Text classification by Convolution Networks for Data-Driven Decision Making

RODRIGUES, Marcos <http://orcid.org/0000-0002-6083-1303>

Text classification by Convolution Networks for Data-Driven Decision Making

Marcos A Rodrigues
GMPR—Geometric Modelling and Pattern Recognition Research Group
Sheffield Hallam University, Sheffield, UK
Email m.rodrigues@shu.ac.uk

Abstract

Recent advances in automation and data-driven intelligence from sophisticated Artificial Intelligence (AI) technologies have impacted on all areas of knowledge and economic activity. AI Deep Learning is a method of learning and extracting knowledge from large amounts of data. AI algorithms iteratively learn from data, finding hidden features and providing insights without explicitly programmed features. Text classification can be cast as a generic problem whose solution can have significant impacts on data-driven decision processes and ERP-Enterprise Resource Planning information systems. Normally, classification is carried out from a given taxonomy. The causes for wrong classification may arise from inconsistent taxonomies, incomplete descriptions, wrong interpretation of category, inconsistent language translation, human error, algorithm design and so on. In this paper, we address the issue of automatic product classification from unconstrained textual descriptions using machine learning techniques. Rather than defining words in a vocabulary (as normally is the case for instance, with Google's `word2vec` technique) this research focuses on character-based classification through a temporal convolution network as in Crepe (Character-level Convolutional Networks for Text Classification). The advantage is that instead of defining a vocabulary with tens of thousands of words, the vocabulary is made up of a small character set composed of the letters a-z, numbers 0-9, and special characters. Furthermore, because in any language words are defined by a sequence of characters, the relationships between the characters within a word or words are learned from the temporal convolution. This negates the need to learn words per se. The research used product descriptions from 6 categories: bakery, chilled, dairy, drinks, fruit and vegetables, meat and fish. A total of 8612 samples were used which were separated into a training set (7751 samples corresponding to 90% of the data) and unseen test set (861 samples or 10% of the data). Examples of descriptions for each category are as follows.

bakery	allinson wholemeal batch bread medium sliced 800g
chilled	alf turner sausage roll
dairy	actimel yogurt drink 0.1% fat strawberry 8x100g
drinks	35 south latitude pinot noir 75cl
fruit &veg	birds eye field fresh select mixed vegetable 690g
meat &fish	bernard matthews breaded ham and cheese chicken escalope 285g

The extremely short descriptions are significant challenges to classification. The designed network has 15 convolution layers followed by 2 fully connected layers. The network was implemented using the Torch Framework on a Mac Pro running macOS Sierra 3.5GHz 6-core Intel Xeon E5 processor with 16GB of memory. After 36 hours of training, results for unseen test data are depicted in the confusion matrix below.

Validation:							
	bakery	chilled	dairy	drinks	Fruit &veg	Meat &fish	Total
bakery	**104**	7	2	0	0	1	114
chilled	4	**173**	7	0	5	9	198
dairy	2	3	**171**	0	0	1	177
drinks	0	2	0	**107**	0	0	109
Fruit &veg	5	2	1	0	**105**	4	117
Meat &fish	1	13	4	0	5	**123**	146
Total	116	200	185	107	115	138	861
TP	91%	87%	97%	98%	90%	84%	91%

The columns of the matrix show the predicted values, while the rows are actual values. The percentage accuracies quoted are for TP-true positives only. The largest misclassification is for meat and fish, where 13 samples were classified as chilled. The overall accuracy of 91% is impressive given that the classification features were extracted from character sequences only and that descriptions are extremely short. It is shown that

character-based classification is a valid solution for short descriptions and we are now investigating alternative optimal network designs and the possibility of using a larger training set.