ghostwriter19 @ SardiStance: Generating new Tweets to Classify SardiStance EVALITA 2020 Political Tweets

Mauro Bennici
You Are My Guide
Torino
mauro@youaremyguide.com

Abstract[1]

English. Understanding the events and the dominant thought is of great help to convey the desired message to our potential audience, be it marketing or political propaganda.

Succeeding while the event is still ongoing is of vital importance to prepare alerts that require immediate action.

A micro message platform like Twitter is the ideal place to be able to read a large amount of data linked to a theme and self-categorized by its users using hashtags and mentions.

In this research, I will show how a simple translator can be used to bring styles, vocabulary, grammar, and other characteristics to a common factor that leads each of us to be unique in the way we express ourselves.

Italiano. Comprendere gli eventi e il pensiero dominante è di grande aiuto per veicolare alla nostra potenziale audience il messaggio desiderato sia esso di marketing o di propaganda politica.

Riuscirci mentre l'evento è ancora in corso è di vitale importanza per predisporre alert che richiedono un intervento immediato.

Una piattaforma di micro messaggi come Twitter è il luogo ideale per poter leggere una grande quantità di dati legata ad un tema, e spesso auto categorizzati dai suoi stessi utenti per mezzo di hashtag e menzioni.

In questa ricerca mostrerò come un semplice traduttore può essere usato per portare a fattor comune stili, lessico, grammatica e altre caratteristiche che portano ognuno di noi ad essere unico nel modo di esprimersi.

1 Introduction

Each of us has a unique way of writing. However, the fewer options we have to experience ourselves to express our concept, the more the necessary synthesis leads to the loss of precious information to accurately assess our real intentions.

Furthermore, the more the subject is debated, the more changes in style and tone occur. The conversation becomes full of irony or aggressive. Extrapolating a single line is dangerous without context. The same sentence can have different interpretations depending on the moment in which it is pronounced, the audience it is intended for, the place where you are, in the historical period in which it was composed.

My hypothesis is that we can translate all these different styles into a single "language style" that fully expresses the real intentions of the writer. The challenge is to understand when a user has expressed a comment in favor, against, or neutral towards the Sardines' Italian political movement.

The research was carried out for the SardiStance (Cignarella et al., 2020) task in the EVALITA 2020 (Basile et al., 2020). Two models were created for the Task 1, but they also performed well on the Task 2.

2 Description of the system

The two tasks are similar. In Task A, it is necessary to classify the stance of a tweet based only on the text of the tweet. Task A is divided into two subtasks:

- Constrained. It is allowed to use additional resources such as a Lexicon but no other resources (such as labeled tweets) to help the training process.
- Unconstrained. Where each resource used must be reported in the final report.

In Task B, you can use the context information provided by the post author. Additional information refers:

- to post statistics (favors, retweets, reply, source)
- to the author's information (number of posts, number of followers, emoji in the bio)
- to the author's circle of relationships (friends, replies, retweets, and quotes)

The research focuses on Task A Constrained.

Considering the constraints of Task A, it is not possible to access any additional information other than the text of the tweet, I concentrated on understanding how to clean it up.

The Training dataset contains:

- the tweet ID
- the user ID
- the text
- the label

The labels options are:

- Against
- Favor
- Neutral / None

To be sure to do not use any data except the text, the user id, useful for Task B, was discarded.

In order to validate my hypotheses, I used the AlBERTo model, created from tweets, (Polignano at al., 2019) and an auto training system such as Ktrain[2], a framework that wrap TensorFlow[3], to classify the tweets. To avoid manual error and involuntary optimization, I used the autofit option.

First, I wrote a series of algorithms to make the texts to be compared homogeneous.
The first one was to break up the composed hashtags into sentences and words.
For example, using capital letters as a separator:

- #IoStoConLeSardine has become "io sto con le sardine" ["I'm with sardines"].
- #NessunoTocchiLeSardine has become "nessuno tocchi le sardine"["nobody touches the sardines"].

As a second step, I made sure to remove repeated vowels in a sentence, such as:

- "Svegliaaaa" to get the word "Sveglia" [Wake up!].

I also replaced the word sardines with "PartitoPoliticoS" ["PoliticalPartyS"] to prevent the entity from being mistaken for the fish that is its symbol. I did not remove any stop words because it is useful to create the translation system.

At this point, I made a copy of the dataset to translate it. I used the spaCy[4] language functions of POS tagging, Dependency Parse, and Entity Recognition to have all the essential components of my translator.

The translator is a simple text representation. It is a matter of rewriting the sentence following the scheme:

- subject adjectives
- subjects
- verb in the infinitive form
- adjectives objects
- objects
- exclamations / other words

At this stage, the words are not modified to make the sentence grammatically correct. Words are exchanged places, only the verb are modified to the

[2] https://github.com/amaiya/ktrain
[3] https://www.tensorflow.org/
[4] https://spacy.io/api/annotation

infinitive form. The entities of type person [PER] take precedence over others.

The translator concentrates its attention on the aspect inside the sentences to be sure to do not remove valid sentiment polarity words (Barbieri et al, 2016). And to avoid to lose them in a round-trip translation activity on translation services (Marivate & Sefara, 2020). The attempt to represent the text in a more recognizable and identifiable form for an algorithm passes from the fact that it can still recognize the entities described and the polarity expressed for each of them. For this purpose, the translator makes several attempts to fit words into their suggested position.

Finally, I trained two models with the Ktrain framework. The model 1, which use the translated tweets, was submitted as ghostwriter19_Task_A_1_c. The model 2, trained with the only cleaned tweets, was submitted as ghostwriter19_Task_A_2_c.

2.1 First results

The model will be evaluated with the F1-score. The main score is the average of the F1-score of the Favor tweets and the F1-score of the Against tweets.

When comparing the two models, the first result is that the translated tweets performed worse, albeit by a few percentage points (table 1).

Model	**F1-Score**
ghostwriter19_Task_A_1_c	0.5613
ghostwriter19_Task_A_2_c	0.6004
Estimated Baseline	0.5386

Table 1: First results

Analyzing the results of both the models in detail (table 2 and 3), we have that:

ghostwriter19_Task_A_1_c	**F1-Score**
Against	0.69
Favor	0.43
Neutral	0.42

Table 2: F1-score details of model 1

ghostwriter19_Task_A_2_c	**F1-Score**
Against	0.70
Favor	0.50
Neutral	0.32

Table 3: F1-score details of model 2

The problem is evident. Model 1 has a more challenging time distinguishing the favor tweets from neutral ones. The good news is that both the models overcame the estimated baseline.

2.2 Hashtags and Mentions

Thinking that on Twitter the hashtags are also used for classification purposes, the operation that replaces them was modified. Now the hashtags are added at the end of the new tweets. Also, the mentions are considered and processed as hashtags (table 4).

Model	**F1-Score**
ghostwriter19_Task_A_1_c	0.5822
ghostwriter19_Task_A_2_c	0.6004
Estimated Baseline	0.5386

Table 4: Model 1 with hashtags and mentions in the translated tweets

Analyzing the results in detail (table 5), we can see that:

ghostwriter19_Task_A_1_c	**F1-Score**
Against	0.71
Favor	0.45
Neutral	0.41

Table 5: F1-score details of model 1 with hashtags and mentions in the translated tweets

The model gained two percentage points for both Against and Favor, compared with a one-point loss in Neutral. Unfortunately, it still remains two points below the model 2, with the only cleaned tweets.

2.3 Passive verbs

Analyzing the new texts generated, I noticed that essential information was lost by putting all the verbs in the infinitive. If the verb was in the passive form, the subject and object of the sentence were reversed. At the same time, I noticed that very long tweets contained more than one sentence.

I modified the translator to consider passive and active verbs, swapping the sentence's subject and object if necessary. The hashtags inserted at the end of the tweet only left at the end of the new tweet generated (table 6).

Model	F1-Score
ghostwriter19_Task_A_1_c	0.6306
ghostwriter19_Task_A_2_c	0.6004
Estimated Baseline	0.5386

Table 6: Model 1 with hashtags and mentions in the translated tweets, plus active / passive verbs

Analyzing the results in detail (table 7), we can see that:

ghostwriter19_Task_A_1_c	F1-Score
Against	0.76
Favor	0.50
Neutral	0.40

Table 7: F1-score details of model 1 with hashtags and mentions in the translated tweets, plus active / passive verbs

The model gained five percentage points for Against and Favor tweets, compared with a one-point more loss for Neutral ones. Now the translation model is the best model.

3 Results

Model 1 was ultimately 3 percentage points better than Model 2 with the Training dataset. The best performance of the model was also confirmed with Test datasets, with 2.5 percentage points of advantage.

3.1 Results for Task A

The final results with the Test dataset are:

Model	F1-score
ghostwriter19_Task_A_1_c	0.6257
ghostwriter19_Task_A_2_c	0.6004
Baseline	0.5784

Table 8: Test dataset results for Task A

The model 1 is about 7.5% better than the baseline (table 8).
I remember that both models were trained with the autofit option, so without any particular study, to validate whether a "translation" of the original text could bring apparent advantages.

3.2 Results for Task B

Although no context information was used, I still proposed the predictions for Task A to Task B.
The final results with the Test dataset are:

Model	F1-score
ghostwriter19_Task_A_1_c	0.6257
ghostwriter19_Task_A_2_c	0.6004
Baseline	0.6284

Table 9: Test dataset results for Task B

Even if model 1 was not able to reach the proposed baseline, the difference between the two systems is 0.4% (table 9). The detailed results of the models are showed in the tables 10 and 11.

3.3 Detailed results for Task A

model	f-avg	prec_a	prec_f	prec_n	recall_a	recall_f	recall_n	f_a	f_f	f_n
1_c	0.6257	0.8106	0.4709	0.3226	0.6981	0.5357	0.4651	0.7502	0.5012	0.3810
2_c	0.6004	0.8094	0.4772	0.2921	0.6523	0.4796	0.5349	0.7224	0.4784	0.3778
baseline	0.5784	0.7549	0.3975	0.2589	0.6806	0.4949	0.2965	0.7158	0.4409	0.2764

Table 10: TASK A detailed results of the proposed models compared to the baseline model.

3.4 Detailed results for Task B

model	f-avg	prec_a	prec_f	prec_n	recall_a	recall_f	recall_n	f_a	f_f	f_n
1_c	0.6257	0.8106	0.4709	0.3226	0.6981	0.5357	0.4651	0.7502	0.5012	0.3810
2_c	0.6004	0.8094	0.4772	0.2921	0.6523	0.4796	0.5349	0.7224	0.4784	0.3778
baseline	0.6284	0.7845	0.4506	0.3054	0.7507	0.5357	0.2965	0.7672	0.4895	0.3009

Table 11: TASK B detailed results of the proposed models compared to the baseline model.

4 Conclusion

In a preliminary way, the final results demonstrate that it is possible to obtain an improvement of the predictions by reducing the differences of expression to a predetermined structure.

The system is, however, right now, more efficient in terms of training times and final scores than ensemble systems of Bi-LSTM, which were used successfully up to 2 years ago (Bennici & Portocarrero, 2018).

The next step is also to optimize the model's training to ascertain that the performance gain is maintained and in what percentage. At the same time, the translator can be improved by switching to a sequence-to-sequence system for a meaningful and efficient text representation that will include, among other things, the change of every words forms accordingly with the grammar and the original intention of the writers (Lewis et al., 2019).

References

Barbieri, F., Basile, V., Croce, D., Nissim, M., Novielli, N., & Patti, V. (2016). Overview of the Evalita 2016 SENTIment POLarity Classification Task. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*, Naples, Italy. CEUR-WS.org.

Basile, V., Croce, D., Di Maro, M., & Passaro, L. (2020). EVALITA 2020: Overview of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, CEUR-WS.org.

Bennici, M., & Portocarrero, X. S. (2018). Ensemble for aspect-based sentiment analysis. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR-WS.org.

Cignarella, A. T., Lai, M., Bosco, C., Patti, V., & Rosso, P. (2020). SardiStance@EVALITA2020: Overview of the Task on Stance Detection in Italian Tweets. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*. CEUR-WS.org.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., . . . Zettlemoyer, L. (2019, October 29). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. https://arxiv.org/abs/1910.13461

Marivate, V., & Sefara, T. (2020). Improving Short Text Classification Through Global Augmentation Methods. *Lecture Notes in Computer Science Machine Learning and Knowledge Extraction*, 385-399. doi:10.1007/978-3-030-57321-8_21

Polignano, M., Basile, P., de Gemmis, M., Semeraro, G., & Basile, V. (2019). Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*. CEUR-WS.org.