**Mini Project Report on**



**TEXT CLASSIFICATION USING MACHINE LEARNING WITH PYTHON**



**Submitted in partial fulfilment of the requirement for the award of the degree of**

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**Submitted by:**

**Student Name :Gokul Singh** **University Roll No.:2021215**

***Under the Mentorship of***

**Dr .Surendra Singh Samant**



**Department of Computer Science and Engineering**

**Graphic Era (Deemed to be University)**

**Dehradun, Uttarakhand**

**January-2024**



**CANDIDATE’S DECLARATION**

I hereby certify that the work which is being presented in the project report entitled **“Text Classification Using Machine Learning With Python”** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineeringof the Graphic Era (Deemed to be University), Dehradun shall be carried out by the under the mentorship of **Dr .Surendra Singh Samant** , Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Name :Gokul Singh University Roll no.:2021215

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Description** | **Page No.** |
| Chapter 1 | Introduction | 1-3 |
| Chapter 2 | Literature Survey | 4-5 |
| Chapter 3 | Methodology | 6-9 |
| Chapter 4 | Result and Discussion | 10 |
| Chapter 5 | Conclusion and Future Work | 11 |
|  | References | 12 |

**Chapter 1**

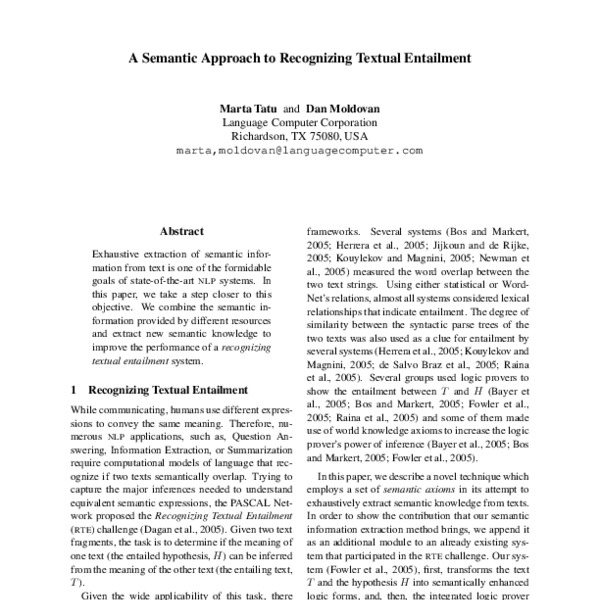
**Introduction**

* 1. **Introduction**

Text is a major part of information present online .It is really important for the text to be precise and short for some specific need, but majority of text online is big and time consuming to process and use.

This is where Text Classification helps us analyze large textual data to provide us concise data according to the need.

Fig. 1.1.1 Textual Data



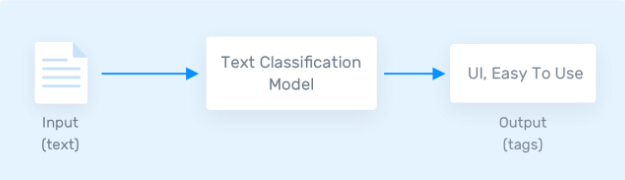
[This Photo](https://aclanthology.org/H05-1047/) by Unknown Author is licensed under [CC BY](https://creativecommons.org/licenses/by/3.0/)

* 1. **Text Classification**

Text Classification is a process of assigning a class or category to the text needed to be processed. It is widely used to categories and organize data present online in areas with high amount of data, since it is likely that the data contains parts which is not are not important for the particular task a company wants to do which often make the analysis of this data very time consuming and the results maybe not be accurate with other non relevant stuff laying around.

Machines can analyze millions of pages of text is a very short time hence preferred by many companies. Text Classification uses NLP and Machine learning algorithms to make text classification easier and scalable.

Fig. 1.2.1 .Text Classification Stages



Text Classification is used to classify all formats of textual data wheatear it is long formatted or short ,docs , contracts etc.

Some examples of Text Classification involve :

1. Text Labeling
2. Sentimental Analysis
3. Language Detection etc.
4. Spam Filtering
   1. **NLP**

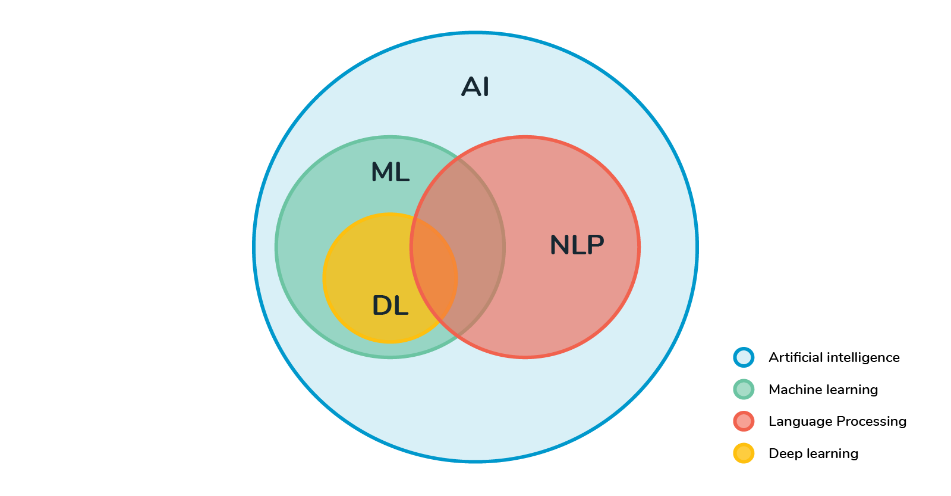
NLP or Natural Language Processing is a subset of AI which teaches a machine how to read ,understand and drive meaning from human languages. It is a part of AI that helps it to work with Linguistics and process human languages. NLP combines computer science ,linguistics and machine learning to study how humans and computer communicate in natural language.

It is widely used in all fields due to its efficiency and effectiveness. It works on a simple rule , break down the data in smaller parts and then analyze it accordingly.

Application of NLP :

1. Text/Speech processing
2. Information extraction
3. Chatbots
4. Language Translation
5. Text Summarization
6. Text Classification

Fig. 1.3.1 NLP Relation



[This Photo](https://devopedia.org/natural-language-processing) by Unknown Author is licensed under [CC BY-SA](https://creativecommons.org/licenses/by-sa/3.0/)

* 1. **Problem Statement**

The main goal of this project is to create a machine learning model that will be able to recognize and determine the label / Category of a resume using concepts like NLP and machine learning .

Though the goal is to create a model which can recognize labels it can further be used to if a candidates resume is recognizable in that category.

The overall goal of the project is to learn NLP and mavhin learning model for classifications.

**Chapter 2**

**Literature Survey**

An earlier noticeable event was when ‘Alen Turing’ published an article ‘computing Machinery and Intelligence’. Since then different approaches of NLP were introduced like Heuristic , ML base , Neural network based etc.

* **Naïve Bayes (1960s) :** It was a probabilistic model that calculated the probability of data to belong to some category with the use of bayes theorem of probability.
* **Support Vector Machine (SVM 1990s) :** It’s main idea is to construct an hyperplane which separates documents into different categories by maximizing the difference between the categories.
* **TF-IDF (1960s) :** The basic idea is that word that occurs frequently in a document but rarely in the entire corpus is more important than others. Having two parts TF(term frequency) and IDF(Inverse document frequency) ,This is a very accurate way of Vectorization of any data.
* **N-grams(1980s) :** Another way of Vectorization it is an extention of Bag of words for N consecutive words. The idea remains the same that the frequency of a word in a document can show similarity between two words. By taking N consecutive words the order of words is also being used.
* **RNN( Reccurent Neural Network early 2000s) :** its core idea is to process sequential data by maintain hidden states that capture meaning and information from the previous inputs, allowing the model to relate the data overtime.
* **CNN( Convolution Neural Network early 2000s) :** shows **i**mproved accuracy and scalability due to use of convolution methods over input sequences to capture local meanings and patterns . Effectively learning hierarchical representation of the text classification task.
* **Transformers(2017)** : Introduced in paper ‘Attention is all you need’ The core idea is the self attention mechanism which allows the model to weigh the parts of the data differently during processes.

**Chapter 3**

**Methodology**

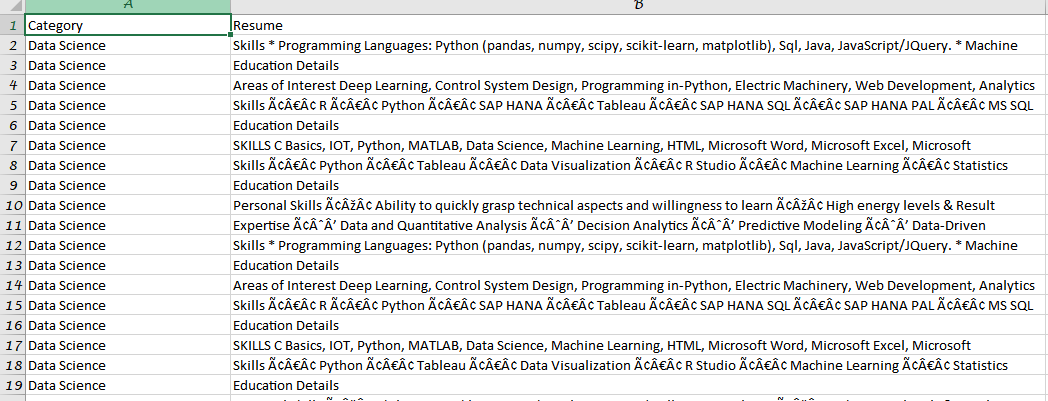
The methodology used follows standard NLP pipeline.

* **Data Acquisition / Preparation** : Acquiring a data is the most important part of the pipeline because a good dataset enhance the performance of our model and a bad dataset can degrade the performance of our model. There are two ways to do so

First is to Acquire data from some online source , this method is usually used since the dataset present online is already compressed is being used .

Second is to create the dataset yourself ,since the data present online may not meet your own needs for the classification , this method is hard to use since it take a lot or tries and time to make a proper dataset.

Fig. 3.1 .Dataset used



* **Text Preprocessing / Preparation** : This process is used to make the data more uniform and suitable for further upcoming steps used in NLP pipeline , an uniform dataset ensures higher accuracy with the predictions.

Some steps used are :

Data lowercasing

Removing punctuations, stopwords emojis html tags etc.

Fig. 3.2 Some preprocessing used



* **Tokenization** : This is the process of breaking down of data in small parts called tokens . it helps analyse and understand data and structure of a text.

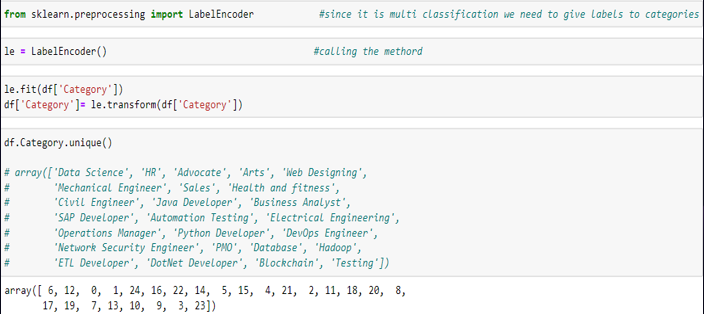
It helps the further upcoming steps to be ,more efficient with the data and predictions because by having smaller units it enable higher accuracy.

Stemming and Lemmetization are also sometimes used to make the tokens more uniform

But is only used for dataset with inflection which in our case is almost non existent.

* **Label Encoding :** Since computer cannot understand human language the categories/label used need to be encoded , hence some unique code or number is given to identify the label .

Fig. 3.3 .Label encoding

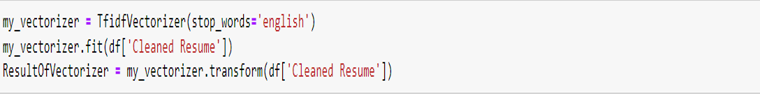


* **Feature Extraction :** Since machine can only understand numbers tokens need to be represented in form of a vector for the model to be able to process it without any problem.

There are many ways to do it like One hot encoding , Bags of words , N grams, TF-IDF vectorization etc.

Using TF-IDF is a better option then others because it is able to consider both the importance of a word for a particular document ass well as for the whole corpus.

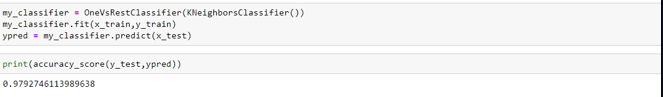
Fig. 3.4 Vectorization



* **Model Training** : A suitable model is chosen for the training and prediction an then the model training Is used using a 80-20 split of our dataset to determine prediction accuracy.

Many model like Naïve bayes, K nearest neighbors , SVM etc can be used in model training depending on the needs and accuracy.

* **Prediction** : Once the model is done training it is tested with the split made earlier and the accuracy is determined

Fig. 3.5 Prediction   


**Chapter 4**

**Result and Discussion**

The training process reveals that the model achieved an accuracy of approximately 97% on the training data . This high accuracy indicated the model is successfully able to recognize the category of a input resume given from the user .

Upon completion of the training the model was saved to a directory using serialization for future use. By doing this it is made sure that the model does not need to train again for future use again and is available to predict it without any delay.

During the prediction the user input is loaded and it goes through all the nlp pipeline steps to ensure accurate prediction and match with our already trained model.

Error handling is also introduced for a smooth experience.

**Chapter 5**

**Conclusion and Future Work**

**5.1 Conclusion :**

The project ‘Text Classification using Machine learning with python’ has shown promising results , the accuracy was high for prediction as well.

TF-IDF for vectorization and K-nearest neighbor for prediction gave high accuracy on our dataset , indicating that the models supervised learning was successful.

**5.2 Future work :**

For future work several factors can be considered **:**

**Dataset :** better dataset can be found with more categories present an distributed evenly this will ensure greater accuracy and effectiveness combine with our model.

**OOV’s :** Although accurate , machine learning algorithms are still susceptible to Out Of Vocabulary **,**there are several steps that can be used to reduce OOVs from a model , this wil ensure grater accuracy with the data that our model has never encountered before.

**NN :** Although accurate , machine learning algorithms are still not as accurate as Neural Network based NLP , so expanding this project to NN will gave greater results.

**References**

[1] GFG NLP article – Introduction and steps [Natural Language Processing - Overview - GeeksforGeeks](https://www.geeksforgeeks.org/natural-language-processing-overview/)

[2] Monkey-Learn Text Classification fundamentals [Text Classification: What it is And Why it Matters (monkeylearn.com)](https://monkeylearn.com/text-classification/).

[3] Resume article based on similarity - [(PDF) Resume recommendation based on text similarity (researchgate.net)](https://www.researchgate.net/publication/372201698_Resume_recommendation_based_on_text_similarity)

[4] NLP playlist on youtube . –

**https://youtube.com/playlist?list=PLKnIA16\_RmvZo7fp5kkIth6nRTeQQsjfX&feature=shared**