Tweet2Check evaluation at Evalita Sentipolc 2016

Emanuele Di Rosa
Head of ML and Semantic Analysis
Finsa s.p.a., Via XX Settembre 14
emanuele.dirosa@finsa.it

Alberto Durante
Research Scientist
Finsa s.p.a., Via XX Settembre 14
alberto.durante@finsa.it

Abstract

English. In this paper we present our Tweet2Check tool, provide an analysis of the experimental results obtained by our tool at the Evalita Sentipolc 2016 evaluation, and compare its performance with the state-of-the-art tools that participated to the evaluation. In the experimental analysis, we show that Tweet2Check is: *(i)* the second classified for the irony task, at a distance of just 0.0068 from the first classified; *(ii)* the second classified for the polarity task, considering the unconstrained runs, at a distance of 0.017 from the first tool; *(iii)* in the top 5 tools (out of 13), considering a score that allows to indicate the *most complete-best performing* tools for Sentiment Analysis of tweets, i.e. by summing up the best F-score of each team for the three tasks (subjectivity, polarity and irony); *(iv)* the second best tool, according to the former score, considering together polarity and irony tasks.

Italiano. *In questo paper presentiamo il nostro sistema Tweet2Check, produciamo un'analisi dei risultati sperimentali ottenuti dal nostro strumento nella valutazione effettuata nell'ambito di Evalita Sentipolc 2016, e confrontiamo la sua performance con quella degli altri sistemi partecipanti. Nell'analisi sperimentale, mostriamo che Tweet2Check è:* (i) *il secondo classificato per il task dedicato alla rilevazione dell'ironia, ad una distanza di appena 0.0068 dal primo classificato;* (ii) *il secondo classificato per il task dedicato alla classificazione della polarità, considerando i sistemi* unconstrained*, ad una distanza di 0.017 dal primo classificato;* (iii) *tra i migliori 5 tool (su 13), considerando un punteggio volto ad individuare gli strumenti più completi e meglio performanti per l'analisi del sentiment dei tweet, cioè sommando la migliore F-score di ogni team per i tre task (soggettività, polarità e ironia);* (iv) *il secondo miglior strumento, secondo lo stesso precedente punteggio, considerando insieme i task di polarità e ironia.*

1 Introduction

In this paper we present Tweet2Check, a machine learning-based tool for sentiment analysis of tweets, in which we applied the same approach that we implemented in App2Check and that we have already validated in Di Rosa and Durante (2016-a; 2016-b), showing that it works very well (the most of the times is the best tool) in the field of analysis of apps reviews; moreover, this approach has been also validated on general product/service reviews, since our tool was classified as second at the International Semantic Sentiment Analysis Challenge 2016 (Sack et al., 2016), related to the polarity classification of Amazon product reviews. Our own research interest in participating to the Sentipolc 2016 evaluation is to apply the methodology that was mainly designed to analyze apps reviews, and thus adapted to analyze tweets, and evaluate its performance on tweets. From a research point of view, it is also interesting, to understand if it is possible to obtain good results by applying the same approach to very different domains such as apps reviews and tweets.

Starting from the results provided by the organizers of the Sentipolc 2016 evaluation, we performed an analysis of the results in which we show that Tweet2Check is: *(i)* the second classified for the irony task, at a distance of just 0.0068 from the first classified; *(ii)* the second classified for the polarity task, considering just the unconstrained

runs, at a distance of 0.017 from the first tool; *(iii)* in the top 5 tools (out of 13), considering a score that allows to indicate the *most complete-best performing* tools for Sentiment Analysis of tweets, i.e. by summing up the best F-score of each team for the three tasks (subjectivity, polarity and irony); *(iv)* the second best tool, according to the former score, considering together polarity and irony task.

Finally, we show that Tweet2Check unconstrained runs are overall always better (or almost equal) than the constrained ones. To support our hypothesis, we provide an evaluation of Tweet2Check also on the Sentipolc 2014 (Basile et al., 2014) datasets. This is very important for an industrial tool, since it allows to potentially predict well tweets coming from new domains, by keeping in the training set a higher number of examples discussing different topics, and thus to generalize well from the perspective of the final user.

2 Tweet2Check description

Tweet2Check is an industrial system using an approach in which supervised learning methods are applied in order to build predictive models for the classification of subjectivity, polarity and irony in tweets. The overall machine learning system is an ensemble learning system which combines many different classifiers, each of which is built by us using different machine learning algorithms and implementing different features: this allows to take advantage of different complementary approaches, both discriminative and generative. To this aim, we considered the most well known machine learning algorithms, considering both the most established and the newest approaches. For each task, every classifier has been trained separately; then, the ensemble combines the predictions of the underlying classifiers. The training of the models is performed by considering only the tweets provided by Sentipolc 2016 for the constrained run, and other tweets discussing other topics for the unconstrained run. While performing the training of the models, many features, which are both Twitter-specific and source-independent, are generated. Moreover, some features allowing to "connect" different tasks are also considered in the pipeline to determine subjectivity, polarity and irony. For example, in the pipeline to determine the polarity of a tweet, a score related to its subjectivity is also included as a feature, thus by reflecting the conceptual connection that there is in reality between subjectivity and polarity: if a tweet can have a polarity assigned is also subjective. The same kind of connection is also applied to the other models.

Tweet2Check does not use just the prediction coming from the predictive model, but it applies also a set of algorithms which takes into account natural language processing techniques, allowing e.g. to also automatically perform topic/named entity extraction, and other resources which have been both handcrafted and automatically extracted. Unfortunately, it is not possible to give more details about the engine due to non-disclosure restrictions.

Tweet2Check is not only constituted by a web service providing access to the sentiment prediction of sentences, but it is also a full user-friendly web application allowing, between other features, to:

- Perform queries on Twitter
- Show the main topics discussed in tweets which are both comment-specific, associated to a specific month or evaluated to the overall results obtained by the query
- Show the polarity, subjectivity and irony associated to each tweet under evaluation
- Show the sentiment of the former extracted topics

A demo of Tweet2Check and its API can be available only for research purposes, by sending a request by email to the first author of the paper. Thus, the results of all of the experiments are repeatable.

3 Experimental Analysis

Considering the Sentipolc 2016 results, we can see that:

- some tools performed very well in one task and very bad in other one (e.g. team2 was the second team for subjectivity and the last one for polarity, team7 was the seventh for subjectivity and the first one for polarity, etc.);
- some other tools show a much better performance on the unconstrained run than on the constrained run (e.g. team1 shows for the subjectivity-unconstrained task a score that is 4% higher than the constrained run).

However, if the goal is to find which are overall the most complete-best performing tools, i.e. performing well considering the contribution that each tool provided on all of the tasks, an overall score/indicator is needed. To this aim, we propose the following score that takes into account, for each team, overall the best run per task. Thus, we introduce formula 1 showing that we consider, given a team and a task, the highest value of F-score between the available runs (considering also constrained and unconstrained runs). Then, in formula 2, we introduce a score per team, calculated as the summation of each contribution provided by each team for the tasks under evaluation (even a subset of them).

$$S_{team,task} = \max_{run}(F_{team,task,run}) \quad (1)$$

$$S_{team} = \sum_{task} S_{team,task} \quad (2)$$

Thanks to this score, it is possible to have an idea of overall the best available tools on: *(i)* each single task; *(ii)* a collection of tasks (couple of tasks at a time in our case), or *(iii)* all of the tasks

Please consider also that this score can be even more restrictive for our tool: we perform better on the unconstrained runs than on the constrained ones, and there are more tools for the constrained runs and performing better than our unconstrained version, so that they would gain positions in the chart (e.g. team3, team4 and team5 for the polarity task perform better on the constrained version). Moreover, we are giving the same equal weight to all of the tasks, even if we focused more on the polarity and irony task which are more related to the original App2Check approach, i.e. more useful and related the evaluation of apps reviews.

Tables 1, 2 and 3 show the results of each single task sorted by the score obtained. The columns contain (from left to right): ranking, team name, the score obtained with formula 1, and a label reporting whether the best run for the team was constrained (c) or unconstrained (u). In Tables 1 and 2 we consider the F-score value coming from the Tweet2Check amended run, representing the correct system answer. For the subjectivity task in Table 1, Tweet2Check does not show good results compared to the other tools, and there is clearly room for further improvements. For all of the other results, Tweet2Check shows good results:

- in Table 2 related to Polarity classification, it is very close to the best result, at a distance of just 0.0188, and it is the second tool considering only the results for the unconstrained run (which are directly comparable)

- in Table 3 related to Irony detection, it is the second best tool, at a distance of just 0.0068 from the first classified.

Tables 4 and 5 show the results obtained using formula 2 considering, respectively, polarity and irony together, and all of the three tasks together[1].

	Team	$\mathbf{S_{team}}$	**con/uncon**
1	team1	0.7444	u
2	team2	0.7184	c
3	team3	0.7134	c
4	team4	0.7107	c
5	team5	0.7105	c
6	team6	0.7086	c
7	team7	0.6937	c/u
8	team8	0.6495	c
9	**Tweet2Check**	**0.6317**	u
10	team10	0.5647	c
11	team11	-	-
12	team12	-	-
13	team13	-	-

Table 1: Subjectivity task at Sentipolc 2016.

In Table 4, Tweet2Check is the second best tool, at a distance of 0.0014 from team4, which is the best tool according to this score. This is clearly our best result at Sentipolc 2016, considering more tasks together, thus highlighting that polarity classification and irony detection are the best tasks performed by Tweet2Check in the current version. In Table 5, we can see that Tweet2Check is the fifth classified, at a distance of 0.0930 from team4, where we consider also the impact of the subjectivity task on the results. In this last case, Tweet2Check is in the top 5 tools chart, over 13 tools. Finally, Tables 6, 7 and 8 report the results obtained training and evaluating Tweet2Check on Evalita Sentipolc 2014 (Basile et al., 2014) datasets. The second and third columns

[1]Since some teams did not participate to all of the tasks, their results are marked as follow:
* The tool did not participate to the Irony task
** The tool participated only to the Polarity task
*** The tool participated only to the Irony task

	Team	S_{team}	con/uncon
1	team7	0.6638	c
2	team1	0.6620	u
3	team4	0.6522	c
4	team3	0.6504	c
5	team5	0.6453	c
6	**Tweet2Check**	**0.6450**	u
7	team10	0.6367	c
8	team11	0.6281	c
9	team12	0.6099	c
10	team6	0.6075	u
11	team8	0.6046	c
12	team2	0.5683	c
13	team13	-	-

Table 2: Polarity task at Sentipolc 2016.

	Team	S_{team}	con/uncon
1	team4	0.5480	c
2	**Tweet2Check**	**0.5412**	c
3	team13	0.5251	c
4	team5	0.5133	c
5	team3	0.4992	c
6	team8	0.4961	c
7	team1	0.4810	u
8	team2	-	-
9	team6	-	-
10	team7	-	-
11	team10	-	-
12	team11	-	-
13	team12	-	-

Table 3: Irony task at Sentipolc 2016.

of the these tables contain, respectively, the F-score of the constrained and the unconstrained runs (in bold the best results). We can see in Table 6 that Tweet2Check ranks first for subjectivity in the unconstrained run, and second for the constrained run. In Tables 7 and 8 Tweet2Check is the best tool for both polarity and irony. Moreover, since we think that Tweet2Check is always better on the unconstrained settings, we decided to further experimentally confirm this observation, and we trained Tweet2Check on the training set of Sentipolc 2014 with the same approach we used for the 2016 edition; thus, we tested it on the test set of the former Sentipolc 2014 evaluation. We show that, also in this case, Tweet2Check unconstrained runs perform better than the constrained

	Team	S_{team}
1	team4	1.2002
2	**Tweet2Check**	**1.1862**
3	team5	1.1586
4	team3	1.1496
5	team1	1.1430
6	team8	1.1007
7	team7*	0.6638
8	team10*	0.6367
9	team11**	0.6281
10	team12**	0.6099
11	team6*	0.6075
12	team2*	0.5683
13	team13***	0.5251

Table 4: The best performing tools on the Polarity and Irony tasks.

	Team	S_{team}
1	team4	1.9109
2	team1	1.8874
3	team5	1.8691
4	team3	1.8630
5	**Tweet2Check**	**1.8179**
6	team8	1.7502
7	team7*	1.3575
8	team6*	1.3161
9	team2*	1.2867
10	team10*	1.2014
11	team11**	0.6281
12	team12**	0.6099
13	team13***	0.5251

Table 5: The best performing tools on the three tasks.

ones, and that our tool is the best tool compared to the tools that participated in 2014.

4 Conclusion

In this paper we presented Tweet2Check and discussed the analysis of the results from Sentipolc 2016, showing that our tool is: *(i)* the second classified for the irony task, at a distance of just 0.0068 from the first classified; *(ii)* the second classified for the polarity task, considering the unconstrained runs, at a distance of 0.017 from the first tool; *(iii)* in the top 5 tools (out of 13), considering a score that allows to indicate the *most complete-best performing* tools for Sentiment Analysis of tweets, i.e. by summing up the best F-score of

Team	F(C)	F(U)
uniba2930	**0.7140**	0.6892
Tweet2Check	0.6927	**0.6903**
UNITOR	0.6871	0.6897
IRADABE	0.6706	0.6464
UPFtaln	0.6497	-
ficlit+cs@unibo	0.5972	-
mind	0.5901	-
SVMSLU	0.5825	-
fbkshelldkm	0.5593	-
itagetaruns	0.5224	-

Table 6: Tweet2Check ranking on the Sentipolc 2014 subjectivity task.

Team	F(C)	F(U)
Tweet2Check	**0.7048**	**0.7142**
uniba2930	0.6771	0.6638
IRADABE	0.6347	0.6108
CoLingLab	0.6312	-
UNITOR	0.6299	0.6546
UPFtaln	0.6049	-
SVMSLU	0.6026	-
ficlit+cs@unibo	0.5980	-
fbkshelldkm	0.5626	-
mind	0.5342	-
itagetaruns	0.5181	-
Itanlp-wafi*	0.5086	-
*amended run	0.6637	-

Table 7: Tweet2Check ranking on the Sentipolc 2014 polarity task.

each team for the three tasks (subjectivity, polarity and irony); *(iv)* the second best tool, according to the former score, considering together polarity and irony tasks.

Team	F(C)	F(U)
Tweet2Check	**0.5915**	-
UNITOR	0.5759	**0.5959**
IRADABE	0.5415	0.5513
SVMSLU	0.5394	-
itagetaruns	0.4929	-
mind	0.4771	-
fbkshelldkm	0.4707	-
UPFtaln	0.4687	-

Table 8: Tweet2Check ranking on the Sentipolc 2014 irony task.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503–512.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Frencesco Barbieri and Valerio Basile and Danilo Croce and Malvina Nissim and Nicole Novielli and Viviana Patti. 2016. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro and Rachele Sprugnoli, editors, *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)*. Associazione Italiana di Linguistica Computazionale (AILC)

Emanuele Di Rosa and Alberto Durante LREC 2016 2016. *App2Check: a Machine Learning-based system for Sentiment Analysis of App Reviews in Italian Language* in Proc. of the 2nd International Workshop on Social Media World Sensors, pp. 8-11. http://ceur-ws.org/Vol-1696/

Emanuele Di Rosa, Alberto Durante. *App2Check extension for Sentiment Analysis of Amazon Products Reviews*. In Semantic Web Challenges Vol. 641-1, CCIS Springer 2016

Diego Reforgiato. Results of the Semantic Sentiment Analysis 2016 International Challenge https://github.com/diegoref/SSA2016

ESWC 2016 Challenges http://2016.eswc-conferences.org/program/eswc-challenges

Harald Sack, Stefan Dietze, Anna Tordai. *Semantic Web Challenges*. 2016. CCIS Springer 2016. Third SemWebEval Challenge at ESWC 2016.

Valerio Basile and Andrea Bolioli and Malvina Nissim and Viviana Patti and Paolo Rosso. *Overview of the Evalita 2014 SENTIment POLarity Classification Task*. 2014.

Filipe N. Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, Fabrício Benevenuto. *SentiBench - a benchmark comparison*

of state-of-the-practice sentiment analysis methods
- In EPJ Data Science 2016. 2014.