



finsa

TECHNOLOGY FOR PEOPLE

Comparing Sentiment Engine Performance on Reviews and Tweets

Emanuele Di Rosa, PhD
CSO, Head of Artificial Intelligence

Finsa s.p.a.

emanuele.dirosa@finsa.it

www.app2check.com

www.finsa.it



Motivations and Goals



Motivations

- Computing *accurately* a sentiment expressed in a text is a task largely needed in the market, and ready-to-use APIs with pre-trained sentiment classifiers are available.
- However, sentiment engines asked to classify a text as positive, negative or neutral, do not reach a 100% of accuracy. They show **misclassifications** in multiple cases, even in cases that are straightforward for humans: this involves both research and industrial tools.



Motivations

- Computing *accurately* a sentiment expressed in a text is a task largely needed in the market, and ready-to-use APIs with pre-trained sentiment classifiers are available.
- However, sentiment engines asked to classify a text as positive, negative or neutral, do not reach a 100% of accuracy. They show **misclassifications** in multiple cases, even in cases that are straightforward for humans: this involves both research and industrial tools.
- On the one hand, *academic research* advances are visible and international challenges are organized each year, asking researchers to train/fine-tune their engines to *work well on specific tasks* (e.g. polarity classification, subjectivity or irony detection), *on specific sources/domains* (e.g tweets about politics), and *specific languages* (English, Italian, Arabic, etc.)



Motivations

- Computing *accurately* a sentiment expressed in a text is a task largely needed in the market, and ready-to-use APIs with pre-trained sentiment classifiers are available.
- However, sentiment engines asked to classify a text as positive, negative or neutral, do not reach a 100% of accuracy. They show **misclassifications** in multiple cases, even in cases that are straightforward for humans: this involves both research and industrial tools.
- On the one hand, *academic research* advances are visible and international challenges are organized each year, asking researchers to train/fine-tune their engines to *work well on specific tasks* (e.g. polarity classification, subjectivity or irony detection), *on specific sources/domains* (e.g tweets about politics), and *specific languages* (English, Italian, Arabic, etc.)
- On the other hand, tools needed by industry have to face the need of the market, asking for engines that can receive as input *any textual source* (tweets, reviews, etc) and being *applied to general purpose applications*.



Motivations

- Computing *accurately* a sentiment expressed in a text is a task largely needed in the market, and ready-to-use APIs with pre-trained sentiment classifiers are available.
- However, sentiment engines asked to classify a text as positive, negative or neutral, do not reach a 100% of accuracy. They show **misclassifications** in multiple cases, even in cases that are not handled by current industrial tools.
- On the other hand, tools needed by industry have to face the need of the market, asking for engines that can receive as input *any textual source* (tweets, reviews, etc) and being *applied to general purpose applications*.

We *ideally* need industrial sentiment engines providing high average performance on multiple sources and domains

Industrial

challenges
to *work well*
on specific
Italian,



Goals

1. Sentiment engine performance: *Perceived by humans* VS *experimentally measured*
2. What's the **performance gap** between industrial “general purpose” engines and research **engines**, since the latter are built to show high performance on *specific settings* (source, domain, language, task, etc)? Are there differences in performance analyzing tweets or reviews **in different languages** (e.g. English and Italian)?

Outline

1. Motivations and goals

2. Sentiment Engine (*mis*)classifications

- On *simple* cases
- On *difficult* cases: «*Cross-domain*» Sentiment Classification

3. Experimental Evaluation of Research and Industrial Engines

- Results on Tweets
- Results on Product Reviews

4. Conclusions

Sentiment Engine (mis)classifications



Sentiment Engines on simple classifications

We consider some industrial and research sentiment engines providing an online demo:

- Research engines:
 - [iFeel Platform](#) (running 18 research tools implementing different methods)
 - [Stanford Deep Learning](#)
- Industrial tools:
 - [IBM Watson](#)
 - [Google Cloud Natural Language API \(Google CNL\)](#)
 - [Finsa X2Check](#)

We test 3 simple sentences with «clear» sentiment classification:

- A negative sentence
- A positive sentence
- A negative («difficult») sentence



Engines on simple classifications: iFeel Platform

Methods Results

“I hate this game”

Your input: I hate this game

Method Name	Status	Method Score	Polarity
OPINIONLEXICON	Completed	-1	Negative
SENTISTRENGTH	Completed	-0.75	Negative
SOCAL	Completed	-6	Negative
HAPPINESSINDEX	Completed	-0.11249999999999982	Negative
SANN	Completed	1	Positive
EMOTICONS	Completed	1	Positive
SENTIMENT140	Completed	-9.882	Negative
STANFORD	Completed	-1	Negative
AFINN	Completed	-3	Negative
MPQA	Completed	-1	Negative
NRCHASHTAG	Completed	-15.064999999999998	Negative
EMOLEX	Completed	-1	Negative
EMOTICONS	Completed	0	Neutral
PANAST	Completed	0	Neutral
SASA	Completed	1	Positive
SENTIWORDNET	Completed	-0.7575258926544899	Negative
VADER	Completed	-0.5719	Negative
UMIGON	Completed	-1	Negative

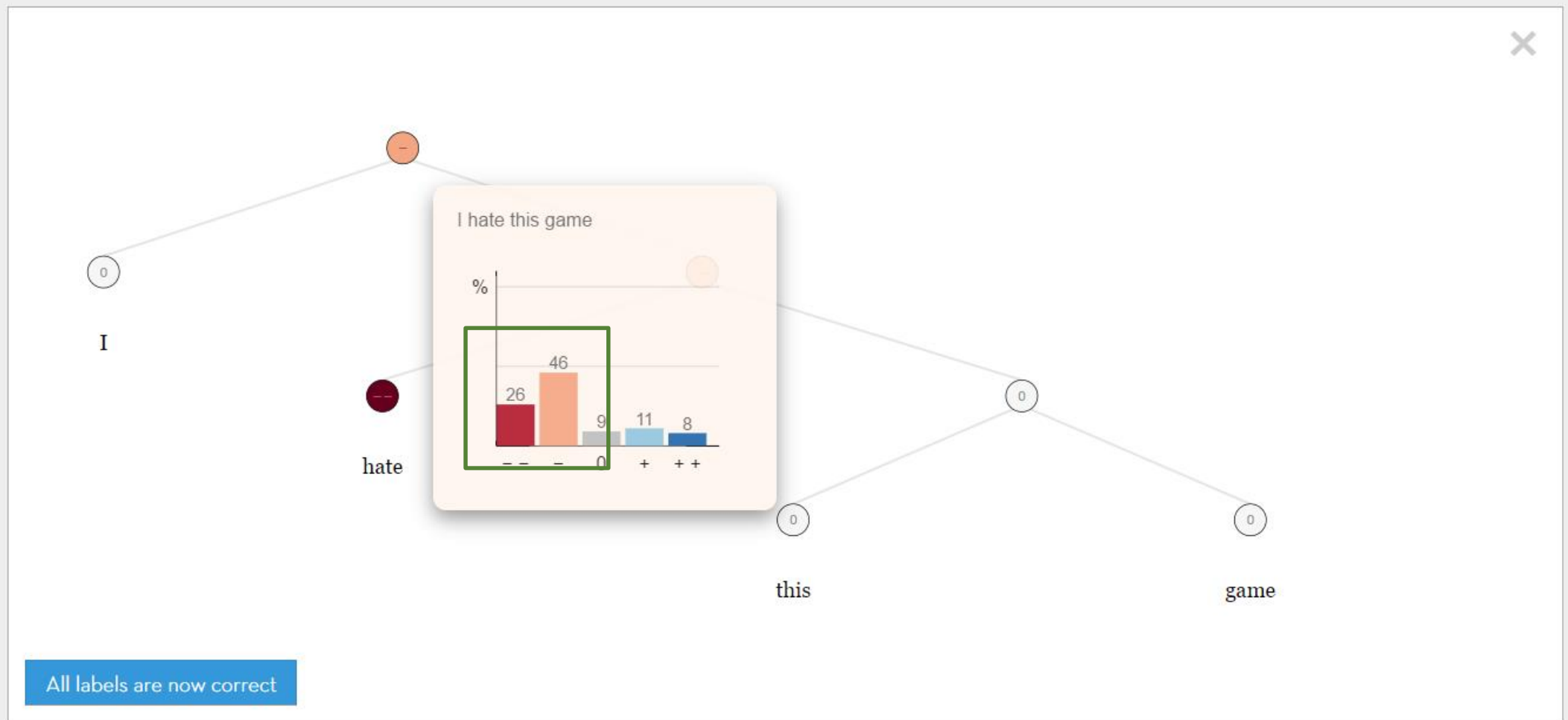


Engines on simple classifications: StanfordDL

“I hate this game”

Sentiment Trees

You can double-click on each tree figure to see its expanded version with greater details. There are 5 classes of sentiment classification: **very negative**, **negative**, neutral, **positive**, and very positive.





Engines on simple classifications: IBM Watson

“I hate this game”

Text	URL
i <u>hate</u> this game	

For results unique to your business needs consider building a [custom model](#).

English

Analyze

Sentiment

Emotion

Keywords

Entities

Categories

Concept

Semantic Roles

Review the overall sentiment and targeted sentiment of the content.

[JSON](#) ^

```
{
  "sentiment": {
    "document": {
      "score": -0.872248,
      "label": "negative"
    }
  }
}
```

Overall Sentiment

Negative

 -0.87



Engines on simple classifications: iFeel Platform

Methods Results

“I like this game”

Your input: I like this game

Method Name	Status	Method Score	Polarity
OPINIONLEXICON	Completed	1	Positive
SENTISTRENGTH	Completed	0.25	Positive
SOCAL	Completed	1	Positive
HAPPINESSINDEX	Completed	0.4950000000000001	Positive
SANN	Completed	1	Positive
EMOTICONS	Completed	1	Positive
SENTIMENT140	Completed	-1.115	Negative
STANFORD	Completed	0	Neutral
AFINN	Completed	2	Positive
MPQA	Completed	1	Positive
NRCHASHTAG	Completed	-2.657	Negative
EMOLEX	Completed	0	Neutral
EMOTICONS	Completed	0	Neutral
PANAST	Completed	0	Neutral
SASA	Completed	1	Positive
SENTIWORDNET	Completed	0.3729485599002547	Positive
VADER	Completed	0	Neutral
UMIGON	Completed	1	Positive



Engines on simple classifications: StanfordDL

“I like this game”

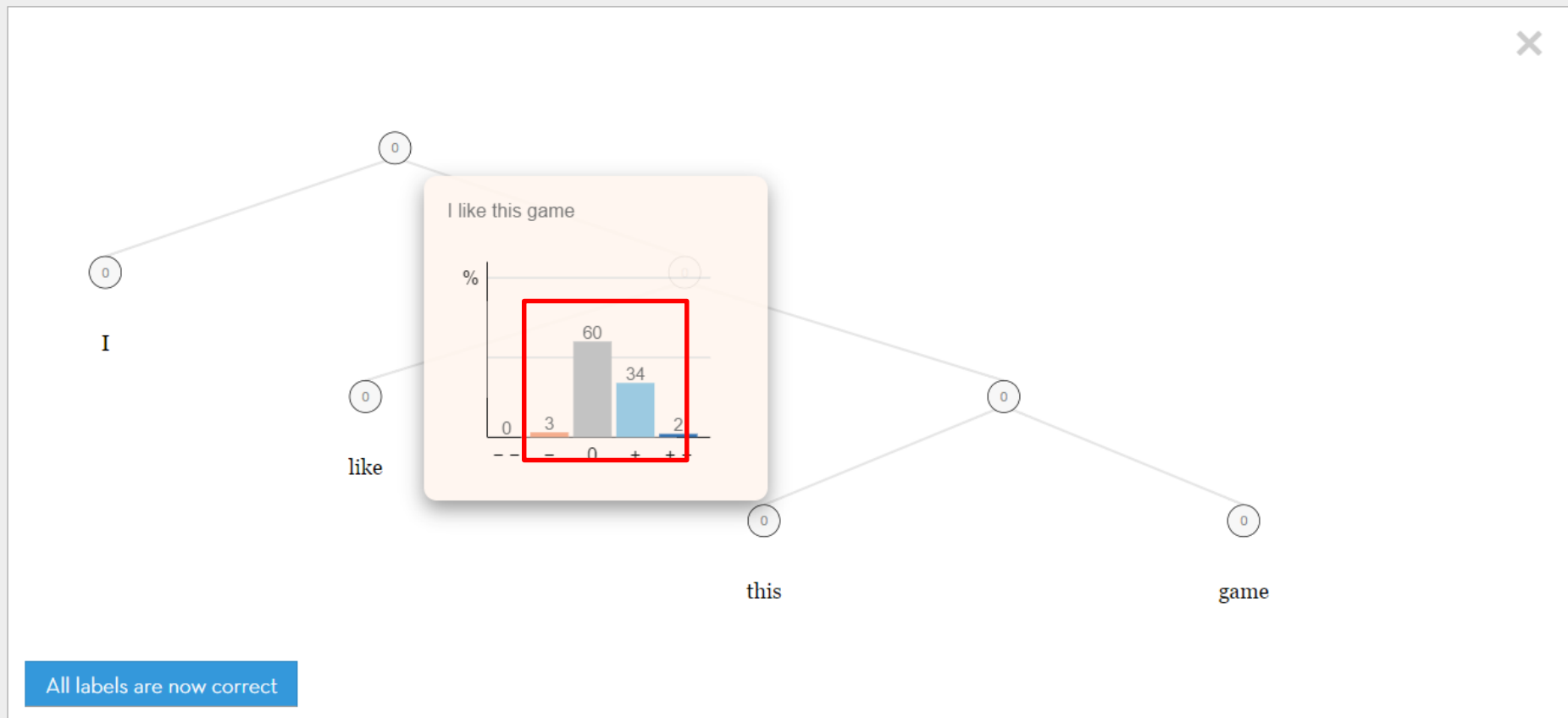


Sentiment Analysis

[Information](#)[Live Demo](#)[Sentiment Treebank](#)[Help the Model](#)[Source Code](#)

Sentiment Trees

You can double-click on each tree figure to see its expanded version with greater details. There are 5 classes of sentiment classification: **very negative**, **negative**, **neutral**, **positive**, and **very positive**.





Engines on simple classifications: IBM Watson

“I like this game”

Text

URL

i like this game

For results unique to your business needs consider building a [custom model](#).

English

Analyze

Sentiment

Emotion

Keywords

Entities

Categories

Concept

Semantic Roles

Review the overall [sentiment](#) and targeted sentiment of the content.

[JSON](#) ^

```
{
  "sentiment": {
    "document": {
      "score": 0,
      "label": "neutral"
    }
  }
}
```

Overall Sentiment

Positive 0.00



Engines on simple classifications: iFeel Platform

“I just connected my game with my facebook account and instead of saving the progress I have lost all my progress and it came on Level 1 although I was on lvl 98 Please help!!!!”

Method Name	Status	Method Score	Polarity
OPINIONLEXICON	Completed	1.6666666666666667	Positive
SENTISTRENGTH	Completed	0.25	Positive
SOCAL	Completed	0.8	Positive
HAPPINESSINDEX	Completed	0.3287500000000001	Positive
SANN	Completed	0	Neutral
EMOTICONS	Completed	1	Positive
SENTIMENT140	Completed	-350.6739999999999	Negative
STANFORD	Completed	-1	Negative
AFINN	Completed	0.8	Positive
MPQA	Completed	1	Positive
NRCHASHTAG	Completed	-152.72899999999993	Negative
EMOLEX	Completed	1	Positive
EMOTICONS	Completed	0	Neutral
PANAST	Completed	0	Neutral
SASA	Completed	1	Positive
SENTIWORDNET	Completed	0.16028867864857324	Positive
VADER	Completed	0.7762	Positive
UMIGON	Completed	-1	Negative



Engines on simple classifications: iFeel Platform

“I just connected my game with my facebook account and instead of saving the progress I have lost all my progress and it came on Level 1 although I was on lvl 98 Please help!!!!”

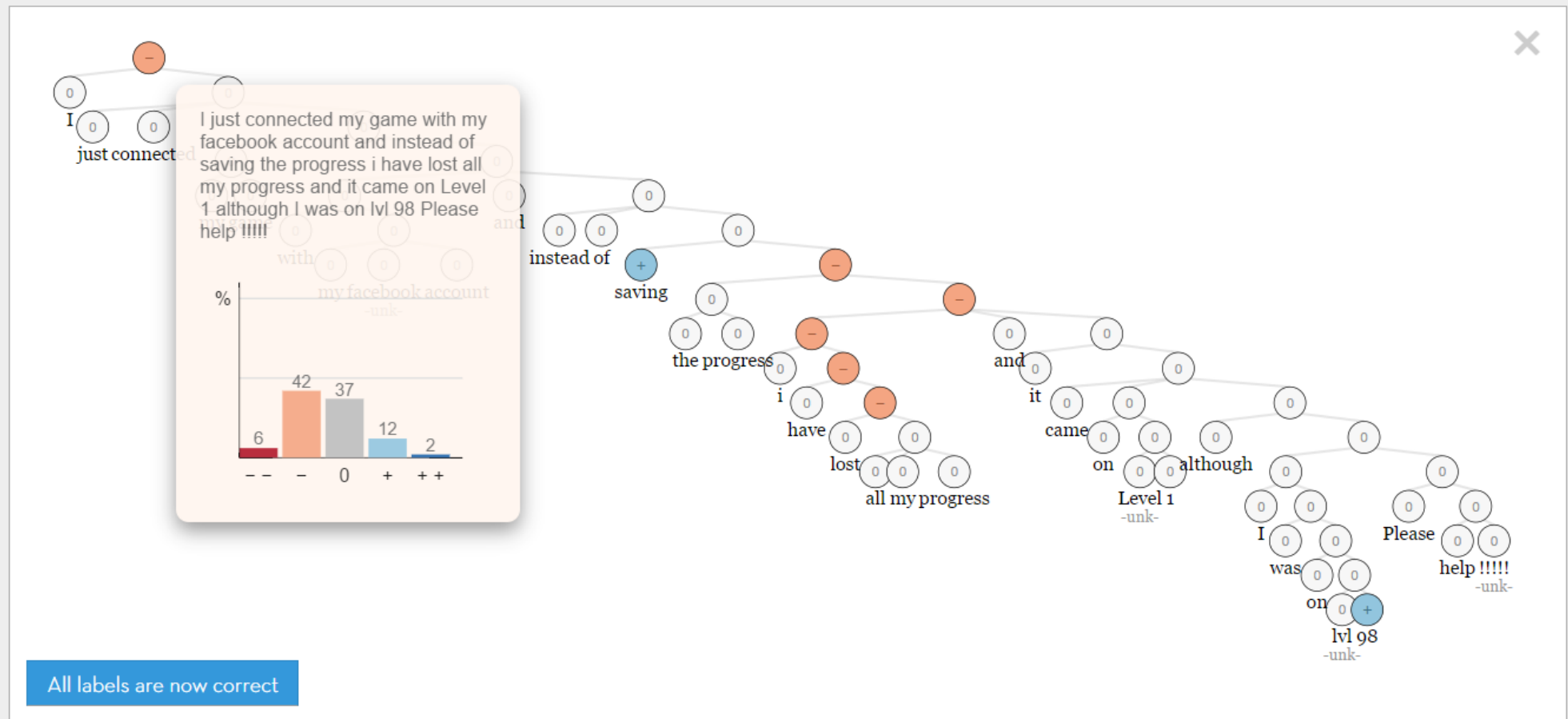
Method Name	Status	Method Score	Polarity
OPINIONLEXICON	Completed	1.666666666666667	Positive
SENTISTRENGTH	Completed	0.25	Positive
SOCAL	Completed	0.8	Positive
HAPPINESSINDEX	Completed	0.3287500000000001	Positive
SANN	Completed	0	Neutral
EMOTICONS	Completed	1	Positive
SENTIMENT140	Completed	-350.6739999999999	Negative
STANFORD	Completed	-1	Negative
AFINN	Completed	0.8	Positive
MPQA	Completed	1	Positive
NRCHASHTAG	Completed	-152.72899999999993	Negative
EMOLEX	Completed	1	Positive
EMOTICONS	Completed	0	Neutral
PANAST	Completed	0	Neutral
SASA	Completed	1	Positive
SENTIWORDNET	Completed	0.16028867864857324	Positive
VADER	Completed	0.7762	Positive
UMIGON	Completed	-1	Negative



Engines on simple classifications: Stanford DL

Sentiment Trees

You can double-click on each tree figure to see its expanded version with greater details. There are 5 classes of sentiment classification: **very negative**, **negative**, neutral, **positive**, and **very positive**.



[Download Results](#)



Engines on simple classifications: IBM Watson

Text

URL

I just connected my game with my facebook account and instead of saving the progress i have lost all my progress and it came on Level 1 although I was on lv 98 Please help!!!!

For results unique to your business needs consider building a custom model. English

Analyze

Sentiment

Emotion

Keywords

Entities

Categories

Concept

Semantic Roles

Review the overall sentiment and targeted sentiment of the content. JSON ^

```
{
  "sentiment": {
    "document": {
      "score": 0.0,
      "label": "neutral"
    }
  }
}
```



Observations

- We showed objective examples of sentiment *misclassification* performed by popular research and industrial engines, even on cases that are *straightforward for humans*
- However, *it is not possible to make any kind of generalization of these results* or let us somehow rank the engines involved in the previous examples. In order to do that, *a wide experimental analysis is needed*.
- Performance in sentiment polarity classification depends on many factors, involving the classifier's training (*source*) set and test (*target*) set. Some sentiment classifiers are built to perform better on a specific:
 - Topic domain (e.g. movies, politics)
 - Textual source (tweets, reviews, etc.)
 - Language
 - ...



Cross-domain classification and domain-adaptation

- How do we classify the polarity of the following text?

"Candy crush is my **addiction**, I love it!"

This is a case of domain-dependent sentiment. Moreover, It is well known in literature that:

- *Users often use some different words when they express sentiment in different domains* [Pan S.J., et al 2010]
- *Classifiers trained on one domain may perform poorly on another domain* [Pang, et al. 2008].
 - Cross-domain sentiment analysis research area works on *domain-adaptation techniques* [Blitzer, et al 2007], [Pan S.J., et al 2010], [Liu B., 2012], [Wu F., et al, 2016], [Wu F., et al, 2017].
 - Sometimes domain-adaptation may also lead to worse performance [Pan, S.J., et al 2010].



Document-level VS Sentence-level VS Entity level SA

"I like this game but after the iOS update I get a crash when the app starts. Please do something!! "

- It is probably impossible to agree about its overall overall (document-level) sentiment classification
- It is known in literature [1] that group of humans, when evaluating sentiment (the polarity in three classes), agree in about the 80% of the cases since there can be controversial cases due to the subjective qualitative evaluation.

[1] T. Wilson, J. Wiebe, P. Hoffmann. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In proc. of HLT 2005.



Document-level VS Sentence-level VS Entity level SA

"I like this game but after the iOS update I get a crash when the app starts. Please do something!!"

Try the API ✕

I like this game but after the iOS update I get a crash when the app starts. Please do something!! ANALYZE

[See supported languages](#)

Entities Sentiment Syntax

Document & Sentence Level Sentiment

Entire Document

	Score	Magnitude
I like this game but after the iOS update I get a crash when the app starts.	-0.1	0.2
Please do something!!	-0.2	0.2
	0	0

Score Range: -1.0 — -0.25 (red) -0.25 — 0.25 (yellow) 0.25 — 1.0 (green)

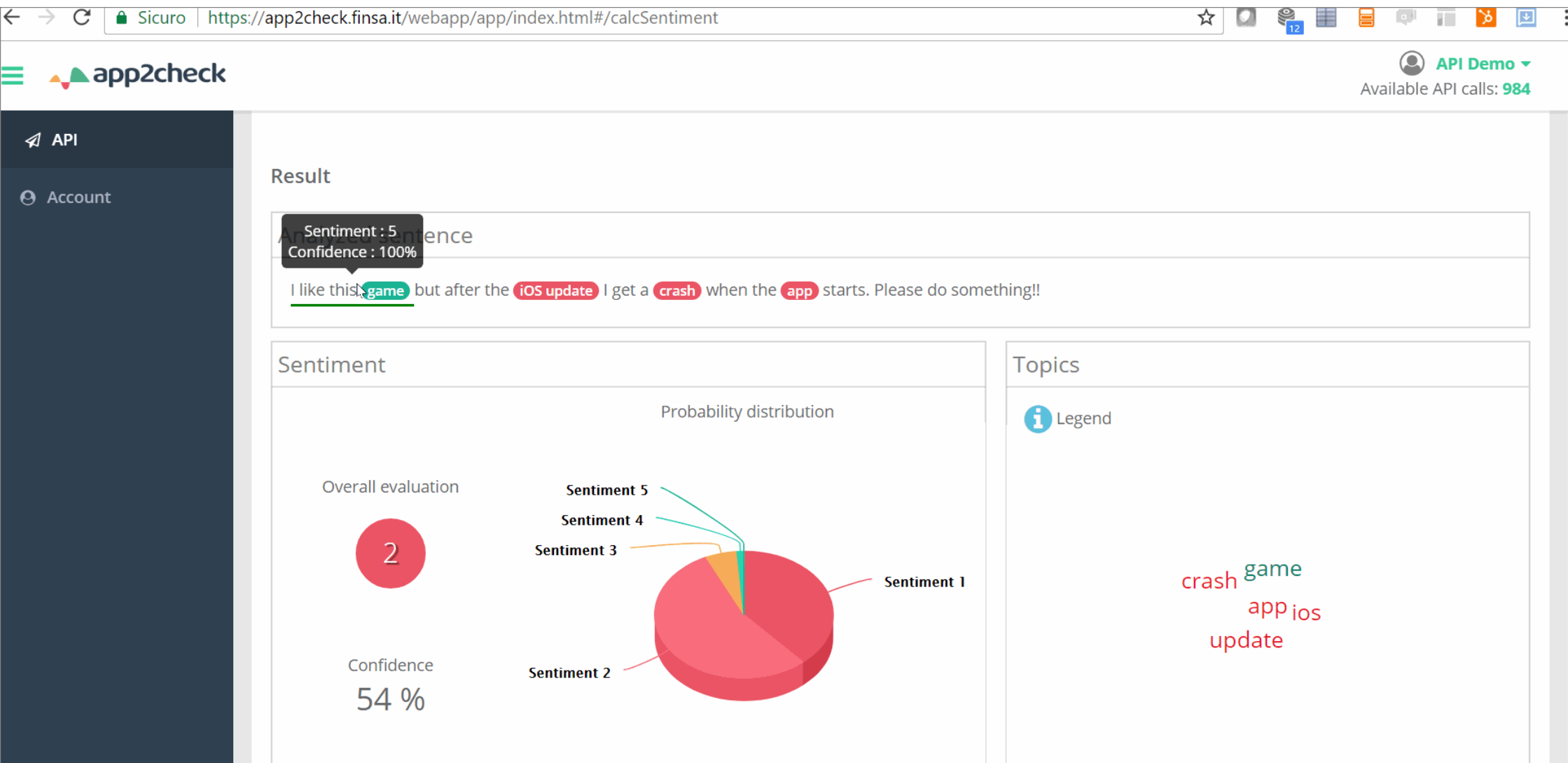
Entity Level Sentiment

1. game Sentiment: Score 0 Magnitude 0	WORK OF ART	2. crash Sentiment: Score -0.1 Magnitude 0.1	EVENT
3. update Sentiment: Score -0.1 Magnitude 0.1	OTHER	4. iOS Sentiment: Score 0 Magnitude 0	CONSUMER GOOD
5. something Sentiment: Score 0 Magnitude 0	OTHER	6. app Sentiment: Score -0.3 Magnitude 0.3	CONSUMER GOOD



Document-level VS Sentence-level VS Entity level SA

"I like this game but after the iOS update I get a crash when the app starts. Please do something!!"





Experimental Evaluation of Research and Industrial Engines



Experimental Evaluation

- In order to fairly compare engines performance, we need:
 - a gold standard reference
 - benchmarks on multiple sources and mixed domains
 - benchmarks in more than one language
- Tweets → we see a **worst case for *industrial* engines**
 - Benchmarks and engines from Evalita SentiPolC 2016 for Italian language
 - Benchmarks and engines from SemEval 2017 for English language
- Reviews → we see a **worst case for *research* engines**
 - Amazon Product Reviews: Benchmarks from ESWC Semantic Sentiment Analysis 2016



Experimental Evaluation

- About pre-trained, ready-to-use industrial Sentiment APIs: most of the commercial engines for SA, in terms of service, **do not allow to use their APIs to perform an experimental comparative analysis**.
- The goal of such tools is to measure user opinion and, as per every measurement tool, being aware of its accuracy is fundamental.
- This is even more important in sentiment analysis since, as we recalled, **pre-trained engines may in general show a significant different performance depending on the target test set**.
- We considered industrial engines, having a public sentiment API and ***without explicit restrictions in the terms of service to make a comparative analysis***

General purpose APIs:

- ✓ *Google CNL*
- ✓ *Finsa X2Check*

X2Check adaptations, specifically trained on the target source:

- ✓ App2Check specifically trained on apps reviews.
- ✓ Tweet2Check specifically trained on tweets.
- ✓ Amazon2Check is specifically trained on amazon reviews.

Evaluation on Tweets in Italian

	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_05-2017	u	0.5426	0.5530	0.5478
16	<i>Baseline</i>		<i>0.4518</i>	<i>0.3808</i>	<i>0.4163</i>

Tab 1: Evaluation on 2K tweets in Italian from Evalita SentiPolC 2016. Industrial engines added to the official results. Industrial engines VS research engines *specifically trained/tuned* on the given domain/source.

Evaluation on Tweets in Italian

	System	Const/unc	Pos	Neg	F
1	SwissCheese	c	0.6529	0.7128	0.6828
2	UwDI	c	0.6850	0.6426	0.6638
3					0.662
4					.6569
5					.6504
6					.6491
7					.6453
8					.6367
9					.6281
10					.6099
11					.6075
12					.6046
13					.5857
14					.5683
15	Google CNL_05-2017	u	0.5426	0.5530	0.5478
16	<i>Baseline</i>		<i>0.4518</i>	<i>0.3808</i>	<i>0.4163</i>

$$Pos = \frac{(F1_0^{pos} + F1_1^{pos})}{2}$$


$$Neg = \frac{(F1_0^{neg} + F1_1^{neg})}{2}$$

$$F = \frac{(Neg + Pos)}{2}$$


Tab 1: Evaluation on 2K tweets in Italian from Evalita SentiPolC 2016. Industrial engines added to the official results. Industrial engines VS research engines *specifically trained/tuned* on the given domain/source.



Evaluation on Tweets in Italian

$$\Delta_F = 3.4\%$$



	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_05-2017	u	0.5426	0.5530	0.5478
16	<i>Baseline</i>		<i>0.4518</i>	<i>0.3808</i>	<i>0.4163</i>



$$\Delta_F = 2.6\%$$



Tab 1: Evaluation on 2K tweets in Italian from Evalita SentiPolC 2016. Industrial engines added to the official results. Industrial engines VS research engines *specifically trained/tuned* on the given domain/source.

Evaluation on Tweets in Italian

$$\Delta_F = 1.3\%$$



	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_05-2017	u	0.5426	0.5530	0.5478
16	<i>Baseline</i>		<i>0.4518</i>	<i>0.3808</i>	<i>0.4163</i>



$$\Delta_F = 0.5\%$$

Tab 1: Evaluation on 2K tweets in Italian from Evalita SentiPolC 2016. Industrial engines added to the official results. Industrial engines VS research engines *specifically trained/tuned* on the given domain/source.

Evaluation on Tweets in Italian

	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_05-2017	u	0.5426	0.5530	0.5478
16	<i>Baseline</i>		<i>0.4518</i>	<i>0.3808</i>	<i>0.4163</i>

$$\Delta_F = 10.1\%$$

$$\Delta_F = 3.8\%$$

Tab 1: Evaluation on 2K tweets in Italian from Evalita SentiPolC 2016. Industrial engines added to the official results. Industrial engines VS research engines *specifically trained/tuned* on the given domain/source.

Evaluation on Tweets in English

	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
4	Senti17	0.674	0.665	0.652
5	NNEMBs	0.669	0.658	0.664
...
28	ej-za-2017	0.571	0.539	0.582
	LSIS	0.571	0.561	0.521
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
38	xiwu	0.479	0.365	0.547
39	SSN_MLRG1	0.431	0.344	0.439
40	YNU-1510	0.340	0.201	0.387
41	WarwickDCS	0.335	0.221	0.382
	Avid	0.335	0.163	0.206

Tab 2: Evaluation on 12,284 tweets in English from SemEval 2017, Task 4, subtask A. Industrial engines added to the official results. . Industrial engines VS research engines *specifically trained/tuned* on the given domain/source.

Evaluation on Tweets in English

	System	AvgR	AvgF1-PN	Acc
1	DataStories	0.681	0.677	0.651
	BB_twtr	0.681	0.685	0.658

$$AvgRec = \frac{1}{3} (R^P + R^N + R^U)$$
$$AvgF_1^{PN} = \frac{1}{2} (F_1^P + F_1^N)$$

39	SSN_MLRG1	0.431	0.344	0.439
40	YNU-1510	0.340	0.201	0.387
41	WarwickDCS	0.335	0.221	0.382
	Avid	0.335	0.163	0.206

Tab 2: Evaluation on 12,284 tweets in English from SemEval 2017, Task 4, subtask A. Industrial engines added to the official results. . Industrial engines VS research engines *specifically trained/tuned* on the given domain/source.

Evaluation on Tweets in English

$$\Delta_{AvgF1} = 12.4\%$$

	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
4	Senti17	0.674	0.665	0.652
5	NNEMBs	0.669	0.658	0.664
...
28	ej-za-2017	0.571	0.539	0.582
	LSIS	0.571	0.561	0.521
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
38	xiwu	0.479	0.365	0.547
39	SSN_MLRG1	0.431	0.344	0.439
40	YNU-1510	0.340	0.201	0.387
41	WarwickDCS	0.335	0.221	0.382
	Avid	0.335	0.163	0.206

$$\Delta_{AvgF1} = 4.7\%$$

Tab 2: Evaluation on 12,284 tweets in English from SemEval 2017, Task 4, subtask A. Industrial engines added to the official results. . Industrial engines VS research engines *specifically trained/tuned* on the given domain/source.

Evaluation on Amazon Product 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_05-2017	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

Tab 5: Evaluation on about 200,000 generic amazon product reviews in English from ESWC Semantic Sentiment Analysis 2016. Industrial engines VS research engines not specifically trained on the target domain/source.

Evaluation on Amazon Product Reviews in English

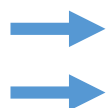
$$F_1^P = 2 \frac{R^P * P^P}{R^P + P^P}$$

$$\mathbf{M}\mathbf{F}_1 = \frac{1}{2}(\mathbf{F}_1^P + \mathbf{F}_1^N)$$

Tab 5: Experiments on **ESVC Semantic Sentiment Analysis 2016**. Industrial engines VS research engines *not* specifically trained on the target domain/source.

Evaluation on Amazon Product Reviews in English

$$\Delta_{MF1} = 4.1\%$$



	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_05-2017	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



$$\Delta_{MF1} = 23.2\%$$



Tab 5: Evaluation on about 200,000 generic amazon product reviews in English from ESWC Semantic Sentiment Analysis 2016. Industrial engines VS research engines not specifically trained on the target domain/source.



Overall Results

In our experimental evaluation, we showed that:

- considering the *best performing research tool specifically trained on the target source* as a reference (worst case for industrial APIs – tweets from SemEval 2017 and Evalita SentiPolc 2016):
 - X2Check is lower than 3.4% of F-score on Italian and 11.6% of Avg-F1 on English benchmarks
 - Google CNL is lower than 13.5% of F-score on Italian and 16.3% of Avg-F1 on English benchmarks
 - App2Check [*not tuned on tweets*] is lower than 9.7% of F-score on Italian and 16.9% on English benchmarks
- considering the *best performing research tool not specifically trained on the target source* as a reference (worst case for research engines – amazon product reviews from ESWC SSA 2016):
 - on Amazon Product Reviews in English
 - ✓ X2Check shows a macro-f1 score of 23.2% higher than the best research tool
 - ✓ Google CNL shows a macro-f1 score of 19.1% higher than the best research tool
 - ✓ App2Check [*not tuned on amazon reviews*] is lower than 13.3% of MF1 on English benchmarks from Amazon product reviews



Conclusions

- Sentiment Analysis is still a very complex task and evaluating the engines results on individual examples, counting just on the «human perception», is not a scientific approach and lead to wrong conclusions about engine performance.
- However, such «manual inspection» may help to focus on the engine's defects, understand the reasons why some misclassifications occur and better design/improve the engine.
- It is necessary evaluate the performance of a «general purpose» (pre-trained) sentiment engine APIs, through an extensive experimental analysis on multiple textual sources and domains, taking into account the overall average KPIs (accuracy, macro-F1 score, etc).
- Since sentiment engines are measurement tools, it would be better if companies provided, together with the pre-trained models, also some performance indicators on specific settings (source, topic domains, language, etc), or at least let buyers perform a comparative analysis.
- Domain/source-specific models show in general better results compared to pre-trained «general purpose» classifiers. However, applying domain-adaptation techniques or recognizing the best specialized model to apply, may reduce misclassifications on the target domain.



finsa

TECHNOLOGY FOR PEOPLE

Thank you

Emanuele Di Rosa, PhD

CSO, Head of Artificial Intelligence

Finsa s.p.a.

emanuele.dirosa@finsa.it

www.app2check.com

www.finsa.it



Engines on simple classifications: X2Check

“I hate this game”

Analyzed sentence

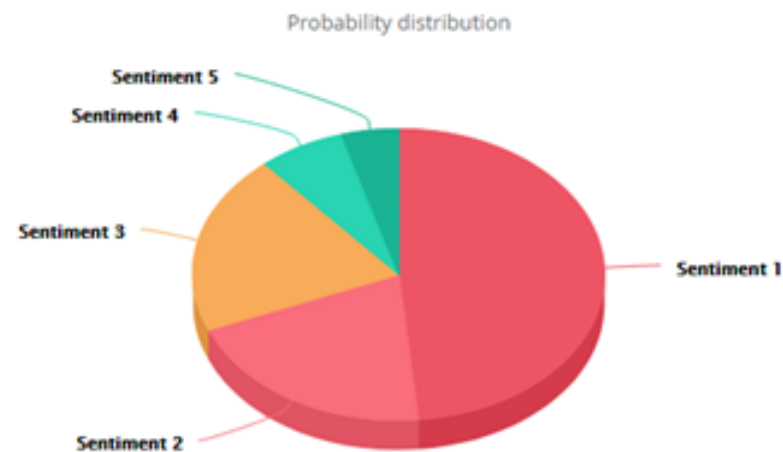
I hate this **game**

Sentiment

Overall evaluation

1

Confidence
48 %



Topics

Legend

game



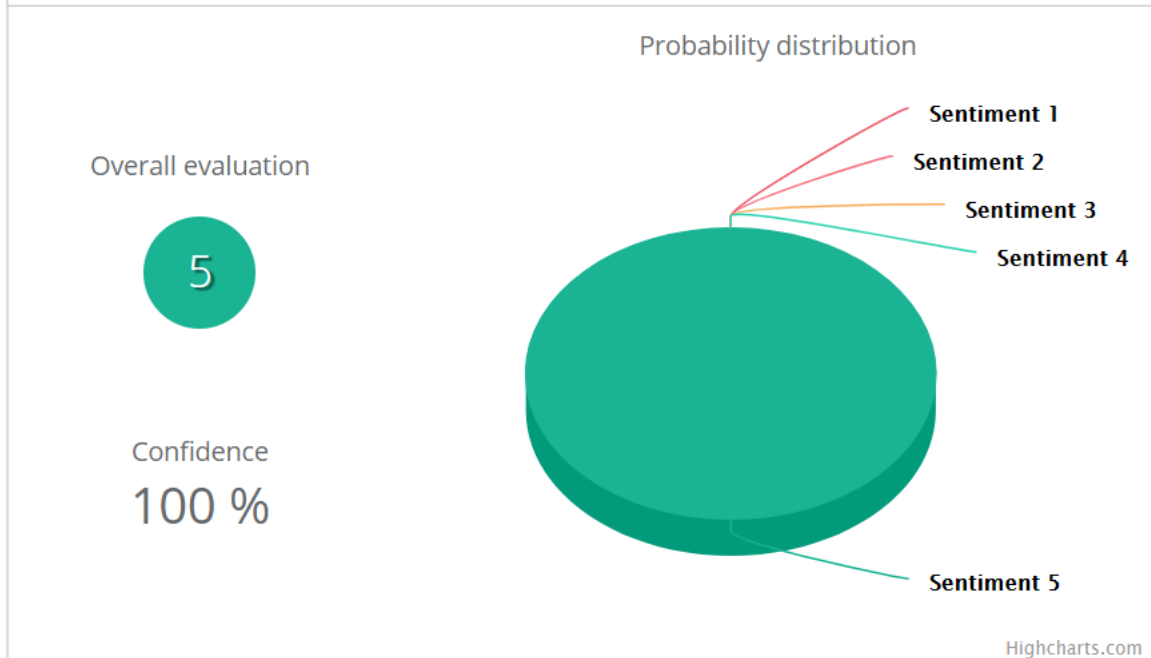
Engines on simple classifications: X2Check

“I like this game”

Analyzed sentence

i like this **game**

Sentiment



Topics

 Legend

game

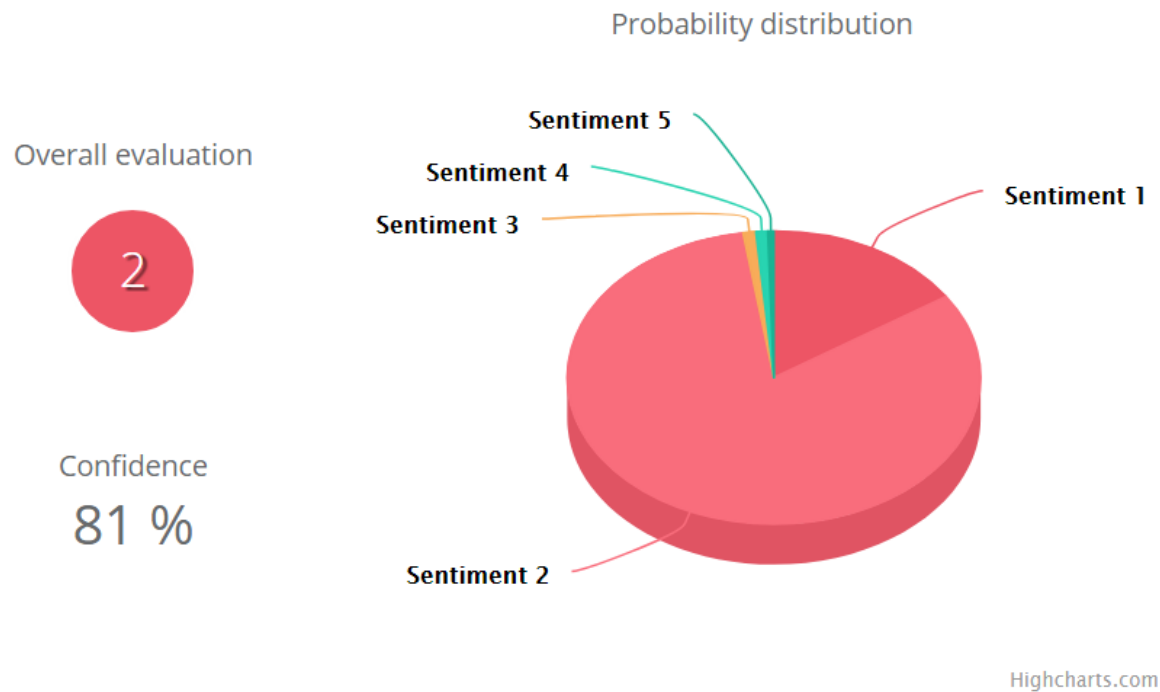


Engines on simple classifications: X2Check

Analyzed sentence

I just **connected my game** with my **facebook account** and instead of saving the **progress** i have lost all my **progress** and it came on Level 1 although I was on lvl 98
Please **help!!!!**

Sentiment



Topics

 Legend

account help facebook
connected
game progress