Classificazione di Tweet a tema politico

Simone Robutti

21 luglio 2014

Matricola: 823523

Indice

1 Introduzione ed Obiettivi

L'esperimento si propone di generare un classificatore in grado di discernere l'orientamento politico di un utente Twitter basandosi sul contenuto dei suoi Tweet.
Il lavoro si sviluppa partendo dal lavoro e dalle considerazioni di Malouf e Mullen, "Taking sides: User classification for informal online political discourse". La differenza fondamentale con questo lavoro è la natura dei testi classificati: non Tweet ma post di un forum di politica, quindi più lunghi e argomentati e con uno spettro sintattico e semantico più ampio.
Altri lavori simili presi in considerazione sono Pennacchiotti e Popescu, "A Machine Learning Approach to Twitter User Classification." e Durant e Smith, "Mining sentiment classification from political web logs" che però per complessità e obiettivi sono lontani da quanto realizzato.

2 Metodologia e approccio

Per rappresentare i tweet nello spazio delle features si è scelto di usare un modello bag of words associando ad ogni parola presente o assente un valore binario.

Utilizzare il numero di occorrenze di una parola all'interno del tweet è un approccio valutato da Malouf e Mullen, "Taking sides: User classification for informal online political discourse" e sconsigliato perché dava risultati leggermente peggiori.
Lo spazio delle etichette è stato poi suddiviso in 4 classi: Sinistra,Destra, Centro e Populisti. Una divisione in due sole classi (Sinistra e Destra o Conservatori e Progressisti) mal si prestava a rappresentare lo scenario politico e mediatico italiano odierno.
Essendo l'informazione presente nel singolo tweet insufficiente, si è scelto per raggruppare i tweet in gruppi di dimensione fissata e determinata empiricamente come documentato in seguito. Inoltre il raggruppamento nel training set non viene fatto in base all'account Twitter che ha generato il Tweet ma semplicemente in base alla classe di appartenenza, in maniera casuale.
Come algoritmo per generare il classificatore viene usata una SVM lineare implementata nel pacchetto SciKit. Per valutare la correttezza del classificatore viene usata la cross validazione interna 10-fold.

3 Dataset

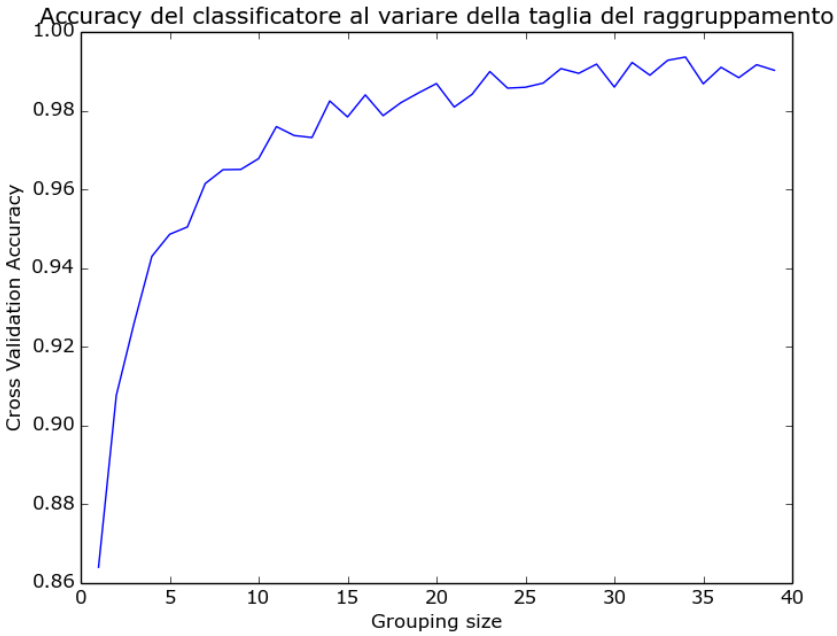
Il dataset è costituito dai tweet pubblicati da account personali di politici italiani e da account di partito. Sono stati esclusi quegli account che pur essendo di politici, trattano temi non strettamente connessi alla politica italiana. Una possibile raffinazione ulteriore potrebbe essere fatta scremando i singoli tweet non strettamente relativi alla politica utilizzando un classificatore in grado di discernere il topic del tweet.
L'estrazione e la normalizzazione delle features prevede prima una divisione dei testi in unigrammi su cui viene poi applicato un processo di stemming e di rimozione delle stop words. Per svolgere questa attività è stata utilizzata la libreria NLTK (Natural Language Toolkit).
Gli account scelti per la composizione del dataset per l'esperimento sono i seguenti:

- PierferdinandoCasini
- Lega Nord 2.0
- Angelino Alfano
- Mario Monti
- Unione di Centro
- Matteo Renzi
- forzasilvio.it
- Scelta Civica

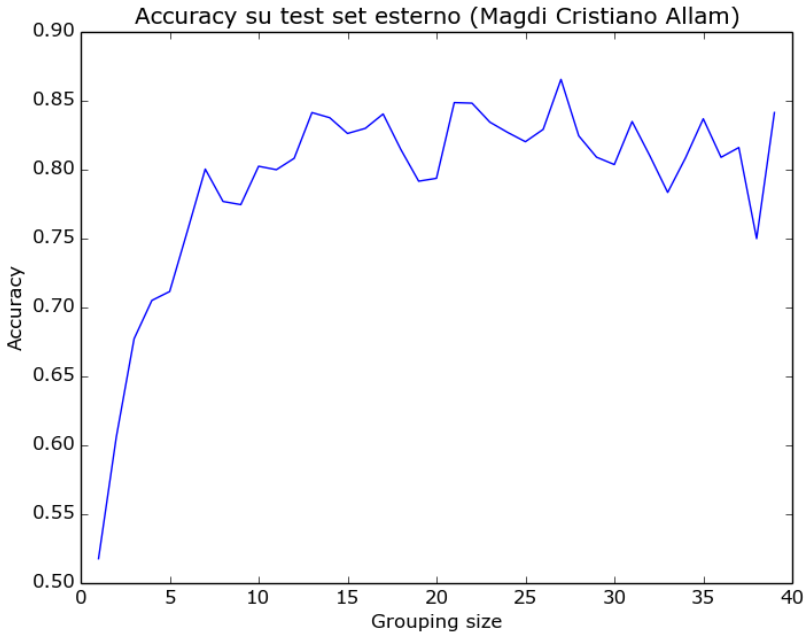
- matteo salvini
- Nichi Vendola
- Partito Democratico
- Sinistra E. Libertà
- Movimento 5 Stelle
- Nuovo Centrodestra
- Beppe Grillo
- Forza Italia
- Civati
- Ferrero Paolo
- Alessandro Di Battista

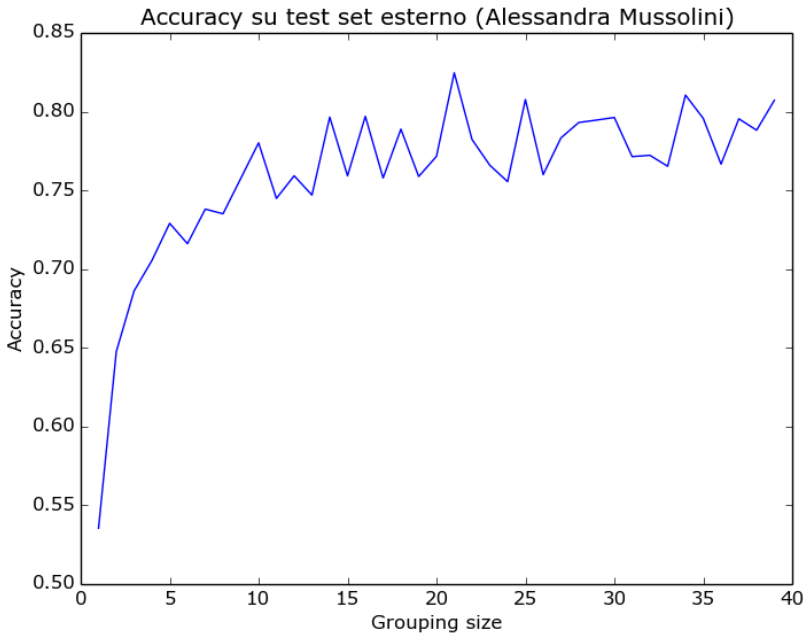
4 Risultati

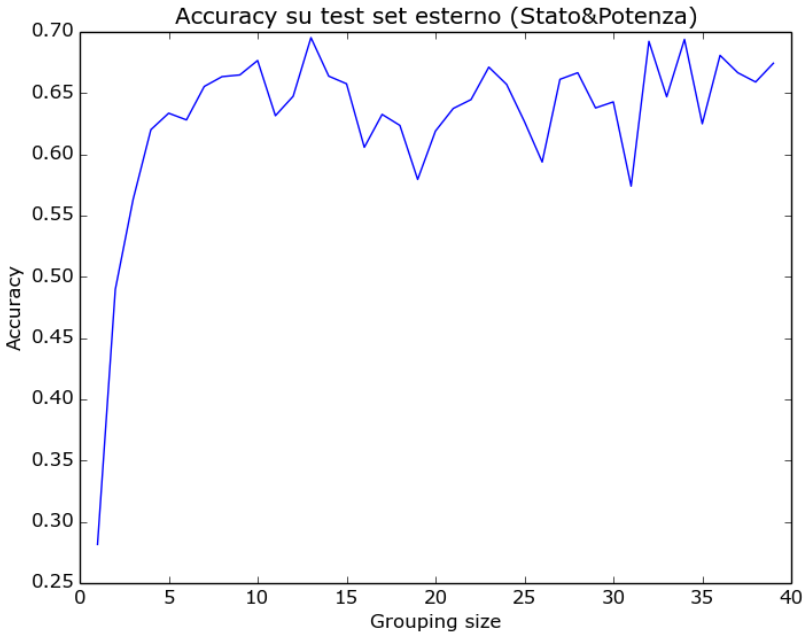
Per valutare la precisione del classificatore sul dataset sono state eseguiti dei test utilizzando la cross validazione interna 10 fold. Il risultato si è dimostrato pesantemente dipendente dalla dimensione del raggruppamento dei tweet.
Per valori compresi tra 1 e 10 si rileva una crescita della precisione del classificatore che tende a stabilizzarsi per valori superiori, assestandosi sull'intervallo compreso tra il 98% e il 99%.

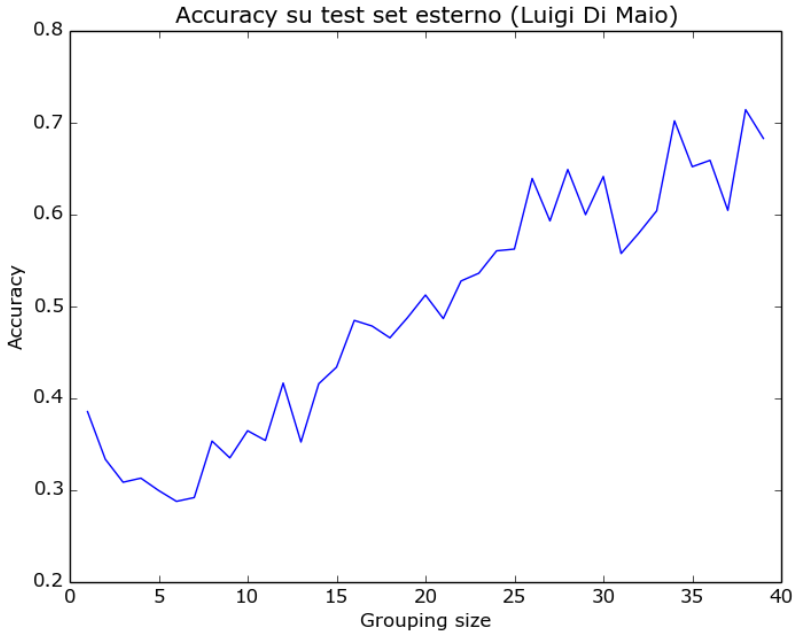


Meno precisione viene invece rilevata andando a classificare gruppi di tweet di account non presenti nel dataset su cui si apprende. La precisione comunque in molti casi è più che soddisfacente.


Accuracy su test set esterno (Magdi Cristiano Allam)
Accuracy
Grouping size
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0
5
10
15
20
25
30
35
40


Accuracy su test set esterno (Alessandra Mussolini)
Accuracy
Grouping size
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0
5
10
15
20
25
30
35
40


Accuracy su test set esterno (Stato&Potenza)
Accuracy
Grouping size
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0
5
10
15
20
25
30
35
40


Accuracy su test set esterno (Luigi Di Maio)
Accuracy
Grouping size
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
5
10
15
20
25
30
35
40

5 Spunti di ampliamento

Il lavoro sperimentale svolto è limitato all'applicazione di una singola tecnica in maniera estremamente mirata e questo limita notevolmente il campo applicativo e la precisione del classificatore generato.
Un approccio noto ed usato in diversi studi su questo tema, come in Pennacchiotti e Popescu, "A Machine Learning Approach to Twitter User Classification.", consiste nell'ampliare il dominio delle features inserendo non solo i testi dei tweet ma anche le informazioni personali e la rete sociale dell'utente da cui è possibile inferire l'orientamento politico. Nel caso di Twitter sarebbe semplice avendo a disposizione pubblicamente tutte queste informazioni in un formato strutturato.
Un altro elemento interessante potrebbe essere l'uso di un classificatore in grado di discernere il topic di un tweet, che vada a migliorare la composizione del dataset per eliminare parte del rumone nel training set e per migliorare le performance nel caso si cercasse di classificare account generici di utenti che trattano altri argomenti diversi dalla politica.

Riferimenti bibliografici

Durant, Kathleen T e Michael D Smith. "Mining sentiment classification from political web logs". In: *Proceedings of Workshop on Web Mining and Web Usage Analysis of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (WebKDD-2006), Philadelphia, PA*. 2006.

Malouf, Robert e Tony Mullen. "Taking sides: User classification for informal online political discourse". In: *Internet Research* 18.2 (2008), pp. 177–190.

Pang, Bo e Lillian Lee. "Opinion mining and sentiment analysis". In: *Foundations and trends in information retrieval* 2.1-2 (2008), pp. 1–135.

Pennacchiotti, Marco e Ana-Maria Popescu. "A Machine Learning Approach to Twitter User Classification." In: *ICWSM* 11 (2011), pp. 281–288.