Un Approccio per la Valutazione della Credibilita` del Contenuto Generato dagli Utenti nei Siti di Recensioni

Marco Viviani, Alessio Serafino, and Gabriella Pasi

Università degli Studi di Milano-Bicocca / DISCo / IR Lab
Viale Sarca, 336 – Edificio U14 – 20126 Milano
a.serafino@campus.unimib.it,{marco.viviani,pasi}@disco.unimib.it
WWW home page: http://www.ir.disco.unimib.it/

Sommario Attraverso la diffusione dei Social Media è stata data agli utenti la possibilità di pubblicare contenuto sul Web senza intermediari e la possibilità di costituire relazioni a molteplici livelli. Ciò comporta il rischio di dare spazio a fonti poco affidabili e falsa informazione. In questo scenario è interessante verificare la bontà del contenuto generato dagli utenti nei siti di recensioni. Utenti non affidabili possono infatti fornire recensioni fasulle al fine di migliorare/affossare l'immagine di un'attività commerciale, ingannando in questo modo altri utenti che non hanno strumenti oggettivi per valutare l'affidabilità delle recensioni stesse. In questo articolo si propone perciò un approccio basato su tecniche di apprendimento automatico supervisionato per il supporto all'utente nella verifica della credibilità dell'informazione nei siti di recensioni. In base alle valutazioni ottenute si è deciso di investigare, per sviluppi futuri, l'apporto che lo studio del linguaggio utilizzato dagli utenti potrebbe avere nel giudicare correttamente le recensioni in esame.

1 Introduzione

Negli ultimi anni si è assistito ad una esplosione dell'informazione disponibile online, in particolare di contenuto generato dagli utenti, comunemente noto come *User Generated Content* (UGC). Con questa espressione si indicano le varie forme di contenuto multimediale e non, liberamente accessibile e pubblicato dagli utenti senza intermediazioni [1]. Complice di questa diffusione è stato lo sviluppo di siti Web e di applicazioni mobili che rientrano nella categoria dei cosiddetti *Social Media*; essi, oltre a consentire all'utente un'enorme libertà espressiva, permettono lo scambio diretto di informazione e la costituzione di relazioni tra utenti a molteplici livelli.

In questo scenario si rischia di dare spazio a fonti non riconosciute, informazione non affidabile o non verificabile. Le motivazioni che spingono gli utenti a generare contenuto falso possono essere molteplici: da quelle economiche (ad esempio attività commerciali interessate a dare di sè un'immagine migliore o a denigrare la concorrenza) a quelle di natura più strettamente personale (come

ad esempio l'attività di *self-promotion* per aumentare la popolarità del proprio profilo/pagina Web). Venendo a mancare i tradizionali intermediari che potevano verificare l'affidabilità di una fonte o di una informazione e lasciati gli utenti giudicare autonomamente sulla base delle proprie esperienze personali e sul buon senso, emerge quindi il bisogno di studiare e sviluppare sistemi automatici che li possano aiutare nella verifica della credibilità delle informazioni e delle fonti in base a criteri oggettivi.

Un ambito in cui lo studio del problema di valutare la veridicità del contenuto generato dagli utenti è particolarmente interessante è quello dei siti di recensioni, o *review site*. In questo tipo di siti, gli utenti forniscono recensioni su attività commerciali e servizi, senza controlli da terze parti. Essi sono caratterizzati da un'ampia mole di contenuto testuale, da innumerevoli metadati collegati sia alle recensioni sia agli utenti e da una forte interazione tra questi ultimi, grazie all'aspetto *social* che completa nella maggior parte dei casi questi siti. Scopo di questo articolo è innanzitutto l'analisi delle principali caratteristiche (le cosiddette *feature*) collegate ai siti di recensioni che possono essere prese in considerazione come criteri oggettivi per la valutazione della credibilità dell'informazione (ad esempio il numero di amici di un utente, la valutazione fornita dall'utente rispetto alla valutazione media, la lunghezza della *review*, ecc.). Come caso di studio è stato considerato il sito di recensioni Yelp e successivamente, sulla base delle caratteristiche (testuali e non testuali) estratte da tale piattaforma e attraverso l'applicazione di metodi di apprendimento automatico supervisionato ad un insieme di recensioni, è stata fornita una classificazione di queste ultime sulla base della loro (possibile) veridicità/falsità. L'approccio proposto è stato valutato in termini di efficacia; nonostante i risultati ottenuti siano incoraggianti, è emerso come ci sia spazio per possibili miglioramenti. Si è dunque voluto investigare come l'utilizzo di *language model* che rappresentano il contenuto delle recensioni possa avere una influenza sulla valutazione della credibilità delle stesse. Questa ipotesi è stata presentata come *proof of concept* e verrà sviluppata in futuri lavori di ricerca.

L'articolo è strutturato come segue: in Sezione 2 vengono illustrati gli approcci proposti in letteratura per l'identificazione di falsa informazione, in particolare nei siti di recensioni. La Sezione 3 descrive il funzionamento e le caratteristiche principali della piattaforma prescelta. In Sezione 4 viene illustrato l'approccio proposto per la classificazione delle recensioni, mentre in Sezione 5 viene illustrato il possibile contributo dell'analisi del linguaggio nella valutazione della credibilità. Infine, in Sezione 6, vengono illustrate le conclusioni e discussi gli sviluppi futuri.

2 Lavori correlati

Nel corso degli anni sono stati proposti differenti approcci in letteratura per la valutazione automatica della credibilità di UGC in diversi Social Media [1] [2] [3]. Come illustrato precedentemente, in questo articolo si affronta in particolare il problema dell'identificazione di *fake review* (recensioni false) nei *review site*. Esso

riguarda spesso lo studio di caratteristiche linguistiche (*language features*) legate al contenuto della recensione, caratteristiche comportamentali (*behavioural features*) relative al comportamento dei recensori o a loro peculiarità, caratteristiche sociali (*social features*) rispetto alle interazioni tra utenti, utilizzate in associazione con tecniche di apprendimento automatico (sia supervisionato che non supervisionato). Tra i lavori che considerano caratteristiche testuali, Liu *et al.* in [4] propongono un sistema di rilevamento automatico di recensioni false su Amazon.com, basato unicamente sulla rilevazione di recensioni duplicate nei profili degli utenti. In [5], Ott *et al.* sviluppano un approccio basato su caratteristiche psicolinguistiche e *n-grammi* definendo al tempo stesso il primo *dataset* di grandi dimensioni da utilizzare come *gold standard*. Altri approcci recenti che utilizzano caratteristiche linguistiche e techiche di *machine learning* supervisionato sono [6,7,8]. In [9,10,11] vengono illustrati alcuni interessanti approcci che utilizzano tecniche di apprendimento automatico semi-supervisionato e *feature* esclusivamente linguistiche. Li *et al.* in [12] propongono un approccio basato su tecniche di apprendimento automatico supervisionato in cui le *feature* coinvolte comprendono sia attributi legati al contenuto testuale della recensione sia al profilo dell'utente. Il lavoro di letteratura che ha ottenuto i migliori risultati e che si avvicina maggiormente all'approccio proposto in questo articolo è quello proposto da Mukherjee *et al.* in [13]. Vengono utilizzate caratteristiche sia comportamentali sia linguistiche in associazione con tecniche di apprendimento supervisionato e viene presa come *gold standard* la classificazione delle recensioni effettuata da Yelp.

2.1 Il problema della valutazione della classificazione

La maggior parte degli approcci sopracitati che considerano caratteristiche multiple e che forniscono migliori risultati sono basati principalmente su tecniche di apprendimento automatico di tipo supervisionato. Le valutazioni vengono condotte su *dataset* etichettati (rispetto alla credibilità), ovvero insiemi di dati la cui credibilità sia nota. Da questo punto di vista è necessario sottolineare come nel caso dei siti di recensioni non esistano *dataset* di *review* in cui la classificazione in veritiere o meno possa essere garantita affidabile al 100% a priori. Per questa ragione in letteratura sono state considerate diverse soluzioni in cui, nella maggior parte dei casi, ci si basa su *pseudo-recensioni* etichettate anzichè su vere recensioni. In [4] ad esempio si assume che le recensioni duplicate siano false, non considerano che anche gli utenti onesti a volte producono delle recensioni simili, e che quelli malintenzionati spesso attingono dai testi di recensioni vere per produrne di false. In [12] le recensioni vengono classificate manualmente; ciò non permette di creare un *gold standard* di dimensioni significative per le analisi. Ott. *et al.* in [5] costruiscono il primo *dataset* di dimensioni maggiori, attraverso l'impiego dagli *Amazon Mechanical Turkers*. È stato tuttavia dimostrato in [13] che anche questa soluzione non è efficace nel momento in cui viene utilizzata per classificare recensioni vere. Nello stesso lavoro viene proposta la soluzione di sfruttare il sistema di categorizzazione fornito da Yelp. L'algoritmo è proprietario e non pubblicamente accessibile; è tuttavia in grado di fornire una suddivisione

in recensioni *recommended* e *not recommended*. Seppur non si possano conoscere i criteri con cui viene attuata questa suddivisione e consci del fatto che essa possa fornire solo un riscontro parziale, si è deciso di considerare questa soluzione nell'utilizzo di tecniche di apprendimento automatico supervisionato, in quanto in grado di fornire un grande numero di recensioni (reali) classificate.

3 Il sito di recensioni Yelp

Yelp.com è un'azienda multinazionale americana che opera attualmente in 32 paesi nel mondo; alla fine del 2015 i suoi utenti hanno prodotto circa 90 milioni di recensioni relative ad attività commerciali o servizi. Per la pubblicazione di una recensione è richiesta la registrazione al sito; ciò non è necessario per la consultazione, che è libera. Tre sono le principali *entità* che caratterizzano Yelp: l'*attività commerciale* (*business*), la *recensione* (*review*) e l'*utente* (*user*).

Business Per ogni attività commerciale Yelp permette di visualizzare una serie di informazioni organizzate in diverse sezioni della pagina dedicata, tra cui:

- il nome dell'attività commerciale,
- l'indirizzo,
- il recapito telefonico e l'indirizzo e-mail,
- il numero di recensioni pubblicate dagli utenti e il numero medio di stelle[1],
- l'orario di apertura e la possibilità di consultare il menù online,
- la fascia di prezzo (tra "$" e "$$$$").
- le cosiddette *more business info*, ad esempio la possibilità di ordinare take-away, se il ristorante accetta prenotazioni o se è disponibile un parcheggio.

Review A sua volta, ogni recensione è caratterizzata da:

- il testo della recensione,
- la valutazione espressa in stelle (per la singola recensione),
- la data di quando è stata effettuata la recensione,
- le foto eventualmente scattate dall'utente che ha redatto la recensione,
- il *check-in*, che identifica la propria presenza nel locale attivando il GPS se la recensione viene effettuata tramite dispositivo mobile,
- i *compliments*, attraverso i quali chi consulta una *review* ha la possibilità di attribuire un complimento alla stessa, nel caso in cui sia stata utile (*useful*), divertente (*funny*) o bella (*cool*). Tutti gli utenti possono attribuire un complimento ad una recensione, anche quelli non registrati al sito.

User Ogni utente è caratterizzato da una pagina profilo su Yelp. La quantità di informazioni presenti su ogni profilo può variare sia in base al livello di arricchimento che lo *user* stesso vuole dare alla propria pagina, sia al livello di "partecipazione" dell'utente sul sito. Attraverso questa pagina è possibile visualizzare:

[1] Le stelle (da 1 a 5) vengono utilizzate nelle *review* per valutare il *business*.

– il nome o il *nickname* dell'utente,
– la lista e il numero di amici dell'utente e dei suoi *follower*,
– tutte le sue recensioni,
– le foto scattate in un *business* e allegate alle recensioni,
– i *compliments* (relativi all'utente) ricevuti dagli altri utenti.
– i *tips*, vale a dire tutti i suggerimenti pubblicati[2],
– i *bookmark*, cioè la lista di tutte le attività a cui si è assegnato un segnalibro,
– le *list*, che raccolgono tutte le attività per le quali l'utente ha scritto almeno una recensione, suddivise per categoria,
– l'eventuale appartenenza al gruppo di utenti *Elite Squad*: utenti molto attivi sul sito ed in qualche modo "certificati" (le cui recensioni sono molto spesso affidabili).

3.1 Il processo di estrazione e memorizzazione dei dati

Le informazioni che caratterizzano le tre entità sono state estratte dal sito di Yelp attraverso un processo di *crawling* e memorizzate in un database. Per effettuare il *crawling* si è adottata una tecnica di *Web scraping*, ovvero il reperimento dei dati direttamente dalle pagine HTML. In particolare si è utilizzata la libreria software JSoup[3] e le informazioni estratte sono state memorizzate in un DBMS relazionale SQLite[4]. In particolare, impostando la seguente ricerca: "Ristoranti italiani nell'area geografica di New York City", è stato possibile estrarre e memorizzare le informazioni sopracitate per il seguente numero di entità:

– 1.000 ristoranti;
– 4.604 utenti;
– 278.692 recensioni (di cui 13.905 per le quali si conosce la classificazione di Yelp: 7.514 *recommended* e 6.391 *not recommended*).

La differenza tra il numero totale di recensioni scaricate e quelle di cui si conosce la classificazione è determinata dalle limitazioni imposte da Yelp. Nella fase di definizione dell'approccio per la classificazione delle recensioni, le informazioni (e quindi le *feature*) prese in considerazione per ogni entità sono state ridotte sulla base della loro significatività rispetto alla credibilità. Il processo di selezione delle *feature* sarà illustrato nella prossima sezione.

4 L'approccio di classificazione proposto

In letteratura, gli approcci che hanno dato sinora i migliori risultati nella classificazione di recensioni in base al loro livello di veridicità utilizzano tecniche di apprendimento automatico supervisionato [14], in cui si dispone di una classificazione nota a priori. In questo articolo si è scelto di utilizzare le stesse tecniche,

[2] I *tips* sono delle brevissime recensioni in cui ci si focalizza su un aspetto particolare del *business* recensito.
[3] http://jsoup.org
[4] http://www.sqlite.org

tenendo in considerazione un numero maggiore di *feature* rispetto a quelle utilizzate in letteratura, e con configurazioni diverse, come verrà illustrato in Sezione 4.1. Nel caso di Yelp è stato selezionato un sottoinsieme di *feature* estratte dal sito relative alle entità *user*, *business* e *review*. Come classificazione nota è stata considerata quella fornita direttamente da Yelp, che identifica recensioni *recommended* e *not recommended*.

Tra i diversi algoritmi di apprendimento automatico supervisionato, in questo articolo si è considerato il modello basato su macchine a vettori di supporto, meglio conosciute con il nome di *Support Vector Machine* (SVM), utilizzate con successo in [12,13,9] per la valutazione della credibilità di recensioni. In particolare, come dimostrato in [5] e [13], l'impiego di SVM lineari rappresenta la scelta migliore per un problema di classificazione come quello affrontato nel contesto della credibilità. Per lo sviluppo del sistema di apprendimento automatico si è scelto di utilizzare il linguaggio Python, in particolare la libreria *scikit-learn*[5], basata su altre due librerie molto note per operazioni su dati, calcoli statistici e matematici: *NumPy*[6] e *SciPy*[7]. Le recensioni memorizzate nel *dataset* descritto in Sezione 3.1 sono state suddivise tra *training-set* e *test-set* come segue:

- *training-set*: 9.908 recensioni etichettate, di cui 5.514 recensioni *recommended* e 4.394 recensioni *not recommended*;
- *test-set*: 3.997 recensioni etichettate, di cui 2.000 recensioni *recommeded* e 1.997 recensioni *not recommended*.

4.1 Selezione delle feature

Come illustrato in Sezione 2, altri approcci in letteratura si sono occupati della scelta di quali *feature* utilizzare per la valutazione della credibilità delle recensioni. Nell'approccio proposto, a differenza dei lavori precedenti, è stato aumentato il numero e il tipo delle caratteristiche coinvolte, tenendo in considerazione (*i*) sia le caratteristiche non testuali collegate alle entità prese in esame (*business*, *review* e *user*), (*ii*) sia le caratteristiche che si possono estrarre dal testo delle recensioni. Di quest'ultimo sono stati analizzati sia aspetti strutturali (il modo in cui è stato redatto il testo, le sua dimensione e la suddivisione in periodi), sia aspetti semantici, legati al significato e al messaggio che il testo vuole trasmettere. In particolare, alcune di queste *feature* sono state prese dalla letteratura, altre sono state scelte sulla base dell'analisi del loro valore medio riespetto alla classificazione delle recensioni in *recommended* e *not recommended* fornita da Yelp, come illustrato in Tabella 1. Per gli *utenti* sono state scelte:

- *nr_review*: il numero di review scritte da un utente, che può indicare il livello di "attività" di un utente all'interno del social media;
- *nr_friends*: il numero medio di legami di "amicizia" con altri utenti, proporzionale al livello di affidabilità della fonte. Chi crea un profilo appositamente

[5] http://scikit-learn.org/stable/
[6] http://www.numpy.org/
[7] http://www.scipy.org/

Tabella 1. I valori medi delle *feature* prese in considerazione rispetto alla classificazione effettuata da Yelp.

User			**Review**		
Features	*Rec*	*Not Rec*	*Features*	*Rec*	*Not rec*
nr_review	108,8	11,8	*rating*	3,7	3,8
nr_friends	73,7	7,1	*extreme_rating*	28%	43%
nr_followers	0,04	0	*check-in*	43%	8%
nr_photos	84,3	3,9	*polarity*	0,22	0,28
nr_tips	14,1	0,4	*subjectivity*	0,36	0,58
nr_compliments	425	77	*text_length*	664 car.	385 car.
elite_squad	23,2%	0,04%	*nr_sentences*	8,8	5,4
profile_picture	78,9%	44%			

per promuovere o denigrare un *business* non detiene infatti un alto numero di relazioni;
- *nr_followers*: vale lo stesso discorso fatto per il numero di amici;
- *nr_photos*: il numero di foto amatoriali scattate da un utente nel *business* recensito, che dovrebbe provare il fatto di averlo effettivamente visitato;
- *nr_tips*: il numero di suggerimenti pubblicati;
- *nr_compliments*: il numero di "complimenti" ricevuti dagli altri utenti (*useful*, *funny* e *cool*). Tale valore può dare un'idea dell'opinione che altre persone hanno sull'utente della cui recensione si sta analizzando la credibilità;
- *elite_squad*: l'appartenenza ad un gruppo *Elite Squad* può essere un buon indicatore del livello di reputazione di un utente e della veridicità delle sue opinioni sui ristoranti;
- *profile_picture*: la presenza della foto di profilo. Chi crea un account per produrre contenuto falso solitamente non vuole rivelare la propria identità con una foto, o comunque non si preoccupa nemmeno di metterne una fittizia. Ciò è stato confermato dall'analisi del valore medio di questa caratteristica nella classificazione di Yelp per le recensioni etichettate come *not recommended*.

Le *feature* selezionate per le *recensioni* sono:

- *rating*: il numero di stelle attribuite dall'utente alla recensione;
- *extreme_rat*: l'indicazione che il voto attribuito alla recensione ha un valore estremo (1 stella o 5 stelle). Le recensioni false sono spesso caratterizzate da un giudizio eccessivamente positivo o negativo;
- *check-in*: l'indicazione che l'utente ha effettuato il *check-in* nell'attività commerciale tramite la geolocalizzazione (e quindi dovrebbe attestare l'effettiva presenza nel *business* recensito);
- *polarity*: è un indice che rientra nella disciplina della *sentiment analysis*. Essa racchiude tutte quelle tecniche di elaborazione del linguaggio per estrarre informazioni di carattere semantico. La polarità, in particolare, indica se il testo ha un'accezione negativa, neutra o positiva;
- *subjectivity*: tale valore di *sentiment analysis* indica il livello di "oggettività" e "soggettività" di un testo e può essere utile per distinguere descrizioni

oggettive di ciò che si sta recensendo (un'opinione avvalorata da fatti) da giudizi puramente personali o con uno scarso riscontro nella realtà;
- *text_length*: la lunghezza del testo, intesa come numero di caratteri;
- *nr_sentences*: il numero di frasi, dove ogni frase è rappresentata da una porzione di testo delimitata da punti.

Sulla base di queste *feature* sono state condotte valutazioni preliminari attraverso due sperimentazioni distinte, scegliendo due gruppi di caratteristiche:

1. Nel prima sperimentazione si sono considerate quelle *feature* legate più ai metadati collegati alle recensioni e agli utenti piuttosto che al testo delle recensioni, vale a dire: *rating*, *avg_rating* (il numero medio di stelle assegnate da un utente, calcolato da tutte le recensioni da lui pubblicate), *extreme_rat*, *check-in*, *nr_review*, *nr_friends*, *nr_followers*, *nr_photos*, *nr_tips*, *nr_compliments*, *elite_squad*, *profile_picture*.
2. Nella seconda sperimentazione, alle *feature* considerate precedentemente, sono state aggiunte le caratteristiche legate al testo della recensione: *text_length*, *nr_sentences*, *polarity*, *subjectivity*.

Per il calcolo della polarità e della soggettività è stata impiegata la libreria *TextBlob*[8] di Python, che fornisce un set di funzioni di *sentiment analysis* (`sentiment.polarity` e `sentiment.subjetity`), basate su tecniche di *natural language processing* (NLP), atte a tale scopo. Le funzioni restituiscono dei valori compresi nell'intervallo $[-1, 1]$; nel caso della polarità, 1 rappresenta il massimo valore di positività, e -1 di negatività; nel caso della soggettività, 1 ne rappresenta il massimo valore, mentre -1 indica un'alta oggettività del testo. Ad esempio, un documento con *polarity* = -0.86 e *subjetivity* = 0.7 è caratterizzato da polarità molto negativa e da un livello di soggettività molto alto.

4.2 Valutazioni preliminari

Per la valutazione dell'efficacia del sistema di apprendimento automatico, e delle *feature* utilizzate, sono stati presi in esame una serie di parametri noti in Information Retrieval, chiamati indici di prestazioni (*performance evaluation indexes*). In particolare sono stati presi in considerazione: *precisione*, *richiamo*, *accuratezza*, *specificità* e *f1-score*. In Tabella 2 vengono confrontati i valori ottenuti dall'approccio presentato in questo articolo a quelli ottenuti dai principali lavori di letteratura che hanno applicato algoritmi di apprendimento automatico per la classificazione di recensioni, da Yelp in particolare [13,9,15]. È necessario considerare che in [15] vengono impiegate tecniche di apprendimento non supervisionato, al contrario che negli altri due approcci. Per questo motivo i risultati sono meno soddisfacenti rispetto a [13,9]. Risulta chiaro come a livello di precisione, i risultati ottenuti dall'approccio proposto siano apprezzabili. Al contrario, il livello di richiamo è inferiore rispetto agli altri approcci (in particolare rispetto a [13]). Di conseguenza anche il valore *f1-score* è inferiore. Nessuno dei tre

[8] https://textblob.readthedocs.org/en/dev/

approcci precedenti ha preso in considerazione l'indice di specificità dei risultati, mentre soltanto in [13] è stato calcolato il livello di accuratezza. Quest'ultimo valore risulta, anche in questo caso, inferiore. Pertanto, malgrado la precisione ottenuta sia in generale al di sopra della media, non sono stati raggiunti globalmente risultati soddisfacenti per quanto riguarda gli altri indici. La causa del problema è imputabile all'elevato numero di *falsi negativi* generati, ovvero tutti quei campioni positivi del *test-set* giudicati negativi dalla macchina. Per campioni positivi si intendono le recensioni classificate come *not recommended*, mentre i campioni negativi sono costituiti dalle recensioni *recommended*.

Tabella 2. Confronto degli indici di prestazioni dell'approccio proposto con quelli prodotti dai lavori in letteratura.

	Precisione	*Richiamo*	*F1-score*	*Accuratezza*
Mukherjee et al. 2013	0,84	0,87	0,85	0,86
Li et al. 2014	0,59	0,89	0,71	-
Mukherjee et al. 2014	0,56	0,63	0,59	-
Approccio proposto	**0,9**	**0,53**	**0,66**	**0,73**

5 L'analisi del linguaggio nella valutazione della credibilità delle recensioni

In questa sezione viene illustrato un esperimento che cerca di "raffinare" il giudizio del classificatore SVM e quindi diminuire il numero di *falsi negativi* grazie all'analisi del linguaggio degli utenti delle recensioni classificate erroneamente. Pur essendo consapevoli che i risultati prodotti dall'algoritmo di classificazione debbano essere unicamente utilizzati per valutare l'approccio proposto, l'esperimento svolto ha avuto come unico scopo la rapida verifica della potenziale importanza dell'analisi del linguaggio utlizzato nelle recensioni per valutarne la credibilità. Un approccio sistematico ed esteso all'incorporazione di vari tipi di analisi effettuate sul linguaggio sarà l'oggetto di futuri lavori.

L'esperimento è basato su *language model*, in particolare sul modello descritto da Croft, Metzler e Strohman in [16], ovvero il *query likelihood ranking*: un modello di linguaggio è costruito per ogni *documento* in una collezione e l'ordinamento dei documenti è effettuato rispetto alla probabilità di questi ultimi di generare una determinata *query*. Formalmente, per calcolare la probabilità $P(q|d)$ dove $q = \{q_1, q_2, \ldots, q_n\}$ rappresenta la query e le parole che la compongono e d rappresenta il documento, è stata utilizzata la seguente formula:

$$\log P(q|d) = \sum_{i=1}^{n} \log \frac{f_{q_i,d} + \mu \frac{c_{q_i}}{|C|}}{|d| + \mu}$$

dove f_{q_i}, d rappresenta la frequenza con cui il termine q_i appare nel documento d, $|d|$ è il numero di termini in d, $\frac{c_{q_i}}{|C|}$ è la stima della probabilità del *language*

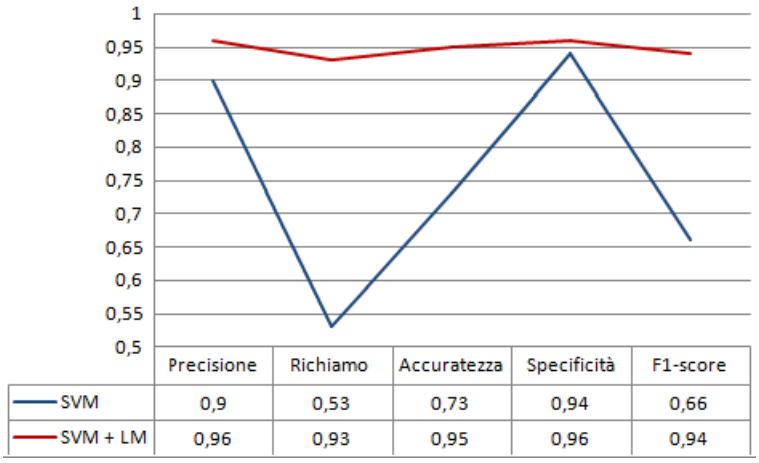
model della collezione per il termine q_i, dove c_{q_i} è il numero di volte che il termine della query appare nella collezione C di documenti, e $|C|$ è il numero totale di occorrenze dei termini nella collezione. Infine, μ è una costante il cui valore viene settato empiricamente tra 1,000 e 2,000 (in base ai risultati migliori ottenuti negli esperimenti TREC).

In questo esperimento si considera come *query* una recensione nel *test-set*. In questo modo è possibile effettuare un confronto tra il modello di una recensione nel *training-set* e una recensione nel *test-set*. In particolare è possibile calcolare la probabilità che una recensione nel *test-set* possa essere generata da una recensione nel *training-set*, sia che essa sia etichettata come *recommended* o *not recommended*. Come precedentemente descritto, il problema principale del basso di livello di richiamo ed accuratezza manifestato nel fase di analisi descritta in Sezione 4.2 è da imputare alla presenza di un elevato numero di *falsi negativi*. L'idea è quindi quella di prendere in considerazione i casi di test etichettati come *falsi negativi* e, tramite i *language model*, verificare la possibilità di diminuire gli "errori" generati dalla SVM. Lo scopo è quindi quello di portare i campioni erroneamente giudicati come *recommended* ad essere attribuiti alla classe *not recommended* (aumentando il valore dei *veri positivi*). Lo stesso approccio è stato applicato ai *falsi positivi*, ovvero quei campioni del *test-set* che la macchina a vettori di supporto ha giudicato *not recommended*, malgrado non lo fossero.

5.1 Valutazione dell'esperimento

Date 100 recensioni dal *training-set* (50 *recommended* e 50 *not recommended*), di ognuna di queste ne viene fornita una rappresentazione formale (un file contenente le parole presenti nel testo con la relativa frequenza). In seguito, per ognuna delle recensioni del *test-set* che in seguito alla classificazione sono rientrate nei *falsi negativi* o nei *falsi positivi*, è stata calcolata la probabilità di essere generata da ognuna delle cento recensioni prese a campione dal *training-set*. Si ottengono in questo modo, per ogni recensione del *test-set*, cento *score*, uno per ognuna delle cento recensioni nel *training-set*. Questi valori sono stati impiegati per *confermare* o *smentire* il giudizio della classificazione prodotta dalla SVM riguardo alla recensione in esame. Specificatamente, vengono aggregati separatamente gli *score* prodotti dal calcolo della probabilità della recensione di essere stata generata dalle cinquanta recensioni *recommended*, e dalle cinquanta *not recommended*. Il valore più alto prodotto dalle distinte somme determina l'appartenenza della recensione presa in esame ad una delle due classi, andando in questo modo a smentire o confermare il responso della *support vector machine*.

Tramite questo esperimento di "raffinamento" della classificazione basato sui *language model* è stato possibile correggere 33 falsi positivi e 813 falsi negativi, ottenendo così 1.865 *veri positivi*, 1.922 *veri negativi*, 78 *falsi positivi* e 132 *falsi negativi*. In questo modo i valori degli indici di prestazioni hanno subito un netto miglioramento, anche rispetto ad *accuratezza* e *richiamo*. In Figura 1 viene mostrato, a mero titolo di *proof of concept*, il confronto tra i risultati ottenuti con la tecnica di *machine learning*, e quelli dell'esperimento basato sul raffinamento della classificazione SVM tramite *language model*.



	Precisione	Richiamo	Accuratezza	Specificità	F1-score
SVM	0,9	0,53	0,73	0,94	0,66
SVM + LM	0,96	0,93	0,95	0,96	0,94

Figura 1. Confronto degli indici di prestazioni dell'approccio basato su SVM e dell'esperimento basato su SVM + *language model*.

6 Conclusioni e sviluppi futuri

Con lo sviluppo del Web sociale e la diffusione dei Social Media agli utenti è stata data la possibilità di generare contenuto – il cosiddetto *User Generated Content* (UGC) – e diffonderlo senza il controllo di intermediari o terze parti ritenute affidabili, come succedeva prima dell'avvento del Web 2.0. Ciò comporta che pochi siano i mezzi oggettivi a disposizione di chi si trovi a dover/voler valutare la credibilità dell'informazione sui Social Media. Emerge quindi il problema di fornire strumenti automatici che aiutino gli utenti in questo compito. Tra le diverse piattaforme che permettono la diffusione di UGC, i siti di recensioni rappresentano uno scenario particolarmente interessante in cui investigare il problema della veridicità delle *review* che vi sono pubblicate. In questo articolo, prendendo in considerazione il sito di recensioni Yelp, è stato presentato un approccio basato su tecniche di apprendimento automatico supervisionato per la classificazione di recensioni. Rispetto ad altri approcci in letteratura, è stato utilizzato un numero maggiore di caratteristiche (*features*) per la fase di apprendimento; tali caratteristiche prendono in considerazione sia il contenuto delle recensioni, sia i metadati associati alle recensioni e agli utenti che le hanno prodotte. È stato inoltre illustrato un esperimento basato sui *language model* per verificare l'influenza del linguaggio utilizzato nelle recensioni nella valutazione della credibilità, raffinando i risultati della classificazione prodotti dall'approccio *machine learning*. Per il futuro si è pensato di applicare l'approccio proposto ad altri contesti e Social Media, di sperimentare tecniche di apprendimento automatico non supervisionato o semi-supervisionato per superare il problema legato ad un *dataset* di dati preclassificati, o di applicare i modelli di linguaggio al modello stesso, utilizzando per esempio i valori di probabilità ottenuti come ulteriori *feature* da utilizzare durante il processo di apprendimento.

Riferimenti bibliografici

1. Kaplan, A.M., Haenlein, M.: Users of the world, unite! the challenges and opportunities of social media. Business Horizons **53**(1) (2010) 59 – 68
2. Abbasi, M.A., Liu, H.: Measuring user credibility in social media. In: Social Computing, Behavioral-Cultural Modeling and Prediction. Springer (2013) 441–448
3. Westerman, D., Spence, P.R., Van Der Heide, B.: Social media as information source: Recency of updates and credibility of information. Journal of Computer-Mediated Communication **19**(2) (2014) 171–183
4. Jindal, N., Liu, B.: Opinion spam and analysis. In: Proceedings of the 2008 International Conference on Web Search and Data Mining, ACM (2008) 219–230
5. Ott, M., Choi, Y., Cardie, C., Hancock, J.T.: Finding deceptive opinion spam by any stretch of the imagination. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Association for Computational Linguistics (2011) 309–319
6. Ott, M., Cardie, C., Hancock, J.T.: Negative deceptive opinion spam. In: HLT-NAACL. (2013) 497–501
7. Shojaee, S., Murad, M.A.A., Azman, A.B., Sharef, N.M., Nadali, S.: Detecting deceptive reviews using lexical and syntactic features. In: Intelligent Systems Design and Applications (ISDA), 2013 13th International Conference on. (Dec 2013) 53–58
8. Li, J., Ott, M., Cardie, C., Hovy, E.H.: Towards a general rule for identifying deceptive opinion spam. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers, The Association for Computer Linguistics (2014) 1566–1576
9. Li, H., Liu, B., Mukherjee, A., Shao, J.: Spotting fake reviews using positive-unlabeled learning. Computación y Sistemas **18**(3) (2014) 467–475
10. Ren, Y., Ji, D., Zhang, H.: Positive unlabeled learning for deceptive reviews detection. In: EMNLP. (2014) 488–498
11. Hernández Fusilier, D., Montes-y Gómez, M., Rosso, P., Guzmán Cabrera, R.: Detecting positive and negative deceptive opinions using pu-learning. Inf. Process. Manage. **51**(4) (July 2015) 433–443
12. Li, F., Huang, M., Yang, Y., Zhu, X.: Learning to identify review spam. In: IJCAI Proceedings-International Joint Conference on Artificial Intelligence. Volume 22. (2011) 2488
13. Mukherjee, A., Venkataraman, V., Liu, B., Glance, N.S.: What yelp fake review filter might be doing? In: ICWSM. (2013)
14. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of machine learning. MIT press (2012)
15. Mukherjee, A., Venkataraman, V.: Opinion spam detection: An unsupervised approach using generative models. Technical report, UH-CS-TR-2014 (2014)
16. Croft, W.B., Metzler, D., Strohman, T.: Search engines: Information retrieval in practice. Volume 283. Addison-Wesley Reading (2010)