A Web-based fast and reliable text classification tool

Jānis Kapenieks
Riga Technical University, Distance Education Study Centre

INTRODUCTION

Opinion analysis in the big data analysis context has been a hot topic in science and the business world recently. Social media has become a key data source for opinions generating a large amount of data every day providing content for further analysis.

In the Big data age, unstructured data classification is one of the key tools for fast and reliable content analysis. I expect significant growth in the demand for content classification services in the nearest future.

There are many online text classification tools available providing limited functionality -such as automated text classification in predefined categories and sentiment analysis based on a pre-trained machine learning algorithm. The limited functionality does not provide tools such as data mining support and/or a machine learning algorithm training interface.

There are a limited number of tools available providing the whole sets of tools required for text classification, i.e. this includes all the steps starting from data mining till building a machine learning algorithm and applying it to a data stream from a social network source. My goal is to create a tool able to generate a classified text stream directly from social media with a user friendly set-up interface.

METHODS AND MATERIALS

The text classification tool will have a core based modular structure (each module providing certain functionality) so the system can be scaled in terms of technology and functionality.

The tool will be built on open source libraries and programming languages running on a Linux OS based server. The tool will be based on three key components: frontend, backend and data storage as described below:

- backend: Python and Nodejs programming language with machine learning and text filtering libraries: TensorFlow, and Keras,
- for data storage Mysql 5.7/8 will be used,
- frontend will be based on web technologies built using PHP and Javascript.

EXPECTED RESULTS

The expected result of my work is a web-based text classification tool for opinion analysis using data streams from social media. The tool will provide a user friendly interface for data collection, algorithm selection, machine learning algorithm setup and training.

Multiple text classification algorithms will be available as listed below:

- Linear SVM
- Random Forest
- Multinomial Naive Bayes
- Bernoulli Naive Bayes
- Ridge Regressio
- Perceptron
- Passive Aggressive Classifier
- Deep machine learning algorithm.

System users will be able to identify the most effective algorithm for their text classification task and compare them based on their accuracy.

The architecture of the text classification tool will be based on a frontend interface and backend services. The frontend interface will provide all the tools the system user will be interacting with the system. This includes setting up data collection streams from multiple social networks and allocating them to pre-specified channels based on keywords.

Data from each channel can be classified and assigned to a pre-defined cluster. The tool will provide a training interface for machine learning algorithms.

This text classification tool is currently in active development for a client with planned testing and implementation in April 2019.

KEY WORDS

Text analysis, machine learning, deep learning, text classification, social media analytics