Classification of Medical Transcriptions

Anastasia Piskunova, Evgeniy Shevlyakov, Arina Pirogova

December 2022

Abstract

In this paper, we have implemented a model to correctly classify the medical diagnosis based on the given medical transcriptions. The main goal is to correctly classify the medical specialties based on the transcription text and suggest our approach to solving this problem.

You can read the program at this link: `https://github.com/evgainy/HMLP`.

1 Introduction

Medical transcription (MT) is part of the healthcare industry that renders and edits doctor dictated reports, procedures, and notes in an electronic format in order to create files representing the treatment history of patients. Health practitioners dictate what they have done after performing procedures on patients, and MTs transcribe the oral dictation, edit reports that have gone through speech recognition software, or both. Clinical text or biomedical text literature can be seen as a large unstructured data repository, which makes text mining come into play.

1.1 Team

Anastasia Piskunova created a text corpus, developed the programm code, prepared this document.

Evgeniy Shevlyakov developed the programm code and prepared this document.

Arina Pirogova developed the programm code and prepared this document.

2 Related Work

With the data collected from PubMed and Medline, authors have created a literature database from that literature they used to take one of each literature

[1]. They used part-of-speech (POS) tagging, phrase block's formulation, and designed VWIA algorithm. Then they have used a model called conditional random fields (CRF) model. This combines the best of both HMM and MEMM [2]. Dynamic biomedical information is extracted, namely association between biomedical entities which is often extracted based on entity co-occurrence analysis with statistics theory [3]. For that purpose, they were using an algorithm called mining multiclass entity association (MMEA) [4].

Another set of researchers have collected information Medline and ScienceDirect [5], and used NLP methods are based on prior knowledge on how language is structured and on specific knowledge on how biological information is mentioned in the literature [6]. The analysis results show that pre-training BERT on biomedical corpora helps it to understand complex biomedical texts [7].

There is research on the natural language processing task for Chinese electronic medical records [8].

At the moment, we have not found similar studies in Russia. More details on the results can be found in table 1.

Authors	Technology/algorithm	F1-score
Lee	BERT	0.79
Fleuren	SVM	0.42
Gong	Conditional random fields (CRF) model	0.74
Qinghui Zhang	LSTM and GRU models	0.73

Table 1: Literature review table

3 Model Description

The general approach to solving the problem is presented on fig 1.

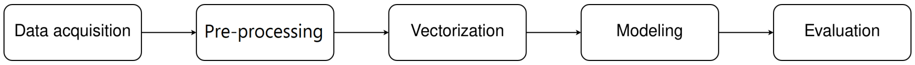


Figure 1: Block diagram of the proposed model framework

A feature of our work is the comparison (and finding the best among them) of different methods for the task of classifying medical transcriptions.

Detailed description of the dataset and its pre-processing are presented in section 4. After pre-processing the dataset, we moved on to the vectorization stage.

In order to extract the features from the dataset, we have used TF-IDF vectorizer. TF-IDF stands for Term Frequency — Inverse Document Frequency and is a statistic that aims to better define how important a word is for a transcription, while also taking into account the relation to other documents from the same corpus. TF-IDF is a score which is applied to every word in every transcription in our dataset. And for every word, the TF-IDF value increases with every appearance of the word in a transcription, but is gradually decreased with every appearance in other transcription.

Let's take a look at the simple formula behind the TF-IDF statistical measure. First let's define some notations:

- N is the number of transcription we have in our dataset
- d is a given transcription from our dataset
- D is the collection of all transcription
- w is a given word in a transcription

The TF-IDF formula is

$$TF - IDF(w, d, D) = TF(w, d) * IDF(w, D),$$

$TF(w, d) = log(1 + f(w, d))$ - term frequency formula, $f(w, d)$ is the frequency of word w in transcription d,
$IDF(W, d) = log(\frac{N}{f(w, d)})$ - inverse transcription frequency formula.

We visualized the TF-IDF features using t-sne plot (fig. 2). We have extracted close to 1000 features and tried to visualize that in a two-dimensional space. So, the data points are quite close to each other.

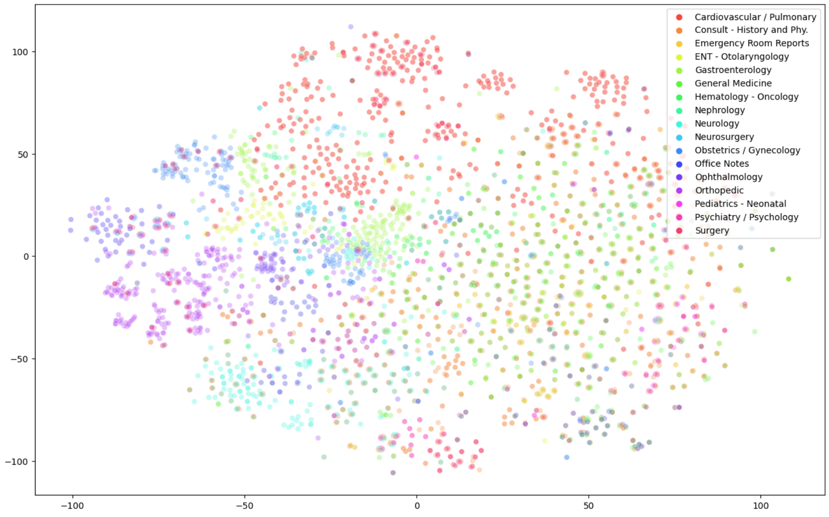


Figure 2: T-sne plot

Then we have performed PCA for reducing the dimensionality in the features for further processing. PCA [9] is commonly used for dimensionality reduction by projecting each data point onto only the first few principal components to obtain lowerdimensional data while preserving as much of the data's variation as possible. We have performed PCA in TF-IDF matrix, so after doing PCA the number of features reduced from 1000 to 439. While doing PCA, we retained the components which has variance more than 0.90.

Since some classes are in minority, we can use synthetic minority oversampling technique (SMOTE) to generate more sample form minority class to solve the data imbalance problem. In our task, we came across a term called imbalanced data distribution, which generally happens when observations in one of the class are much higher or lower than the other classes. As machine learning algorithms tend to increase accuracy by reducing the error, they do not consider the class distribution. Synthetic minority oversampling technique (SMOTE) is one of the most commonly used oversampling methods to solve the imbalance problem. It aims to balance class distribution by randomly increasing minority class examples by replicating them. SMOTE helps to generate new samples from the existing minority classes of data. It generates the virtual training records by linear interpolation for the minority class. These synthetic training records are generated by randomly selecting one or more of the k-nearest neighbors for each example in the minority class. After the oversampling process, the data is reconstructed, and several classification models can be applied for the processed data.

The first classification method we used is SVM. Support Vector Machine(SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. In the SVM algorithm, we plot each data item as a point in n-dimensional space. The algorithm creates a line or a hyperplane which separates the data into classes. We chose the linear type of the kernel because it is more optimal for multiclass classification.

The second classification method we used is multiclass logistic regression. In this case $y \in Y = 1, ..., M$, each class has its own weight vector $w_y, y = 1, 2, ..., M$. We choose the class for which the scalar product is the largest:

$$a(x) = \underset{y \in Y}{argmax}(w_y, x)$$

Then, to calculate the probability of a correct choice, you can use the following formula:

$$P(y|x, w) = \frac{exp(w_y, x)}{\sum_{z \in Y} exp(w_z, x)} = \underset{y \in Y}{softmax}(w_y, x)$$

Learning outcomes do not exceed 0.45 (F1-score) taking into account the described methods. There is a need to improve the model.

First we used a pre-processed dataset. Let us apply some domain knowledge and see if we can improve the results. The surgey category is kind of superset as there can be surgeries belonging to specializations like cardiology,neurolrogy etc. Similarly other categories like SOAP/Chart/Progress Notes, Office Notes, Consult - History and Phy., Discharge Summary, Pain Management also overlap with specialities. Hence we remove them. Two pairs of categories have been merged: Neurosurgery and Neurology, Nephrology and Urology (fig. 3).

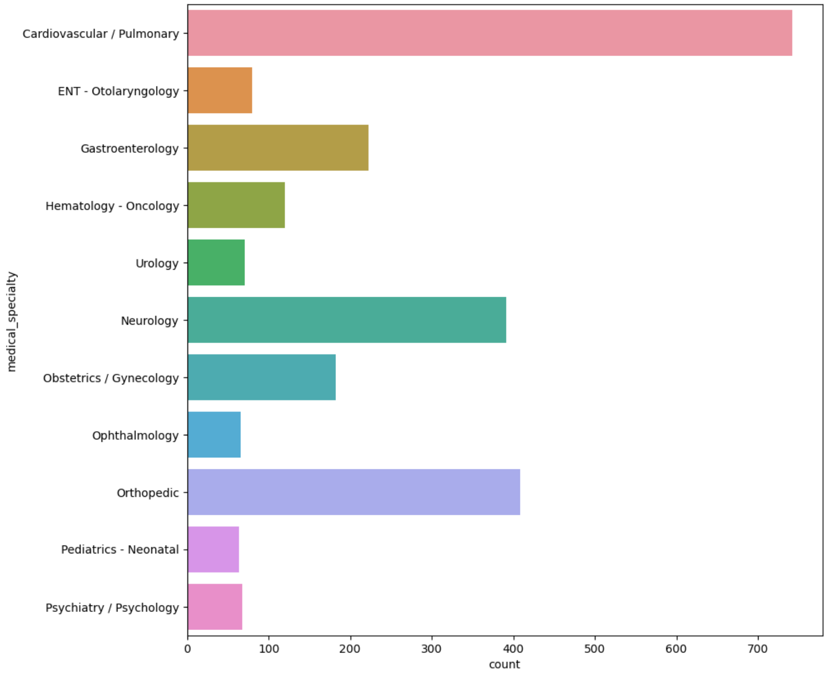


Figure 3: Histogram of updated categories

Our further steps were similar: vectorization -> dimensionality reduction -> SMOTE -> training with SVM and multiclass logistic regression.

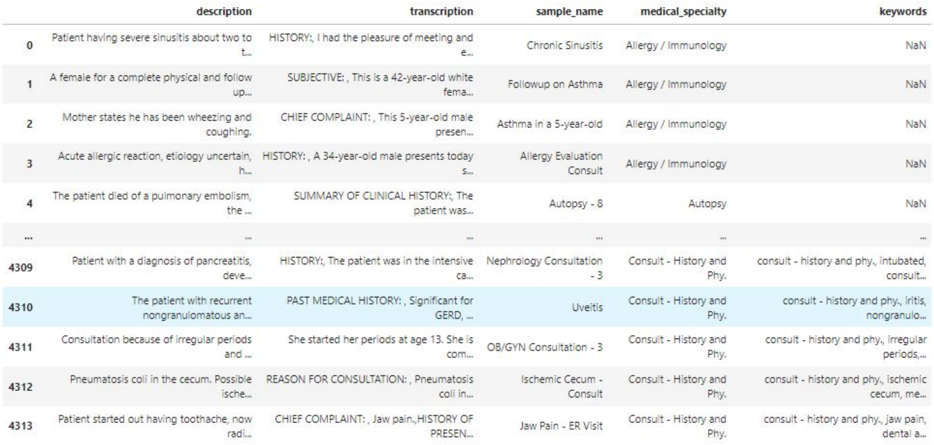
4 Dataset

Medical data is extremely hard to find due to HIPAA privacy regulations. But `https://www.mtsamples.com/` is contained sample transcription reports for many specialties and different work types. This webcite is designed to give access to a large collection of transcribed medical reports. These reports can be used by learning, as well as working medical transcriptionists for their daily transcription needs.

Data collection was carried out manually by copying data from the site. Thus, 4314 medical transcriptions from various fields of medicine were obtained.

This dataset contains six columns: 'description', 'medical specialty', 'sample name', 'transcription', and 'keywords' as shown in fig. 4.

The primary analysis of the hull is shown in table 2.



	description	transcription	sample_name	medical_specialty	keywords
0	Patient having severe sinusitis about two to t...	HISTORY:, I had the pleasure of meeting and e...	Chronic Sinusitis	Allergy / Immunology	NaN
1	A female for a complete physical and follow up...	SUBJECTIVE: , This is a 42-year-old white fema...	Followup on Asthma	Allergy / Immunology	NaN
2	Mother states he has been wheezing and coughing.	CHIEF COMPLAINT: , This 5-year-old male presen...	Asthma in a 5-year-old	Allergy / Immunology	NaN
3	Acute allergic reaction, etiology uncertain, h...	HISTORY: , A 34-year-old male presents today s...	Allergy Evaluation Consult	Allergy / Immunology	NaN
4	The patient died of a pulmonary embolism, the ...	SUMMARY OF CLINICAL HISTORY:, The patient was...	Autopsy - 8	Autopsy	NaN
...	...	...	...	...	...
4309	Patient with a diagnosis of pancreatitis, deve...	HISTORY:, The patient was in the intensive ca...	Nephrology Consultation - 3	Consult - History and Phy.	consult - history and phy., intubated, consult...
4310	The patient with recurrent nongranulomatous an...	PAST MEDICAL HISTORY: , Significant for GERD, ...	Uveitis	Consult - History and Phy.	consult - history and phy., iritis, nongranulo...
4311	Consultation because of irregular periods and ...	She started her periods at age 13. She is com...	OB/GYN Consultation - 3	Consult - History and Phy.	consult - history and phy., irregular periods,...
4312	Pneumatosis coli in the cecum. Possible ische...	REASON FOR CONSULTATION: , Pneumatosis coli in...	Ischemic Cecum - Consult	Consult - History and Phy.	consult - history and phy., ischemic cecum, me...
4313	Patient started out having toothache, now radi...	CHIEF COMPLAINT: , Jaw pain.,HISTORY OF PRESEN...	Jaw Pain - ER Visit	Consult - History and Phy.	consult - history and phy., jaw pain, dental a...

Figure 4: Sample data description of medical transcription dataset

Number of categories	33
Number of sentences in transcriptions	151094
Number of unique words in transcriptions	33858

Table 2: Counting words, sentences and categories

As part of pre-processing, we have filtered out the categories which have more than 50 samples, so the number of categories got reduced from 33 to 17.

If we look at the histogram (fig. 5), we can clearly state that it is a data imbalance problem. There are a huge number of records belonging to the class 'Consult - History and Phy.', which is almost thrice when compared with some of the other classes in the dataset. Since we are trying to classify the medical specialities based on medical transcriptions, we need only the 'transcription' and 'medical specialty' columns in the dataset.

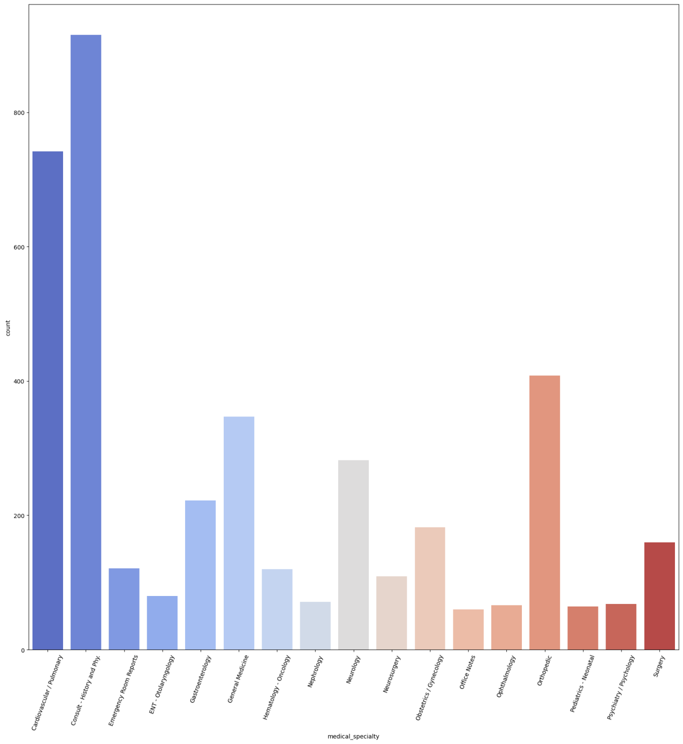
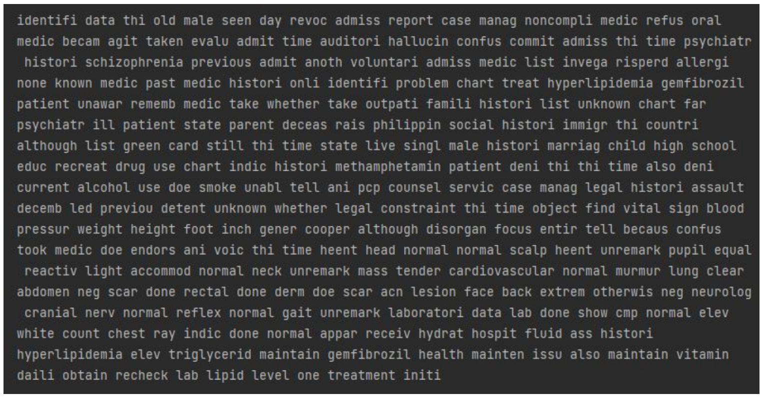


Figure 5: The histogram after category filtering

Then we transformed all the texts to lower case, deleted punctuations, removed stop words, performed lemmatization and tokenization. The result of pre-processing is shown in the Figure 6.



```
identifi data thi old male seen day revoc admiss report case manag noncompli medic refus oral
medic becam agit taken evalu admit time auditori hallucin confus commit admiss thi time psychiatr
 histori schizophrenia previous admit anoth voluntari admiss medic list invega risperd allergi
none known medic past medic histori onli identifi problem chart treat hyperlipidemia gemfibrozil
patient unawar rememb medic take whether take outpati famili histori list unknown chart far
psychiatr ill patient state parent deceas rais philippin social histori immigr thi countri
although list green card still thi time state live singl male histori marriag child high school
educ recreat drug use chart indic histori methamphetamin patient deni thi thi time also deni
current alcohol use doe smoke unabl tell ani pcp counsel servic case manag legal histori assault
decemb led previou detent unknown whether legal constraint thi time object find vital sign blood
pressur weight height foot inch gener cooper although disorgan focus entir tell becaus confus
took medic doe endors ani voic thi time heent head normal normal scalp heent unremark pupil equal
 reactiv light accommod normal neck unremark mass tender cardiovascular normal murmur lung clear
abdomen neg scar done rectal done derm doe scar acn lesion face back extrem otherwis neg neurolog
 cranial nerv normal reflex normal gait unremark laboratori data lab done show cmp normal elev
white count chest ray indic done normal appar receiv hydrat hospit fluid ass histori
hyperlipidemia elev triglycerid maintain gemfibrozil health mainten issu also maintain vitamin
daili obtain recheck lab lipid level one treatment initi
```

Figure 6: Transcription example after pre-processing

5 Experiments

5.1 Metrics

First, we introduce the notation:

- TP - True Positive (when the actual label (true label) is positive (1) and machine learning model also predicts that label as positive (1))
- TN - True Negative (when the actual label (true label) is negative (0) and machine learning model also predicts that label as negative (0))
- FP - False Positive (when the actual label (true label) is negative (0) but machine learning model predicts that label as positive (1))
- FN - False Negative (When the actual label is positive (1) but machine learning model predicts that label as negative (0))

Precision and Recall are the two most common metrics that take into account class imbalance.

Precision is the first part of the F1 score. It can also be used as an individual machine learning metric:

$$Precision = (\#\ of\ TP)/(\#\ of\ TP + \#\ of\ FP)$$

Recall is the second component of the F1 Score, although recall can also be used as an individual machine learning metric. The formula for recall is shown here:

$$Recall = (\#\ of\ TP)/(\#\ of\ TP + \#\ of\ FN)$$

Precision and Recall are the two building blocks of the F1 score, and in this project, we use this metric. The goal of the F1 score is to combine the precision and recall metrics into a single metric. At the same time, the F1 score has been designed to work well on imbalanced data. In the F1 score, we compute the average of precision and recall:

$$F1 = 2 * (Precision * Recall)/(Precision + Recall)$$

In our work, we give preference to macro-average of F1 score representation. Macro is preferred over micro as the former gives equal importance to each class whereas the later gives equal importance to each sample (which means the more the number of samples, the more say it has in the final score thus favoring majority classes much).

5.2 Experiment Setup

As mentioned earlier, the vectorization is done with TF-IDF. As part of the hyperparameters were chosen:

- ngram_range=(1, 3). This means that unigrams, bigrams, and trigrams will be taken into account while creating features;

- max_features=1000 - max number of features considered for splitting a node.

We used PCA which will reduce the dimension of features by creating new features which have most of the varience of the original data. We have passed the parameter n_components=0.9 which is the percentage of feature in final dataset.

We have carried out the alignment of the smallest classes using parametr sampling_strategy='minority' in SMOTE.

The data was split using stratified Train Test split strategy. For this we use the function train_test_split() and choose stratify=y (stratified train test split of the dataset), random_state=1 (it controls the shuffling applied to the dataset before splitting it into two sets).

One of the methods of model training is logistic regression. Its hyperparameters are:

- penalty='elasticnet' - is a combination of the two most popular regularized variants of linear regression: ridge and lasso;

- solver='saga' - is a variant of SAG that also supports the non-smooth penalty L1 option (i.e. L1 Regularization);

- l1_ratio=0.5 - the penalty will be a combination of L1 & L2, 0.5 will define the weight of L1 in the mix;

- random_state=1 - adjusts randomness seed, which will be generated by random number generator.

Another method used to train the model is SVM. Its hyperparameters are:

- C=1.0 - regularization parameter;

- kernel='linear' - choice of kernel in favor of linear;

5.3 Baselines

The simplest approaches for solving the problem are the SVM and the method of multiclass logistic regression without SMOTE. Vectorization was carried out by TF-IDF, and dimensionality was reduced using PCA in both methods. These approaches have not shown good results. Let's analyze various metrics for a deeper analysis. Table 3 compares the simplest basic approaches with different representations.

Model	Precision	Recall	F1 score	Accuracy(%)
Support vector classifier	0.50	0.38	0.40	0.53
Multiclass logistic regression	0.52	0.37	0.40	0.54

Table 3: Performance metrics for based models

6 Results

After applying the methods of SVM and multiclass logistic regression, we get the following results (tab. 4). As we can see, the best classification result is achieved by improving the dataset.

Model	Precision	Recall	F1 score	Accuracy(%)
SVC+SMOTE	0.56	0.45	0.47	0.63
LR+SMOTE	0.54	0.42	0.45	0.62
model-impr.+SVC	0.79	0.75	0.76	0.81
model-impr.+LR	0.78	0.67	0.71	0.78
model-impr.+SMOTE+SVC	0.82	0.76	0.79	0.85
model-impr.+SMOTE+LR	0.81	0.72	0.75	0.84

Table 4: Final results

There is a study that used data similar to ours [10]. Most likely, the dataset was collected from the same site as ours. In any case, the result of this study is much worse than ours (the best result is 0.65). In addition, they used completely different classification methods. Thus, we can claim that our model shows the best training outcomes on medical transcription data.

7 Conclusion

The dataset collected from the site turned out to be very noisy and with a significant data imbalance. These are the main problems that we encountered in the process of training the model. To achieve the best results, we had to improve the dataset, i.e. combine some categories and delete them using knowledge of the subject area. Then using SMOTE and different training methods (SVM and multiclass logistic regression) we managed to achieve better results (0.79) than in studies with similar datasets.

References

1. Rijo R, Martinho R, Pereira L, Silva C, "Text mining applied to electronic medical records", 2015. Int J E-Health Med Commun 6(3):1–18.

2. Blog G, "Complete tutorial on text classification using conditional random fields model (in Python)". Analytics Vidhya, 2018. Available: https://www.analyticsvid.hya.com/blog/2018/08/nlp-guide-conditional-random-fields-text-classification/.

3. Gong L, "Application of biomedical text mining", 2018. IntechOpen I:427–428.

4. Thabtah F, Cowling P, Peng Y, "MCAR: multi-class classification based on association rule", 2005.

5. ScienceDirect. Available: https://www.sciencedirect.com/.

6. Fleuren WW, Alkema W, "Application of text mining in the biomedical domain", 2015. Sci Direct 74(1):97–106.

7. Lee J, YoonW, Kim S, Kim D, Kim S, So CH, Kang J, "BioBERT: a pretrained biomedical language representation model for biomedical text mining", 2019. Cornell University, p. 10–11.

8. Qinghui Zhang, Yuan Qihao, Pengtao Lv, Mengya Zhang, Lei Lv, "Research on Medical Text Classification Based on Improved Capsule Network", 2022.

9. Jolliffe IT, Cadima J, "Principal component analysis: a review and recent developments", 2016. Royal Soc Publishing 374(2065)

10. Abdul Razak Zakieh, Adil Alpkocak, "Classification of Medical Transcriptions with Explanations", 2021.