SENTIMENT ANALYSIS USING DEEP LEARNING FOR SPANISH SHORT TEXTS

Alvaro Rojas Israel Chaparro

Abstract— **The present work is about the implementation of a sentiment analysis system using recurrent neural networks, which were trained with tweets obtained from the TASS. The TASS dataset consist in tweets divided in four classes: positive, negative, neutral and none that our implementation must achieve to classify. The implementation probes multiple data preprocessing techniques and architectures for the network getting a good performance in the general classification.**

I. INTRODUCTION

Sentiment analysis is a type of data mining that measures the inclination of peoples opinions through natural language processing (NLP), computational linguistics and text analysis, which are used to extract and analyze subjective information from the Web - mostly social media and similar sources.

Deep learning has emerged as a powerful machine learning technique that learns multiple layers of representations or features of the data and produces state-of-the-art prediction results. Along with the success of deep learning in many other application domains, deep learning is also popularly used in sentiment analysis in recent years.

The TASS Dataset is a corpus of texts (mainly tweets) in Spanish tagged for Sentiment Analysis related tasks. It is divided into several subsets created for the various tasks proposed in the different editions through the years. All the information on these datasets can be found in the TASS website at http://www.sepln.org/workshops/tass. In TASS Data set we have four different setiment classes: Positive, Negative, Neural and NONE

Word embedding is one of the most popular representation of document vocabulary. It is capable of capturing context of a word in a document, semantic and syntactic similarity, relation with other words, etc. They are vector representations of a particular word.

In this work, we use a deep learning principal model for improve the task of Sentiment Analysis with the TASS Dataset using word embedding (Word2Vec) and another two proposals:

- Universal Language Model Fine-tuning for Text Classification. - Smooth Inverse Frequency

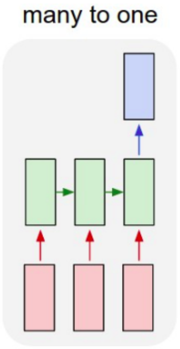
A. PIPELINE:

- Get the 2017 and 2018 TASS Dataset.
- Define principal model.
- Process the Dataset with Word2Vec and another two word embedding proposals.
- Train the principal model with the different word embeddings.
- Compare the word embedding proposals.

II. LSTM MODEL WITH WORD2VEC

We choose as a principal model a Recurrent Neural Network (LSTM) model, in a many-to-one architecture, on the LSTM we use a Word embeddings like inputs and a class of TASS Dataset like a output.

We use word2vec to enconde the tweets and.



III. UNIVERSAL LANGUAGE MODEL FINE-TUNING

This is a recent proposal on Word Embeddings, uses a trained CNN with Wiki Named Entity Recognition Corpus, that assigns vector tokens in a specific context.

Uses a pre-trained CNN, load weights and uses a Fine-tuning process like Transfer Learning: unfrezze all layers except the last two, all layers except the last and unfrezze all to train the CNN with the new input and output data.

The code of implementation are based in FASTAI course: https://github.com/fastai/fastai/blob/master/courses/dl2/imdb.ipynb

IV. SMOOTH INVERSE FREQUENCY:

Uses a pre-trained Glove Word Embeddings and generate a Sentence Embedding using a SIF weighting scheme.

The Sentence vectors are the singular vectors of the matrix whose columns are the word vectors of a Sentence; this method use the weighted average for calculate de word vectors.

The code of implementation are based in: https://github.com/PrincetonML/SIF.

V. RESULTS:

All the results presented are with Development Data.

A. TASS 2017:

	UMLFiT	SIF	Our
ML Model	CNN	MLP	LSTM
Accuracy	0.3951	0.2600	0.4762
F1	0.1946	0.1395	0.3637

B. TASS 2018:

	UMLFiT	SIF	Our
ML Model	CNN	MLP	LSTM
Accuracy	0.3457	0.2578	0.4540
F1	0.2134	0.1446	0.3515

C. TASS 2018 TEST:

Group	Macro-P	Macro-R	Macro-F1	Accuracy
LARMINCC	0.348	0.353	0.350	0.403

REFERENCES

[1] Howard, Jeremy and Ruder, Sebastian. Universal Language Model Fine-tuning for Text Classification In Proceed- ings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 328339, July 2018.

[2] Sanjeev Arora, Yingyu Liang and Tengyu Ma. A Simple But Tough-To-Beat Baseline For Sentence Embeddings. In International Conference on Learning Representations(ICLR), 2017.