Cairo University

Faculty of Computers and Artificial Intelligence

Graduation Project

1st Semester 2020-2021 Project

Text Classification

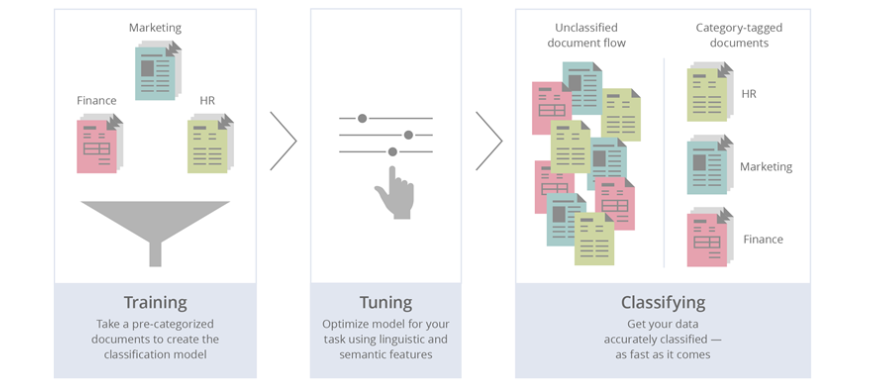
|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **Name** | **E-mail** | **Group** |
| 20170022 | Ahmed Sayed Mansour | ahmed.mans20719@gmail.com | CS-IS-1 |
| 20170021 | Ahmed Sayed Ibrahim | ahmed111522@gmail.com | CS-DS-1 |
| 20170136 | Atef Magdy Mitwally | atefmagdy12@gmail.com | CS-IS-1 |
| 20170053 | Ashraf Samir Ali | ashrafsamer423@gmail.com | CS-DS-1 |
| 20170002 | Ibrahim Ramadan Abdou | ibrahemramadan130@gmail.com | CS-IS-1 |

**Introduction**

**What is classification?**

Classification is a supervised machine learning technique, Classification is the process of categorizing a given set of data into classes, and it can be performed on both structured and unstructured data. The process starts with predicting the class of given data points. The classes are often referred to as target, label or categories.

**What is text classification?**

Text classification is the process of categorizing a text into organized groups. By using Natural Language Processing (NLP), text classifiers can automatically analyze text and then assign a set of pre-defined tags or categories based on its content.

**Variants of text classification**

* **Binary categorization:** only two categories
* Retrieval: {relevant-doc, non-relevant-doc}
* Span filtering: {spam, not-spam}
* Opinion: {positive, negative}
* **K-category categorization:** more than two categories
* Topic categorization: {sports, science, travel, business}
* Email routing: {folder1, folder2, folder3 …}
* **Hierarchical categorization:** categories from hierarchy.
* **Joint categorization:** multiple related categorization tasks done in joint matter.

**Why text classification is important?**

It’s estimated that around 80% of all information is unstructured, with text being one of the most common types of unstructured data. Because of the messy nature of text, analyzing, understanding, organizing, and sorting through text data is hard and time-consuming, so most companies fail to use it to its full potential.

This is where text classification with machine learning comes in. Using text classifiers, companies can automatically structure all manner of relevant text, from emails, legal documents, social media, Chabot’s, surveys, and more in a fast and cost-effective way. This allows companies to save time analyzing text data, automate business processes, and make data-driven business decisions.

**The uses of text classification**

* **Language detection**: the procedure of detecting the language of a given text (e.g., know if an incoming support ticket is written in English or Spanish for automatically routing tickets to the appropriate team).
* **Topic Labeling classification**: the task of identifying the theme or topic of a piece of text (e.g., know if a product review is about Ease of Use, Customer Support, or Pricing when analyzing customer feedback).
* **Sentiment Analysis**: the process of understanding if a given text is talking positively or negatively about a given subject (e.g., for brand monitoring purposes).
* **Intent classification**: is another great use case for text classification that analyzes text to understand the reason behind feedback.

**Objectives**

* Building single label text classification models.
* Building multilabel text classification models.
* Trying different techniques for classification models and differentiate between the different techniques based on (Accuracy, Performance, Memory handling).
* Building web application using react library as a front-end and Django as a back-end, to let people to use our trained models.
* Posting our research on the web app to be public for developers and who is interested in text classification techniques.

**Layout**

Chapter 2: Background (more details about text classification, algorithms and techniques)

Chapter 3: Related work (analysis about the current text classification projects)

Chapter 4: Proposed text classification (project workflow, architecture and algorithms)

Chapter 5: Experiments and Discussions (comparing different algorithms we used)

* Conclusion
* References

**Background**

In this chapter, we first review the methodologies for text classification. Concretely, we

Illustrate the **Preprocessing-data**, **NLP** of text classification, **algorithms** and **technologies**.

Preprocessing-data includes text representation. For text representation, we review how to extract features from text, how to clean text and in NLP we review most known techniques for Text-NLP**.** Lastly, we review some algorithms and technologies used for classification implementation.

**Preprocessing Data:** In natural language processing, text preprocessing is the practice of cleaning and preparing text data.

**Noise Removal:** noise removal is a text preprocessing task devoted to stripping text of formatting.

**Tokenization:** tokenization is the text preprocessing task of breaking up text into smaller components of text (known as tokens).

**Text Normalization:** normalization encompasses many text preprocessing tasks including stemming, lemmatization, upper or lowercasing, and stop-words removal.

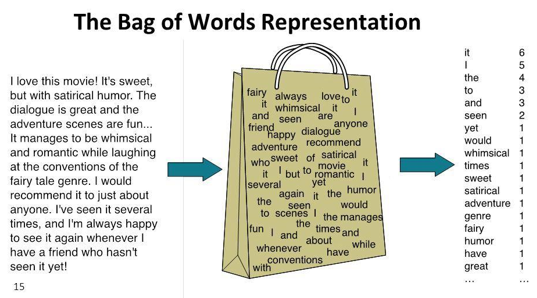
**Stemming:** stemming is the text preprocessing normalization task concerned with bluntly removing word affixes (prefixes and suffixes).

**Lemmatization:** lemmatization is the text preprocessing normalization task concerned with bringing words down to their root forms.

**Stop-word Removal:** stop-word removal is the process of removing words from a string that don’t provide any information about the tone of a statement.

**Nature Language Processing:**

Natural Language Processing (**NLP**) is a field of Artificial Intelligence (AI) that makes human language intelligible to machines. NLP combines the power of linguistics and computer science to study the rules and structure of language, and create intelligent systems (run on machine learning and NLP algorithms) capable of understanding, analyzing, and extracting meaning from text and speech. There are many approaches to NLP text classification, which fall into three types of systems:

* **Rule-based systems:** In the rule-based approach, texts are separated into an organized group using a set of handicraft linguistic rules. Those handicraft linguistic rules contain users to define a list of words that are characterized by groups. For example, words like Donald Trump and Boris Johnson would be categorized into politics. People like LeBron James and Ronaldo would be categorized into sports**.**
* **Machine learning-based systems:** Machine-based classifier learns to make a classification based on past observation from the data sets. User data is prelabeled as train and test data. It collects the classification strategy from the previous inputs and learns continuously. Machine-based classifier usage a **bag of a word** for feature extension. It’s preferred to use **tf–idf algorithm** to normalize the number of repeated words.
* **Hybrid systems:** Hybrid approach usage combines a rule-based and machine Based approach. Hybrid based approach usage of the rule-based system to create a tag and use machine learning to train the system and create a rule. Then the machine-based rule list is compared with the rule-based rule list. If something does not match on the tags, humans improve the list manually. It is the best method to implement text classification.

**Algorithms:** We separated our problem into two categories:

* **Single label text classification:** Is the task of classifying the elements of a set into two groups on the basis of a classification rule. Some of the most popular machine learning algorithms for creating text classification models include the **Naive Bayes** family of algorithms, **support vector machines (SVM),** and **deep learning**.
* **Multi label text classification:** is a generalization of multiclass classification, which is the single-label problem of categorizing instances into precisely one of more than two classes; in the multi-label problem there is no constraint on how many of the classes the instance can be assigned to. Some of the most popular techniques **OneVsRest** to use with any of the mentioned algorithms above**.**

**Related work:** for our research we will discuss some researches that talking about the idea and showing the differences.

**Resource 1:**

Shaikh, J., 2017. Machine Learning, NLP: Text Classification using scikit-learn, python and NLTK... [Online] Medium. Available at: <https://towardsdatascience.com/machine-learning-nlp-text-classification-using-scikit-learn-python-and-nltk-c52b92a7c73a> [Accessed 7 February 2021].

**Problem:** Binary classification on **The 20 Newsgroups data set.**

**Solution:** Using scikit-learn and NLTK to load the data and for the NLP process using sklearn.feature\_extraction library to apply Count Vectorization and TF-IDF Transformer to the data. Building the model using Naive Bayes algorithm with accuracy of 69% and using SVM with accuracy of 68%.

**Advantages:**

* + Using different algorithms from the sklearn library.
  + Explaining all steps and reporting information on each experiment.

**Disadvantages:**

* The data set is small and only covers a little number of classes.
* Not removing all unimportant data like headers and footers.
* Not using the new trends of text classification like Neural Networks algorithms.

**Resource 2:**

HUILGOL, P., 2021. Pretrained Models For Text Classification | Deep Learning Models. [Online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2020/03/6-pretrained-models-text-classification/> [Accessed 7 February 2020].

**Problem:** Discussing best pretrained models for text Classification

**Solution:** Using pretrained models like (XLNet, ERNIE, T5…) on different datasets.

**Advantages:**

* The datasets meet industry-accepted standards.
* The pretrained models have already been vetted on the quality aspect.
* Reporting a summary on each model using different datasets.

**Disadvantages:** Using the most new trends of ML and deep learning which needs a lot of studying to understand.

**The main differences between related works and our Project:**

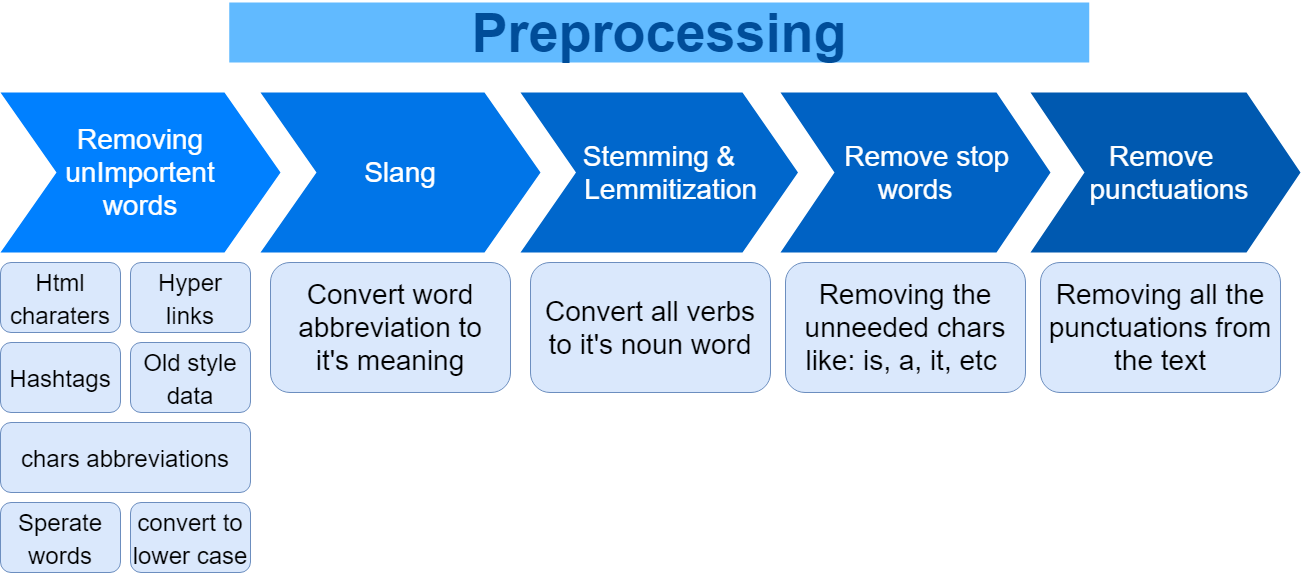
* All related work working on small datasets while we using the Arxiv data (3 GB).
* Our training on many number of labels to cover a wider area of subjects.
* Letting users to try our models on our website.
* Our training for Multilabel classification for better performance not only single label on the related work.

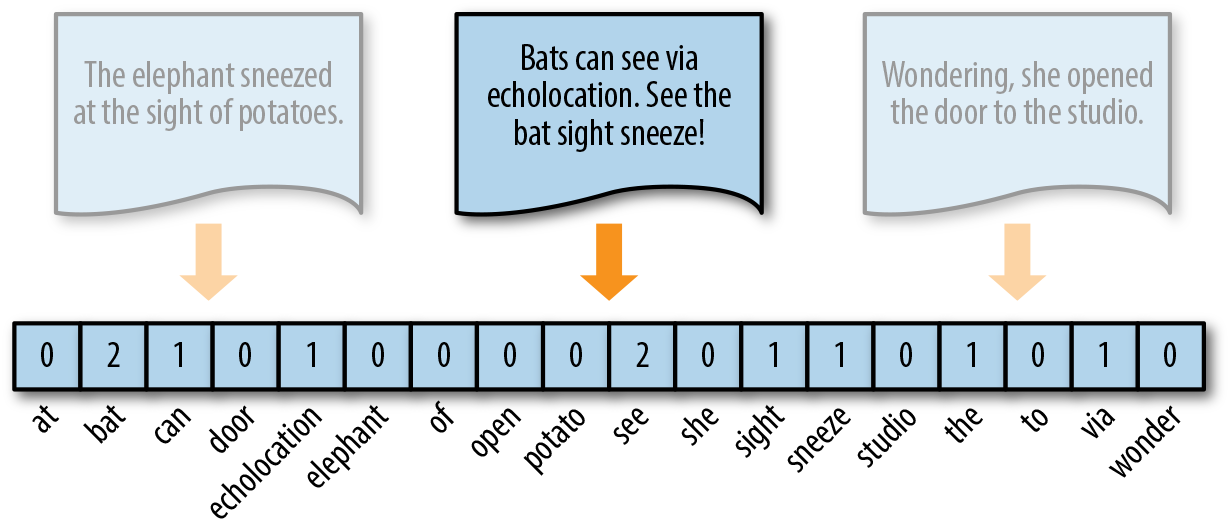
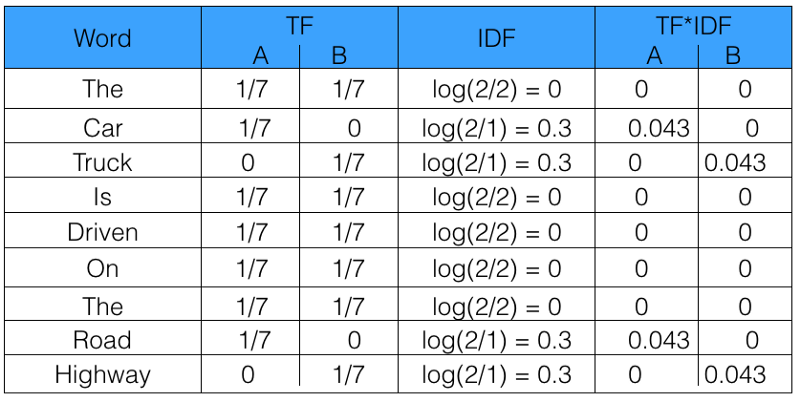
**Proposed text classification:**

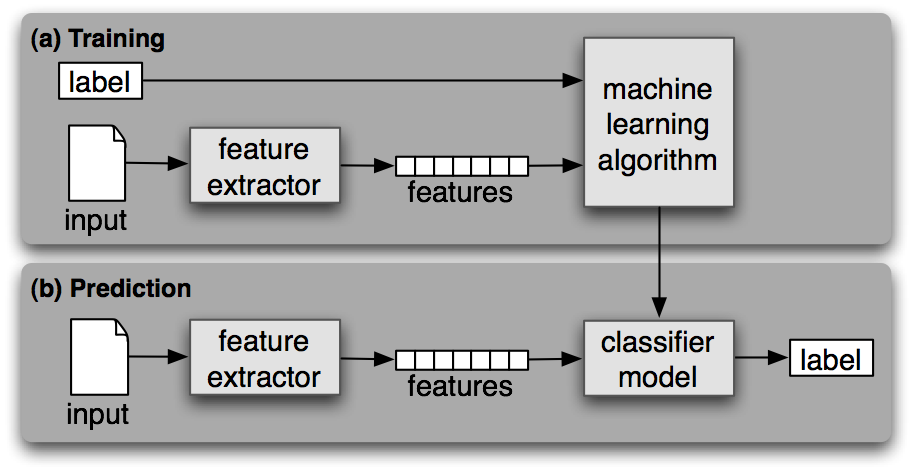
The steps for our model and how to prepare the data.

**Workflow:**

1. **Preprocessing data:**



1. **Natural Language Processing(NLP):**
   1. **Text Vectorization:**
   2. **TF-IDF Normalization:**
2. **Training and Prediction:**



**Architecture**

**Component**

**Algorithms**

**Experiments and Discussions**

**Conclusion**

**References**