

# TabText: a Systematic Approach to Aggregate Knowledge Across Tabular Data Structures

Dimitris Bertsimas  
Operations Research Center  
Massachusetts Institute of Technology  
Cambridge, MA 02139  
dbertsim@mit.edu

Kimberly Villalobos Carballo  
Operations Research Center  
Massachusetts Institute of Technology  
Cambridge, MA 02139  
kimvc@mit.edu

Yu Ma  
Operations Research Center  
Massachusetts Institute of Technology  
Cambridge, MA 02139  
midsummer@mit.edu

Liangyuan Na  
Operations Research Center  
Massachusetts Institute of Technology  
Cambridge, MA 02139  
lyna@mit.edu

Léonard Boussieux  
Operations Research Center  
Massachusetts Institute of Technology  
Cambridge, MA 02139  
leobix@mit.edu

Cynthia Zeng  
Operations Research Center  
Massachusetts Institute of Technology  
Cambridge, MA 02139  
czeng12@mit.edu

Luis R. Soenksen  
Abdul Latif Jameel Clinic for Machine Learning in Health  
Massachusetts Institute of Technology  
Cambridge, MA 02139  
soenksen@mit.edu

Ignacio Fuentes  
Abdul Latif Jameel Clinic for Machine Learning in Health  
Massachusetts Institute of Technology  
Cambridge, MA 02139  
ifuentes@mit.edu

**Abstract**—Processing and analyzing tabular data in a productive and efficient way is essential for building successful applications of machine learning in fields such as healthcare. However, the lack of a unified framework for representing and standardizing tabular information poses a significant challenge to researchers and professionals alike. In this work, we present TabText, a methodology that leverages the unstructured data format of language to encode tabular data from different table structures and time periods efficiently and accurately. We show using two healthcare datasets and four predictions tasks that features extracted via TabText outperform those extracted with traditional processing methods by 2-5%. Furthermore, we analyze the sensitivity of our framework against different choices for sentence representations of missing values, meta information and language descriptiveness, and provide insights into winning strategies that improve performance.

**Index Terms**—BERT, Tabular Data, Healthcare, Time Series.

## I. INTRODUCTION

Tabular data is arguably one of the most used and available data formats across different data science domains. Especially in healthcare systems, tabular data plays a crucial role in recording information for each patient such as vitals, comorbidities, diagnoses and treatments. However, it remains a challenge to find an approach that can systematically encode information from different tabular data structures and across hospital systems. As noted in several literature and previous industry attempts [1]–[3], due to the lack of standardization and harmonization in current healthcare data recording systems, significant amounts of expert knowledge as well as manual

labor are dedicated for selection, encoding and imputation of the data.

The source of the problem is that, since most of the data in healthcare systems is recorded in tabular format, machine learning models for healthcare applications often stick to the tabular structure even though a lot of the data is not tabular in nature. For instance, data attributes involving language, such as *diagnosis*, can have thousands of different values that are commonly processed as categorical features with extremely large number of categories. Another difficulty regarding the tabular approaches is that patients with different diagnoses require specific tests and treatments, and therefore a lot of the data attributes are recorded only for a very selected group of patients. Predictive models either exclude such attributes, potentially ignoring valuable data, or they impute the many missing values with the very few recorded instances.

In contrast to tabular approaches, language is a very flexible data modality that can easily represent information about different patients without imposing any structural similarity between them. Furthermore, there exist very successful off-the-shelf models for biomedical language representation that can be readily applied. In particular, deep learning methods such as BERT [4] have received a lot of attention in recent years due to their superior performance in tasks ranging across different areas of expertise, including but not limited to question answering and sentence completion.

Several previous works have shown the potential of using natural language processing (NLP) models to systematically and efficiently process tabular data in the form of language [5]–[8]. However, these works have largely relied on training of fixed BERT-based models that are not flexible to changes of tabular structures, and thus in practice they cannot be used once multiple healthcare institutions enter the picture. These works have mostly assumed that encoding data using advanced deep learning models leads to better performance in comparison to traditional data processing methods, however, concrete evidence hasn't been provided. In addition, language models are considered sensitive to its input representations [9], and most previous works do not thoroughly investigate how the choice of language affects their results, and they did not explore different language representations for missing values. Lastly, previous approaches do not consider how to cohesively combine static information from tabular data with time-series components, in which several observations may correspond to the same patient at different time periods. In our work, we build and evaluate a new data processing methodology and address the aforementioned questions.

The main contributions of this work are as follows:

- We develop a BERT-based method for unified data processing that uses language to systematically process 1) tabular data from different table structures, 2) time-series features and 3) additional non tabular data such as meta information.
- We analyze how the exact choice of language utilized by our model for data representation affects its performance.
- We show that our processing method is compatible with standard machine learning models and outperforms traditional processing approaches by 2-5% AUC across two different data sets.

## II. TABTEXT

In standard data processing methods, categorical and binary features are handled using label-encoding or one-hot encoding, time-series data is represented by including features with statistical summaries of the data (e.g. mean, min, max, variance, average change, etc.), and the remaining numerical features are processed with their original values. In addition, missing values are often imputed with 0, although more sophisticated methods can also be used.

In this section we describe our method TabText; a new data processing framework for machine learning applications in healthcare. We explain how TabText is able to process data from different tabular structures without imposing any structural constraints, as well as how it handles time-series features and missing data.

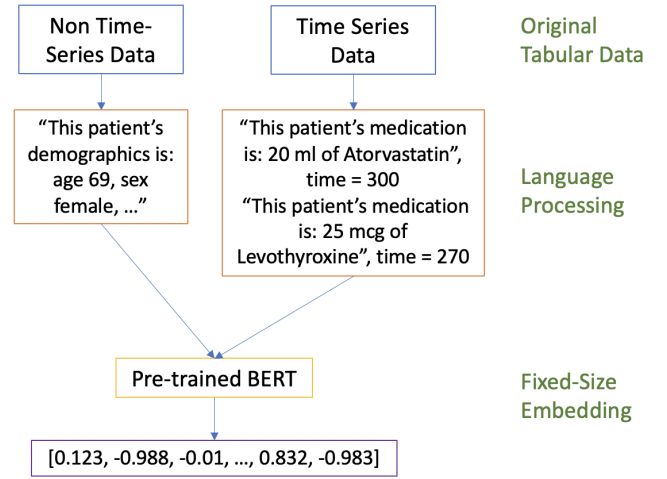


Figure 1. General Framework of the TabText Processing

### A. From Tabular Data to Language

When interpreting tabular data, humans read each row by scanning not only the specific values in it, but also the corresponding column headers that provide context for each cell. In addition, they take into account the meta information of the table like the title or any other available descriptions. Inspired by these facts, in TabText we process tabular data by creating a sentence for each row that contains the column attribute with its corresponding value, as well as any available meta information (See Figure 1). We then pass this sentence into a pre-trained, publicly available BERT model to generate embeddings that can later be used to conduct downstream tasks with any standard machine learning model (e.g. neural networks, trees, etc). Since we focus on healthcare applications, we use a pretrained Bio+ClinicalBERT [10], [11] model from HuggingFace that generates embeddings for biomedical language. These embeddings have fixed dimension (768) and can be produced for any text with less than 510 characters. Whenever the sentence for a particular row exceeds this limit, we separate it into smaller sentences that satisfy the length requirement; we generate embeddings for each sub-sentence and average them all to obtain the final embedding.

### B. Time Series Evaluation

Time-series data is an essential component for healthcare applications, since medical decisions take into account the entire medical record of the patients. However, it remains a challenge to systematically and holistically combine non-time-series with time-series data into a cohesive predictive framework. To process rows that correspond to the same patient at different points in time, TabText generates embeddings for each one of those rows separately and then combines them using a weighted sum, where the weight corresponding to a particular row is the respective time-stamp. This approach allows TabText to give more importance to recent observations while still obtaining a fix size embedding that represents the

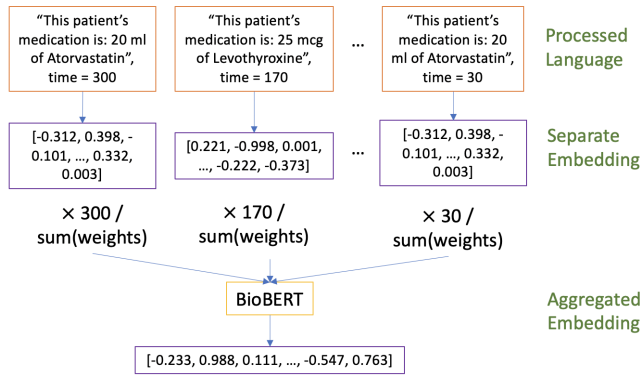


Figure 2. Weighted Average of Time-Series Data

medical history of each patient. We demonstrate this approach in Figure 2.

### C. Sentence Representation

We consider several ways of representing our data as language that we think could potentially impact the results of our performance. Specifically, we investigate the following:

1) *Missing Values*: Biomedical data usually has significant amounts of missing values, since specific medical records are tracked for specific patients. Machine learning models then rely on preprocessing methods such as imputation [12] to fill missing entries with synthetically generated values that are computed as to approximate the real ones. However, this approach has important limitations. For example, missingness itself could be a representation of certain bias in the original dataset, and by imputing missing values we could artificially worsen the existing bias. Furthermore, traditional methods often convert non-numerical features to numerical values before imputation, which further induces another layer of error propagation and information loss. We thus investigate four possible ways of directly encoding missing values into our TabText framework:

- **Exclusion**: we simply do not include this information at all in our generated language.
- **Encode Missing**: we encode the information by adding in the sentence that the column "is missing".
- **Zero Padding**: we encode the information as "is 0".
- **Maintain Original Information**: we encode the information with the original missing value, be it an empty string or "NaN".

2) *Inclusion of Meta Information*: Meta information such as a high-level descriptions of the table content is often available and can be helpful for downstream models to distinguish

features generated from different data sources. Thus, for each embedding we test the impact of including or excluding this information.

3) *Descriptiveness of Language Representation*: Creating sentences that follow the natural form of language could potentially help the BERT models to understand the content of the model better. We thus consider two possibilities: to either directly concatenate column name with column value into the sentence, or to manually write descriptive language for encoding the information.

4) *Process Languages from Different Sources*: BERT models may benefit from texts whose sentences share the same context. We then consider the impact of maintaining separate embeddings for each tabular source via concatenation, versus combining the languages from different sources into a single paragraph for BERT embedding.

Figure 3 shows an example of each representation choice.

Age	Sex	BMI	Zip Code	State
39	Female	26.4	02139	MA

Pulse	Blood Pressure	Weight
70	110	170

#### Missing Values

VS  
 "age is missing"  
 "age is 0"  
 "age is NaN"

#### Descriptiveness of the Columns

VS  
 "age:39, Sex: Female, BMI: 26.4, Zip Code: 02139, State: MA"  
 "This person has an age of 39, Sex is Female, BMI is 26.4, Zip Code is 02139, State is MA"

#### Inclusion of Table Description

VS  
 "The following is the demographics information of this patient, which describes information such as name, date of birth and address, along with insurance information. age:39, Sex: Female, BMI: 26.4, Zip Code: 02139, State: MA"  
 "age:39, Sex: Female, BMI: 26.4, Zip Code: 02139, State: MA"

#### Languages from Different Sources

VS  
 "age:39, Sex: Female, BMI: 26.4, Zip Code: 02139, State: MA"  
 "Pulse: 70, Blood Pressure: 110, Weight: 170"  
 "age:39, Sex: Female, BMI: 26.4, Zip Code: 02139, State: MA, Pulse: 70, Blood Pressure: 110, Weight: 170"

Figure 3. Different representations of the data

## III. RESULTS

### A. TabText for Classification Problems

We study the effect of TabText on classification problems using two confidential hospital data sets. The first data set (A) concerns the prediction of patient mortality after the dispatch of a rapid response team, in which event an imminent clinical deterioration summons immediate provider attention.

The second data set (B) concerns the prediction of length of stay as well as mortality at end of stay for each inpatient in the hospital.

Data set A for rapid response mortality prediction uses 1590 samples with 121 positive labels and include information on demographics, encounters (non-time-series), medication, problems, signs and socials (time-series). Data set B contains medical records of 47345 patients, with 30 different attributes on demographics, vitals, treatment and diagnosis. We use data set B for three different binary classification tasks: end of stay mortality (whether patients die or not at the end of their stay), discharge in the next 24 hours (whether patients are discharged or not within the next 24 hours) and similarly discharge in the next 48 hours.

We first split the original data into 80/20 training and testing, then train a gradient boosted tree model using only the training set and select the best hyperparameter combination that gives the best validation performance using gridsearch. We evaluate our performance using area under the receiver operating curve (ROC AUC) on the test set. Below we report the best performing AUC on the test set achieved by any TabText sentence representation, as well as the best performing AUC achieved by traditional processing through either of the time-series approaches.

Table I  
TABTEXT VS. TRADITIONAL PROCESSING RESULTS

Out Of Sample AUC			
Data Set	Task of Interest	Traditional	TabText
A	Rapid Response Mortality Prediction	0.693	0.741
B	End of Stay Mortality Prediction	0.791	0.830
B	Discharge in the next 24h Prediction	0.730	0.744
B	Discharge in the next 48h Prediction	0.726	0.743

We see that on average across the 4 binary classification tasks, TabText is able to improve on traditional tabular processing by AUC of 2-5%.

### B. Sentence Representation Sensitivity

In Table II we use the rapid response example to demonstrate that different sentence representations can substantially change the outcome of the prediction tasks. We note that in comparison to the baseline model, the best performing combination outperforms by 5%, whereas the worst performing combination underperforms by 6%.

Table II  
TABTEXT ON RAPID RESPONSE TEAM MORTALITY PREDICTION

Missing Handling	Meta Info	Descriptiveness	Test AUC
Is missing	Include	Yes	0.741
Exclusion	Does not Include	Yes	0.721
Is missing	Does not Include	No	0.715
Original	Include	No	0.713
Exclusion	Does not Include	No	0.712
Is 0	Include	No	0.712
Is 0	Include	Yes	0.704
Is 0	Does not Include	Yes	0.688
Original	Include	Yes	0.681
Exclusion	Include	No	0.679
Is 0	Does not Include	No	0.675
Is Missing	Does not Include	Yes	0.653
Is Missing	Include	No	0.651
Exclusion	Include	Yes	0.650
Original	Does not Include	No	0.639
Original	Does not Include	Yes	0.635

Below we also report the aggregated performance of the above table averaged by each type of representation that we are interested in.

1) *Descriptiveness of Language Representation*: We see that writing the sentences in a more formal and human-readable manner on average does not outperform a naive machine-generated sentence that includes the same content.

Descriptiveness	Test AUC
Yes	0.687
No	0.684

2) *Missing Values*: Overall, we see that exclusion of all missing information or encoding the missingness of the information directly into the language representation improves upon keeping the original column information by around 2.5%. However, the exact representation method for missing value does not play an important role in improving the performance.

Missing Handling	Test AUC
Exclusion	0.691
Is Missing	0.695
Is 0	0.690
Original	0.667

3) *Inclusion of Meta Information*: The inclusion of a meta information shows to improve the AUC by 1%, implying that giving the model high-level conceptual information helps its performance.

Inclusion of Meta Information	Test AUC
Include	0.691
Does not include	0.680

### C. Aggregate Knowledge Across Tabular Data Structures

In comparison to traditional methods of manual processing, a crucial practical advantage our framework provides is its

ability to easily integrate tabular data of different sources and structures in a standardized and harmonized format. Leveraging the unstructured data format of language, our method provides unparalleled flexibility to encode all types of information in a holistic manner. For example, traditional data processing includes:

#### Data cleaning:

- Anything that is not of standardized format need to be processed, for example, height of 5'3" cannot be used directly.
- Information with no way of standardization, for example, a column of dosage unit may have both ml and capsule, without specifying the exact definitions and conversion rules between these possible values.

#### Categorical variables:

- Requires us to convert to either ordered numerical levels (consecutive integers) or binary categories using one-hot encoding.
- If number of categories is too large, we need to restrict to a smaller subgroup of categories and discard the rest of the categories.

#### Missing values:

- Filter population samples by discarding samples into complete-case study.
- Compute missing entries with existing imputation methods.

However, with TabText, all of the above issues can be easily resolved by directly adding this information as part of the generated language paragraph. Doing so first provides immense data cleaning and processing speedups as it requires minimal human labor as opposed to traditional methods. Secondly, it does not sacrifice any information for the purpose of constructing them into a standardized format, and maintains all information in their most native form.

## IV. DISCUSSIONS AND LIMITATIONS

We introduce a new framework for processing tabular data via natural language and we use BERT-based methods to generate embedding that can be used to make downstream predictions. We demonstrate that our method offers both, substantial performance improvements as well as practical advantages for industry deployment. We show that the different choice of representing missing values, meta-information, as well as the descriptiveness of the generated language form are crucial to TabText's success. Lastly, we provide approaches to cohesively combine information from both non time-series data as well as time-series data.

We note that the current approach should be tested on more benchmark data sets as well as other machine learning tasks

such as regression and clustering. Furthermore, we wish to investigate possible approaches such as BioNumQA-BERT that improve on previous methods for encoding of numerical values [13]. Lastly, further research is needed for settings in which the final feature space is very high dimensional, which can affect predictive power as well as model interpretability.

## REFERENCES

- [1] Lester Darryl Geneviève et al. "Factors influencing harmonized health data collection, sharing and linkage in Denmark and Switzerland: A systematic review". In: *PLOS ONE* 14.12 (Dec. 2019). Ed. by Danuta Wasserman, e0226015. DOI: 10.1371/journal.pone. 0226015. URL: <https://doi.org/10.1371/journal.pone. 0226015>.
- [2] Yang Nan et al. "Data harmonisation for information fusion in digital healthcare: A state-of-the-art systematic review, meta-analysis and future research directions". In: *Information Fusion* 82 (June 2022), pp. 99–122. DOI: 10.1016/j.inffus.2022.01.001. URL: <https://doi.org/10.1016/j.inffus.2022.01.001>.
- [3] Evan Sweeney. *Experts say IBM Watson's flaws are rooted in data collection and interoperability*. Sept. 2017. URL: <https://www.fiercehealthcare.com/analytics/ ibm-watson-s-flaws-trace-back-to-data-collection- interoperability>.
- [4] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018. DOI: 10.48550/ARXIV.1810.04805. URL: <https://arxiv. org/abs/1810.04805>.
- [5] Jonathan Herzig et al. "TaPas: Weakly Supervised Table Parsing via Pre-training". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 4320–4333. DOI: 10.18653/v1/ 2020.acl-main.398. URL: <https://aclanthology.org/2020. acl-main.398>.
- [6] Pengcheng Yin et al. "TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 8413–8426. DOI: 10.18653/v1/2020.acl-main.745. URL: <https://aclanthology.org/2020.acl-main.745>.
- [7] Inkit Padhi et al. "Tabular Transformers for Modeling Multivariate Time Series". In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021, pp. 3565–3569. DOI: 10.1109/ICASSP39728.2021.9414142.
- [8] Gowthami Somepalli et al. *SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training*. 2021. DOI: 10.48550/ARXIV. 2106.01342. URL: <https://arxiv.org/abs/2106.01342>.
- [9] Aamir Miyajiwala et al. *On Sensitivity of Deep Learning Based Text Classification Algorithms to Practical Input Perturbations*. 2022. eprint: arXiv:2201.00318.
- [10] Jinhyuk Lee et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: (2019). DOI: 10.48550/ARXIV.1901.08746. URL: <https://arxiv.org/abs/1901.08746>.
- [11] Emily Alsentzer et al. *Publicly Available Clinical BERT Embeddings*. 2019. DOI: 10.48550/ARXIV.1904.03323. URL: <https://arxiv.org/abs/1904.03323>.
- [12] Stef Van Buuren and Karin Oudshoorn. *Flexible multivariate imputation by MICE*. Leiden: TNO, 1999.
- [13] Ye Wu et al. "BioNumQA-BERT: Answering Biomedical Questions Using Numerical Facts with a Deep Language Representation Model". In: *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics. BCB '21, Gainesville, Florida: Association for Computing Machinery, 2021*. ISBN: 9781450384506. DOI: 10.1145/ 3459930.3469557. URL: <https://doi.org/10.1145/ 3459930.3469557>.