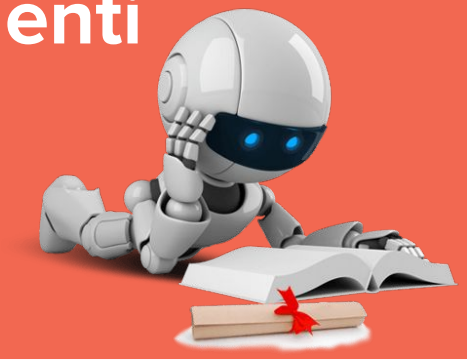
*ideato

MACHINE LEARNING

per la moderazione di commenti



EX MACHINA

Bologna, 25 gen 2019

Marco Radossi

mr@ideato.it

Sviluppatore software in ideato

Scenario

- Piattaforma editoriale
- 9000-11000 commenti al giorno
- 5 moderatori
- 300 litri di caffè al mese *
- ~~€€€€€~~

Obiettivi del progetto

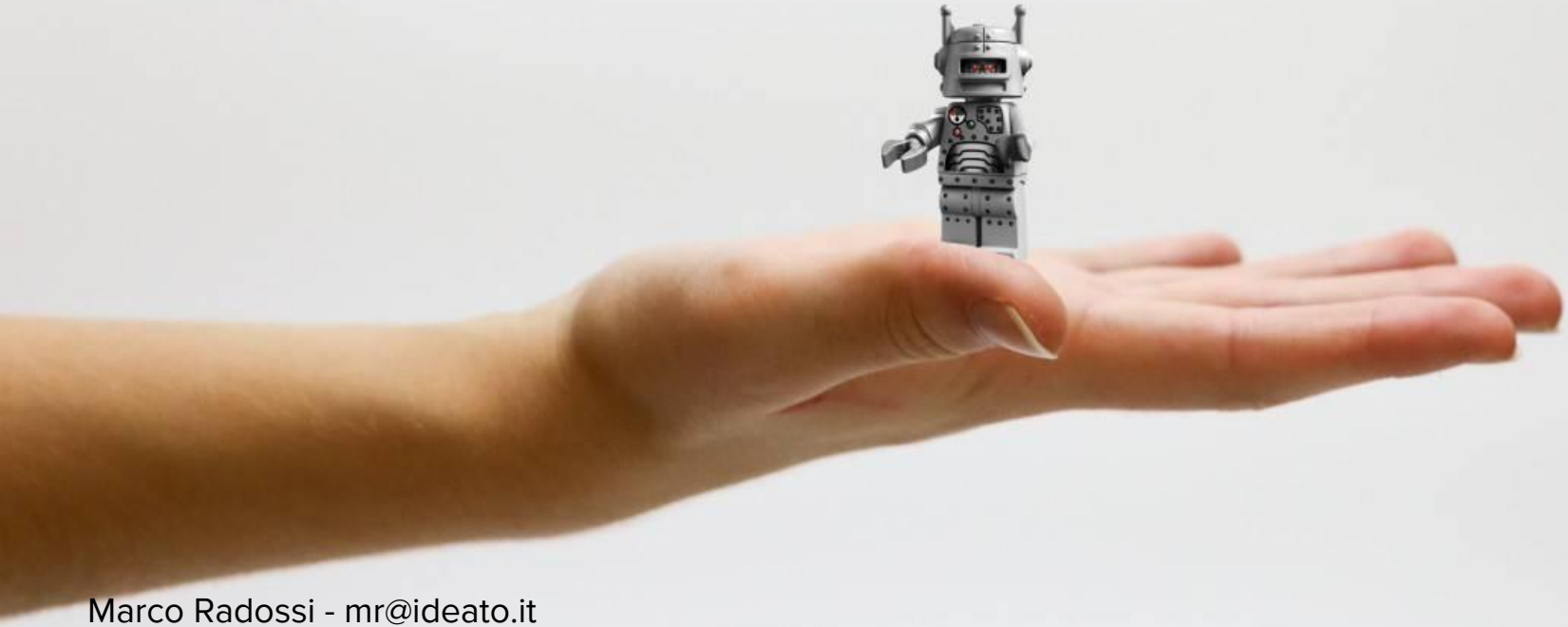
- **errori** ⬇
- **costi** ⬇⬇⬇

Proposta

semplificare e **velocizzare**
l’operato dei **moderatori**

Marco Radossi - mr@ideato.it

Come fare?



Come procedere

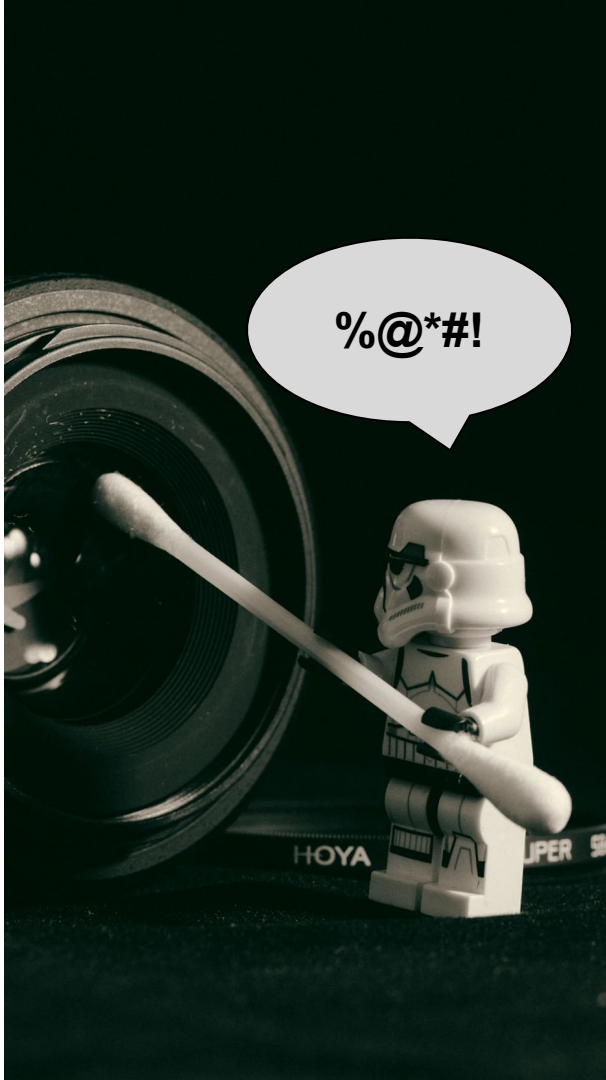
1. Recuperare i dati
2. Preparare i dati
3. Scegliere il modello
4. Addestramento
5. Valutazione
6. Tuning dei parametri

1. Recuperare i dati

- Cercare fonti di dati
- Curare **quantità** e **qualità**
- Preparare statistiche

2. Preparare i dati

- Pulire i dati
  - rimozione duplicazione
  - correzione degli errori
  - valutazione dei buchi
- Feature engineering
- Salvare i dati in maniera utile



3. Scegliere il modello

- analizzare il tipo di problema
  - classificazione binaria
  - classificazione multipla
  - regressione
  - anomaly detection
- verificare chi ha risolto problemi simili quale modello ha usato

3. Scegliere il modello

- BoW + Regressione lineare
- BoW + Naive Bayes
- TF-IdF + Random Forest

- Word2Vec + RNN LSTM
- Word2Vec + RNN GRU
- Word2Vec + CNN
- Word2Vec + RNN LSTM + CNN

3. Scegliere il modello

Serve una BASELINE:

il problema è risolvibile?

4. Addestramento

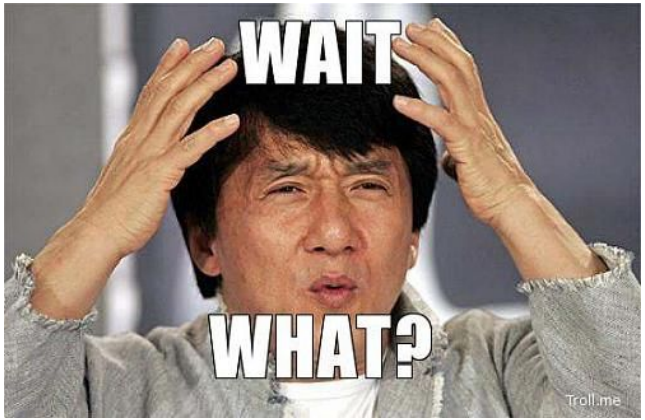
- BoW + Logistic regression:

```
model = Sequential()
model.add(Dense(2,
                activation='softmax',
                input_dim=len(feature_vector))
model.compile(optimizer='sgd',
                loss='categorical_crossentropy',
                metrics=['accuracy'])
model.fit(x_train, y_train, epochs=100,
                validation_data=(x_val, y_val))
```

4. Addestramento

- BoW + Logistic regression:

63% ???



4. Addestramento

- BoW + Logistic regression: **72%**

4. Addestramento

- BoW + Logistic regression: **72%**
- BoW + Naive Bayes: **75%**

4. Addestramento

- BoW + Logistic regression: **72%**
- BoW + Naive Bayes: **75%**
- TF-IdF + Random forest: **87,2%**

4. Addestramento

- BoW + Logistic regression: **72%**
- BoW + Naive Bayes: **75%**
- TF-IdF + Random forest: **87,2%**
- Word2Vec + RNN LSTM: **79,3%**

  (poche epoche, poca potenza)

4. Addestramento

- GPU (molto meglio delle CPU)
- cloud
- grafici di andamento
- **pazienza, tanta, ore, giorni...**

Google offre **CoLab** con le TPU

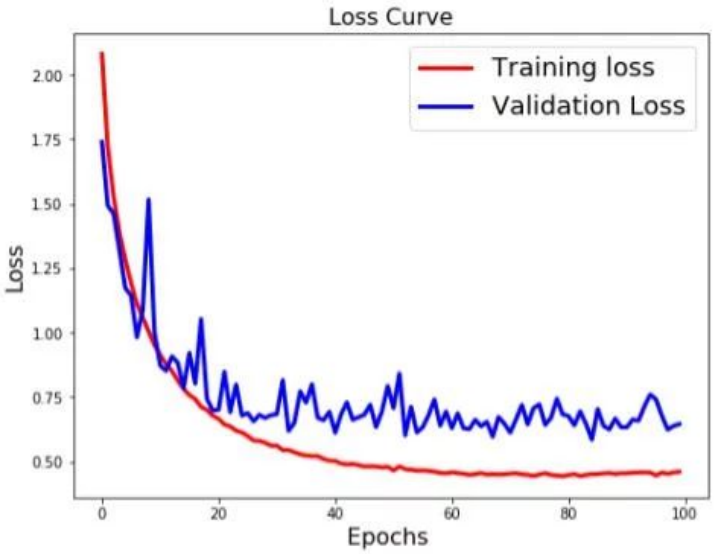
Amazon offre **SageMaker**

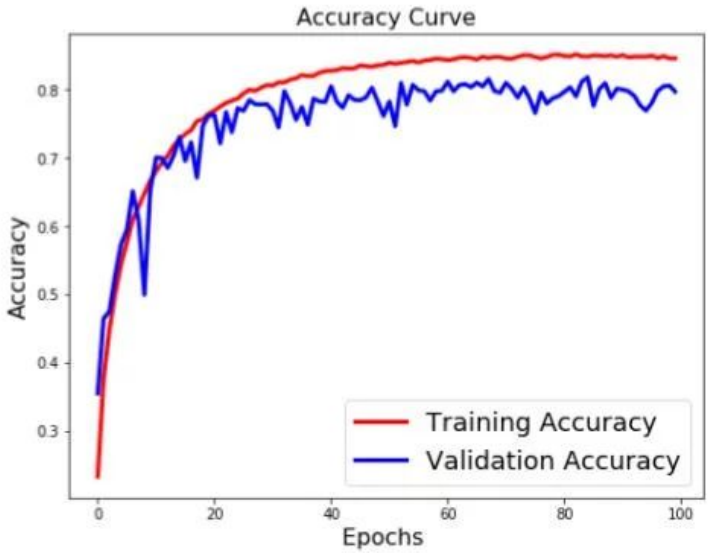
IBM offre **Watson Developer Cloud**

Marco Radossi - mr@ideato.it

5. Valutazione

Come va con **dati mai visti**?





6. Tuning dei parametri

- epoche
- batch
- learning rate
- condizioni iniziali del modello

Chi modera meglio?

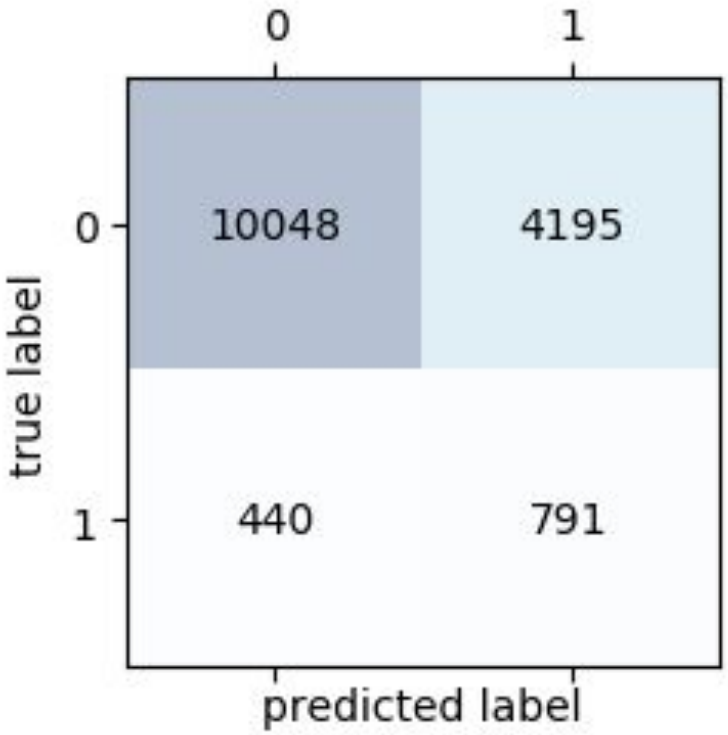
Algoritmo o **Moderatore**?

Difficile da misurare

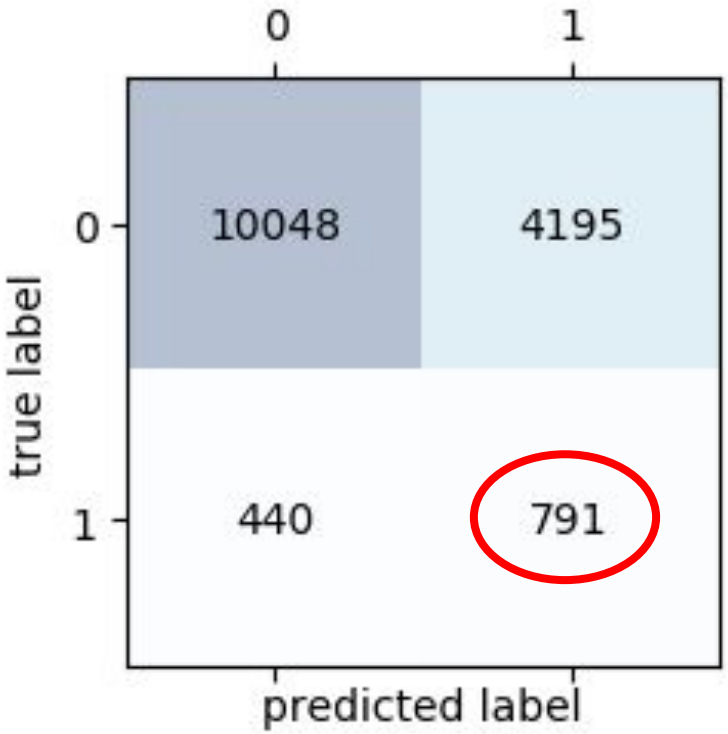
(seconda moderazione)

Marco Radossi - mr@ideato.it

Piccolo esperimento



Piccolo esperimento



Commenti da eliminare

(veri positivi)

Ti prego....buttati nel umido...

MODERATORE?????

Vattene pagliaccio
Ridicolo, pagliaccio e ignorante...
E voi di topacci di fogna!

del tipo?

Perché non intervisti uno sei tanti operai mandati sulla strada con le loro famiglie.Redazione prego pubblicare

beep

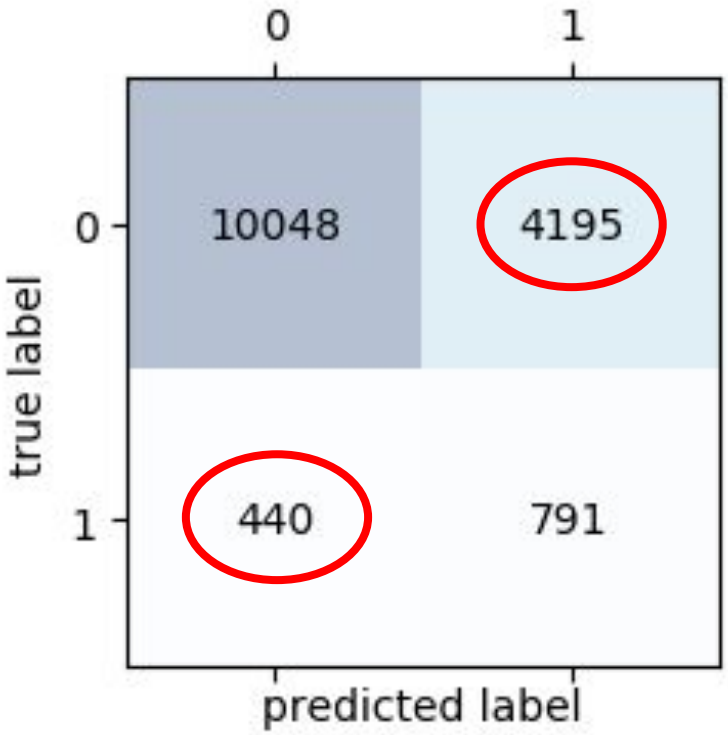
beep è il tuo luogo di nascita

beep

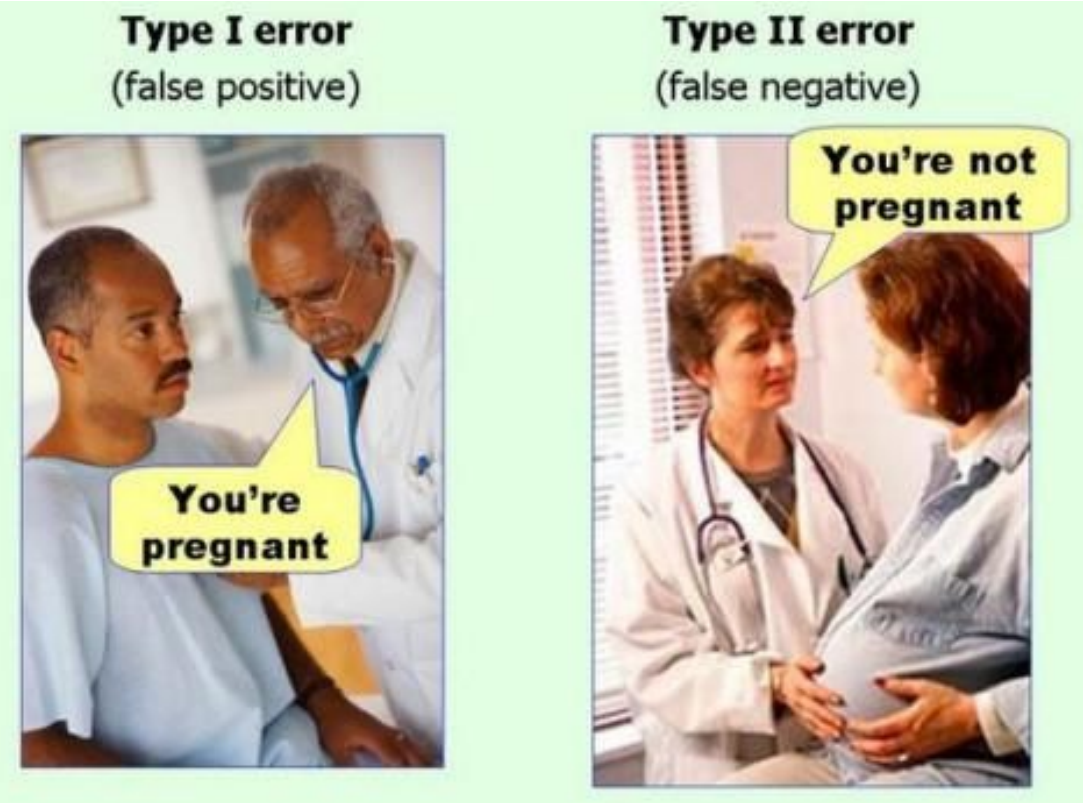
maiale...puzzi anche attraverso il PC
Dave ma IM sta per IMbecille?!?

Marco Radossi - mr@ideato.it

Errori dell’algoritmo



Errori dell’algoritmo



Falsi negativi: il moderatore elimina, l'algoritmo pubblica

scartati da una seconda moderazione

moderator	predizion	Confidence 0=bad 1=ok	body
1	0	0,97	4) arrivava dall beep
1	0	0,95	Ahahahahahahahahaha ahahahahahahahahaha ahahahahahahahahaha ahahahahahahahahaha
1	0	0,93	non penso che arrivi
1	0	0,93	Che squadrone ! Peccato solo x beep un bel cancretto no ?
1	0	0,92	*Malcom pardon
1	0	0,90	ma neanche così grande come vogliono far credere
1	0	0,90	Per come si stanno mettendo le cose si
1	0	0,90	speriamo che beep sia più bello...
1	0	0,90	Oscar si, era molto forte
1	0	0,90	I soldi di cui parla beep servono per gli aumenti di capitale e sono mandatori! Servono per coprire i buchi di bilancio mostruosi de beep he ammonteranno ad altri 80 100 milioni anche per il prossimo esercizio.. altro che beep Vedrete di quanto abbatteranno il monte beep già da questa estate.. aspettate solo 1 mesetto..
1	0	0,90	beep nelle ultime 15 beep nelle ultime 3 a beep è stato beep movibile con ben beep e. Perfetto per talento e personalità e subito pronto.
1	0	0,90	Ahahahahahahahaahahahahhaahaha....si vi siete scordati il beep
1	0	0,90	Intanto noi abbiamo beep voi sto c...
1	0	0,89	Anch'io condivido pienamente
1	0	0,87	certo beep un altro discorso, un'altra categoria
1	0	0,87	È un ottimo prospetto, un affare per beep
1	0	0,87	Però entrambi sono molto testardi e poco inclini ai cambiamenti
1	0	0,87	F beep J. il prezzo non l'ho inventato io .... altrimenti avre beep lle calcagna. E se pensi c beep 8 annui .... beh ... sei ottimista
1	0	0,87	Si ma quelli che hai citato sono tutti ottimi professionisti!
1	0	0,87	Ma beep sa dice?
1	0	0,87	beh, diciamo che e semplicemente copiato da beep

Marco Radossi - mr@ideato.it

Falsi negativi: il moderatore elimina, l'algoritmo pubblica

scartati da una seconda moderazione: <15%

moderator	predizion	Confidence 0=bad 1=ok	body
1	0	0,97	4) arrivava dall beep
1		0,95	Ahahahahahahahahaha ahahahahahahahahaha ahahahahahahahahaha ahahahahahahahahaha
1	0	0,93	non penso che arrivi
1	0		Che squadrone ! Peccato solo x beep un bel cancretto no ?
1	0	0,92	Malcom pardon
1	0	0,90	a neanche così grande come vogliono far credere
	0	0,90	r come si stanno mettendo le cose si
	0	0,90	eriamo che beep sia più bello...
	0	0,90	car si, era molto forte
1		0,90	oldi di cui parla beep ervono per gli aumenti di capitale e sono andatori! Servono per coprire i buchi di bilancio mostruosi de beep he ammonteranno ad altri 80 100 milioni anche per il prossimo esercizio.. altro che beep Vedrete di quanto abbatteranno il monte beep ià da questa estate.. aspettate solo 1 mesetto..
1		0,90	beep nelle ultime 15 beep nelle ultime 3 a beep è stato beep movibile con ben beep e. Perfetto per talento e personalità e subito pronto.
1		0,90	Ahahahahahahahaahahahahhaahaha....si vi siete scordati il beep
1		0,90	Intanto noi abbiamo beep voi sto c...
1		0,89	Anch'io condivido pienamente
1	0	0,87	certo beep un altro discorso, un'altra categoria
1	0	0,87	È un ottimo prospetto, un affare per beep
1	0	0,87	Però entrambi sono molto testardi e poco inclini ai cambiamenti
1	0	0,87	F beep J. il prezzo non l'ho inventato io .... altrimenti avre beep lle calcagna. E se pensi c beep 8 annui .... beh ... sei ottimista
1	0	0,87	Si ma quelli che hai citato sono tutti ottimi professionisti!
1	0	0,87	Ma beep a dice?
1	0	0,87	beh, diciamo che e semplicemente copiato da beep

Marco Radossi - mr@ideato.it

Falsi positivi: il moderatore pubblica, l'algoritmo elimina

scartati da una seconda moderazione

alta indecisione

moderator	predizion	Confidence 0=bad 1=ok	body
0	1	0,00	CHE SCHIFO.
0	1	0,00	Che schifo
0	1	0,00	Attenzione attenzione! Ogni tanto sa anche essere realista
0	1	0,00	Non r beep significa non capire beep ultimi
0	1	0,00	Basta che sia molto lontamo
0	1	0,00	Dovete conquistarla beep modo aspettare le disgrazie degli altri.
0	1	0,00	I beep à stato un incompetente a livello gestionale ma in fatto di stile con c beep beep c e paragone
0	1	0,00	s beep vuole beep lo ama è a lui questo nn piace
0	1	0,00	Ha parlato l'ignorante.... figuriamoci se non doveva dire la sua boiata quotidiana... tra poco sarà il turno del suo amichetto beep . Che tristezza! Veramente, che tristezza!
0	1	0,00	Gran lavoro s beep o il vostro quest'estate, complimenti.
0	1	0,00	Sai cosa dic beep scia che si lamentino
0	1	0,00	Certo, sta spingendo, perche è quel laido del procuratore che intravede un sacco di soldi per se stesso.Devono cambiare le regole altrimenti i beep ogni anno così da poter intascare loro un micchio di quattrini.Laidi che non siet beep resisti e resta con noi !
0	1	0,00	Che ignoranza, e voi che pensate si tratti di fondi affidabili
0	1	0,00	A me piace moltissimo beep che le persone ma quanto sento parlare questo stupido ho voglia che beep
0	1	0,00	Piace anche a me, anche perché è de Roma c beep
0	1	0,00	Non è questione di numeri, è questione di ignoranza allo stato puro. Questi non hanno nient'altro beep e lo utilizzano per scaricare tutte le loro frustrazioni della vita quotidiana. In poche parole, una vita grama la loro.
0	1	0,00	E beep a rotto il rap beep ne perche lo ha riproverato perche non lo beep re a lui... questa e una presa per il c....ulo
0	1	0,00	C'è anche maloox 3 per quelli come te.

Marco Radossi - mr@ideato.it

Falsi positivi: il moderatore pubblica, l'algoritmo elimina

scartati da una seconda moderazione: **<30%**

alta indecisione

moderator	predizion	Confidence 0=bad 1=ok	body
0	1	0,00	CHE SCHIFO.
0	1	0,00	Che schifo
0	1	0,00	Attenzione attenzione! Ogni tanto sa anche essere realista
0	1		Non r beep significa non capire beep ultimi
0	1	0,00	asta che sia molto lontamo
	1	0,00	vete conquistarla beep modo aspettare le disgrazie gli altri.
	1	0,00	beep à stato un incompetente a livello gestionale ma in fatto di stile beep beep c e paragone
	1	0,00	beep vuole beep lo ama è a lui questo nn piace
0		00	a parlato l'ignorante.... figuriamoci se non doveva dire la sua boiata quotidiana... tra poco sarà il turno del suo amichetto beep . Che tristezza! Veramente, che tristezza!
0	1	0,00	Gran lavoro s beep o il vostro quest'estate, complimenti.
0	1	0,00	Sai cosa dic beep scia che si lamentino
0		0,00	Certo, sta spingendo, perche è quel laido del procuratore che intravede un sacco di soldi per se stesso.Devono cambiare le regole altrimenti i beep ogni anno così da poter intascare loro un micchio di quattrini.Laidi che non siet beep resisti e resta con noi !
0	1	0,00	Che ignoranza, e voi che pensate si tratti di fondi affidabili
0	1	0,00	A me piace moltissimo beep che le persone ma quanto sento parlare questo stupido ho voglia che beep
0	1	0,00	Piace anche a me, anche perché è de Roma c beep
0	1	0,00	Non è questione di numeri, è questione di ignoranza allo stato puro. Questi non hanno nient'altro beep e lo utilizzano per scaricare tutte le loro frustrazioni della vita quotidiana. In poche parole, una vita grama la loro.
0	1	0,00	E beep a rotto il rap beep ne perche lo ha riproverato perche non lo beep re a lui... questa e una presa per il c....ulo
0	1	0,00	C'è anche maloox 3 per quelli come te.

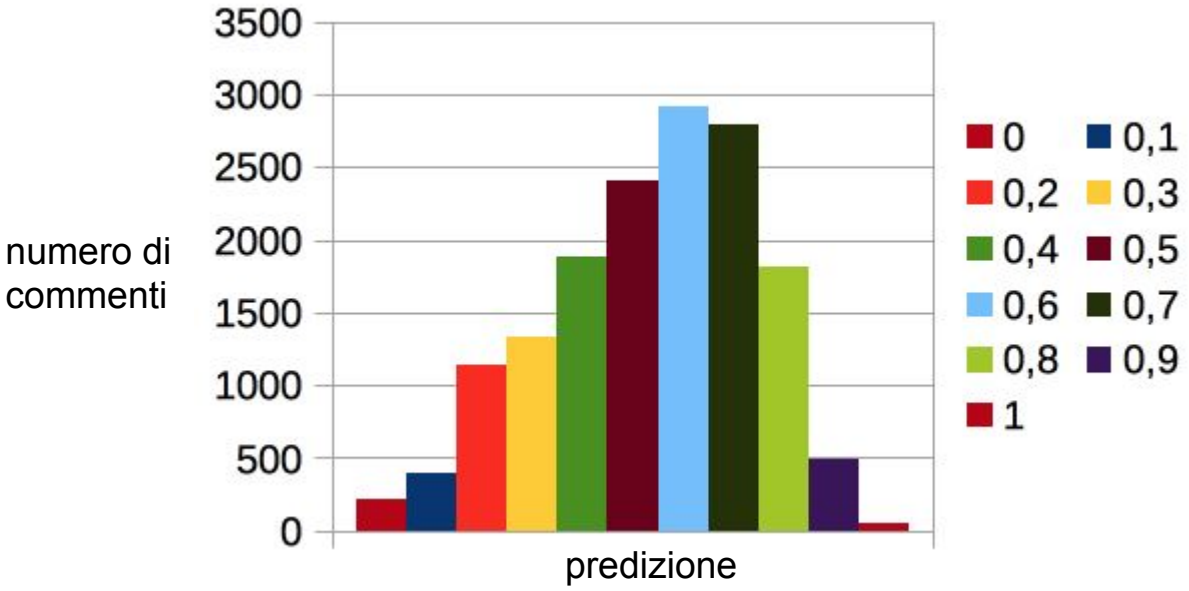
Marco Radossi - mr@ideato.it

I costi si riducono?

Marco Radossi - mr@ideato.it

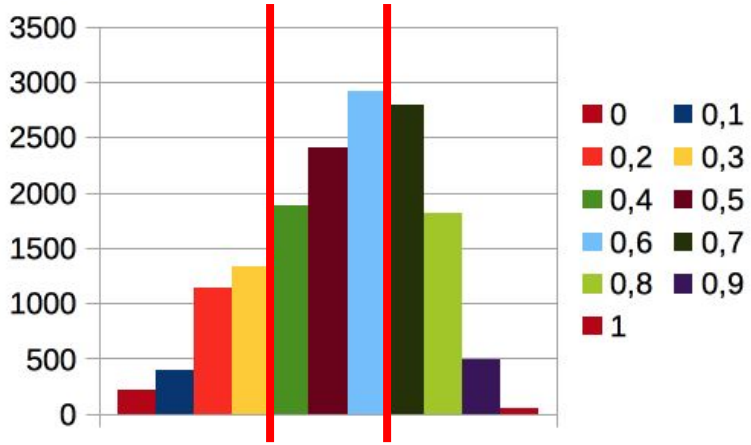
Comportamento dell’algoritmo

0 = elimina **1 = pubblica**



Marco Radossi - mr@ideato.it

Ipotesi di moderazione automatica



15470 commenti:

3095 eliminati (~20%)

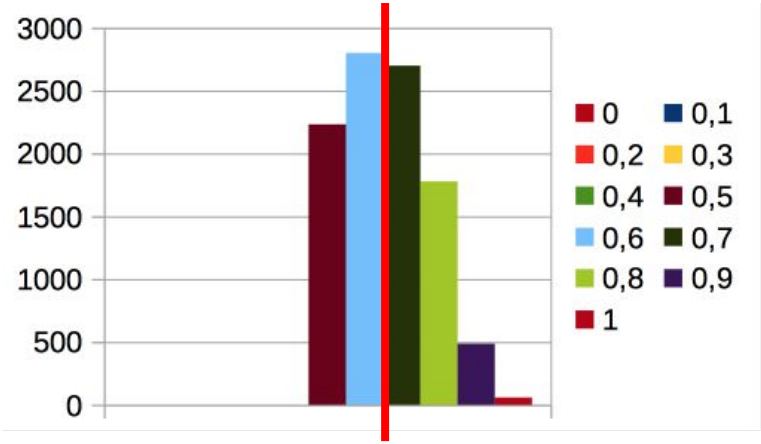
7217 da verificare

5158 pubblicati (~33%)

Sbarramenti a 0,3 e 0,7

Marco Radossi - mr@ideato.it

Pubblicati da Moderatore (Me) e Modello (Mo)



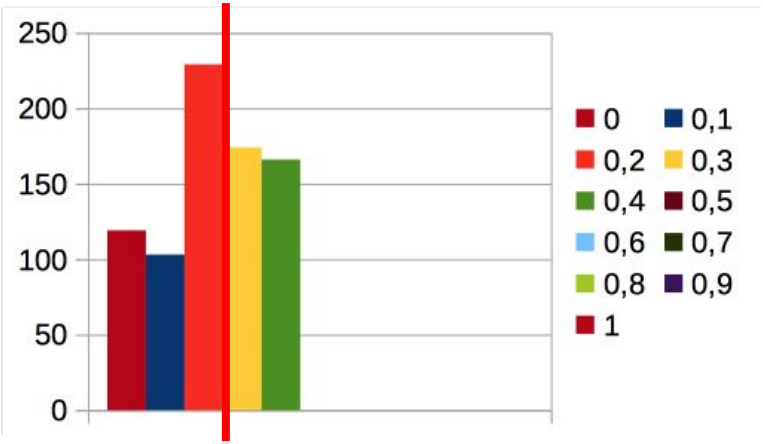
15470 commenti:

10048 pubblicati da Me e Mo

5017 > 0.7 (**30.8%**)

Marco Radossi - mr@ideato.it

Eliminati da Moderatore e Modello



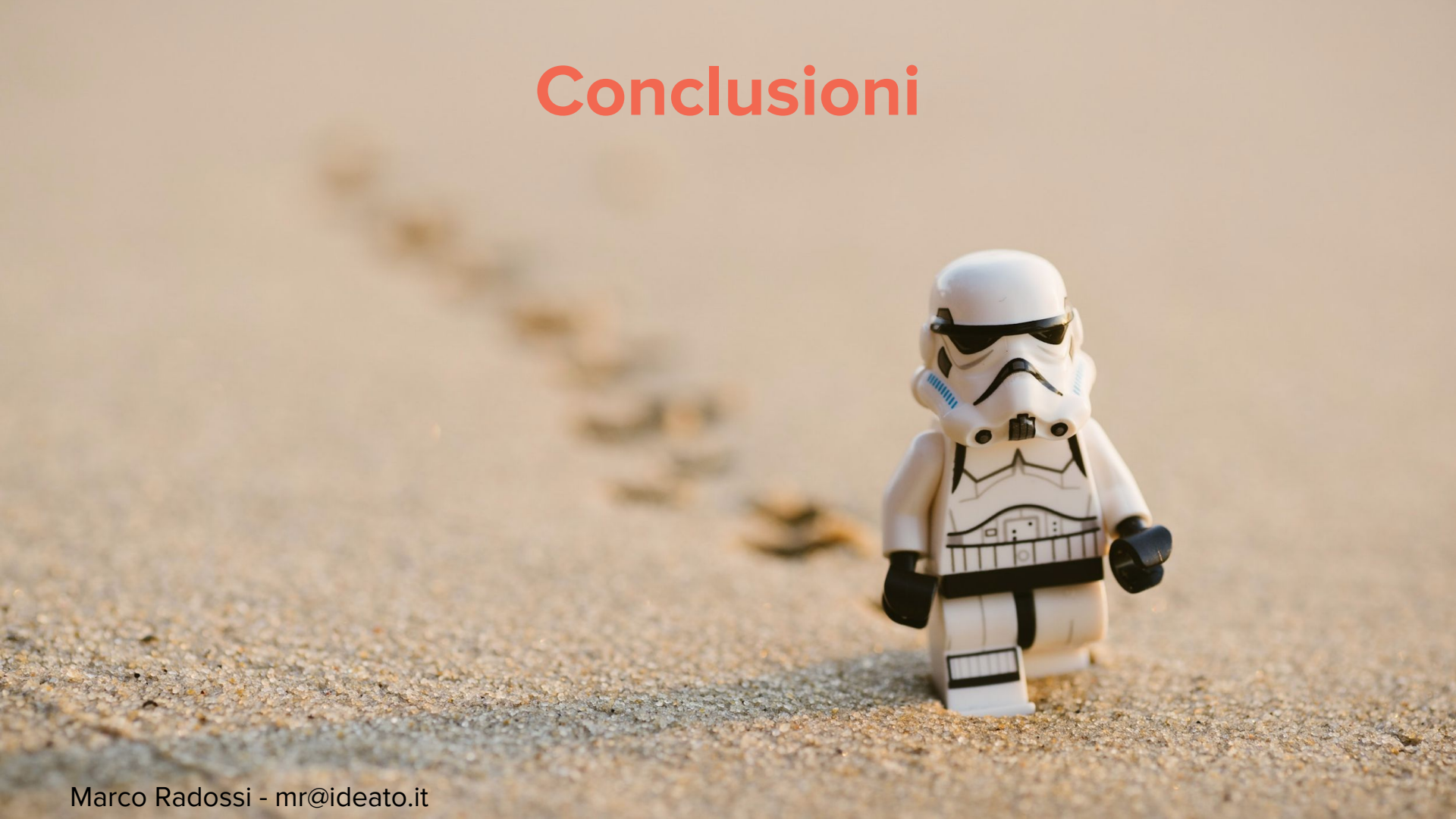
15470 commenti:

791 eliminati da Me e Mo

451 < 0.3 (**2.9%**)

Marco Radossi - mr@ideato.it

Conclusioni



Conclusioni

- Dati poco attendibili
- A volte l’algoritmo modera meglio
- I costi sono riducibili

Marco Radossi - mr@ideato.it

Possibile soluzione

comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
"know the sex of the foetus""""	0	0	0	0	0	0
I think that your a Fagget get a oife and burn in Hell I hate you	1	0	1	1	1	1
86.29.244.57|86.29.244.57]] 04:21, 14 May 2007	0	0	0	0	0	0
Sorry, the link you gave me is dead. And I formatted in harvnb,	0	0	0	0	0	0
REPLY ABOVE: That was me, loser. The UN defines Vietnam to be part of Sout Just admit that you vietnamese are all a bunch of wannabe cra	0	0	0	0	1	0
" 29 August 2006 wow. i understand that the sentance I ""added"" was not crucial	0	0	0	0	0	0
you are a stupid fuck and your mother's cunt stinks	1	1	1	0	1	0
") (ETA: John D. Haynes House. SarekOfVulcan (talk) "	0	0	0	0	0	0

Possibile soluzione: dataset

OBSCENE THREAT INSULT TOXIC

comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
"know the sex of the foetus""""	0	0	0	0	0	0
I think that your a Fagget get a oife and burn in Hell I hate you	1	0	1	1	1	1
86.29.244.57|86.29.244.57]] 04:21, 14 May 2007	0	0	0	0	0	0
Sorry, the link you gave me is dead. And I formatted in harvnb,	0	0	0	0	0	0
REPLY ABOVE: That was me, loser. The UN defines Vietnam to be part of Sout Just admit that you vietnamese are all a bunch of wannabe cra	0	0	0	0	1	0
" 29 August 2006 wow. i understand that the sentance I ""added"" was not crucial	0	0	0	0	0	0
you are a stupid fuck and your mother's cunt stinks	1	1	1	0	1	0
") (ETA: John D. Haynes House. SarekOfVulcan (talk) "	0	0	0	0	0	0

Possibile soluzione: Chargram LSTM + CONV

143613 esempi

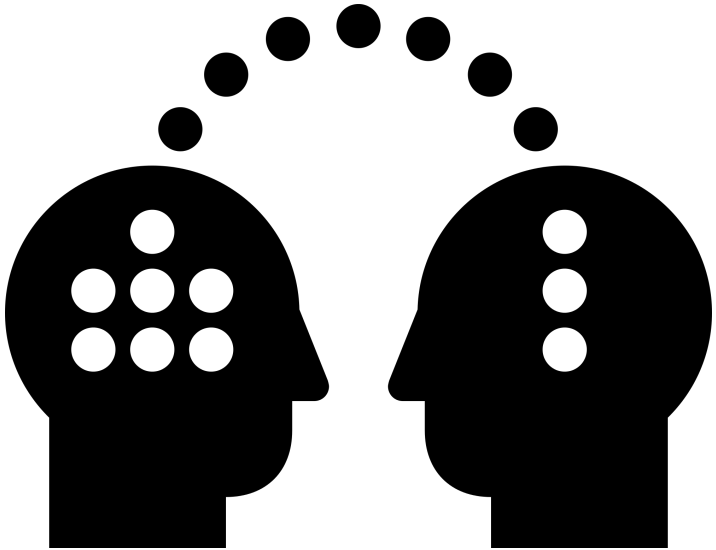
```
Train on 143613 samples, validate on 15958 samples
Epoch 1/6 143613/143613 [==============================] - 2580s 18ms/step - loss: 0.0786 - acc: 0.9763 - val_loss: 0.0585 - val_acc: 0.9806
Epoch 2/6 143613/143613 [==============================] - 2426s 17ms/step - loss: 0.0582 - acc: 0.9804 - val_loss: 0.0519 - val_acc: 0.9816
Epoch 3/6 143613/143613 [==============================] - 2471s 17ms/step - loss: 0.0531 - acc: 0.9816 - val_loss: 0.0489 - val_acc: 0.9823
Epoch 4/6 143613/143613 [==============================] - 2991s 21ms/step - loss: 0.0505 - acc: 0.9821 - val_loss: 0.0484 - val_acc: 0.9829
Epoch 5/6 143613/143613 [==============================] - 3023s 21ms/step - loss: 0.0487 - acc: 0.9826 - val_loss: 0.0463 - val_acc: 0.9829
Epoch 6/6 143613/143613 [==============================] - 2961s 21ms/step - loss: 0.0474 - acc: 0.9830 - val_loss: 0.0463 - val_acc: 0.9831
```

98.3% di accuratezza

Marco Radossi - mr@ideato.it

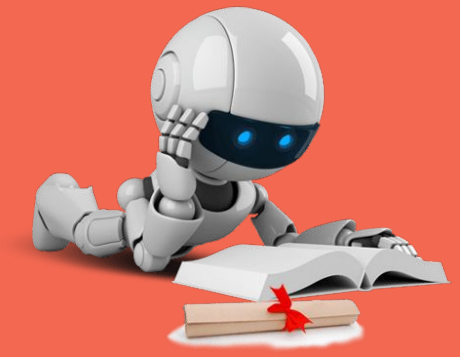
Possibile soluzione: tecnica da applicare

Transfer Learning
+
Fine Tuning



*ideato

grazie



Marco Radossi - mr@ideato.it

Via alle domande...