# Evaluating Industrial and Research Sentiment Analysis Engines on Multiple Sources

Emanuele Di Rosa, Head of AI – Alberto Durante, Research Scientist

Department of Artificial Intelligence Solutions

at Finsa s.p.a. (Genova, Italy)

## Motivation & context

- Building polarity classification engines showing the same high accuracy on different topics domains and on structurally different textual sources (e.g. reviews, tweets, blogs, etc.) is a goal not yet fully accomplished in both research and industry.
- There is a high variation while comparing accuracy reached by the same engine on different kinds of textual sources and domains.

The main factors are:

- Sentiment is very often domain-dependent (*addictive* drug vs *addictive* game)
- Textual sources have different characteristics: length, level of formality, respect of grammar, slang words, irony, objectivity vs subjectivity, etc.

Two different perspectives:

- **Research is focused on specific settings**; international challenges search for the best approach to work on specific tasks, sources (e.g. tweets, product reviews, or blogs), languages or topic domains (e.g. politics).
- **Commercial engines** have to be built to show an overall **high average accuracy on heterogeneous sources and cross-domains**.

## Paper contribution

The paper innovative contribution lies in:

1. An overall experimental comparison between **industrial engines** (Google CNL and X2Check) and state-of-the-art **research tools** that competed in the last international competitions.
2. For each source under evaluation (tweets, apps reviews and general product reviews), the quantification of the performance gap between industrial and research tools when they are **specifically trained on the target source** versus their **not specialized version on the target source**.
3. A multi-language evaluation in which test sets are considered in both English and Italian.
4. Two new testsets for document-level polarity detection, each containing 10 thousand apps reviews in Italian and English, made available to the research community.
5. The results of our X2Check approach in which, given the input source, a metamodel estimates which classification model is better to apply.

## Overall results

- Comparing **industrial engines**, X2Check shows an overall average cross-source F-score that is:
  - 9.1% higher than Google CNL on Italian benchmarks
  - 5.1% higher than Google CNL on English benchmarks
- Comparing **industrial engines with the best research engine** specialized on the target source under evaluation (on 2,000 tweets in Italian and 12,000 tweets in English):
  - X2Check is lower than 3.4% on Italian and 11.6% on English benchmarks
  - Google CNL is lower than 13.5% on Italian and 16.3% on English benchmarks

Moreover, X2Check shows:

- always a higher score than research tools not specialized on the test set under evaluation
- a better performance than Google CNL on 5 out of 6 benchmarks
- a performance always very close to its specialized versions (App2Check, Tweet2Check, Amazon2Check) made to deal with the specific benchmark under evaluation.

## Experimental Analysis

Comparison on 2,000 tweets in Italian from Evalita Sentipolc 2016

| | System | Const/unc | Pos | Neg | F |
|---|---|---|---|---|---|
| 1 | SwissCheese | c | 0.6529 | **0.7128** | **0.6828** |
| 2 | UniPI | c | **0.6850** | 0.6426 | 0.6638 |
| 3 | Unitor | u | 0.6354 | 0.6885 | 0.662 |
| 4 | Tweet2Check | u | 0.6696 | 0.6442 | 0.6569 |
| 5 | ItaliaNLP | c | 0.6265 | 0.6743 | 0.6504 |
| 6 | X2Check | u | 0.6629 | 0.6442 | 0.6491 |
| 7 | IRADABE | c | 0.6426 | 0.648 | 0.6453 |
| 8 | UniBO | c | 0.6708 | 0.6026 | 0.6367 |
| 9 | IntIntUniba | c | 0.6189 | 0.6372 | 0.6281 |
| 10 | CoLingLab | c | 0.5619 | 0.6579 | 0.6099 |
| 11 | INGEOTEC | u | 0.5944 | 0.6205 | 0.6075 |
| 12 | ADAPT | c | 0.5632 | 0.6461 | 0.6046 |
| 13 | App2Check | u | 0.5466 | 0.6250 | 0.5857 |
| 14 | samskara | c | 0.5198 | 0.6168 | 0.5683 |
| 15 | Google CNL_05-17 | u | 0.5426 | 0.5530 | 0.5478 |

Comparison on 12,000 tweets in English from SemEval-2017 Task 4, subtask A

| | System | AvgR | AvgF1-PN | Acc |
|---|---|---|---|---|
| 1 | DataStories | **0.681** | 0.677 | 0.651 |
| | BB_twtr | **0.681** | **0.685** | 0.658 |
| 3 | LIA | 0.676 | 0.674 | 0.661 |
| ... | ... | ... | ... | ... |
| 30 | Tweet2Check | 0.566 | 0.565 | 0.526 |
| 31 | X2Check | 0.563 | 0.561 | 0.523 |
| 32 | XJSA | 0.556 | 0.519 | 0.575 |
| 33 | Neverland-THU | 0.555 | 0.507 | 0.597 |
| 34 | MI&T-Lab | 0.551 | 0.522 | 0.561 |
| 35 | Google CNL_06-2017 | 0.550 | 0.514 | 0.567 |
| 36 | diegoref | 0.546 | 0.527 | 0.540 |
| 37 | App2Check | 0.541 | 0.508 | 0.545 |
| ... | ... | ... | ... | ... |
| 41 | WarwickDCS | 0.335 | 0.221 | 0.382 |
| | Avid | 0.335 | 0.163 | 0.206 |

Comparison on 1,000 tweets in English with no neutral class from [2]

| | Tool | Acc | MF1 | F1(-) | F1(+) |
|---|---|---|---|---|---|
| 1 | X2Check | **0.867** | **0.873** | **0.869** | **0.877** |
| 2 | Tweet2Check | 0.858 | 0.863 | 0.860 | 0.865 |
| 3 | AFINN | 0.709 | 0.772 | 0.796 | 0.748 |
| 4 | Sentistrength | 0.668 | 0.751 | 0.763 | 0.740 |
| 5 | App2Check | 0.692 | 0.716 | 0.730 | 0.702 |
| 6 | Senticnet | 0.660 | 0.684 | 0.642 | 0.725 |
| 7 | Umigon | 0.588 | 0.680 | 0.692 | 0.667 |
| 8 | Vader | 0.542 | 0.671 | 0.653 | 0.690 |
| 9 | Google CNL_05-17 | 0.524 | 0.666 | 0.604 | 0.728 |
| 10 | Sentiment140 | 0.644 | 0.649 | 0.706 | 0.592 |
| 11 | SentiWordNet | 0.614 | 0.637 | 0.616 | 0.658 |
| 12 | Op. Lexicon | 0.508 | 0.625 | 0.653 | 0.598 |
| 13 | SOCAL | 0.509 | 0.621 | 0.655 | 0.588 |
| 14 | NRC Hashtag | 0.596 | 0.591 | 0.678 | 0.505 |
| ... | ... | ... | ... | ... | ... |

Comparison on 10,000 apps reviews in Italian wrt app rating

| | Tool | Acc | MF1 | F1(-) | F1(x) | F1(+) |
|---|---|---|---|---|---|---|
| 1 | App2Check | **0.857** | **0.733** | **0.827** | **0.456** | **0.917** |
| 2 | X2Check | 0.843 | 0.714 | 0.814 | 0.420 | 0.908 |
| 3 | Google CNL_05-17 | 0.791 | 0.634 | 0.799 | 0.217 | 0.888 |
| 4 | SentiWordNet | 0.659 | 0.479 | 0.604 | 0.062 | 0.771 |
| 5 | AFINN | 0.603 | 0.475 | 0.492 | 0.166 | 0.767 |
| 6 | SentiStrength | 0.597 | 0.475 | 0.463 | 0.193 | 0.768 |
| 7 | Stanford DL | 0.540 | 0.456 | 0.565 | 0.135 | 0.668 |
| 8 | Op. Lexicon | 0.553 | 0.449 | 0.451 | 0.175 | 0.722 |
| 9 | Sentiment140 | 0.587 | 0.441 | 0.574 | 0.067 | 0.682 |
| 10 | Umigon | 0.501 | 0.428 | 0.478 | 0.146 | 0.662 |
| 11 | SO-CAL | 0.493 | 0.418 | 0.458 | 0.138 | 0.656 |
| 12 | NRC Hashtag | 0.529 | 0.412 | 0.536 | 0.083 | 0.617 |
| 13 | Senticnet | 0.631 | 0.409 | 0.366 | 0.091 | 0.769 |
| 14 | Vader | 0.462 | 0.385 | 0.295 | 0.197 | 0.663 |
| 15 | Emolex | 0.455 | 0.383 | 0.385 | 0.141 | 0.623 |
| 16 | SASA | 0.488 | 0.377 | 0.296 | 0.164 | 0.671 |
| 17 | SentiStr. Ita | 0.396 | 0.340 | 0.320 | 0.139 | 0.562 |
| ... | ... | ... | ... | ... | ... | ... |

Comparison on 10,000 apps reviews in English wrt app rating

| | Tool | Acc | MF1 | F1(-) | F1(x) | F1(+) |
|---|---|---|---|---|---|---|
| 1 | Google CNL_05-17 | **0.662** | **0.629** | **0.717** | **0.376** | **0.793** |
| 2 | X2Check | 0.633 | 0.539 | 0.698 | 0.193 | 0.726 |
| 3 | App2Check | 0.631 | 0.524 | 0.708 | 0.149 | 0.714 |
| 4 | Umigon | 0.511 | 0.488 | 0.527 | 0.294 | 0.644 |
| 5 | Stanford DL | 0.514 | 0.475 | 0.607 | 0.242 | 0.574 |
| 6 | AFINN | 0.499 | 0.455 | 0.460 | 0.273 | 0.634 |
| 7 | Op. Lexicon | 0.481 | 0.450 | 0.434 | 0.285 | 0.632 |
| 8 | SentiStrength | 0.475 | 0.450 | 0.429 | 0.301 | 0.620 |
| 9 | SO-CAL | 0.447 | 0.424 | 0.465 | 0.250 | 0.557 |
| 10 | Sentiment140 | 0.537 | 0.420 | 0.643 | 0.092 | 0.525 |
| 11 | NRC Hashtag | 0.518 | 0.405 | 0.634 | 0.092 | 0.489 |
| 12 | Vader | 0.421 | 0.405 | 0.268 | 0.343 | 0.604 |
| 13 | SentiWordNet | 0.508 | 0.403 | 0.529 | 0.080 | 0.600 |
| 14 | Emolex | 0.407 | 0.393 | 0.379 | 0.276 | 0.523 |
| 15 | SASA | 0.402 | 0.380 | 0.348 | 0.264 | 0.527 |
| 16 | H. Index | 0.377 | 0.344 | 0.222 | 0.283 | 0.527 |
| 17 | Senticnet | 0.441 | 0.335 | 0.327 | 0.107 | 0.571 |
| ... | ... | ... | ... | ... | ... | ... |

Comparison on 200,000 Amazon Products Reviews in English

| | Tool | M-F1 | Acc | F1(-) | F1(+) |
|---|---|---|---|---|---|
| 1 | Amazon2Check | **0.865** | **0.864** | **0.869** | **0.860** |
| 2 | X2Check | 0.862 | 0.862 | 0.868 | 0.856 |
| 3 | Google CNL_05-17 | 0.821 | 0.827 | 0.853 | 0.790 |
| 4 | App2Check | 0.729 | 0.736 | 0.772 | 0.685 |
| 5 | SentiStrength | 0.630 | 0.552 | 0.568 | 0.692 |
| 6 | StanfordDL | 0.602 | 0.604 | 0.705 | 0.498 |

## References

[1] Di Rosa E., Durante A. LREC 2016 *App2Check: a Machine Learning-based system for Sentiment Analysis of App Reviews in Italian Language* in Proc. of the 2nd Int. Workshop on Social Media World Sensors, Vol. 1696, pp. 8-13, 2016.

[2] Araujo,M., dos Reis, J.C., Pereira, A.M., Benevenuto, F. ACM SAC 2016 *An evaluation of machine translation for multilingual sentence-level sentiment analysis* in Proc. of ACM SAC 2016, pp. 1140-1145, 2016.

AI*IA 2017, Bari 14-17 November 2017