ALC

Luca Ambrosini and Giancarlo Nicolò

Univesitat Politècnica De València

Abstract. This paper describes a methodology to create from scratch a set of baselines for tackling text classification problems using natural language processing and machine learning techniques. Our approach focuses on neural network models taken from state of the art that exploit different corpus preprocessing, tweets representations and features extraction methods. Finally, we will discuss how we applied this methodology to the *Classification Of Spanish Election Tweets (COSET)* task at IberEval 2017 and present the results we obtained

1 Introduction

Intro over nlp and text classification: motivation why this field is important and the actual challenge and open problems (will be top just to list some problem and their way to be solved, for example citing the different workshop that we had during the course)
Text classification is an important task in Natural Language Processing with many applications, such as web search, information retrieval, ranking and document classification (Deerwester et al., 1990; Pang and Lee, 2008). Recently, models based on neural networks have become increasingly popular (Kim, 2014; Zhang and LeCun, 2015; Conneau et al., 2016). While these models achieve very good performance in practice, they tend to be relatively slow both at train and test time, limiting their use on very large datasets.
The
vedi citazioni su word embedding di mark e bag of word da maite
Introduce informally our task and list some way of tackling this problem from the different perspectives, maybe say that we de-construct tthe actual approach and cathegorized their inner process in: representation, preprocessing, model and post processing (that we haven't done).
Finally explain how we structured this report

2 Task definition

Maite PRSOCO example
The main objective proposed by the organizers of the PRSOCO shared task was to predict the personality traits of developers given a collection of their source code. The personality of a developer was determined following the Five Factor Theory or Big Five [5, 11, 3] which is the most widely accepted in psychology. Therefore, five traits define the personality of an author. Those traits are: agreeableness (A), conscientiousness (C), extroversion (E), openness to experience (O), and emotional stability / neuroticism (N). Each trait was labeled within a range between 20 and 80. The models were evaluated by the organizers using two metrics: the average Root Mean Squared Error (RMSE) as well as the Pearson Product-Moment Correlation (PC). For further information about the task, please review the overview paper of the task [16].
COSET webpage example
Political conversation in Twitter increases when a General Election comes close. Analyzing the topics discussed by users provides interesting insights of this growing public conversation on politics.

In COSET, the aim is to classify a corpus of political tweets in 5 categories of classification: political issues, related to the most abstract electoral confrontation; policy issues, about sectorial policies; personal issues, on the life and activities of the candidates; campaign issues, related with the evolution of the campaign; and other issues.
The tweets are written in Spanish and they talk about the 2015 Spanish General Election. In the training phase participants will be provided with Twitter Ids and their manually issue codification.
TODO
problem over the tweet (chiedi alla tipa di torino i tweet challenges)
First draft
The aim proposed by organizers of COSET shared task was to classify tweets from Spanish election in five different categories: (i) political issues, related to the most abstract electoral confrontation; (ii) policy issues, about sectorial policies; (iii) personal issues, on the life and activities of the candidates; (iv) campaign issues, related with the evolution of the campaign; (v) and other issues.
the personality traits of developers given a collection of their source code. The personality of a developer was determined following the Five Factor Theory or Big Five [5, 11, 3] which is the most widely accepted in psychology. Therefore, five traits define the personality of an author. Those traits are: agreeableness (A), conscientiousness (C), extroversion (E), openness to experience (O), and emotional stability / neuroticism (N). Each trait was labeled within a range between 20 and 80. The models were evaluated by the organizers using two metrics: the average Root Mean Squared Error (RMSE) as well as the Pearson Product-Moment Correlation (PC). For further information about the task, please review the overview paper of the task [16].

2.1 Copora statistics

Average train sequence length: 135 chars (in chars)
Average train sequence length: 24 (in words)
Max train sequence length: 49 (words)
si ottimizzato il parametro del numero di parole da usare come input alla rete neurale, arrivando alla misura di 30
numeri superiore risultano in un eccesso di padding e perdita di informazione della rete. Numeri inferiori perdono troppa informazione

3 Methods

Bag of words
Bag of n-gram
Random forest Decision tree Support vecto machines Multilayer perceptron.

4 Representation

vedi citazioni su word embedding di mark e bag of word da maite

4.1 Word embedding

ask paolo lexicon/terminology online vs learnign?

4.2 N-gram embedding

Representation was learn only online

5 Preprocessing

ref used tool opackage (put on a .bib)

5.1 Stemming

5.2 Stop words

5.3 Cleaning

Url.

R-W.

5.4 Tokenizing

tokenizer options

Mentions.

Smiles.

Emoji.

Numbers.

6 Models

6.1 Neural models

Start whith the meaning reason behind this model (e.g. LSTM because in translation is a state of the art).
Long short term memory. Bidirectional long short term memory.
Convolutional neural network. Hybrid convolutional neural network.

Fast text. same and equal of original but changing the number of layer (worstening in increasing the number of layer) + adding of the gaussian noise and batch normalization

KIM. Pippo **?**
pippo style (**?**)
from kim model we optimize ttill the following net configuration/topology: description of the net

7 Evaluation

7.1 Metrics

$$F_{1-macro} = \frac{1}{|L|} \sum_{l \in L} F_1(y_l, \hat{y}_l) \tag{1}$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{2}$$

$$precision = \frac{1}{|L|} \sum_{l \in L} Pr(y_l, \hat{y}_l) \tag{3}$$

$$recall = \frac{1}{|L|} \sum_{l \in L} R(y_l, \hat{y}_l) \tag{4}$$

7.2 Results

table of only the two main models and 6 kind of preprocessing
table of word representation

8 Conclusions

Transfer learning lstm
stemming online learning
expert modelling kim model
leave trainable the vector embedding is always positive even though you already have a trained vector embedding representation.
transfer learning

Bibliography