IRISA at SMM4H 2018: Neural Network and Bagging for Tweet Classification
Anne-Lyse Minard, Christian Raymond, Vincent Claveau

To cite this version:
Anne-Lyse Minard, Christian Raymond, Vincent Claveau. IRISA at SMM4H 2018: Neural Network and Bagging for Tweet Classification. SMM4H 2018 - Social Media Mining for Health Applications, Workshop of EMNLP, Oct 2018, Brussels, Belgium. pp.1-2. hal-01937019

HAL Id: hal-01937019
https://hal.archives-ouvertes.fr/hal-01937019
Submitted on 27 Nov 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Abstract

This paper describes the systems developed by IRISA to participate to the four tasks of the SMM4H 2018 challenge. For these tweet classification tasks, we adopt a common approach based on recurrent neural networks (BiLSTM). Our main contributions are the use of certain features, the use of Bagging in order to deal with unbalanced datasets, and on the automatic selection of difficult examples. These techniques allow us to reach 91.4, 46.5, 47.8, 85.0 as F1-scores for Tasks 1 to 4.

1 Introduction

IRISA has participated in the four tasks of the SMM4H challenge (Weissenbacher et al., 2018). Yet, we have focused on Task 2 and 3, which are the most challenging ones, in particular because they have unbalanced data. Moreover, for Task 2, the three classes have very fuzzy boundaries, which makes some tweets difficult to classify even for humans. Our main contribution is to rely on Bagging (Bootstrap Aggregating) in order to deal with this problem of unbalanced data.

2 Methods

2.1 RNN: BiLSTM

For the four tasks, we have developed classifiers based on recurrent neural networks which consists in one Bidirectional LSTM layer (Graves et al., 2013) and a dense layer with a softmax activation as hidden layer. The input layer takes a representation of a tweet which consists in the word embeddings of each token and, depending of the task, a one-hot vector for each token or a one-hot vector for some medical terms in the tweet. Metamap Lite (Demner-Fushman et al., 2017) is used to extract specific medical terms from the tweets. We restrict the number of semantic types according to the task: for Task 1, we have selected only terms related to drugs or substances; for Task 2, only to procedural terms; and for Task 3, we have selected both terms related to drugs and terms related to symptoms. For Task 1 and Task 2, we observe an improvement while using medical terms, whereas for Task 4 the use of metamap has no influence on the results. We use the word embeddings distributed by Grave et al. (2018). They have been trained with FastText (Bojanowski et al., 2017).

2.2 Bonzaiboost

During the development phase, we have used BONZAIBOOST, an implementation of the boosting algorithm adaboost.MH (Laurent et al., 2014) on decision trees. The results obtained are a bit lower than those of recurrent neural network methods. Yet, the experiments done with BONZAI-BOOST allowed us to extract the most discriminating words, to choose the better features for the RNN, and to select the difficult examples (see Section 2.4). For Task 1, the important words found are drug names, such as xanax. For Task 2, the useful words are verbs indicating the action of taking a drug, the results of its intake, or the fact that a drug is needed (e.g. took, need). For Task 3, the discriminating words include symptom names (e.g. dizzy, headache). Finally for Task 4, no relevant discriminating words have been found. These findings help us to determine the semantic types of the medical terms to be used in the feature set.

2.3 Bagging

Bagging (Breiman, 1996) is a technique that consists in combining the prediction of different learners, where each "learner" uses only a sample of the original training set. We learn several models, with, for each, a subset of the training dataset, different training parameters (number of epochs, number of hidden layers...) and different feature sets. To deal with unbalanced datasets in Tasks
Table 1: Description of the submitted runs and results obtained on the training dataset.

<table>
<thead>
<tr>
<th>Task</th>
<th>Run 1</th>
<th>Run 2</th>
<th>Run 3</th>
<th>Run 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>91.1</td>
<td>91.4</td>
<td>90.6</td>
<td></td>
</tr>
<tr>
<td>T2</td>
<td>43.6</td>
<td>45.5</td>
<td>46.5</td>
<td></td>
</tr>
<tr>
<td>T3</td>
<td>43.9</td>
<td>46.2</td>
<td>47.8</td>
<td></td>
</tr>
<tr>
<td>T4</td>
<td>84.4</td>
<td>85.0</td>
<td>82.4</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Final results in terms of F1-score.

<table>
<thead>
<tr>
<th>Task</th>
<th>Run 1</th>
<th>Run 2</th>
<th>Run 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>91.1</td>
<td>91.4</td>
<td>90.6</td>
</tr>
<tr>
<td>T2</td>
<td>43.6</td>
<td>45.5</td>
<td>46.5</td>
</tr>
<tr>
<td>T3</td>
<td>43.9</td>
<td>46.2</td>
<td>47.8</td>
</tr>
<tr>
<td>T4</td>
<td>84.4</td>
<td>85.0</td>
<td>82.4</td>
</tr>
</tbody>
</table>

References


