Passage-based document representation for text classification

Taxiarchis Papakostas-Ioannidis, UvA-id: 11407387

June 18, 2018

Abstract

The scope of this thesis is to study whether a representation of a document in terms of passages, outperforms in text classification the traditional full text document representation. For that purpose, different text modellings, will be applied on a passage level and be tested, compared and evaluated on different classifiers. The work is going to be carried out in collaboration with the Elsevier company, as an internship during the next three months, starting on the 3^{rd} of April.

1 Personal details

My email `mailto:taxiarchis.papakostas@gmail.com`

My Internals supervisor email `mailto:e.kanoulas@uva.nl`

My Externals supervisor email `mailto:g.tsatsaronis@elsevier.com`

The wiki on my github account `https://github.com/aris-papakos/Thesis.git`

2 Research question

Most of the information retrieval and natural language processing algorithms in literature treat documents as a whole text representation. Only a few attempts have been made to represent a document by means of its passages. In practise, a text classifier is trained on a classified text document treating it as a uniform document. Based on the hypothesis that a paragraph-based representation of text is more effective than the whole text representation in text classification, the main research question of this work can be formulated as follows:

- Can a passage-based representation of text be more effective in text classification than to a whole-text representation?

The research question can be further divided into the following questions, which need to be answered first, in order to go further:

- How to represent a text, based on passages and which numerical representation is more effective?

- In which way can we predict documents class based on paragraph classes?
- Which classification method performs better in a paragraph-based representation of text?
- How a classifier based on paragraphs can be trained?

3 Related Literature

3.1 Text Representation and Document Representation

Due to machine inability of interpreting plain text, words and sentences are converted to probabilities or vectors in the space. In the literature, several techniques which face the problem of representing text for processing have been presented. In this work we will focus more in three robust models: a) the Bag of words (BoW), which is going to be used as the simplest text vector representation model, b) the Latent Dirichlet Allocation (LDA) [1], and c) the paragraph vectors [2], which will be mainly applied. All these models, and especially our focus models (b) and (c), have been used in various research works [3] [4] for modelling sentences and passages for text classification and cover both categories of representations, namely conversion of words and sentences into probabilities and vectors in the space. Furthermore, Jinsuk Kim and Myoung Ho Kim proposed a method [5] for structuring a document based on passages, which is going to be used as a baseline in our research.

3.2 Text Classification

During the last decades, several researches on text classification have been presented in literature. Some widely used models are the Naive Bayes Classification (NBC) [6], the k-nearest neighbors algorithm (KNN) [7], the support vector machines (SVMs)[8] [9] and a number of variations, which generally perform very well in most of the cases and will be used as baseline models in this work. Furthermore, state of the art deep learning classifiers, like the Convolution Neural Network (CNN) [10] text classifier, or a combination of methods, like the Stacking Support Vector machines providing an ensemble of SVMs [11], could be explored and be a potential choice for a deep neural model. The text classifier, which will be built, will be treated as a black box for testing our main hypothesis. Having this classifier as a basis, more than one will be built, trained and tested subsequently.

4 Methodology

4.1 Data

Elsevier is an information and analytics company and one of the world's major providers of scientific, technical, and medical information. As already mentioned in the abstract, Elsevier is the company where the thesis internship will be carried out in the three months period, beginning from the 3rd of April. The company will make available to us almost 14 million benchmarks on pre-classified Science Direct articles and almost 70 million benchmarks from the abstract and citation database Scopus. It should be noted that all available documents are in .xml format and the company has its own parsers for processing and extract the text content. The data will be used to assign classes to documents and to their passages, but also as training and testing data for the text classification task.

4.2 Methods

The methodology that is going to be followed can be modeled in specific steps. At first, a document will be subdivided in passages, based on the internal markup, and in each passage a class will be assigned. Each passage of the document may have a different class or all the passages would be of the same class. All the passages will be modeled after choosing a text representation among topic models, bug of words (BOW), language models and paragraph vectors. After this step, a classifier is needed to be trained on the classified passages. Meanwhile, the case can be extended on training different types of classifiers including a state of the art classifier, as a deep learning convolution neural network, or even further by using an ensemble approach. For the testing phase, documents will be represented again in a passage level and the classifier will predict their classes. Finally, a class will be assigned to the original document , by performing a majority vote on the predicted passage classes. In order to compare and evaluate the results, the whole process needs to be performed again for the full text representation of the same documents.

4.3 Evaluation

The evaluation of the results will be performed by comparing the measures of precision, recall, F1 score as well as the accuracy. The problem will be treated as a binary classification problem and the classifiers will be tested for their ability to assign the correct label/class to the documents, based on the document representation. By taking all these into account, it may safely be said that the above mentioned measures are the most suitable choice for the evaluation process.

5 Risk assessment

It is thought that the above methodology constitutes a complete and realistic design for a master's thesis. However, several risks could be addressed. In case that the company will not provide all the necessary data or if the documents are semi- or unclassified, one potential choice could be an open pre-classified data-set which can be downloaded from the web, for conducting our experiments. Another case scenario is that the company will not give access to all the necessary computational resources. In this case, a personal investment could be done on a cloud infrastructure like AWS. The worst-case scenario is that the initial hypothesis of our research is not working. In this case, we will focus on more to the data analysis procedure.

6 Project plan

The weeks of the project plan has been numbered based on the Calendar of 2018 containing Week numbers.

KW	**Planning**
Week 14:	Complete what is missing in related literature - literature review.
Week 15:	Data exploring, modelling of documents in terms of passages.
Week 16:	Modeling documents in terms of passages – Division and representation of documents in a section level.
Week 17:	Modelling documents in terms of passages - Building code for assigning a class to sections.
Week 18:	Trying different text representations for publications passages.
Week 19:	Trying different text representations for publications passages - Midterm progress report.
Week 20:	Trying different text representations for publications passages - Midterm progress report.
Week 21:	Training classifiers.
Week 22:	Building a classification pipeline for testing - Runnning experiments.
Week 23:	Running experiments and evaluating results.
Week 24:	Making modifications - train, running experiments, evaluating and comparing results.
Week 25:	Summarizing – Time for possible collisions - Writing final thesis report.
Week 26:	Final thesis report – Presentation.

References

[1] Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), pp.993-1022.

[2] Le, Q. and Mikolov, T., 2014. Distributed representations of sen- tences and documents. Proceedings of the 31st International Conference on Machine Learning, Beijing, China, pp. 1188-1196

[3] Li, L. and Zhang, Y., An empirical study of text classification using Latent Dirichlet Allocation.

[4] Hashimoto, K., Kontonatsios, G., Miwa, M. and Ananiadou, S., 2016. Topic detection using paragraph vectors to support active learning in systematic reviews. Journal of biomedical informatics, 62, pp.59-65.

[5] Kim, J. and Kim, M.H., 2004. An evaluation of passage-based text categorization. Journal of Intelligent Information Systems, 23(1), pp.47-65.

[6] Shimodaira, H., 2014. Text classification using naive bayes. Learning and Data Note, 7, pp.1-9.

[7] Yong, Z., Youwen, L. and Shixiong, X., 2009. An improved KNN text classification algorithm based on clustering. Journal of computers, 4(3), pp.230-237.

[8] Chen, K., Zhang, Z., Long, J. and Zhang, H., 2016. Turning from TF-IDF to TF-IGM for term weighting in text classification. Expert Systems with Applications, 66, pp.245-260.

[9] Zhang, W., Yoshida, T. and Tang, X., 2008. Text classification based on multi-word with support vector machine. Knowledge-Based Systems, 21(8), pp.879-886.

[10] Kim, Y., 2014. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

[11] Sebastiani, F., 2002. Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), pp.1-47.