 contains many data series, the two primary data series continue to be ***restrictions*** and ***industry relevance*** – which can be multiplied together to create ***industry restrictions***.

Restrictions is a cardinal proxy of the number of regulatory restrictions contained in regulatory text, devised by counting select words and phrases that are typically used in legal language to create binding obligations or prohibitions. The current words used by RegData U.S. 5.0 are the same words used in all previous RegData versions: *shall*, *must*, *may not*, *required*, and *prohibited*. The database also includes a secondary measure of volume—the total word counts—as an alternative measure of the volume of regulation over time.

As of version 4.0, a new method of counting restrictions is released, which captures restrictions formerly hidden in lists or bullet points in the text. Please see the *New Features Included in Version 4.0* section in the "RegData 4.0 User's Guide" for more details on how restrictions 2.0 works and how it differs from the old count. Similar to the old restrictions count, the industry relevance data series can be multiplied by restrictions 2.0 to create *industry restrictions 2.0*. Our expectation is that **most users will strongly prefer the restrictions 2.0 data series over the restrictions 1.0 data series**.

Industry relevance is the second key variable in RegData, representing estimates of the relevance of a CFR part to the different sectors and industries in the economy.

RegData utilizes the industry definitions in NAICS, which categorizes all economic activity into different industries. For example, in one version of NAICS (the two-digit version), the US economy is divided into approximately 20 industries, whereas in the most granular version of NAICS (the six-digit version), the economy is divided into over 1,000 industries. To illustrate, NAICS code 51 signifies the "Information" industry, while NAICS code 511191 signifies a much more granular subsector of the information industry, the “Greeting Card Publishers” industry. RegData uses NAICS because it allows researchers to easily merge RegData with other datasets that contain information about the United States economy. In the United States, the Bureau of Economic Analysis and Bureau of Labor Statistics are just two examples of organizations that publish several datasets designed around NAICS.

To create the estimates of the relevance of a CFR part to a specific industry, RegData U.S. 5.0 uses custom trained machine-learning algorithms. These algorithms “learn” what words, phrases, and other features can best identify when a unit of text is relevant to a specific industry by analyzing our compilation of training documents. Training documents are documents that are known to be relevant to one or more explicitly named industries. Over the course of the RegData project, we have gathered tens of thousands of training documents from publications in the Federal Register that name the NAICS codes affected by rulemakings. For version 5.0, we have also added thousands of training documents from other sources, including the CFR and Wikipedia.

Combining the probabilistic output of *industry relevance* data series and the *restrictions* data series allows the researcher to create *industry restrictions*. This metric is an estimate of the number of restrictions that are relevant to a particular industry or set of industries in one or more CFR parts—see Al-Ubaydli and McLaughlin (2015) for a discussion and examples. The advent of an industry-specific metric of regulation that is comprehensive (i.e., inclusive of all federal regulations that are in effect in each year), replicable, and transparent has created paths to performing economic research on regulation in ways that were previously infeasible.

Summary of Data

RegData U.S. 5.0 provides the following for each part-level segment of the CFR dating from 1970 to 2022:

- The CFR publication year, title number, and part number.
- A count of regulatory restrictions, denoted by the individual phrases counted: *shall*, *must*, *may not*, *required*, and *prohibited* (restrictions 1.0).
- A new method of counting these restrictions to include those hidden in lists or bullet points (restrictions 2.0).
- A count of words.
- The authoring agency and department.
- A probability that the part is relevant to industries included in the 2007 NAICS at the 2, 3, 4, 5, and 6-digit levels.
- Complexity data (including Shannon Entropy, conditional counts, and sentence length). (Note that the discrete drop off in complexity between 2016 and 2017 is due to a change in document formatting, rather than a change in the actual content of regulatory text. Contact us for more details.)
- The last date the text was updated.

These data elements can be combined to produce the following metrics that have been commonly used by other researchers:

- Regulatory restrictions by agency and department.
- Estimates of regulatory restrictions by industry at any NAICS level.
- Estimates of regulatory restrictions by industry for each agency and department.

- Relative complexity of regulations for each industry, agency, and department.

Modifying RegData U.S. 5.0

Utilizing the open-source QuantGov framework, it is possible to both reproduce RegData and extend it. Modifications could include the following:

- A different definition of regulatory restrictions.
- A different unit of analysis than the CFR part.
- A different set of training documents to train the industry classifier.
- Classification into a different set of categories besides the 2007 NAICS.

The QuantGov framework also allows for analysis of bodies of text unrelated to regulation. Any modification would require some amount of technical effort (specifically in Python) that would not be required to use any of the official RegData datasets. For more information on how to modify RegData U.S. 5.0 see our technical documentation.

Methodology

Sources of Regulatory Text

We use three sources of data to construct a plain-text historical series of the Code of Federal Regulations. For the years 1970–1995, we use scans of book pages that have been processed with Optical Character Recognition (OCR) using the Tesseract-OCR engine. For 1996–2016, we use the HTML versions of the historical CFR made available by the Government Publishing Office (GPO). While GPO does make an XML version of the historical CFR available, this version is produced automatically from the language used to typeset the CFR and is not reliable for our purposes. Finally, the XML version of the Electronic CFR, a current version of the CFR produced by the Office of the Federal Register and GPO, is annualized and used for 2017-2022.

The CFR is divided into titles, which are subdivided into chapters, subchapters, parts, and so forth. We analyze at the part level for several reasons. Parts are present in every CFR title, unlike some division types, and are generally concerned with only one relatively specific topic, without being so specific as to lose important context. Moreover, it is at the part level that authoring agencies and authorizing legislation are identified in the CFR indices. Parts are parsed out of the raw text using regular expressions and extracted into individual files for analysis. Because both the OCR-processed documents and the HTML volumes occasionally suffer from missing data or difficult-to-parse formatting, we employ an error detection and smoothing algorithm to produce the final series. This procedure affected less than 5 percent of all CFR parts analyzed.

NAICS Algorithm

NAICS Codes

For each unit of analysis (e.g., for each CFR part), we estimated the probability that each unit pertains to a specific industry. In our case, we used the NAICS (North American Industry Classification

System) codes as the pool of industries to choose from. This version of regdata uses the 2022 updates to the NAICS codes. More information on these codes can be found here. We chose these codes to represent our industries for a variety of reasons. First, they are widely used in and out of research contexts throughout the U.S., Canada and Mexico. There are a variety of available economic statistics that use these codes and, as such, provide a common classification system for economic analyses. Second, they are maintained by the governments of those three countries and are updated every 5 years to ensure quality and consistency, the last update being in 2022. Third, the codes cover a wide range of industries at various levels of specificity, with 2-digit codes corresponding to the most general industries, and 6-digit codes corresponding to the most specific. In 2022, there were 24 codes at the 2-digit level and 1,012 codes at the 6-digit level. More specific industries are subdivisions of more general ones, so that industries have exactly one parent and one or more children. For example, 11 is “Agriculture, Forestry, Fishing and Hunting”, 111 is “Crop Production”, 1113 is “Fruit and Tree Nut Farming”, 11133 is “Noncitrus Fruit and Tree Nut Farming” and 111331 is “Apple Orchards”. In addition, despite the codes covering an almost exhaustive list of industries at various levels of specificity, each code at the 6-digit level is exclusive of one another and provides a unique classification of each industry without overlap. This provides a nice system for document labeling and machine learning tasks.

Trainer Labeling

To classify documents and estimate probabilities, whose sources range from the Code of Federal Regulations (CFR) to state statutes and regulations, we chose to implement a machine learning model. This was preferable to labeling our entire collection of documents by hand, which would be very time-consuming and would have to be completed every year. To implement this machine learning process, we needed documents that were labeled to be able to train our model. Machine learning methods require a great deal of accurately labeled data to perform well and as will be shown below, the amount of training documents and the resultant accuracy of a model are in direct relationship. For example, at the 6-digit level there are 1,012 codes, which means that, if we expected 100 documents per label, we would have to label at least 100,000 documents. In previous iterations of RegData, we attempted to train a model for each 4, 5 and 6-digit code which required a large amount of data. Since it was difficult to label and find that amount of data, the accuracy of the models suffered. We employed a “filtering” process which excluded results from the poorer performing models (i.e. results from those models with an F1 score 2 or more standard deviations below the median). Unfiltered data had to be specifically requested and came with a disclaimer as to the quality. It is a very difficult task, even for a powerful machine learning model, to have to decide between two very similar labels at the 6-digit level (for example, “Soybean Farming” vs. “Dry Pea and Bean Farming”) than at the 3-digit level (“Crop Production” vs. “Animal Production and Aquaculture”), which can lead to inconsistent and inaccurate model output. Therefore, we decided to focus our efforts on creating an improved 3 digit level model and will explore creating improved models at the 4, 5, and 6 digit levels in future iterations.

To reduce the amount of manual effort in labeling trainers we decided to leverage existing labeled data as much as possible through the Federal Register (FR) API which provides XML versions of proposed and final federal regulations. In some proposed and final rules, agencies use the NAICS codes and descriptions to identify the industries to which their rules are expected to apply. As the primary source of our training documents, we searched proposed and final rules published in the Federal Register for all years. We also searched other sources including Wikipedia and state regulations. In total, we found and labeled 33,792 training documents through searching those sources.

1. We searched for exact matches of the numerical codes combined with the acronym NAICS as well as the numerical codes along with their description. This produced 7,985 training

documents, which was not enough to have, on average, 100 documents per label, which we felt was the minimum to properly train each code.

2. A broader search was then performed where just the description of each code was searched. This provided a great deal more documents (17,488). Some descriptions that were too broad were identified, such as “Real Estate”, and those matches were excluded from the final results. After these searches were completed, we conducted some processing.

    a. All those documents where the agency was listed as Small Business Administration, Small Business Size Standards and Personnel Management Office were excluded as documents with those agencies generally are just lists of NAICS codes and provide no information on regulations of any kind.

    b. Documents with seven or more codes listed were excluded as being too generic in nature. Third, we labeled any document with the codes of its parents. For example, any document labeled as “1113-Fruit and Tree Nut Farming” would also be labeled as “111-Crop Production” as well.

    c. Finally, we also labeled any documents with the codes of their children, if that document had not been labeled by any of the other children within that parent code. For example, if a document was labeled as “11 Agriculture, Forestry, Fishing and Hunting”, then we also labeled the document with the codes 111, 112 113, 114 and 115, if it had not already been labeled with one of those codes.

    d. These trainers were cleaned by removing any XML tags. Some of the longer trainers were identified and truncated to only include summary, overview and background sections of the FR documents to improve model performance and training time.

3. There were still some codes for which there were not enough (bare minimum 100) trainers. These codes were identified and the QuantGov team manually searched the entire QuantGov document collection for possible trainers using the NAICS indices, which are highly granular lists of industries that fall under each NAICS category, and the QuantGov search function. For example, the description for NAICS code 622110 is “General Medical and Surgical Hospitals”, but the indices provide greater detail listing out the various types of hospitals in that category (children’s, micro-hospitals, etc.). This search provided an additional 693 trainers.

4. In addition to searching their vast collection of legal documents, the QuantGov team searched Wikipedia, again using the indices, and manually reviewed the webpages to surface training documents. This produced 7,626 training documents.

Model Selection and Training

Many models were tested during the model selection process including: several transformer models (listed below), random forest, XGBoost and ridge regression The data was split into an 85-15% train-test split that was stratified by each code to ensure the test data set’s distribution of codes mirrored that of the train data. An exact 85-15% split by code was not always possible since many trainers had more than one code and data contamination (a document appearing in both the train and test split) was to be avoided. The number of documents for each code is found in table 1. Note that the total number adds up to far more than 33,972 because a document can be labeled with more than one code.

All training documents went through basic text preprocessing. This included removing extraneous spaces, removing XML and HTML tags and tokenization (breaking up the document into words). Each large language model came with its own tokenizer. After preprocessing, the training documents were fed through a number of different models.

1. RoBERTa: a few variations of the RoBERTa transformer model were tested, specifically the version adapted for multi-label classification. This model was trained for different number of epochs before and after the additional texts were added. In addition, a smaller and faster version known as DistilBERT was also tried. Different preprocessing methods were used including methods to summarize the texts before using them as training documents.
2. RoBERTa Two Layer Approach: This uses the same variety of model as above, but splits the training of the models into 25 different models, first training a model to determine which 2-digit category the texts belong to and then 24 different models to differentiate between different 3-digit codes within each 2-digit (i.e. categorize between 11, 22, 33, etc. and then train 24 different models; one each to differentiate between 111, 112, 113, 114 and 115 once the first layer has determined that the document can be labeled with 11)
3. FastText: The FastText variety of models are published/pretrained by Facebook and usually take less time to train.

Additionally, three different first generation models were trained including variations of the random forest, XGBoost and ridge regression.

Model Performance and Final Selection

The RoBERTa 2-layer approach provided the best results. To measure the performance of these models and compare results, a variety of performance metrics were evaluated, the most-important being F-1 score. These performance metrics are described the table 2. Comparing the mean and median F-1 scores for the highest-performing iteration in each method, we can see in table 3 that the 2-layer RoBERTa approach performs the best with a mean F-1 score of 56.2% and median of 58.4%.

Conclusion

Obtaining Data

With the reghub.ai website, users can customize data downloads by using the API or by using the interactive downloader. Summary data can be downloaded by specifying the year, agency, or industry the are interested in. Document-level data can also be downloaded using the API. A bulk download is also provided for users who want access to all the data. Users can also download the data for past versions (back to version 3.2) of the RegData U.S. project through the API (see here for more info on past versions). Reghub also allows users to download the raw text files used in the RegData U.S. project.

Table 4 covers all the different variables that are available in RegData U.S. 5.0 and the definitions for those variables. For more information about these variables, how to obtain data, or RegData U.S. 5.0, please email info@quantgov.org.

Appendices

Table 1: NAICS Training Documents

Code	Description	Test	Train	Total	Proportion Test
111	Crop Production	1274	2949	4223	30.2%
112	Animal Production and Aquaculture	1273	2842	4115	30.9%
113	Forestry and Logging	23	77	100	23.0%
114	Fishing, Hunting and Trapping	214	1383	1597	13.4%
115	Support Activities for Agriculture and Forestry	57	191	248	23.0%
211	Oil and Gas Extraction	255	869	1124	22.7%
212	Mining (except Oil and Gas)	252	1070	1322	19.1%
213	Support Activities for Mining	252	854	1106	22.8%
221	Utilities	118	478	596	19.8%
236	Construction of Buildings	81	264	345	23.5%
237	Heavy and Civil Engineering Construction	54	162	216	25.0%
238	Specialty Trade Contractors	131	303	434	30.2%
311	Food Manufacturing	1333	3073	4406	30.3%
312	Beverage and Tobacco Product Manufacturing	74	404	478	15.5%
313	Textile Mills	30	102	132	22.7%
314	Textile Product Mills	23	82	105	21.9%
315	Apparel Manufacturing	31	132	163	19.0%
316	Leather and Allied Product Manufacturing	22	70	92	23.9%
321	Wood Product Manufacturing	58	226	284	20.4%
322	Paper Manufacturing	101	369	470	21.5%
323	Printing and Related Support Activities	15	87	102	14.7%
324	Petroleum and Coal Products Manufacturing	205	799	1004	20.4%
325	Chemical Manufacturing	1435	3650	5085	28.2%
326	Plastics and Rubber Products Manufacturing	47	202	249	18.9%
327	Nonmetallic Mineral Product Manufacturing	110	315	425	25.9%
331	Primary Metal Manufacturing	99	282	381	26.0%
332	Fabricated Metal Product Manufacturing	93	509	602	15.4%
333	Machinery Manufacturing	124	576	700	17.7%
334	Computer and Electronic Product Manufacturing	104	459	563	18.5%

Code	Description	Test	Train	Total	Proportion Test
335	Electrical Equipment, Appliance, and Component Manufacturing	90	392	482	18.7%
336	Transportation Equipment Manufacturing	98	438	536	18.3%
337	Furniture and Related Product Manufacturing	19	77	96	19.8%
339	Miscellaneous Manufacturing	81	353	434	18.7%
423	Merchant Wholesalers, Durable Goods	224	764	988	22.7%
424	Merchant Wholesalers, Nondurable Goods	223	722	945	23.6%
425	Wholesale Trade Agents and Brokers	32	104	136	23.5%
441	Motor Vehicle and Parts Dealers	47	246	293	16.0%
444	Building Material and Garden Equipment and Supplies Dealers	25	76	101	24.8%
445	Food and Beverage Retailers	77	332	409	18.8%
449	Furniture, Home Furnishings, Electronics, and Appliance Retailers	13	57	70	18.6%
455	General Merchandise Retailers	21	65	86	24.4%
456	Health and Personal Care Retailers	20	65	85	23.5%
457	Gasoline Stations and Fuel Dealers	52	131	183	28.4%
458	Clothing, Clothing Accessories, Shoe, and Jewelry Retailers	15	39	54	27.8%
459	Sporting Goods, Hobby, Musical Instrument, Book, and Miscellaneous Retailers	25	109	134	18.7%
481	Air Transportation	121	929	1050	11.5%
482	Rail Transportation	50	281	331	15.1%
483	Water Transportation	57	176	233	24.5%
484	Truck Transportation	23	79	102	22.5%
485	Transit and Ground Passenger Transportation	22	78	100	22.0%
486	Pipeline Transportation	43	175	218	19.7%
487	Scenic and Sightseeing Transportation	44	116	160	27.5%
488	Support Activities for Transportation	138	930	1068	12.9%
491	Postal Service	36	78	114	31.6%
492	Couriers and Messengers	34	187	221	15.4%
493	Warehousing and Storage	30	70	100	30.0%
512	Motion Picture and Sound Recording Industries	69	213	282	24.5%
513	Publishing Industries	34	177	211	16.1%
516	Broadcasting and Content Providers	62	204	266	23.3%

Code	Description	Test	Train	Total	Proportion Test
517	Telecommunications	165	904	1069	15.4%
518	Computing Infrastructure Providers, Data Processing, Web Hosting, and Related Services	83	270	353	23.5%
519	Web Search Portals, Libraries, Archives, and Other Information Services	34	144	178	19.1%
521	Monetary Authorities-Central Bank	26	30	56	46.4%
522	Credit Intermediation and Related Activities	105	481	586	17.9%
523	Securities, Commodity Contracts, and Other Financial Investments and Related Activities	111	487	598	18.6%
524	Insurance Carriers and Related Activities	45	357	402	11.2%
525	Funds, Trusts, and Other Financial Vehicles	97	431	528	18.4%
531	Real Estate	21	101	122	17.2%
532	Rental and Leasing Services	25	64	89	28.1%
533	Lessors of Nonfinancial Intangible Assets (except Copyrighted Works)	8	15	23	34.8%
541	Professional, Scientific, and Technical Services	240	1180	1420	16.9%
551	Management of Companies and Enterprises	29	84	113	25.7%
561	Administrative and Support Services	196	1006	1202	16.3%
562	Waste Management and Remediation Services	183	894	1077	17.0%
611	Educational Services	181	793	974	18.6%
621	Ambulatory Health Care Services	108	546	654	16.5%
622	Hospitals	23	85	108	21.3%
623	Nursing and Residential Care Facilities	14	101	115	12.2%
624	Social Assistance	95	317	412	23.1%
711	Performing Arts, Spectator Sports, and Related Industries	37	195	232	15.9%
712	Museums, Historical Sites, and Similar Institutions	39	164	203	19.2%
713	Amusement, Gambling, and Recreation Industries	123	714	837	14.7%
721	Accommodation	24	75	99	24.2%
722	Food Services and Drinking Places	41	154	195	21.0%
811	Repair and Maintenance	124	406	530	23.4%
812	Personal and Laundry Services	39	187	226	17.3%

Code	Description	Test	Train	Total	Proportion Test
813	Religious, Grantmaking, Civic, Professional, and Similar Organizations	162	764	926	17.5%
814	Private Households	6	21	27	22.2%
921	Executive, Legislative, and Other General Government Support	23	145	168	13.7%
922	Justice, Public Order, and Safety Activities	150	891	1041	14.4%
923	Administration of Human Resource Programs	26	74	100	26.0%
924	Administration of Environmental Quality Programs	20	74	94	21.3%
925	Administration of Housing Programs, Urban Planning, and Community Development	28	71	99	28.3%
926	Administration of Economic Programs	20	81	101	19.8%
927	Space Research and Technology	31	75	106	29.2%
928	National Security and International Affairs	4	35	39	10.3%

Table 2: Algorithm Metric Descriptions

Metric	Description	Definition
F1	Balances recall and precision in a combined score.	Geometric mean of precision and recall.
Precision	Measures resistance to false positives.	Percentage of positive-classified documents that are true positives.
Recall	Measures detection ability.	Percentage of true positive documents that are classified positive.
Accuracy	Measures exact correctness but is subject to inflated scores when most observations are false.	Percentage of documents correctly classified.
Receiver Operator Characteristic – Area Under the Curve (ROC-AUC)	Measures that true positives generally have a higher predicted probability than true negatives.	Area under a curve plotting the false positive rate (percentage of true negative documents classified positive) against the recall at every probability threshold from 0 to 1.

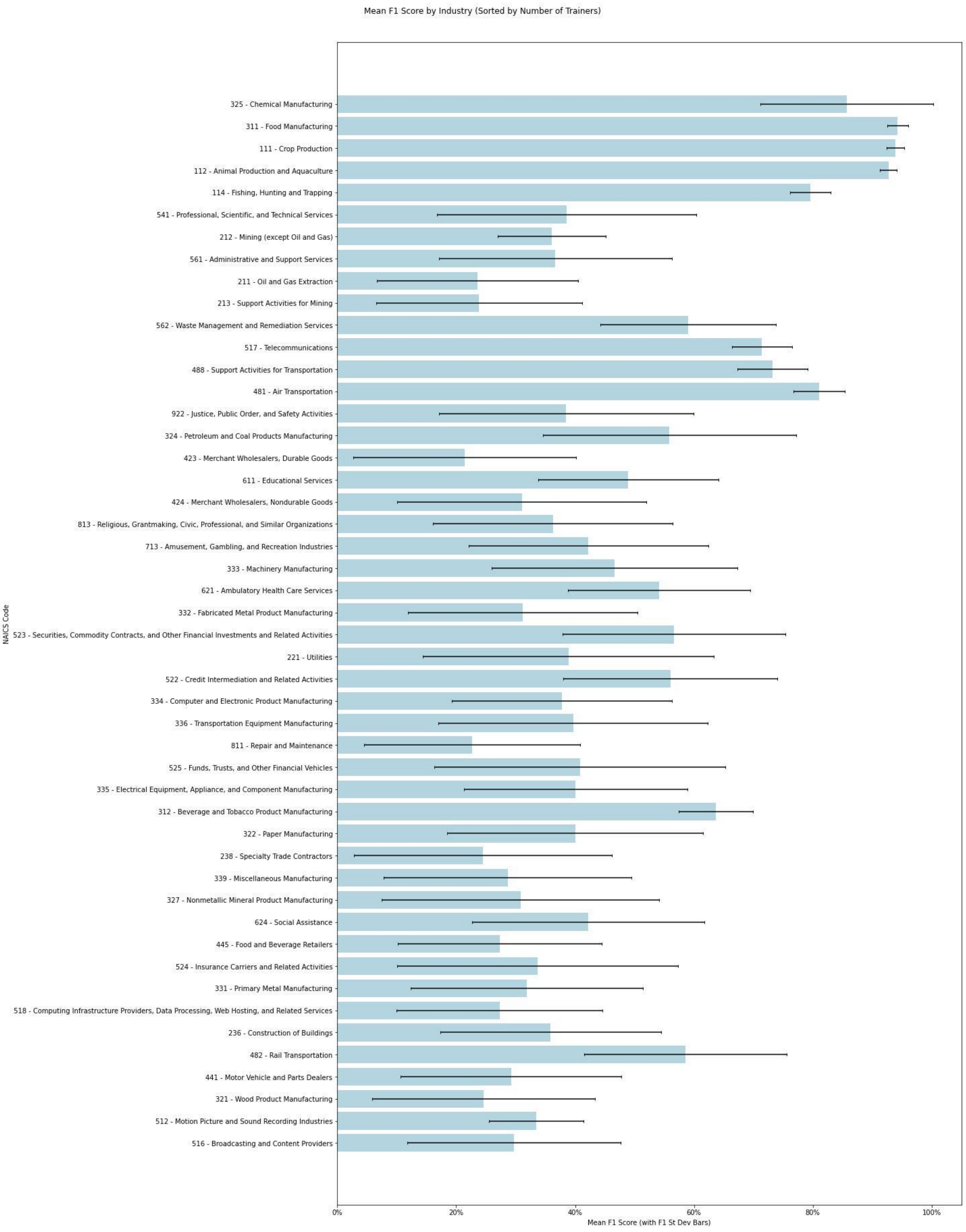
Table 3: NAICS Algorithm Performance

Method	Mean F1	Median F1	Mean ROC-AUC	Mean Precision	Mean Recall	Mean Accuracy
RoBERTa 2-Layers	56.2%	58.4%	74.8%	67.1%	50.0%	98.9%
RoBERTa 1-Layer	53.5%	56.4%	72.1%	72.0%	44.5%	98.9%
FastText	46.4%	49.0%	67.9%	57.3%	40.5%	94.6%
XGBoost	25.8%	21.8%	59.1%	64.9%	18.3%	98.6%
Random Forest	13.7%	4.3%	54.9%	52.0%	9.9%	98.5%
Ridge	13.7%	0.0%	55.3%	34.4%	10.7%	98.5%

Table 4: Variable Descriptions

Name:	**Description:**
year	Year that the part was published in. Ranging from 1970 to 2022.
title	Title containing the CFR part. Ranging from 1 to 50.
part	Part number within the title, range varies by title.
agency	Name of the agency as given in the CFR Index.
department	Name of the department associated with the agency.
words	Total number of words.
shall	Occurrences of the word *shall*.
must	Occurrences of the word *must*.
may not	Occurrences of the words *may not*.
required	Occurrences of the word *required*.
prohibited	Occurrences of the word *prohibited*.
restrictions	Total number of the five restrictive terms.
restrictions 2.0	Total number of the five restrictive terms using the restrictions 2.0 methodology to include lists and bullet points.
naics-code	The NAICS code as defined in the 2007 NAICS.
industry-relevance	Probability that the CFR part is relevant to the industry.
industry-relevant-restrictions	Number of restrictions relevant to a specific industry. Calculated by multiplying *restrictions* by *industry_relevance* at the document level.
shannon entropy	Shannon entropy score for the given part.
conditionals	Number of conditionals for the given part.
conditionals_per_100_sentences	Number of conditionals for the given part, divided by the number of sentences times 100.
acronyms	Number of acronyms for the given part.
acronyms_per_100_sentences	Number of acronyms for the given part, divided by the number of sentences times 100.
unique_acronyms	Number of unique acronyms for the given part.
long_word_score	Average squared word length for the given part.
sentence length	Average sentence length for the given part.
last-updated	Last date that the given part was updated.

Figure 1: Mean F1 Score by Industry



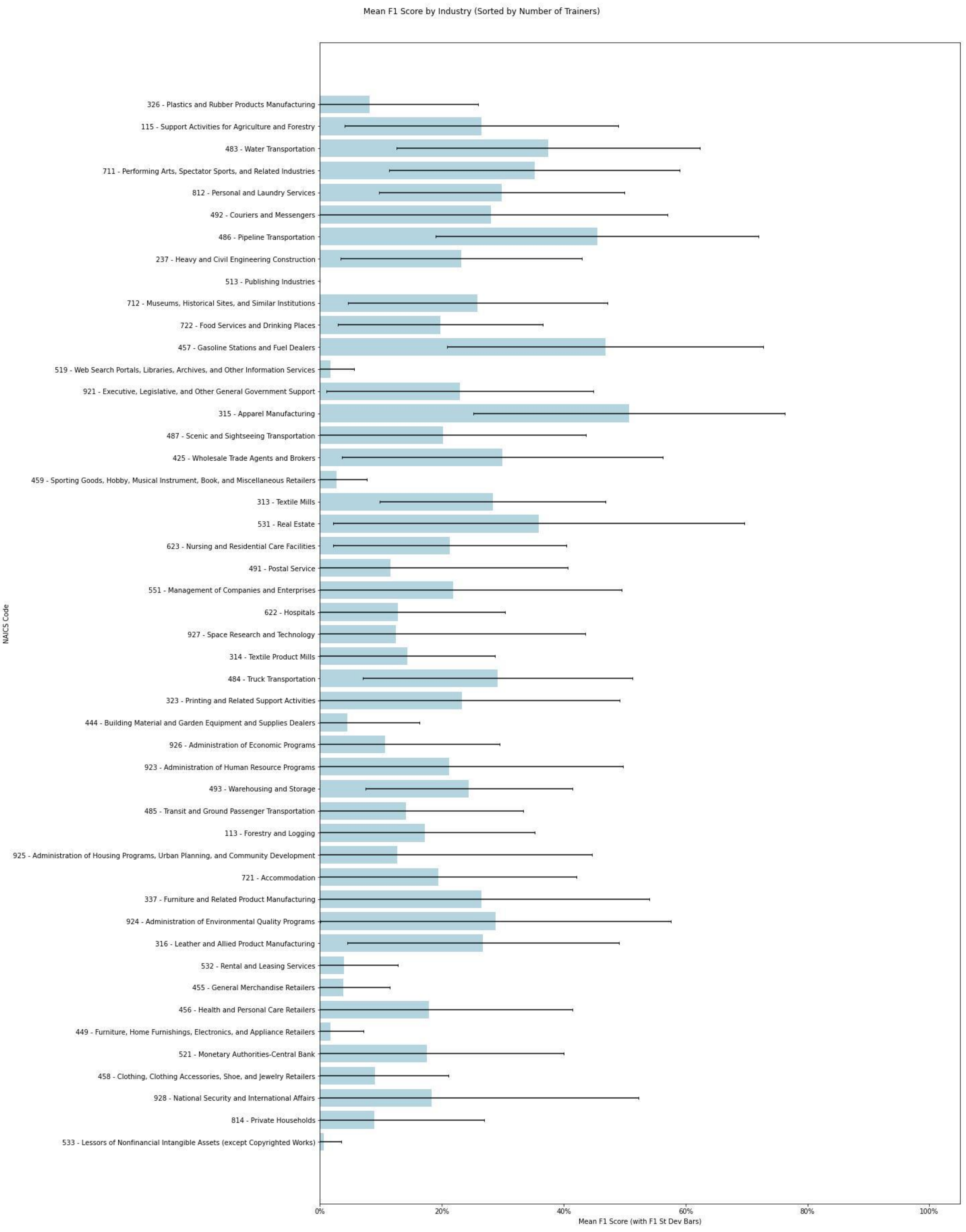
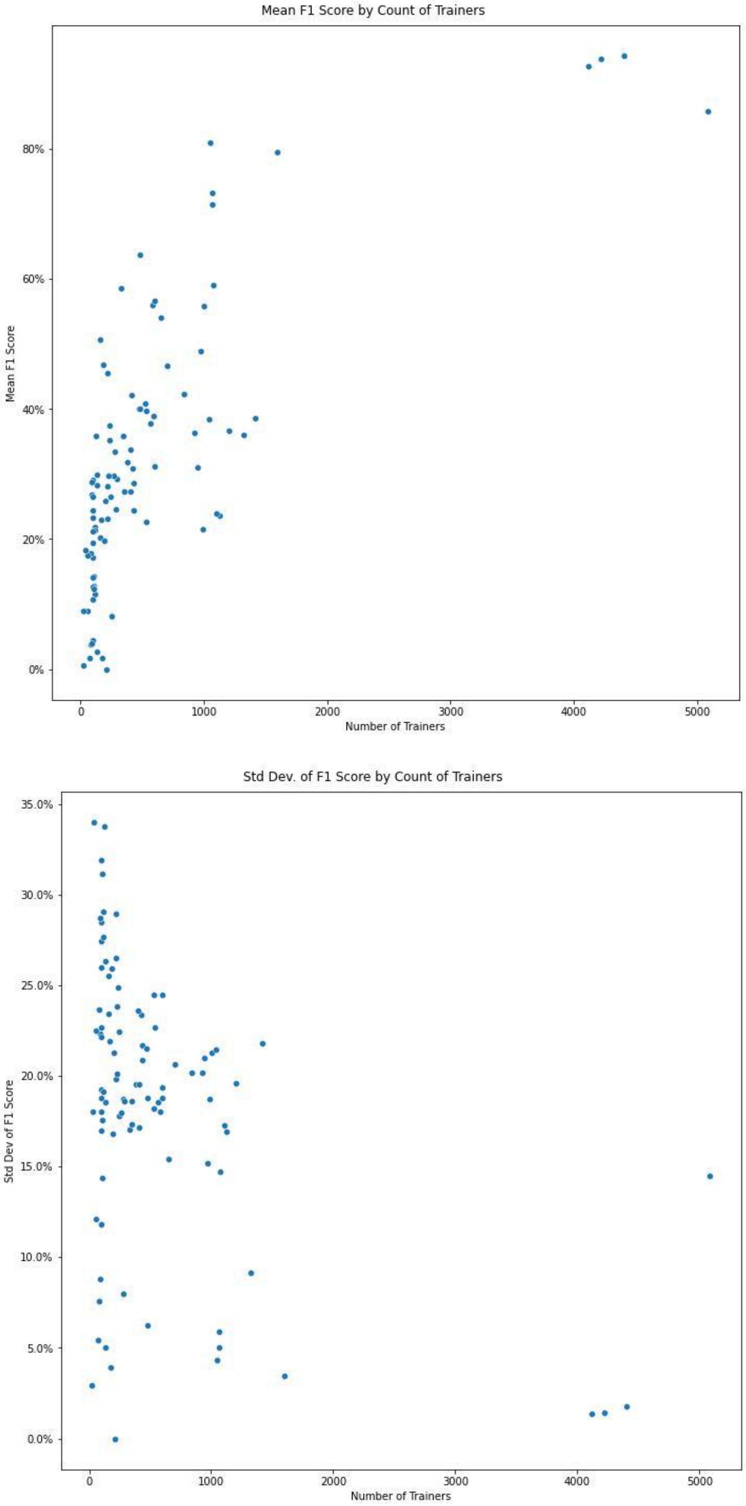

Mean F1 Score by Industry (Sorted by Number of Trainers)
NAICS Code
326 - Plastics and Rubber Products Manufacturing
115 - Support Activities for Agriculture and Forestry
483 - Water Transportation
711 - Performing Arts, Spectator Sports, and Related Industries
812 - Personal and Laundry Services
492 - Couriers and Messengers
486 - Pipeline Transportation
237 - Heavy and Civil Engineering Construction
513 - Publishing Industries
712 - Museums, Historical Sites, and Similar Institutions
722 - Food Services and Drinking Places
457 - Gasoline Stations and Fuel Dealers
519 - Web Search Portals, Libraries, Archives, and Other Information Services
921 - Executive, Legislative, and Other General Government Support
315 - Apparel Manufacturing
487 - Scenic and Sightseeing Transportation
425 - Wholesale Trade Agents and Brokers
459 - Sporting Goods, Hobby, Musical Instrument, Book, and Miscellaneous Retailers
313 - Textile Mills
531 - Real Estate
623 - Nursing and Residential Care Facilities
491 - Postal Service
551 - Management of Companies and Enterprises
622 - Hospitals
927 - Space Research and Technology
314 - Textile Product Mills
484 - Truck Transportation
323 - Printing and Related Support Activities
444 - Building Material and Garden Equipment and Supplies Dealers
926 - Administration of Economic Programs
923 - Administration of Human Resource Programs
493 - Warehousing and Storage
485 - Transit and Ground Passenger Transportation
113 - Forestry and Logging
925 - Administration of Housing Programs, Urban Planning, and Community Development
721 - Accommodation
337 - Furniture and Related Product Manufacturing
924 - Administration of Environmental Quality Programs
316 - Leather and Allied Product Manufacturing
532 - Rental and Leasing Services
455 - General Merchandise Retailers
456 - Health and Personal Care Retailers
449 - Furniture, Home Furnishings, Electronics, and Appliance Retailers
521 - Monetary Authorities-Central Bank
458 - Clothing, Clothing Accessories, Shoe, and Jewelry Retailers
928 - National Security and International Affairs
814 - Private Households
533 - Lessors of Nonfinancial Intangible Assets (except Copyrighted Works)
0%
20%
40%
60%
80%
100%
Mean F1 Score (with F1 St Dev Bars)

Figure 2: F1 Scores by Number of Trainers



Mean F1 Score by Count of Trainers

Std Dev. of F1 Score by Count of Trainers