FACULDADE ENGENHARIA
Departamento de Informática

Content Categories Detect

Proposta de Projeto

Orientador: Sebastião Pais(sebastiao@di.ubi.pt)
Coorientador: João Cordeiro(jpaulo@di.ubi.pt)

Context

Text can be an extremely rich source of information, but extracting insights from it can be complex and time-consuming due to its unstructured nature. However, thanks to advances in natural language processing and machine learning, which fall under the vast umbrella of artificial intelligence, sorting text data is getting easier. Text analysis, as a whole, is an emerging field of study. Fields such as Marketing, Product Management, Academia, and Governance are already leveraging the process of analyzing and extracting information from textual data.

Text classification is one of the fundamental tasks in natural language processing with broad applications such as sentiment analysis, topic labelling, spam detection, and intent detection. It is a machine learning technique that assigns a set of predefined categories to open-ended text. Text classifiers can be used to organize, structure, and categorize pretty much any text – from documents, medical studies and files, and all over the web. For example, new articles can be organized by topics; support tickets can be organized by urgency; chat conversations can be organized by language; brand mentions can be organized by sentiment; and so on.

Rule-based approaches classify text into organized groups by using a set of handcrafted linguistic rules. These rules instruct the system to use semantically relevant text elements to identify relevant categories based on its content. Each rule consists of an antecedent or pattern and a predicted category.

Say that you want to classify news articles into two groups: Sports and Politics. First, you will need to define two lists of words that characterize each group (e.g., words related to sports such as football, basketball, LeBron James, etc., and words related to politics, such as Donald Trump, Hillary Clinton, Putin, etc.).

Objectives

Text classification or Text Categorisation is the activity of labelling natural language texts with relevant categories from a predefined set. In laymen terms, text classification is a process of extracting generic tags from unstructured text. These generic tags come from a set of predefined categories. Classifying your content and products help users to search and navigate within a website or application easily.

Thus, this project proposes to develop a new approach to categorising text, with a base in BERT.

Workplan

T1 Review the State-of-the-art, write the survey about this problematic;

T2 Propose a new Unsupervised and Language Independent Approach to Content Categories Detect;

T3 Implementation;

T4 Testing and evaluation;

T5 The writing of the report.

Academic Prerequisites

Interest about Artificial Intelligence and Natural Language Processing; Web programming.

Assessment elements to deliver.

Source code and documentation of all code development; Project report.

Expected Results

* New Unsupervised and Language Independent Approach to Content Categories Detect;
* Open Web Platform;
* Report.

Contacts

Sebastião Pais (sebastiao@di.ubi.pt)