TWITA

Sentiment analysis on Italian tweets [2013]

Valerio Basile - Malvina Nissim

Pelucchi Mauro

Obiettivi

Creare un **corpus** di tweets in Italiano. L'insieme deve essere creato attraverso una **procedura automatica**, **esportabile** anche in altre lingue.

Formare un **polarity lexicon** per l'italiano: un dizionario dove ad ogni parola corrisponde la sua **polarità lessicale** (negativa, neutra, positiva). La metodologia usata deve essere automatica e replicabile verso altre lingue.

Applicare il dizionario creato a **due subset del corpus di tweets** e valutare i risultati rispetto alle annotazione manuali. Il primo sottoinsieme è costituito estraendo in modo casuale dal corpus di tweets, il secondo invece contiene i testi di uno specifico topic.

TWITA

Il paper ha portato alla creazione di **TWITA**, il primo corpus italiano di tweets.

La procedura automatica con cui è stato creato il corpus è PORTABILE verso qualsiasi linguaggio o fenomeno. Il sistema anche se sembra molto semplice e superficiale ottiene performance classificative.

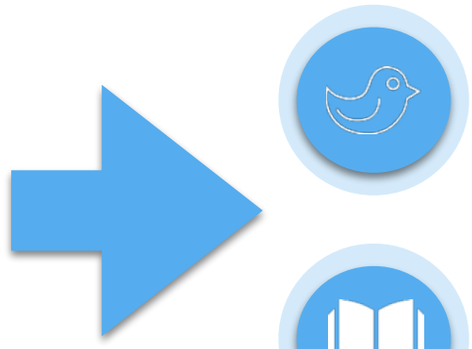
TWITA e il polarity lexicon sono la base per i sistemi di "opinion mining", cioè l'analisi semantica dei messaggi e dei testi per capire se l'opinione espressa verso un prodotto o un servizio, è positiva, negativa o neutra.

Contesto

Twitter è un servizio gratuito di social networking e microblogging basato sullo scambio pubblico di messaggi (**tweets**) lunghi 140 caratteri. Ad ogni tweets sono associati dei metadati (mittente, coordinate geografiche, utenti menzionati, …).
Nel 2013 **non esiste alcun corpus di tweets in italiano** e la maggior parte dei sistemi di **opinion mining** sono per la lingua inglese.

Passi della procedura

Cosa serve per creare un sistema di **opinion mining** basato su Twitter?



1 - Corpus di tweets nella lingua desiderata

2 - Un lessico polarizzato

Creazione del set di termini per effettuare l'operazione di Language Detection

Estrazione dei Tweets e creazione del corpus

Processing dei testi ed estrazione dei termini

Sentiment Analysis (creazione del lessico polarizzato)

Polarizzazione dei tweets e valutazione sui 2 subset

Language Detection

Per creare un corpus di tweets serve partire dall'insieme dei **top termini** che caratterizzano l'italiano (in generale in un certo linguaggio).

I top terms permettono di estrarre solo i tweets di un certo linguaggio e quindi svolgono l'operazione di **LANGUAGE DETECTION.**

L'operazione di language detection consiste nell'utilizzare una lista di parole chiave rappresentative (creata automaticamente) per estrarre i tweets. Per costruire questa lista utilizziamo tre passi (**replicabili in qualsiasi lingua**):

Estrazione delle 1000 parole più frequenti da ItWaC (un insieme di 2 miliardi di parole costruito scagionando i domini .it).

Estrazione dei tweets che contengono le parole chiave e detect della lingua attraverso software free (obiettivo è eliminare le sequenze di caratteri comuni con altre lingue).

Filtro dei termini che hanno un'alta frequenza anche in altre lingue (usando Google N-Grams).

Creazione del corpus

I **top terms** generati dalla procedura di language detection vengono usati come input delle **Twitter API.**

TWITA è un corpus di **100 milioni di tweets**, relativi all'anno 2012. Ogni tweet è caratterizzato dal testo, dal timestamp, dalle coordinate geografiche (se presenti) e dalle informazioni dell'utente.

Problema: difficile misurare le performance (in termini di precision e recall). Non possiamo sapere quanti e quali tweets in Italiano non stiamo considerando.
Vantaggi: moltissimi dati!

Valutazione: approccio empirico. Check manuale di 1000 tweets casuali (**precision a 99.7%**) e plot di 1 milioni di tweets su una mappa.

Processing

Obiettivo della fase di processing è arricchire ogni tweet con i token, i pos-tags, e i lemmi.

Sostituzione degli hashtag, degli url e delle menzioni

Tokenize

Aggiunta dei POS-tags

Lemmatizzazione

Re-inserimento di tutte le informazioni originali

Creazione del lessico polarizzato

Come ottenere un **lessico polarizzato** in automatico?

1 - Possiamo usare le emoticons inserite nel testo

2 - Uso di semi polarizzati

TWITA utilizza semi polarizzati in automatico attraverso **SentiWordNet**. SentiWordNet è disponibile solo in inglese, ma attraverso **MultiWordNet** possiamo tradurre e mappare i termini italiani sui termini in inglese.

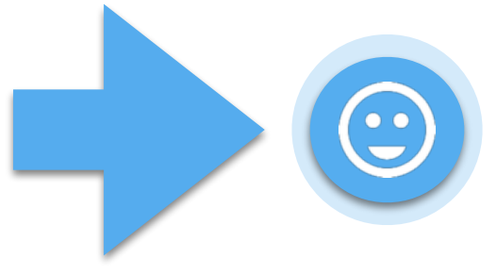
MultiWordNet è un database lessicale multi-lingua. Include termini in italiano e inglese. In poche parole: è possibile cercare un termine in italiano su MultiWordNet e trovare i corrispondenti synset in inglese.

WordNet è un database semantico-lessicale per la lingua inglese. L'organizzazione del lessico si avvale di raggruppamenti di termini con significato affine, chiamati "synset". All'interno dei synset le differenze di significato sono numerate e definite.

SentiWordNet è una risorsa lessicale per opinion mining: assegna ad ogni synset di WordNet tre punteggi di sentimento: positività, negatività e neutralità. I termini trovati attraverso MultiWordNet vengono ricercati su SentiWordNet: ad ogni lemma viene associata la polarità e viene costruito il lessico polarizzato.

Calcolo del sentiment di un tweet

Cosa è il **sentiment** di un tweet?



Il sentiment è la combinazione della polarità semantica delle parole nel testo del tweet. Può essere **positiva**, **neutrale** o **negativa**.

Polypathy è la deviazione standard dell'indice di polarità di un termine rispetto ai vari significati che può assumere.
La polypathy permette di escludere dal calcolo le parole che compaiono più volte nel lessico con polarità molto diversa.

Ogni tweet è ridotto in token. Possiamo trovare più volte un termine nel nostro lexicon con significato diverso.
I termini con alta polypathy (>0.5) vengono esclusi.

Ogni lemma viene cercato nel lexicon in modo tale da avere la sua polarità (che è pari alla media dei vari significati del termine), se non viene trovato il valore usato è zero.

La polarità del tweet è data dalla somma delle polarità dei suoi token.

Risultati

2 gold standard set valutati da 3 osservatori indipendenti (la polarità del tweet è assegnata attraverso il majority vote)

ogni osservatore ha classificato (attraverso una piattaforma web) i tweet come positivi, neutrali, negativi

1000 tweets casuali

70.1% **recall** (tweet positivi, valutati con nomi, verbi, avverbi e aggettivi) valutata senza filtro su polypathy

71.4% **recall** (tweet positivi, valutati con nomi, verbi, avverbi e aggettivi) valutata con filtro su polypathy

1000 tweets di un topic specifico (esempio “Beppe Grillo”)

59.3% **recall** (tweet positivi, valutati con nomi, verbi, avverbi e aggettivi) valutata senza filtro su polypathy

59.3% **recall** (tweet positivi, valutati con nomi, verbi, avverbi e aggettivi) valutata con filtro su polypathy

Lavori futuri

Valutare i tweets tenendo in considerazione fenomeni come l'**ironia** ed il **sarcasmo**.

Utilizzo di **analisi statistiche** e di metadati per correggere il valore della polarity: ad esempio è possibile sfruttare la correlazione fra sentiment e giorno della settimana, ora del giorno, posizione geografica…

Problemi

Un termine può far parte di più synset in WordNet. L'opzione più precisa sarebbe quella di prendere solo la polarità relativa al significato corretto (con un'operazione di **disambiguazione**). Questo non è possibile attualmente per l'italiano.

Recall e precision del sub-set dei tweets creato per topic sono più basse rispetto al primo sub-set: questo è dovuto alla difficoltà di assegnare la polarità quando nella testa si parla di **due argomenti**.

Referendum Costituzionale 2016

Obiettivo: creare un barometro in tempo reale del sentiment sul Referendum Costituzionale 2016

Utilizzo dell'approccio di TWITA per valutazione del sentiment relativo al Referendum Costituzionale del 2016

1 - Creazione di una top terms list relativa al Referendum Costituzionale del 2016 per estrarre i dati da Twitter (da vari quotidiani sono state estratte le 20 parole più usate negli articoli trattanti il referendum)

2 - Estrazione dei dati da twitter in base alla top terms list in tempo reale (attraverso e API Stream di Twitter)

3 - Processing dei twitter ed estrazione dei termini (tokenize, filtro stop words in italiano, generazione degli n-grammi)

4 - Traduzione da italiano ad inglese dei termini estratti (attraverso le API di Google Translate)

5 - Creazione di un lessico polarizzato (in inglese) da WordNet

6 - Valutazione del sentiment del tweet (è la somma della polarità dei termini che fanno parte del tweet)

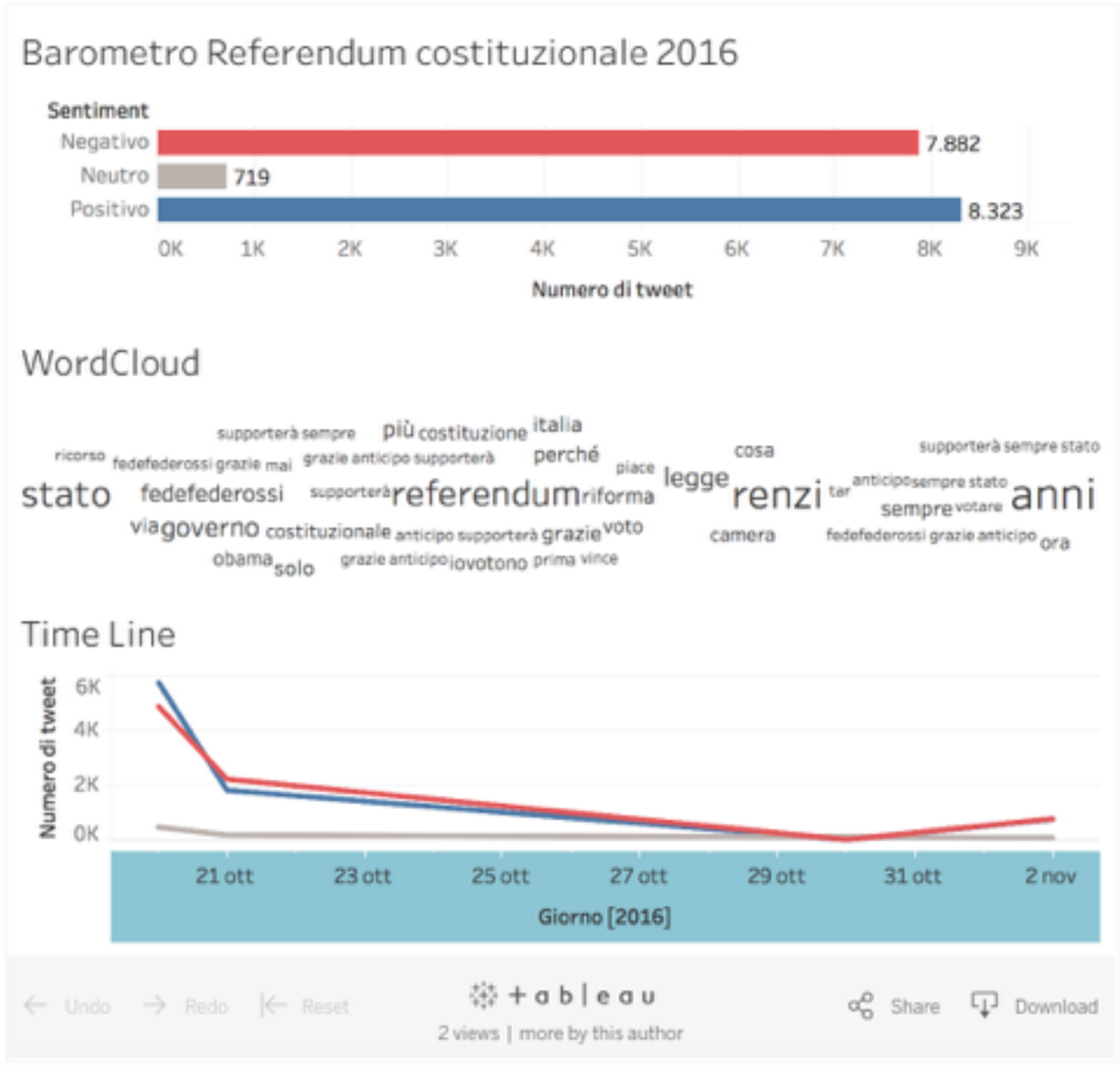
Consulta il barometro su Tableau

Referendum Costituzionale 2016

Valutazione (manuale) su 50 tweet
(rispetto alla positività)

75 % Precision

31% Recall



Consulta il barometro su Tableau