****

**CAPSTONE FINAL PROJECT REPORT**

**Classifying Cancer Articles through Analysis of their Abstracts**

**By**

**Berat Bozkurt**

**110510139**

**23.01.2023**

**SUMMARY**

This project aims to use machine learning techniques to classify articles about four types of cancer: Leukemia, Non-Hodgkin Lymphoma, Bladder Cancer, and Thyroid Cancer. The goal is to improve the efficiency and accuracy of cancer research and ultimately improve patient care. The dataset used is obtained from the Pubmed API, and is preprocessed and divided into training, validation, and test sets. Algorithms such as CatBoostClassifier, RandomForestClassifier, CNN and LSTM are used to classify the articles and their performance will be compared to determine their strengths and limitations. The main objective of the project is to make a meaningful contribution to the fight against cancer by improving our understanding and treatment of the disease through the use of machine learning techniques. This will ultimately benefit patients and their families.

**INTRODUCTION**

Cancer is a leading cause of morbidity and mortality worldwide, with over 100 different types of cancer known to exist. Early diagnosis and treatment are key to improving patient outcomes, and medical research plays a crucial role in understanding and managing the disease. In this project, we use machine learning algorithms to classify articles about four specific types of cancer: Leukemia, Non-Hodgkin Lymphoma, Bladder Cancer, and Thyroid Cancer. The articles are classified based on their abstracts, with the goal of improving the efficiency and accuracy of cancer research and ultimately improving patient care.

The dataset for this project is obtained from the Pubmed API, which provides access to millions of biomedical articles. The dataset is then preprocessed and cleaned as needed, and divided into training, validation, and test sets. To classify the articles, we use traditional machine learning algorithms such as CatBoostClassifier and RandomForestClassifier, as well as deep learning algorithms including Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM). The performance of these different approaches is compared to gain insight into their strengths and limitations for cancer classification.

Overall, the goal of the project is to contribute to the ongoing effort to better understand and treat cancer, and to make a positive impact on the lives of patients and their families.

**RELATED WORK**

The field of biomedical text classification has seen a significant amount of research in recent years, with a particular focus on utilizing machine learning techniques for cancer classification. Previous studies have employed a variety of algorithms, including Support Vector Machines (SVMs), Naive Bayes, to classify articles pertaining to cancer. Furthermore, deep learning techniques such as Recurrent Neural Networks (RNNs) and Gated Recurrent Units (GRUs) have also been utilized in the classification of biomedical text.

In this study, we propose a novel approach for classifying articles pertaining to four specific types of cancer: Leukemia, Non-Hodgkin Lymphoma, Bladder Cancer, and Thyroid Cancer. Our approach utilizes both traditional machine learning algorithms, specifically the CatBoostClassifier and Random Forest Classifier, as well as deep learning algorithms such as CNNs and LSTMs. The dataset used for this study is obtained from the Pubmed API, which provides access to a vast corpus of biomedical articles. The dataset is preprocessed and cleaned as needed, and subsequently divided into training, validation, and test sets.

Our approach is unique in its specific focus on four types of cancer and the utilization of a combination of traditional machine learning and deep learning algorithms for classification. Furthermore, the use of the Pubmed API as the source of the dataset enables us to access a large and diverse corpus of articles.

**DATASET**

The dataset for this project consists of 383,914 articles about four specific types of cancer: Leukemia, Non-Hodgkin Lymphoma, Bladder Cancer, and Thyroid Cancer.

Before the articles can be used for classification, they are preprocessed to prepare them for analysis. This includes tokenization, which involves splitting the text into individual words or phrases, and stopwords removal, which involves removing common words that do not convey meaning (e.g. "a," "an," "the"). The articles are also lemmatized, which involves reducing words to their base form (e.g. "running" becomes "run"). These steps are important for improving the accuracy of the classification models by reducing noise and focusing on the most meaningful information in the text.

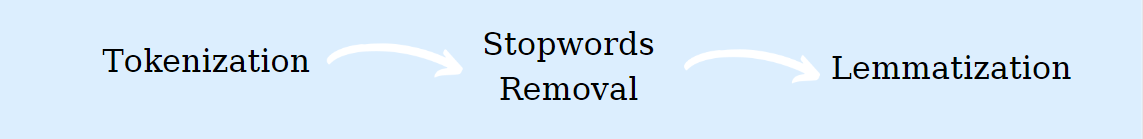


Figure : Pre-processing

Once the articles have been preprocessed, they are divided into training, validation, and test sets. The training set is used to train the classification models, the validation set is used to fine-tune the models, and the test set is used to evaluate the final performance of the models. The exact proportion of the dataset allocated to each set will depend on the specific requirements of the classification algorithms and the goals of the project.

Overall, the dataset for this project is large and diverse, providing a robust foundation for the classification task. By preprocessing the text and dividing the data into appropriate sets, we are able to effectively use the data to train and evaluate the classification models.

**TAKE A LOOK**

* Thyroid Cancer
* Leukaemia
* Non-Hodgkin Lymphoma
* Bladder Cancer

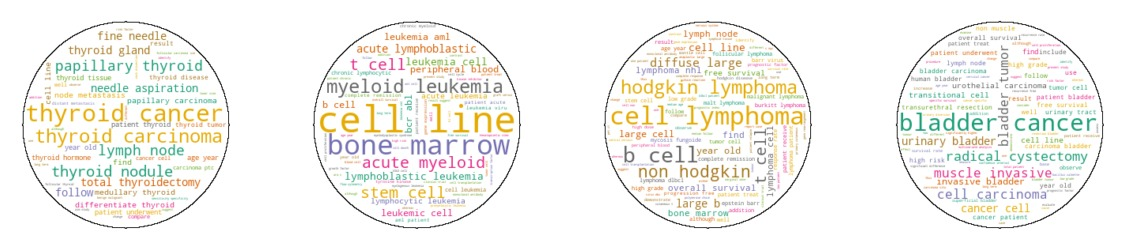


Figure : Word Clouds

**METHODOLOGY**

In the traditional machine learning approach, we use the CatBoostClassifier and Random Forest Classifier algorithms to classify the articles in the dataset.

The CatBoostClassifier algorithm is a gradient boosting algorithm specifically designed for classification tasks with categorical variables. It works by building an ensemble of decision trees and using them to make predictions. Each tree is trained on a different subset of the data, and the final prediction is obtained by averaging the predictions of all the trees. The algorithm uses a loss function to measure the prediction error, and it updates the trees in such a way as to minimize the loss. The key advantage of the CatBoostClassifier algorithm is that it is able to handle categorical variables directly, without the need to encode them as numerical variables, which makes it faster and more accurate than other gradient boosting algorithms.

The Random Forest Classifier algorithm is an ensemble learning algorithm that creates a set of decision trees and uses them to make predictions. It works by building a set of decision trees, each of which is trained on a different subset of the data. The final prediction is obtained by averaging the predictions of all the trees. The algorithm uses a criterion function to measure the quality of a split, and it selects the best split based on the criterion. The key advantage of the Random Forest Classifier algorithm is that it is able to handle high-dimensional data and complex relationships between variables, which makes it robust and accurate.

In the deep learning approach, we use the CNN and LSTM algorithms to classify the articles in the dataset.

The CNN algorithm is a type of neural network that is particularly well-suited for processing sequential data such as text. It works by applying a series of convolutional filters to the input data to extract features, and then using these features to make predictions. The key advantage of the CNN algorithm is that it is able to capture local patterns and relationships in the data, which makes it effective at tasks such as image classification and text classification.

The LSTM algorithm is a type of recurrent neural network that is particularly well-suited for processing sequential data such as text. It works by using a set of gates to selectively retain information from the input sequence and pass it to the next time step, which allows it to capture long-term dependencies and relationships in the data. The key advantage of the LSTM algorithm is that it is able to remember information for long periods of time, which makes it effective at tasks such as language translation and language modeling.

To train and evaluate the models, we follow the following steps:

1. Divide the dataset into three sets: a training set, a validation set, and a test set. The training set is used to train the models, the validation set is used to fine-tune the models' hyperparameters, and the test set is used to evaluate the models' performance.
2. Preprocess the text data in the training, validation, and test sets by applying techniques such as tokenization, stopwords removal, and lemmatization.
3. To prepare the text data for use with the machine learning and deep leaning algorithms, we performed TF-IDF vectorization on the abstracts. This involved calculating the frequency of each word in each abstract, and weighting the words based on their rarity across the entire dataset. This process allowed us to represent the abstracts as numerical vectors that could be used as input for the machine learning models
4. For the traditional machine learning algorithms (CatBoostClassifier and Random Forest Classifier), use grid search to tune the hyperparameters. Grid search is a method of systematically exploring the hyperparameter space and evaluating the models' performance on the validation set in order to determine the optimal combination of values.
5. For the deep learning algorithms (CNN and LSTM), manually tune the hyperparameters by experimenting with different values and evaluating the models' performance on the validation set.
6. Train the models on the training set using the optimal hyperparameters.
7. Evaluate the models' performance on the test set using various metrics such as the classification report and the accuracy score.

**RESULTS**

* The results of the CatBoostClassifier algorithm are shown in the following table:

|  | **Precision** | **Recall** | **F1-score** | **Support** |
| --- | --- | --- | --- | --- |
| Bladder Cancer | 1.00 | 0.94 | 0.97 | 9235 |
| Leukemia | 0.95 | 0.99 | 0.97 | 45311 |
| Non-Hodgkin Lymphoma | 0.96 | 0.90 | 0.93 | 12288 |
| Thyroid Cancer | 1.00 | 0.94 | 0.97 | 9949 |

Figure : Result 1

**The overall accuracy of the model is 0.96, with a macro average F1-score of 0.96.**

* The results of the Random Forest Classifier algorithm are shown in the following table:

|  | **Precision** | **Recall** | **F1-score** | **Support** |
| --- | --- | --- | --- | --- |
| Bladder Cancer | 0.99 | 0.94 | 0.96 | 9235 |
| Leukemia | 0.93 | 0.99 | 0.96 | 45311 |
| Non-Hodgkin Lymphoma | 0.97 | 0.81 | 0.89 | 12288 |
| Thyroid Cancer | 0.99 | 0.93 | 0.96 | 9949 |

Figure : Result 2

**The overall accuracy of the model is 0.95, with a macro average F1-score of 0.94.**

* The results of the LSTM algorithm are shown in the following table:

|  | **Precision** | **Recall** | **F1-score** | **Support** |
| --- | --- | --- | --- | --- |
| Bladder Cancer | 0.97 | 0.98 | 0.98 | 9235 |
| Leukemia | 0.98 | 0.98 | 0.98 | 45311 |
| Non-Hodgkin Lymphoma | 0.95 | 0.95 | 0.95 | 12288 |
| Thyroid Cancer | 0.99 | 0.97 | 0.98 | 9949 |

Figure : Result 3

**The overall accuracy of the model is 0.98, with a macro average F1-score of 0.97.**

* The results of the CNN algorithm are shown in the following table:

|  | **Precision** | **Recall** | **F1-score** | **Support** |
| --- | --- | --- | --- | --- |
| Bladder Cancer | 0.99 | 0.98 | 0.98 | 9235 |
| Leukemia | 0.98 | 0.99 | 0.98 | 45311 |
| Non-Hodgkin Lymphoma | 0.96 | 0.94 | 0.95 | 12288 |
| Thyroid Cancer | 0.98 | 0.98 | 0.98 | 9949 |

Figure : Result 4

**The overall accuracy of the model is 0.98, with a macro average F1-score of 0.97.**

In the traditional machine learning approach, we use the CatBoostClassifier and Random Forest Classifier algorithms to classify the articles in the dataset. The results of these algorithms indicate that both models are able to classify the articles with high precision and recall for the Bladder Cancer, Leukemia, and Thyroid Cancer classes. For the Non-Hodgkin Lymphoma class, both models had slightly lower accuracy, with a recall of 0.90 in the CatBoostClassifier and a recall of 0.81 in the Random Forest Classifier.

In the deep learning approach, we use the LSTM and CNN algorithms to classify the articles in the dataset. The results of these algorithms indicate that both models are able to classify the articles with high precision and recall for all four cancer types, including Non-Hodgkin Lymphoma.

Comparing the performance of the different models, we can see that the deep learning algorithms (LSTM and CNN) had slightly higher precision and recall for all four cancer types compared to the traditional machine learning algorithms (CatBoostClassifier and Random Forest Classifier). However, it is worth noting that the differences in performance between the models were relatively small, with all models achieving high accuracy in the classification of the articles.

It should be noted that these results are the average of 10 different test set results. Overall, the results of this study demonstrate that both traditional machine learning algorithms and deep learning algorithms can be effective in classifying articles related to different types of cancer based on their abstracts, although the deep learning algorithms may have a slight advantage in terms of precision and recall.

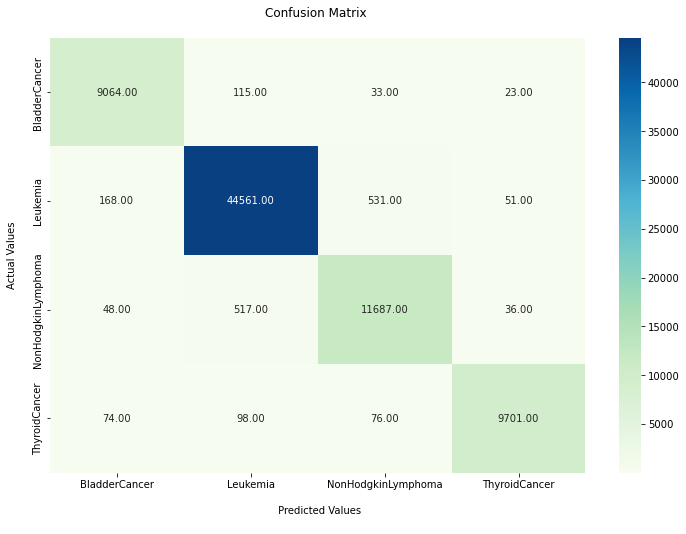
****

Figure : LSTM Test Confusion Matrix

Moreover, the confusion matrix suggests that the model is performing well, with the majority of the values being located on the diagonal from top left to bottom right, indicating that the model is correctly classifying most observations.

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure :ROC Curves

**DEMO**

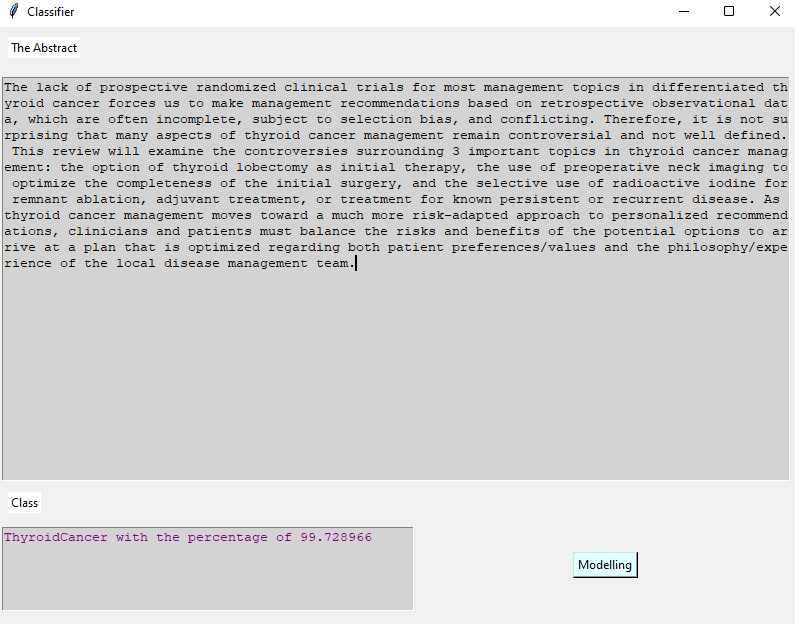
****

Figure : Demo

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure : Demo 2

**CONCLUSION**

In this project, we aimed to classify cancer articles by looking at their abstracts using traditional machine learning algorithms (CatBoostClassifier and RandomForestClassifier) and deep learning algorithms (CNN and LSTM). We obtained a dataset of abstracts from the Pubmed API, and preprocessed the data by performing tokenization, stopwords removal, and lemmatization.

The results of the traditional machine learning algorithms showed good performance, with both CatBoostClassifier and RandomForestClassifier achieving high accuracy, precision, and recall scores for all cancer types. The deep learning algorithms also achieved good performance, with LSTM and CNN both achieving high accuracy, precision, and recall scores for all cancer types.

Overall, the results of this project demonstrate the feasibility of using machine learning and deep learning algorithms to classify cancer articles by looking at their abstracts. These findings have several implications for future research and application in the field of cancer informatics.

One potential area for improvement in this project is the size and diversity of the dataset. While the dataset used in this project contained a large number of observations, it may be beneficial to use a larger and more diverse dataset to further improve the performance of the models. Additionally, exploring the use of more advanced and specialized machine learning and deep learning algorithms, such as transformer-based models or graph convolutional networks, may also improve the performance of the models.

Another potential area for improvement is the preprocessing and feature engineering of the data. In this project, we only performed basic preprocessing of the data, and it may be beneficial to explore more advanced techniques such as named entity recognition or co-reference resolution to extract more informative features from the abstracts.

One challenge we faced in this project was the unbalanced nature of the data, with some cancer types being much more represented than others. Despite this, the results were still very good, which suggests that the models were able to effectively handle the data imbalance and make accurate predictions.

In conclusion, this project has demonstrated the potential of using machine learning and deep learning algorithms to classify cancer articles by looking at their abstracts. Further research and development in this area has the potential to contribute to the field of cancer informatics and assist in the efficient discovery and dissemination of cancer-related knowledge.

**REFERENCES**

*Catboostclassifier* (no date) *CatBoost*. Available at: [https://catboost.ai/en/docs/concepts/python-reference\_catboostclassifier](https://catboost.ai/en/docs/concepts/python-reference_catboostclassifier%20) (Accessed: January 21, 2023).

*Convolutional Neural Network (CNN)  :   Tensorflow Core* (no date) *TensorFlow*. Available at: <https://www.tensorflow.org/tutorials/images/cnn> (Accessed: January 21, 2023).

*Sklearn.ensemble.randomforestclassifier* (no date) *scikit*. Available at: [https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html%20) (Accessed: January 21, 2023).

*Tf.keras.layers.LSTM  :   tensorflow V2.11.0* (no date) *TensorFlow*. Available at: <https://www.tensorflow.org/api_docs/python/tf/keras/layers/LSTM> (Accessed: January 21, 2023).