Human Language Technlogies
Twitter Sentiment Analysis

Gianmarco Ricciarelli
gianmarcoricciarelli@gmail.com

Abstract

Nowadays, when delving into the Social Network landscape, a user can choose among different paths, depending on the type of experience he/she is searching. Regardless the type of platform that is chosen by the user, either Facebook, or Instagram or Twitter, the amount of textual data that is produced every day is massive. With this report, I describe the project I developed for the *Human Language Technologies* class hosted by the University of Pisa's Master's Degree in Computer Science, that is, a prediction-oriented analysis of a dataset composed by almost 40000 labeled tweets via a series of algorithms like SVM, NB, CNN and LSTM.

1 Corpus Collection

The amount of textual data produced by the users on the various Social Networks on a day-by-day basis is massive, and, for this reason, it is not too difficult to scrape the web in order to build a corpus composed by an acceptable amount of documents. Among all the available platforms, I choosed to work with Twitter, since the documents, that is, the tweets, produced by the users are limited to be 280 characters long at most, and since the API that is used for communicating with the server is very simple and inuitive to use. In order to properly train the algorithms that I wanted to compare, I assembled a corpus of almost 40000 labeled tweets by requesting Twitter to download the documents used for the *International Workshop on Semantic Evaluation* competitions from 2013 to 2017. I choosed to work with the documents provided by the SemEval competitions because every tweet was hand-labeled by a human, that is, guaranteeing a sound catalogation of the sentiments expressed by the tweets. A tweet can be labeled either as positive, negative, or neutral. Despite the tweets' labels being so well defined, a SemEval dataset is quite small if it is taken on its own. This is due to the fact that, since Twitter is a very dynamic and constantly upgraded platform, some of the documents provided by each dataset were deleted by the users. This is why I decided to merge the datasets provided by the various competitions in order to obtain the final version of the corpus.

2 Corpus Analysis

The original version of the corpus I assembled was composed by more than 40000 labeled tweets. As I said before, some of the tweets were deleted by the users, resulting in a document containing the 'Not Available' string. Moreover, the corpus contained a small percentage of duplicate tweets. By removing the not available tweets and the duplicates, I obtained the final version of the corpus composed by 39308 labeled documents. In particular, the obtained corpus contains:

- 15092 positive tweets;
- 6007 negative tweets;
- 18209 neutral tweets;

As we can see, the majority of the documents composing the corpus are labeled as neutral, and also the positive tweets are very well represented. Conversely, the negative ones are poorly represented, and, as we will see, this fact will be crucial for their classification in the learning and testing phases of the algorithms I will apply. The type of subdivision used for the documents allows the corpus to be considered as containing both subjective and objective tweets. By considering this new subdivision, we can see that the corpus contains:

- 21099 subjective tweets;
- 18209 objective tweets;

This time, both the two sets of documents are well represented. In order to provide a richer analysis, I decided to apply the classification algorithms on three types of problems; the positive-vs-negative-vs-neutral problem, the positive-vs-negative problem and finally the subjective-vs-objective problem. For starting my analysis, I tokenized the tweets contained in the corpus and I observed the distribution of the word frequencies.

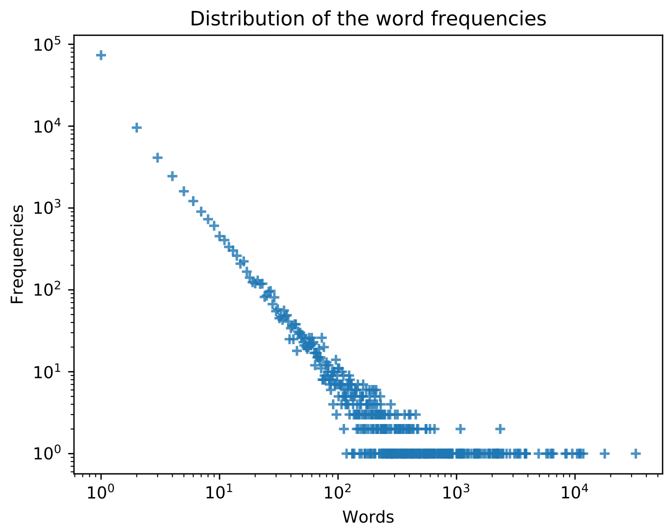

Figure 1: *Word Frequencies Distribution.*

As we can see in Figure 1, the words' frequencies follow the well known Zipf's law, with a small number of words that appears frenquently, the so called stop words, and the majority of the remaining words that appears less and less frenquently. Following the analysis proposed in [1], I updated the tokens obtained by the early stage of the analysis by adding the tags from the application of a Part of Speech Tagging procedure, and I observed the tags' distribution between the corpus' documents when considering only the positive/negative tweets and the subjective/objective tweets.

As I said before, [1] proposes the following formula to perform a pairwise comparison of tags distributions for each tag and two sets:

$$P^T_{1,2} = \frac{N^T_1 - N^T_2}{N^T_1 + N^T_2}$$

where N^T_1 and N^T_2 are numbers of tag T occurrences in the first and second sets respectively. As we can see from Figure 2, POS tags for positive and negative tweets are not distributed evenly, hence we can infer

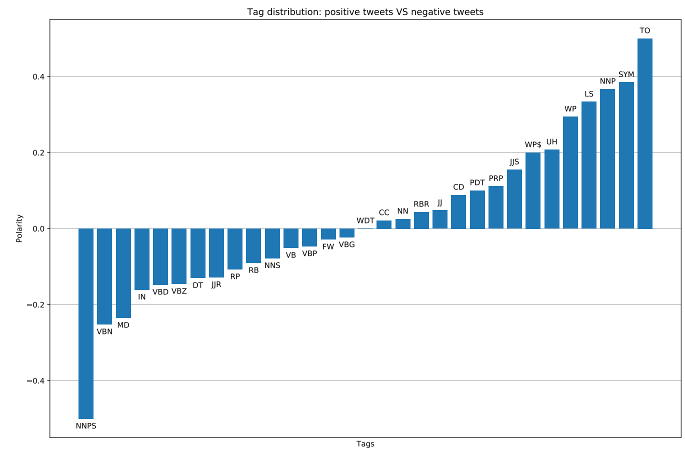

Figure 2: *Tag Distribution for Positive and Negative tweets.*

the sentiment behing a document by looking at its tags. We can see that the presence of a plural proper noun (NNPS) is a strong indicator for the positive label as well as past participle verbs (VBN) and modal (MD). For the negative label, we can see that 'to' (TO) is a strong indicator, as well as symbol (SYM) and singular proper nouns (NNP).

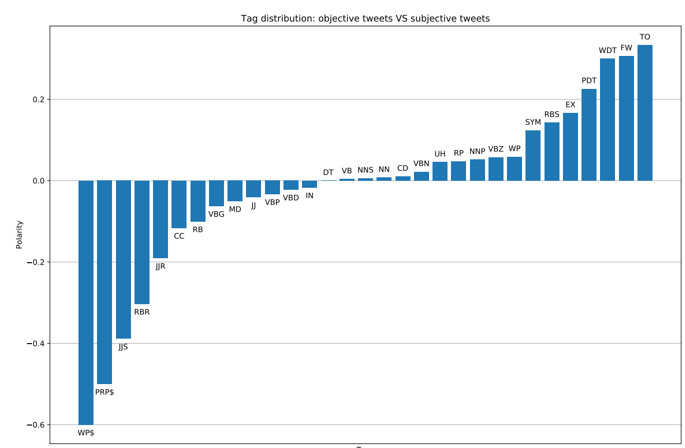

Figure 3: *Tag Distribution for Objective and Subjective tweets.*

In Figure 3 we can observ the tags' distribution for the subjective and objective tweets. We can see that possessive wh-pronouns (WP$), possessive pronouns (PRP$) and superlative adjectives (JJS) are all indicators for objective tweets, while 'to' (TO), foreign words (FW) and wh-determiners (WDT) are indicators for subjective tweets. Finally, I have collected some of the most used words for positive and negative tweets and represented them via wordclouds, in order to provide a better understanding of the sentiments that labels the documents composing the corpus, and for providing some visual representation of the informations showed with Figure 2 and Figure 3.

Figure 4: *Positive tokens wordcloud.*

In Figure 4 we can see the wordcloud for the tokens contained in the positive labeled tweets, while in Figure 5 we can see the tokens for the negative labeled ones.

Figure 5: *Negative tokens wordcloud*

3 Classification

The Human Language Technologies research's field is very popular at the moment, and, in particular, the Sentiment Analysis branch has a rich literature and many tools and resoruces availables for learning. Having analysed the obtained corpus, as described in Section 1 and Section 2, the next step for my research was to implement a Classification procedure, in which I decided to apply some popular Machine Learning algorithms like Support Vector Machines, Naïve Bayes, Convolutional Neural Networks and Long Short-term Memory Recurrent Neural Networks, as described in [2, 3, 4, 5, 6]. Being the corpus composed of documents labeled with three distinct labels, that is, positive, negative and neutral, I decided to direct the Classification phase towards the takling of three different tasks:

- Classification of positive tweets vs negative tweets;
- Classification of positive tweets vs negative tweets vs neutral tweets;
- Classification of subjective tweets vs objective tweets;

As we might expect, for the first and third tasks the entire corpus was used, while for the second one I removed from the corpus all the documents labeled as neutral. As in a typical Machine Learning procedure, the documents composing the corpus cannot be fed directly to the algorithms, since they are texts, and not numbers, so, before applying the algorithms, I developed a preprocessing routine, in which I developed the conversion from texts understandable by humans to texts understandable by machines. After preprocessing the corpus I validated each one of the learners via the application of well-known validation techniques like Cross Validation and Random Grid Search. Finally, for each one of the three tasks, I collected and compared the results provided by the application of the various learning algorithms.

3.1 Preprocessing Routine

As I said before, there are plenty of resources for learning about Human Language Technologies, and, in general, Sentiment Analysis. After reading about the subject, I developed a Preprocessing Routine that is composed by the following steps:

1. Tokenization: the original documents are splitted into tokens via a tokenizer that is specifically trained to work on tweets;
2. Part Of Speech Tagging: each token list is tagged via the application of a POS tagger;
3. Cleaning: for each token list, the token that are not alphanumeric and that contains stop words are removed;
4. Lemmatization: each one of the token that came out of the Cleaning phase is transformed in its original lemma;

As an example, if we apply the Preprocessing Routine to the document:

`April has to live. Sharknado wouldn't be the same without her tiny chainsaw. #aprillives`,

we obtain the token list:

`['april', 'live', 'sharknado', 'without', 'tiny', 'chainsaw']`.

3.2 Validation Routine

Every Machine Learning algorithm is characterized by a number of parameters, that is, the so-called hyperparameters, that the user can modify accordingly to the type of task he/she is takling. Being the task of tuning every single hyperparameter a cumbersome one, I decided to apply the well-know Random Grid Search technique, as described in [7], in order to discover the most effective combination of hyperparameters for the models I want to compare. I used the 80% of the corpus as a training set, while the remaining 20% was used as testing set. Moreover, since the training/validation set extracted from the corpus presents a reasonable amount of examples, for each "point" produced by the Random Grid Search, I applied the K-Fold Cross Validation technique, setting the value for the K parameter to 3.

3.3 Positive vs Negative Classification

References

[1] Pak, Alexander & Paroubek, Patrick. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Proceedings of LREC. 10.

[2] Nakov, P. & Rosenthal, S. & Kiritchenko, S. et al. Developing a successful SemEval task in sentiment analysis of Twitter and other social media texts. Lang Resources & Evaluation 50, 35–65 (2016). https://doi.org/10.1007/s10579-015-9328-1

[3] Mohammad, Saif & Kiritchenko, Svetlana & Zhu, Xiaodan. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. Second Joint Conference on Lexical and Computational Semantics, Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation, June 2013, Atlanta, Georgia, USA.

[4] Nakov, Preslav & Ritter, Alan & Rosenthal, Sara & Sebastiani, Fabrizio & Stoyanov, Veselin. SemEval-2016 Task 4: Sentiment Analysis in Twitter. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), June 2016, San Diego, California.

[5] Esuli, Andrea. ISTI-CNR at SemEval-2016 Task 4: Quantification on an Ordinal Scale. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), June 2016. Association for Computational Linguistics, San Diego, California.

[6] Attardi, Giuseppe & Sartiano, Daniele. (2016). UniPI at SemEval-2016 Task 4: Convolutional Neural Networks for Sentiment Classification. 220-224. 10.18653/v1/S16-1033.

[7] James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 13, 1 (February 2012), 281–305.