Секция «Цифровые технологии как фактор трансформации общественно-политического устройства в современных государствах»

How to perform automatic labelling of texts for text classification tasks. Heuristics and experiences.

Научный руководитель – Казаринова Дарья Борисовна

Де Лука Габриэле

Аспирант

Российский университет дружбы народов, Факультет гуманитарных и социальных наук, Москва, Россия

E-mail: gabriele.deluca@mail.ru

This thesis will discuss the heuristics and experiences acquired in solving the problem of automatic labelling of texts for the purpose of performing text classification, by using a set of keywords, uniquely associated with the issue being studied by the analyst. [1] The method proposed is preliminary to the development of a machine learning model for the classification of political texts.

The labelling of the observations in a dataset is a fundamental process in the preparation of the dataset for the performing of supervised learning tasks. In the context of text classification, the labelling of texts consists in the attribution of each text to any of a finite number of classes, which then allows the training of a machine learning algorithm on the features of the dataset, the words of the texts, in order to extract the rules accordingly to which each text is assigned to its respective class.

The manual labelling of texts is a significantly expensive task both in terms of time required, number of humans to involve, and financial resources spent. In fact, the process necessitates many hours of human time performed by specifically trained experts, competent in the subject on which the classification task should be performed. As a consequence, the labelling of texts by humans often ends up accounting for a significant portion of the total costs associated with the development of a machine learning model, and to the conduct of data mining on a text corpus.

This paper will discuss an heuristic method that, while simple, can be applied to solve the problem of the automatic labelling of text data when insufficient resources, either financial or human, are available to the data scientist. While better methods for labelling texts exists, and while they end up providing higher accuracy in the subsequent training of the model, this method is cheap and effective, and can be implemented with minimal *a priori* knowledge on the issue being studied. Before the labelling can be performed, it is necessary to preprocess the text corpus accordingly to the usual procedures for natural language processing tasks. The texts will have to be stemmed and tokenized and, finally, the stopwords will need to be removed. While text stemming it is not a strict requirement for a NLP pipeline, in this case it must be performed, lest the automatic labelling we propose would not work effectively.

After the text is stemmed and tokenised, a list of keywords associated with the issue being studied has to be developed. In absence of *a priori* knowledge on the issue being studied, the preliminary list of keywords can be determined by taking the name or description of the issue itself, and then looking up in a dictionary of synonyms, relevant for the language in which the text corpus is written, all synonyms associated with either the issue itself or its description. A user, competent in the language used in the text, will need to rank the keywords in order of importance accordingly to how strict the association of said keywords is with the issue being studied. This process requires minimal training for the user, and can be performed in a matter

of minutes. This will allow creating an ordered list of words, somehow associated with the issue being studied.

The list of keywords previously extracted will then need to be subject to the same preprocessing tasks that were applied to each document in the text corpus, as described above. The result is a list of stems of keywords associated with the issue being studied.

A simplified version of the automatic labelling can then be performed on the basis of whether or not the keywords identified are present in each text. The documents in the corpus can be checked to determine whether the stems of their associated tokens contain any of the stems of keywords being identified. If they do, the texts will be marked with the positive label for a binary classification task. If they don't, they will be marked with the negative label for the same classification task. In this manner, all texts in the collection are labelled with either a positive or a negative label, and none is left unmarked.

It should be noted that, lest the output of the learning process be banal, in the training phase of the process the ML model should not be allowed to see the keywords used to select the texts. If that were the case, the model is extremely likely to predict the rules used to label the texts (that is, the keywords), and this would be of poor significance to the analyst: it is therefore imperative to blind the algorithm to those keywords. This can easily be done by adding a step in the tokenization function for the removal of all tokens containing the stem of one of the keywords previously identified.

This method is however likely to produce more positives than it is necessary. Manual inspection of the texts labelled as positive, should this be feasible, is therefore still advised. This inspection will however require a significantly shorter time than the one required to fully inspect and label the dataset. In fact, the objective of this inspection is only to determine whether the texts labelled as positive have, indeed, "something" to do with the issue being studied, and to this regard it is sufficient to inspect the immediate neighborhood of the tokens identified for labelling. If the texts positively selected do not relate to the issue, then the labelling process produced false positives, and this must be taken into account. If that is the case, dropping the keywords associated with the texts falsely marked as positive is therefore advised.

Источники и литература

1) Bing Liu, Xiaoli Li, Wee Sun Lee, and Philip S. Yu, Text Classification by Labeling Words, 2004/7/25, AAAI, vol. 4, pages 425-430. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.66.572&rep=rep1&type=pdf