CriCa team: MultiModal Stance Detection in tweets on Catalan 1Oct Referendum (MultiStanceCat)

Almendros Cuquerella, Carlos[1 [0000−0001−6042−7856]] and Cervantes Rodríguez, Cristóbal[1 [0000−0003−2581−4468]]

Universitat Politècnica de València, Valencia, Spain
{calmendrosc,cristobalcerv}@gmail.com

Abstract. This paper describes the process that we followed to develop our stance analysis tool for IberEval 2018 on MultiModal Stance Detection in tweets on Catalan 1Oct Referendum (MultiStanceCat) task. Our approach is based on the tools provided by the scikit-learn toolkit[3] to develop a system capable of detect the stance of some tweets about the Catalan 1Oct Referendum, using only the text written in the body of the tweet and also using the context formed by the previous and the next tweet of the one we are analyzing.

Keywords: Stance Detection · Catalan 1Oct Referendum · Twitter · IberEval 2018.

1 Introduction

Nowadays, the large amount of new data produced by services like Twitter brings the opportunity to get useful and interesting information about people opinion and feelings on a wide variety of topics. This high volume of information could be difficult or nearly impossible to handle and process by human operators, requiring to make use of advanced algorithms to obtain the relevant information hidden on the data with low time and economic costs.

This information can be used for tasks that improve the service quality and security (among others) offered by companies, which are becoming more interested on the text classification field.

For the purpose of obtaining the opinion and feelings of the society about popular topics, Twitter has become an attractive tool, being used for many studies [2]. One important area of study is text classification, whose aim is to label natural language texts into a fixed number of predefined categories (eg. favor, against, neutral, ...).

For the MultiStanceCat [1] task, we found that using Twitter as a source of data adds the difficulty of having to use text that in addition of being written in informal natural language, has a limited amount of characters available, thing that makes people use contractions or remove letters that doesn't need a human to understand the message, but could get some trouble for our algorithm, which would understand the correct word and the reduced one as different, if no countermeasure is taken against this. Also, in tweets we can also find grammatical errors, vulgar vocabulary or slang, abbreviations, hashtags, links and images, which requires a more complex analyzer to be developed to be able to get advantage of them.

In this task, we had to classify a set of tweets related to the Catalan 1Oct Referendum based on the stance (Favor, Against or Neutral) of the person who wrote it, concerning the independence of Catalonia.

The paper is structured as follows. Section 2 describes our system and the preprocessing we made to the dataset. Next, in Section 3, the obtained results are discussed. Finally, we present our conclusions in Section 4 with a summary of our findings.

2 Dataset and system

The dataset is formed by 6 files, distributed in the following way: 2 set of tweets for training, 2 files with the labels for these tweets, and 2 files for test. One of each of those three types of files is for Spanish language and the other for Catalan.

We joined the two train files (Catalan tweets and Spanish tweets) to compose a bigger corpus and shuffled it, to obtain an homogeneous distribution of the tweets, before dividing it in two parts. The first one, containing 80% of the texts, used for the training phase and the second one, with the remaining 20%, used to validate the analyzer among training iterations.

Our first approach was to tokenize the tweets to separate the different words before passing them to the vectorizer. Then we obtained a features matrix based on a Tf-idf vectorizer, where we can find a row for each tweet and a column for the weight of each feature, related with the frequency of each token in the tweet and in the corpus. With this features matrix and the tags obtained from the labels files, we trained a LinearSVC classifier and got a model to classify new tweets based on their stance.

Finally, we used the development set to evaluate the classifier, vectorizing the tweets with the same vectorizer used before and classified them with the LinearSVC trained with the training set. The results of this approach can be found in the next section (see Table 1).

Using the previous approach as base, we tried to add new features, with the intention of improving the obtained results. Our first hypothesis was that we could use the similarities between Spanish and Catalan to increase the number of occurrences of each feature in the training case and to decrease the quantity of different features that could appear while training or classifying.

Spanish and Catalan languages have many words that share the same stem but have different ending, for example the Spanish word "televisión" is treated as different from the same Catalan word "televisió" in our first approximation. Our intention is to make our analyzer capable to detect that type of situations.

We started using one of the stemmers included in the NLTK toolkit [4], the Snowball stemmer, that is capable of stem for Spanish language (see Table 4).

Next, we tried fixing the maximum length of the words, having launched the analyzer with a maximum length of 3, 4, 5 (Table 2). In this case, a word like "televisión" will be transformed to "tel", "tele" and "telev" respectively. This gives us the advantage of making the vectorizer to treat this word and its Catalan translation as if they were the same work. This approach had the disadvantage of generalizing too much, causing that words like "casa" (house) and "castell" (castle) to be recognized as the same one, when using a maximum length of 3. This could happen for the other length too, with other words.

After that, we tried a variation of the last experiment, using a fixed length of characters at the end of the word to be removed. The results of this experiment with length 1, 2 and 3 are reported in Table 3. This way of processing each word has the disadvantage of having a big impact on short words and an almost negligible one on long words. For this reason we defined in Table 4 ranges of lengths and a fixed quantity of characters to be removed, causing the deletion length to be proportional to the word length.

For the second part of the MultiStanceCat task, we had to study if adding the previous and the next tweets to the current tweet that we are analyzing could help to increase the stance classification accuracy. We used additional information in 3 different way:

1. Concatenating these three text bodies, adding a space between them (prev + text + next).
2. Concatenating these three text bodies, trying to give more weight to the text of the current tweet, duplicating it in the concatenated text (prev + text + text + next).
3. Vectorizing each of these three tweets in an independent way and joining their features matrix before training Linear SVC.

Results are given in Table 5

3 Results

Experiments using both languages at the same time and using them in an independent way have been carried out (Table 1) having obtained better results in the first case. Using a too short prefix length showed to worsen the obtained results (see Table 2), but no significant variations on F1-macro have been seen when trying different values for the fixed length suffix removal approach. The best results have been obtained with the ranged suffix removal length, because the length of the removed suffix is proportional to the word's length, resulting in a bigger removal on long words and shorter in the short ones.

Type	F1-macro	precision	recall
Both languages together	0.69171	0.71692	0.67917
Only Catalan language	0.51662	0.86154	0.46529

Table 1. Only text without stemmer

Length	F1-macro	precision	recall
3	0.60584	0.63833	0.59843
4	0.65126	0.66710	0.64103
5	0.65662	0.67052	0.65201

Table 2. Fixed prefix length stemmer

Length	F1-macro	precision	recall
1	0.65684	0.68332	0.64528
2	0.66167	0.67276	0.65805
3	0.65604	0.67163	0.64865

Table 3. Fixed suffix removal length stemmer

Type	F1-macro	precision	recall
Best fixed prefix length	0.65662	0.67052	0.65201
Best fixed suffix removal length	0.66167	0.67276	0.65805
NLTK Snowball (Spanish)	0.66156	0.67753	0.65522
Ranged suffix removal length	0.69118	0.71940	0.67721

Table 4. Stemmer comparative

When using the context as additional information, better results have been achieved when duplicating the tweet's text, to increase the weight of this information in relation with the context information.

Type	F1-macro	precision	recall
Prev + Text + Next	0.66813	0.74715	0.64524
Prev + Text + Text + Next	0.67536	0.75615	0.65073
Vectorizers concatenation	0.63634	0.71835	0.61710

Table 5. Stance analyzer with context

As can be seen comparing the previous results with the ones in Table 6 and Table 7 the macro-F's obtained by us for our results were far better than the ones provided by the organization. This happens because we used a different way to evaluate the correction of our stance detector, checking the quantity of right classifications for all three classes, instead of only looking for the positive and negative stances (discarding the neutral one) as done by organizers.

Team	Run	Macro-F
CriCa	text+context	0.3068
Casacufans	text+context	0.2933
Casacufans	text+context+images	0.2913
uc3m	text+context	0.2876
CriCa	text	0.2315
Casacufans	text	0.2247
uc3m	text	0.2195

Table 6. General results for Catalan language

Team	Run	Macro-F
uc3m	text+context	0.2802
CriCa	context	0.2715
Casacufans	text+context+images	0.2709
Casacufans	text+context	0.2698
ELiRF	text (run1)	0.2274
uc3m	text	0.2247
CriCa	text	0.2206
Casacufans	text	0.2194
ELiRF	text (run2)	0.2132

Table 7. General results for Spanish language

4 Conclusions

For this task, we have developed two models to classify tweets according to their stance in two similar languages (Spanish and Catalan). Both were trained with a LinearSVC classifier and the difference was that on the second one we also made use of the context to extract features.

In the first model (without context) we observed that using the stem of the words, instead of the whole word, improves the accuracy of results, if these stems are long enough. For this reason, one approach that could be tested, is to translate one of the two corpus in order to work with the same language and extract the features on the basis of a single language.

Looking at the results of the second model, it is proved that the use of context is relevant for this task. This can be seen in the table of results (Table 6 and Table 7) where this executions take a Macro-F of 0.3068 versus the 0.2315 obtained without context.

The good results obtained in IberEval denote that the use of this classifier along with the data preprocessing techniques that we tried are a good election for this task, but there is still a wide range of improvements that can be added to increase the accuracy of our analyzer.

References

1. Taulé M., Rangel F., Martí M.A., Rosso P. 'Overview of the Task on MultiModal Stance Detection in Tweets on Catalan #1Oct Referendum'. In Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), Seville, 18 September 2018.
2. Taulé, M., Martí, M.A., Rangel F., Rosso M., Bosco C., Patti, V. (2017) Overview of the task on Stance and Gender Detection in Tweets on Catalan Independence at IberEval 2017. Notebook Papers of 2nd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), Murcia, Spain, September 19, CEUR Workshop Proceedings: 157-177. CEUR-WS.org.
3. scikit-learn: Machine Learning in Python
   `http://scikit-learn.org`
4. NLTK: Natural Language Toolkit
   `https://www.nltk.org/`