A Hierarchical Neural Network Approach for Bots and Gender Profiling

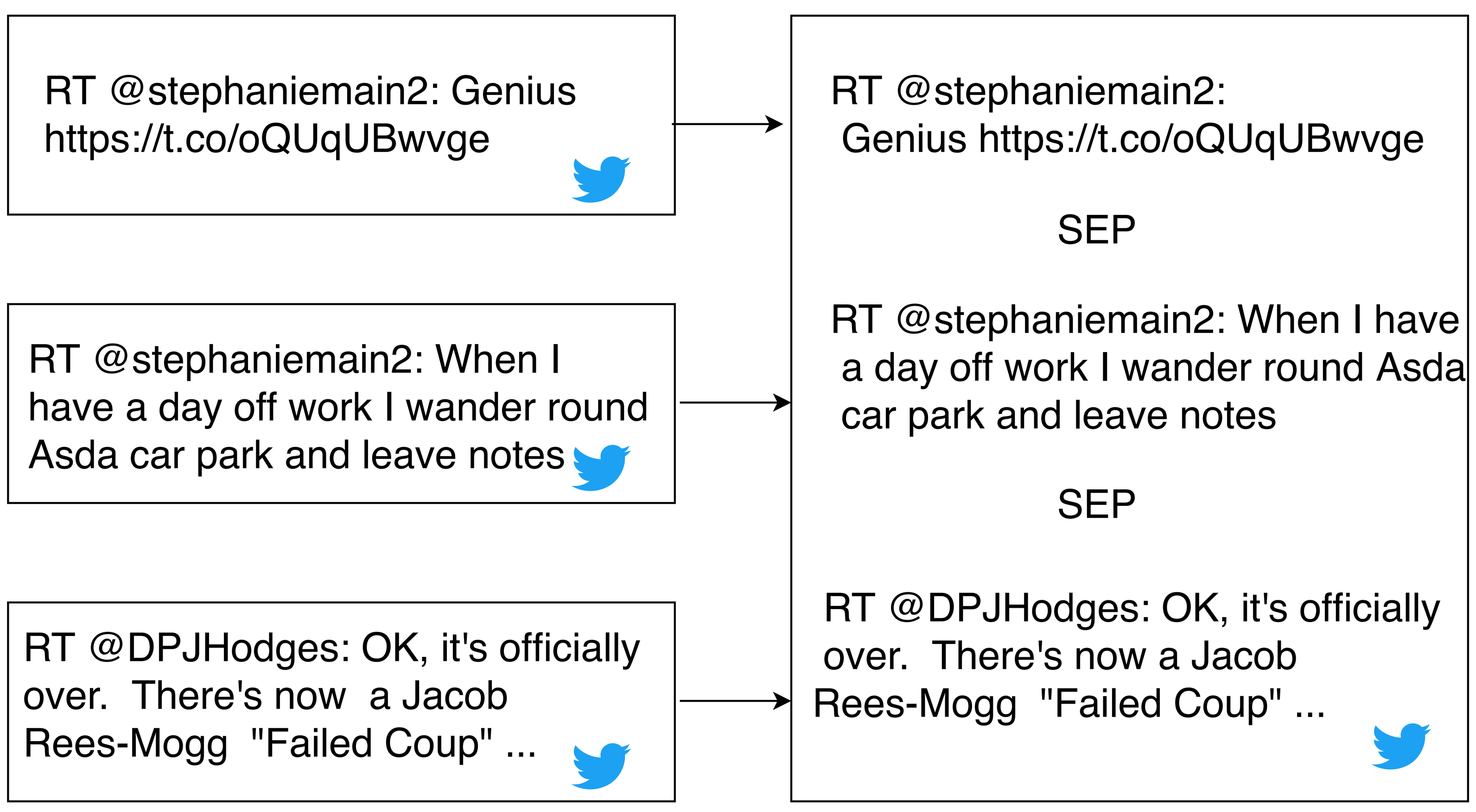
Andrea Cimino and Felice Dell'Orletta

Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR) Pisa
Italia**NLP** Lab – *www.italianlp.it*
{andrea.cimino,felice.dellorletta}@ilc.cnr.it

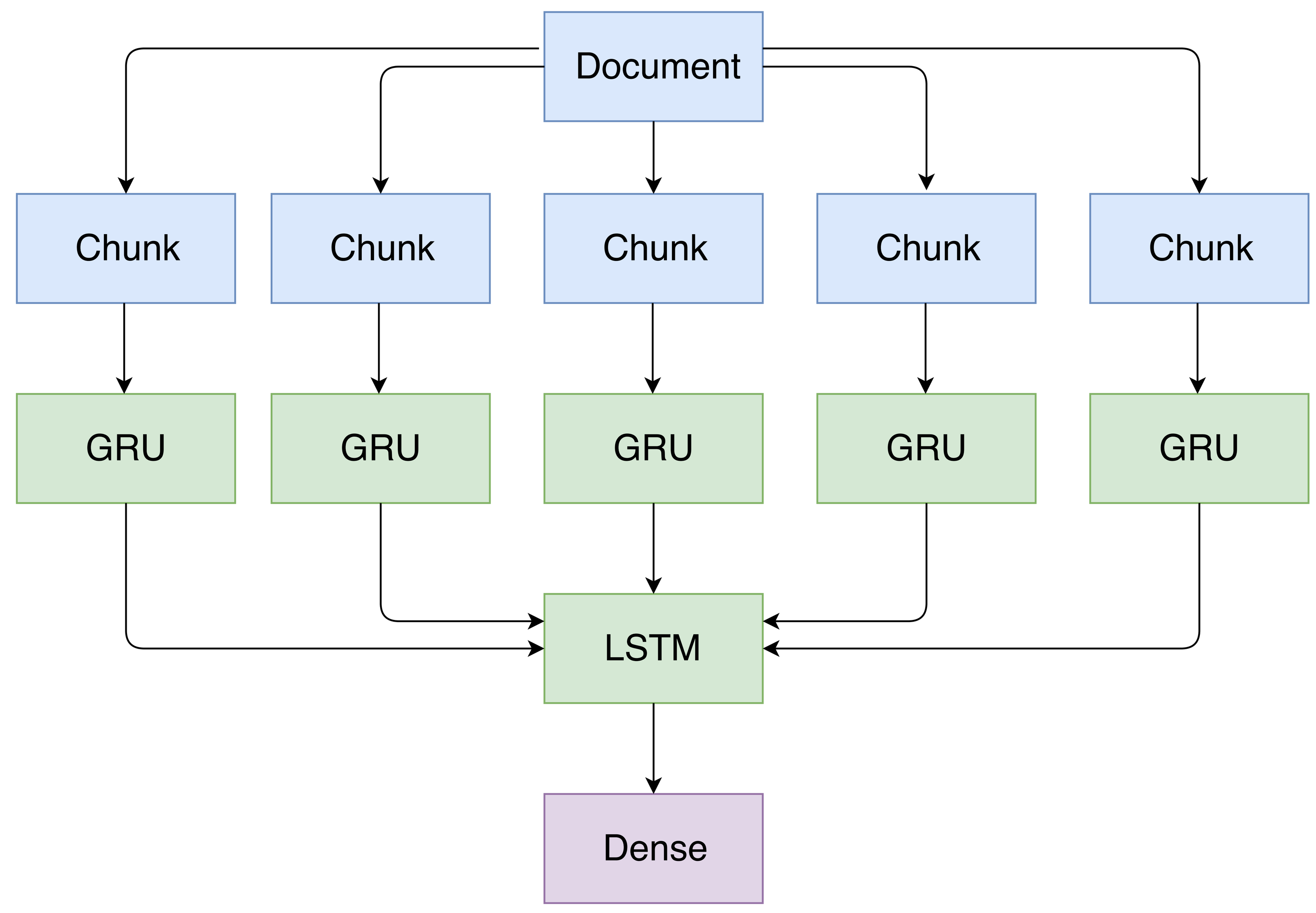
Our approach

We describe our participation in the Bots and Gender Profiling shared task of PAN@CLEF2019 for the English language. We tested three approaches based on three different document classification algorithms. The first approach is based on a SVM classifier with handcrafted features using a wide set of linguistic information. The second and the third approaches exploit recent advances in Natural Language Processing using a Hierarchical GRU-LSTM Neural Network using word embeddings trained on Twitter and finally an adaptation of the BERT system.
After an in-house evaluation, we submitted the final run with the Hierarchical Neural Network model, which achieved a final accuracy of 0.9083 in the Bots Profiling task and a score of 0.7898 in the Gender Profiling task.

Training Document Construction



GRU/LSTM Hierarchical Architecture



Models and Features

The SVM Model

The SVM classifier exploits a wide set of features ranging across different levels of linguistic description. The features are organised into three main categories: *raw and lexical text features*, *morpho-syntactic features* and *lexicon features*. All the calculated features are the input of the linear SVM algorithm implemented in the liblinear library which finally generates the final statistical model used then to classify unseen documents. The SVM classifier exploits a wide set of features ranging across different levels of linguistic description.

The BERT Model

Following the latest advances in NLP, we wanted to test how well pretrained language model representations behave on the Bot and Gender stylistic profiling shared task. We used BERT since Google provides pretrained models, which need only to be fine-tuned with an inexpensive procedure. More precisely, we choosed the recommended model: BERT-Base Multilingual Cased which is trained on 104 languages with 110M parameters.

One of the limitations of this pretrained model is that such model was trained on sentences not longer than 512 tokens, which made the standard fine-tuning procedure not suitable for our case, since the training documents (the concatenation of the tweets) were much longer than 512 tokens. For this reason, we generated 5 different fined tuned downstream tasks models by considering 5 chunks of 500 tokens each.

In testing phase, each document was still divided in 5 chunks. Each chunk was then classified by the previously 5 fine tuned models. We then choose as winning class among BOT, male and female, the majority class resulting by all the predictions of the 5 models on the 5 chunks.

The Hierarchical GRU/LSTM Model

GRU units are able to propagate important features that came early in the input sequence over a long distance, thus capturing potential long-distance dependencies. Unfortunately, it has been shown that long dependencies are lost in case of very long sequences. For this reason, since we treat the batch of tweets to be classified as single document, we resorted to a two-layer hierarchical GRU/LSTM architecture.
In addition, each document containing the set of tweets to be analyzed is first truncated to the first 2500 tokens. This operation is done since in-house experiments have not shown a significant drop in performance w.r.t. analyzing all the tweets contained in a single example. Each document is then split in 5 chunks of 500 tokens, which are the input of five different GRU unit (48 dimensions), which produce 5 "chunk" embeddings. Finally, all the chunk embeddings are the input of a final LSTM (48 dimensions) layer.
We applied a dropout factor to both input gates and to the recurrent connections in order to prevent overfitting. Furthermore, we performed a 5-fold training approach.
Each input word is represented by a vector which is composed by:
Word embeddings: the word embedding extracted by the available word embedding lexicon (32 dimensions), and for each word embedding an extra component was added to handle the "unknown word" (1 dimension).
Word polarity: the corresponding word sentiment polarities obtained using SentiWordnet
Is capitalized word: whether the word is capitalized.
Is uppercased word: whether the word is uppercased.
Is URL: whether the word is an URL.
Is hashtag: whether the word is an hashtag.
Is mention: whether the word contains a mention.
Is separator: whether the word is the "SEP" reserved token

Results

Configuration	Bot F-score	Male F-score	Female F-score	Avg F-score
Linear SVM	**0.94**	0.71	0.59	0.746
Hierarchical GRU/LSTM	0.92	**0.76**	0.74	**0.806**
BERT Multi	0.90	0.72	**0.71**	0.776

Table 1: *Classification results of the proposed models on the official development set.*

Models	Bot vs Human	Male vs Female
Best participants	0.9595	0.8432
GRU-LSTM model	0.9083	0.7898
Char nGrams baseline	0.9360	0.7920
Word nGrams baseline	0.9356	0.7989
W2V baseline	0.9030	0.7879

Table 2: *Classification results of Hierarchical GRU/LSTM models on the official test set.*

Takehome messages and conclusions

We presented three systems for the Bots and Gender Profiling shared task. According to our experiments, the SVM algorithm seems to perform better on the bot detection task, while the GRU-LSTM architecture showed good results in the gender prediction task. Surprisingly, BERT did not show a performance gain w.r.t. non language model based architectures. Finally, it would be interesting to add char based features to further improve our GRU/LSTM model.