Università degli studi di Pisa

Corso di laurea magistrale in

Informatica Umanistica

Relazione finale del Seminario di Cultura Digitale

TECNICHE LINGUISTICO-COMPUTAZIONALI PER IL RILEVAMENTO DI GIUDIZI FALSI

Marzia Giardiello

n. Matricola: 305654

1. INTRODUZIONE

Nel corso degli ultimi dieci anni, le tecniche di Natural Language Processing (NLP) combinate con gli algoritmi di apprendimento automatico hanno iniziato ad essere utilizzate per indagare sulla "forma" di un testo piuttosto che sul contenuto. La gamma di compiti, che condividono questo approccio all'analisi dei testi, è ampia, variando da task come il riconoscimento del genere testuale, l'analisi della leggibilità (ossia definire se un testo è o meno difficile da comprendere), il riconoscimento della lingua madre di chi ha prodotto uno scritto (in inglese per ora) in una lingua differente dalla propria L1, il rilevamento di giudizi falsi nelle recensioni di attività commerciali e turistiche, la valutazione delle competenze linguistiche dello scrivente e il riconoscimento dello stile di un autore. Tutti questi obiettivi derivano da un macro task che consiste nella ricostruzione del profilo linguistico di un testo.

I problemi tipicamente trattati in questo tipo di studi possono essere riassunti in due principali domande di ricerca intese a indagare 1) quali caratteristiche linguistiche funzionano meglio per un determinato compito, e 2) quale tipo di algoritmo di apprendimento automatico è più adatto per un certo task.

Al cuore del compito di accedere alla struttura linguistica del testo c'è comunque la catena di analisi linguistica, i cui passaggi principali sono:

- **La segmentazione delle frasi, la tokenizzazione e la lemmatizzazione**
  [Tokenizzare un testo significa dividere le sequenze di caratteri in unità minime di analisi dette "token": parole, punteggiatura, date, numeri, sigle, ecc. I token possono essere anche entità strutturalmente complesse (es. date), ma sono comunque assunte come unità di base per i successivi livelli di elaborazione (morfologico, sintattico ecc.). La nozione di token è distinta da quella di parola: le parole sono solo un sottoinsieme di token].
  [Lemmatizzare un testo significa attuare il processo di riduzione di una forma flessa di una parola alla sua forma canonica, detta "lemma"]

- **L'annotazione morfosintattica**
  [Ad ogni token viene associata l'informazione della categoria grammaticale che la parola ha nel contesto, più il relativo lemma]

- **L'annotazione sintattica a dipendenze**
  [Analisi della struttura sintattica della frase in termini di relazioni di dipendenza (es. soggetto, oggetto, etc.)]

Il profilo linguistico costituisce, dunque, il punto di partenza per accedere alla struttura linguistica del testo: a partire da un corpus annotato linguisticamente in maniera automatica è possibile estrarre una serie di caratteristiche che sono rappresentative dell'informazione linguistica. Sulla base di queste caratteristiche viene ricostruito il profilo linguistico. Tra queste caratteristiche, esistono le caratteristiche basilari, come:

- Il calcolo della lunghezza della frase calcolata in numero medio di token per frase;
- Il calcolo della lunghezza delle parole calcolate in numero medio di caratteri per parola.

Caratteristiche più complesse come quelle lessicali:

- Gli indici di ricchezza lessicale, (Type/Token Ratio) ossia il rapporto tra numero di parole tipo in un testo (dizionario) e il numero di occorrenze totali di parole (unità del dizionario).
- Il calcolo della percentuale del vocabolario del testo appartenente al vocabolario di base[1].

Caratteristiche morfosintattiche:

- La distribuzione delle categorie morfosintattiche (grammaticali).
- La densità lessicale: ossia rapporto tra parole di contenuto (verbi, sostantivi, aggettivi e avverbi) e numero totale di token lessicali in un testo.
- Il modo, il tempo e la persona dei verbi (distribuzione dei verbi in base a queste caratteristiche).

Caratteristiche sintattiche:

- La distribuzione di link di dipendenza sintattica (distribuzione dei diversi tipi di dipendenze sintattiche ad es. soggetto, oggetto diretto, modificatore, ecc).
- Le strutture sintattiche: altezza dell'albero sintattico del testo; catene di complementi preposizionali (complementi che modificano il nome).
- La distribuzione di frasi subordinate rispetto alla distribuzione di frasi principali.
- La distribuzione delle subordinate rispetto alla principale.

Dal profilo linguistico, dunque, è possibile estrarre caratteristiche utili a svolgere una serie di compiti che sono di grande uso in vari contesti.

In questa relazione verrà posta attenzione sull'**identificazione di plagi**, in particolare di recensioni false presenti sui siti web commerciali e di promozione turistica. Questa relazione, nello specifico, si focalizza sul contributo dei ricercatori dei dipartimenti di informatica e di comunicazione della Cornell University di Ithaca, New York. [2]

[1] Vocabolario di base della lingua italiano De Mauro (2000) per le ricerche condotte dall'Istituto di Linguistica Computazionale del CNR di Pisa.

[2] Myle Ott; Yejin Choi; Claire Cardie. [Department of Computer Science Cornell University Ithaca, NY 14853]
Jeffrey T. Hancock. [Department of Communication Cornell University Ithaca, NY 14853]

2. RECENSIONI FALSE

Negli ultimi anni, l'individuazione delle recensioni false ha attirato notevole attenzione sia da parte delle imprese che dalle comunità di ricerca. Per recensioni che riflettono esperienze e pareri genuini da parte di utenti, appare ovvio che la rilevazione dei giudizi falsi rappresenta un fenomeno negativo importante.

I consumatori, sempre più frequentemente, ricercano, giudicano e recensiscono prodotti online. Di conseguenza, i siti web come Yelp[3] o il famosissimo TripAdvisor[4], portale web di viaggi che pubblica le recensioni degli utenti su alberghi, ristoranti e attrazioni turistiche che contengono appunto i giudizi dei consumatori, stanno diventando obiettivi del cosiddetto *opinion spam*. Per opinion spam si intende l'attività illegale attraverso cui si cerca di trarre in inganno i lettori o formulando pareri positivi immeritevoli nei confronti di alcune attività al fine di promuoverle, o dando false opinioni negative ad altre strutture al fine di danneggiare la loro reputazione.

L'obiettivo di questa relazione è di mettere in luce le tecniche linguistico-computazionali per stanare le opinioni fittizie e ingannevoli che sono state deliberatamente scritte, per apparire autentiche.

2.1 CONTROVERSIE E INCHIESTE

Prima di esplorare le tecniche linguistico-computazionali utili ad individuare in maniera automatica i giudizi fasulli è interessante soffermarsi sulle tante controversie susseguitesi negli ultimi anni nei confronti di questa pratica, ormai, diffusa.

Molti quotidiani italiani e internazionali hanno dedicato spazio a varie notizie, riguardanti casi di forte richiamo.

Il 4 ottobre 2011, il quotidiano francese "Le Monde" ha pubblicato la notizia[5] di una condanna ai danni dei siti *Expedia* (altro sito web di viaggio), *TripAdvisor* e *Hotels.com* da parte del Tribunale di Parigi. La condanna riferiva di una maxi-multa di 430mila euro da pagare al *Synhorcat*, sindacato di ristoratori e albergatori e a due alberghi che avevano presentato istanza contro i siti sopra citati per aver messo in atto pratiche sleali e ingannevoli. La notizia fu riportata sui canali italiani attraverso un comunicato dall'associazione Federalberghi (associazione di categoria che rappresenta gli interessi delle imprese alberghiere in Italia. Il Presidente è Bernabò Bocca e il Direttore Generale di Federalberghi è Alessandro Massimo Nucara)[6] che commentò positivamente l'azione francese. L'associazione italiana ha approfittato di questa vicenda giudiziaria per scagliarsi

[3] http://www.yelp.com
[4] http://www.tripadvisor.com
[5] http://www.lemonde.fr/technologies/article/2011/10/04/expedia-condamne-pour-pratiques-deloyales-envers-les-hoteliers_1582202_651865.html
[6] www.Wikipedia.it *Federalberghi*

contro l'anonimato, garantito dai siti, che potrebbe danneggiare gravemente gli esercizi commerciali presenti sui portali. Questa battaglia verrà ripresa nel 2014 dal vicepresidente della Fipe (*Federazione Italiana Pubblici Esercizi*), Aldo Cursano, e riportata in un articolo del quotidiano "Repubblica".

Procedendo, però, in ordine cronologico è interessante anche riportare i risultati dell'esperimento di Marco Camisani Calzolari, professore di Comunicazione allo Iulm, raccontati ai cittadini italiani in un articolo[7] di *Repubblica.it.*

Lo spunto è arrivato da uno studio, pubblicato poche settimane prima, condotto proprio dai ricercatori della Cornell University che hanno scoperto l'esistenza di un vero e proprio "mercato" di recensioni sul web, pagate profumatamente dalle aziende in causa. Attraverso questo meccanismo, Marco Camisani Calzolari, è riuscito ad ottenere recensioni positive come idraulico senza mai aver svolto questo lavoro in vita sua, pagando soltanto cinquanta dollari. Sul suo sito personale, sono presenti opinioni brillanti: "*Ha risolto il mio problema giusto in tempo*" dice Himanshu; "*ti ringrazio per l'eccezionale servizio che mi ha offerto la scorsa settimana*" aggiunge Teresa e insieme a loro altre decine di finti utenti. I giudizi sono tutti di alto livello, originali e differenti in stile. La verità è che nessuno in realtà conosce il professore; ai recensori è stato semplicemente raccontato che Marco è un bravo idraulico che vuole incrementare la propria visibilità. Senza naturalmente nessun tipo di verifica.

Come funziona questo "gioco"? Si prendono contatti con i siti che forniscono recensioni ad hoc, si racconta la propria esigenza e si effettua il pagamento. Il risultato sono una serie di giudizi, guarda caso, tutti positivi. Nel caso specifico del professor Calzolari, i giudizi per la maggior parte arrivavano da falsi profili Facebook creati da società specializzate a sviluppare traffico.

Anche il "Fatto Quotidiano" ha voluto realizzare un esperimento del genere, realizzando due prove: scrivere recensioni senza limiti e scriverne a pagamento su TripAdvisor, perché il problema non riguarda solo gli utenti, ma anche i siti e le organizzazioni, facilmente individuabili con una semplice ricerca su Google, che pubblicano il costo di un pacchetto di recensioni positive. La notizia è del 26/08/2014 pubblicata sul sito della testata giornalistica.[8]

La prima prova condotta dal "Fatto Quotidiano" è stata quella della cosiddetta "recensione fotocopia", ossia un'opinione uguale data a strutture differenti nello stesso giorno. Si è trattato di un vero e proprio gioco da ragazzi: è bastato creare un account; scegliere, in questo caso specifico, quattro ristoranti in quattro città diverse e recensirli tutti nella stessa maniera, dichiarando di aver mangiato la sera precedente. E il controllo qui dov'è?

La seconda prova, quella di fare pubblicità ad altri a pagamento, è stata altrettanto portata a termine facilmente. Sui portali di annunci ci si può tranquillamente imbattersi in annunci di

---

[7] http://www.repubblica.it/economia/2012/09/28/news/recensioni_false_internet-43464174

[8] http://www.ilfattoquotidiano.it/2014/08/26/tripadvisor-dagli-annunci-al-pagamento-ecco-come-e-facile-barare-sulle-recensioni/1099337/

selezione di “collaboratori” a cui assegnare “micro-lavori di inserimento testi su alcuni portali internet”. Il “Fatto Quotidiano” ha risposto a uno di questi annunci ottenendo così l’incarico, il primo dei quali consisteva nel votare con 4 stelle un ristorante in una località toscana e recensirlo in lingua italiana. Oltre all’indicazione dell’incarico, sono seguite poi una serie di informazioni per rendere più reale il testo, come alcune caratteristiche della struttura e i punti forte del menù. Dopo qualche giorno, sul conto PayPal dell’utente, alias “Il Fatto Quotidiano”, arrivavano i 3 euro pattuiti per il servizio.

Sempre del 2014 sono altre due notizie riportale dal sito web del quotidiano “Repubblica”: la prima[9] riporta l’accusa lanciata da Federalberghi a TripAdvisor per aver pubblicato una recensione che riguarda l’hotel “Regency” di Roma. Il commento del cliente recitava più o meno così “*Buone caratteristiche dell’ascensore della struttura, impianto wi-fi perfetto, personale competente*”. Non risulterebbe nulla di strano se non fosse per la dichiarazione del cliente, che affermava di aver soggiornato in quel hotel nel marzo del 2013, senza considerare che questa struttura sia in realtà chiusa dal 2007. Questo confermerebbe proprio ciò di cui si lamentano, sempre più aspramente, le associazioni alberghiere, ossia che il sito americano non effettua neppure un minimo controllo né qualche verifica sulle opinioni; al contrario continua a ricevere responsi e giudizi “impossibili”. TripAdvisor, al tempo, si è difeso sostenendo di effettuare rigidi controlli sulle frodi adottando sia rigidi sistemi di verifica sia un team specializzato a individuare i truffatori. Il sito stesso ha, poi, minimizzato il problema sottolineando come in realtà si tratti soltanto di una percentuale minima di recensioni fasulle, questione tra l’altro smentita da Federalberghi, per il quale citando uno studio dei ricercatori di *Gartner* (centro di ricerca americano), la percentuale non sarebbe così bassa, dato che si parla di un valore compreso tra il 10% e il 14%. Per questo motivo, Federalberghi si è rivolta all'**Autorità Garante della Concorrenza e del Mercato** per chiedere che TripAdvisor adotti con urgenza misure idonee a prevenire gli abusi a danno dei consumatori.

Una soluzione a questo genere di problema l’ha proposta, come anticipato, il vicepresidente della Fipe (*Federazione Italiana Pubblici Esercizi*), Aldo Cursano, che è riuscito dopo anni di protesta ad ottenere un accordo tra TripAdvisor e i ristoratori toscani: ossia mandare un consulente a verificare, con il loro aiuto, l’integrità de contenuti online. E non solo, Cursano, come riportato anche nella seconda notizia[10] anticipata, aveva anche avanzato una proposta forte e, forse, davvero garante di onestà ossia che chiunque voglia scrivere una recensione nei confronti di un esercizio dovrebbe inizialmente identificarsi, magari caricando una fotografia dello scontrino. La società però non è disposta a mettere a rischio la privacy degli utenti e ha rifiutato. La risposta del sito di recensioni è sempre la stessa, ossia che i numeri parlano chiaro e, con 260 milioni di utenti al mese, TripAdvisor può ritenersi garante di successo e trasparenza, anche se circa 90 recensioni al minuto sono comunque difficili da controllare con precisione assoluta.

---

[9] http://www.repubblica.it/cronaca/2014/08/26/news/tripadvisor_nella_bufera_le_accuse_di_federalberghi_e_inaffidabile-94459807/

[10] http://www.repubblica.it/dal-quotidiano/r2/2014/03/17/news/la_santa_alleanza_chef-tripadvisor_smascheriamo_le_recensioni_truffa-81178743/

Le proteste del vicepresidente della Fipe venivano a seguire l'ennesimo caso di opinion spam: questa volta, però, non si è trattato di una recensione "gonfiata" bensì di una stroncatura inventata di sana pianta che, come si può immaginare, risulta essere molto più grave e lesiva della professionalità di chi riceve stroncature false, ovviamente.

"*Un pallino su cinque per le penne al pomodoro? Io non le ho mai servite* ". È quello che ha dichiarato lo chef Amerigo Capria, 35 anni del ristorante il Baccarossa a Firenze, allievo dello chef pluristellato Carlo Cracco.

Che una stroncatura faccia male è normale, ma una stroncatura falsa è di un'ingiustizia disumana soprattutto se fatta nei confronti di un professionista giovane che ha da poco avviato la propria attività. Da non sottovalutare, inoltre, che, purtroppo, quello di chef Capria non è l'unico caso riscontrato: anche altri suoi colleghi hanno sostenuto di aver notato un diffuso malcostume.

3. TECNICHE LINGUISTICHE-COMPUTAZIONALI.

In genere, i pareri ingannevoli non sono né facilmente ignorati né identificati da un lettore umano. Di conseguenza, ci sono poche buone fonti di dati etichettati per questa ricerca. Infatti, in assenza di dati gold-standard, gli studi correlati sono stati costretti a utilizzare procedure ad hoc per la valutazione. Al contrario, il contributo dei ricercatori della Cornell University, presentato in questa relazione, è la creazione del primo data-set a larga scala, a disposizione del pubblico per la ricerca dello spamming di opinioni ingannevoli, contenente 400 recensioni veritiere e 400 recensioni ingannevoli gold-standard. Su questo data-set (training set) poi è stato addestrato un classificatore che è in grado di discriminare una recensione falsa da una veritiera.

Nello specifico di questo studio, qui riportato, i compiti prefissati sono stati: (a) un task standard di categorizzazione del testo, in cui sono stati usati classificatori basati sugli n-grammi, per etichettare opinioni sia ingannevoli che veritiere; (b) un caso di rilevamento psicolinguistico di inganno, in cui ci si aspetta nelle dichiarazioni ingannevoli caratteristiche dell'effetto psicologico del mentire, come ad esempio un incremento dell'emozione negativa e il distanziamento psicologico e, (c) un problema di identificazione di genere, in cui la scrittura ingannevole e veritiera sono viste rispettivamente come sotto-generi della scrittura creativa e della scrittura informativa.

Altri ricercatori, precedentemente, si sono occupati di spam (l'invio di messaggi indesiderati, solitamente commerciali), storicamente studiato nel contesto delle e-mail e del web. Solo recentemente gli studiosi hanno cominciato a guardare allo spam di opinioni scoprendo che non solo è un fenomeno molto diffuso, bensì un fenomeno di natura diversa che si discosta sia dall'email spam che dal web spam. Fino al contributo analizzato in questo testo, venivano utilizzati i dati delle recensioni dei prodotti e, in assenza di dati standard, si addestravano i modelli utilizzando le caratteristiche basate sul testo, sul recensore e sul prodotto, più che altro per distinguere tra opinioni duplicate (considerate recensioni ingannevole) e opinioni non duplicate (considerate recensioni veritiere).

3.1 COSTRUZIONE DEL DATA-SET

Le fasi di costruzione del data-set di riferimento sono state due, una per la raccolta delle recensioni ingannevoli e una per quella delle recensioni veritiere.

3.1.1 RACCOLTA RECENSIONI INGANNEVOLI

Le opinioni ingannevoli sono state recuperate attraverso un noto servizio di crowdsourcing (modello di business nel quale un'azienda affida la progettazione, la realizzazione o lo sviluppo di un progetto, ad un insieme indefinito di persone, non organizzate precedentemente): **Amazon Mechanical Turk**[11]. I lavoratori anonimi, anche detti Turker, accettano gli incarichi a pagamento, ossia gli HIT (human intelligence task), secondo le definizioni del sito.

Per recuperare opinioni ingannevoli "gold-standard" utilizzando AMT, è stato creato un insieme di 400 hits che sono stati allocati uniformemente in 20 hotel selezionati. Per garantire che le opinioni fossero scritte da autori unici, è stata permessa una sola iscrizione per turker. Il task, inoltre, è stato ristretto a turkers che si trovano negli Stati Uniti, e che mantengono una valutazione di approvazione di almeno il 90%. Ai turkers sono stati consentiti massimo 30 minuti per lavorare sull'obiettivo, e sono stati pagati un dollaro per ogni presentazione accettata. Ogni HIT presenta il turker con il nome e il sito web di un hotel. Le istruzioni dell'hit chiedono al turker di supporre di lavorare per il reparto marketing dell'hotel, e far finta che il loro capo voglia che loro scrivano una recensione falsa (come se fossero un cliente), per essere pubblicata su un sito web di recensioni di viaggi; inoltre, la revisione deve suonare quanto più realistica possibile, per ritrarre l'hotel in una luce positiva. Una dichiarazione di non responsabilità indica che qualsiasi presentazione che possa risultare di qualità insufficiente (ad esempio, scritta per l'hotel sbagliato, incomprensibile, troppo breve, plagiata) sarà respinta. Sono stati necessari circa 14 giorni per raccogliere 400 opinioni ingannevoli soddisfacenti. La tabella 1 riassume le descrizioni statistiche. Gli argomenti variano drasticamente in lunghezza, e in tempo impiegato per il lavoro. In particolare, quasi il 12% delle presentazioni sono state completate in meno di un minuto. Sorprendentemente, la lunghezza media dei testi prodotti in meno di un minuto e dei testi prodotti spendendo oltre il minuto non rivela alcuna differenza significativa. Probabilmente, questi utenti "veloci" potrebbero aver iniziato a lavorare prima di avere formalmente accettato l'HIT, presumibilmente per aggirare il limite di tempo imposto. Infatti, la presentazione più veloce ha soli 5 secondi e contiene 114 parole.

[11] www.mturk.com

TEMPO SPESO t (minuti)	
Tutte le recensioni	Conteggio: 400 t_min: 0.08, t_max: 29.78 t_avg: 8.06
LUNGHEZZA L (parole)	
Tutte le recensioni	L_min: 25, L_max: 425 L_avg: 115.75
Tempo speso t < 1	Conteggio: 47 L_min: 39, L_max: 407 L_avg:113.94
Tempo speso t > 1	Conteggio: 353 L_min: 25, L_max: 425 L_avg:115.99

Tabella 1: Statistiche descrittive per 400 opinioni ingannevoli

3.1.2 RACCOLTA RECENSIONI VERITIERE

Per le opinioni veritiere, sono state estratte tutte le 6977 recensioni provenienti dai 20 hotel più popolari di Chicago, su TripAdvisor. Da queste sono state eliminate:

- 3130 recensioni non 5 stelle
- 41 recensioni non in inglese
- 75 recensioni con meno di 150 caratteri in quanto, per la costruzione, le opinioni ingannevoli sono di almeno 150 caratteri
- 1.607 recensioni scritte dai nuovi utenti, cioè che non hanno precedentemente inviato un parere su TripAdvisor, dal momento che queste opinioni più probabilmente contengono opinioni spam, che ridurrebbe l'integrità dei dati sulle recensioni veritiere.

Infine, è stato equilibrato il numero di pareri veritieri e di pareri ingannevoli, selezionando 400 delle rimanenti 2.124 recensioni veritiere, in modo che le lunghezze dei documenti delle recensioni veritiere selezionate venissero distribuite in modo simile a quelle delle recensioni ingannevoli.

3.2 PERFORMANCE UMANE

Valutare le prestazioni di rilevazione umana dell'inganno è necessario per convalidare i pareri ingannevoli raccolti nel data-set. Se le prestazioni umane sono basse, le opinioni fittizie sono convincenti, e quindi, meritevoli di ulteriore attenzione.

L'approccio alla valutazione delle performance umane utilizzato dai ricercatori della Cornell University è stato quello di far giudicare le recensioni sempre attraverso il metodo del crowdsourcing con Mechanical Turk. Purtroppo, però alcuni turkers selezionati a caso, probabilmente per massimizzare i loro guadagni in termini di tempo hanno sicuramente evitato di leggere le recensioni. Perciò, è stato richiesto l'aiuto di tre studenti universitari volontari per dare giudizi su un sottoinsieme dei dati. Questo sottoinsieme equilibrato contiene 40 recensioni di quattro alberghi scelti a caso. A differenza dei turkers, agli studenti volontari non è stata offerta una ricompensa monetaria; di conseguenza, i loro giudizi sono stati considerati più onesti di quelli ottenuti tramite AMT.

Inoltre, per verificare in che misura i singoli giudici umani fossero di parte, sono state valutate le prestazioni di due meta-giudici virtuali. Specificamente, il meta-giudice MAJORITY predice "ingannevole" quando almeno due dei tre giudici umani credono che la revisione sia ingannevole, e il meta-giudice SKEPTIC predice "ingannevole" quando un qualsiasi giudice umano crede che la recensione sia ingannevole. Le prestazioni umane e quelle dei meta-giudici si sono attestate tra il 65% e il 75% di accuratezza. Risulta chiaro quindi che i giudici umani non sono particolarmente efficaci in questo compito.

3.3 APPROCCI LINGUISTICO-COMPUTAZIONALI

3.3.1 PRIMO APPROCCIO: CATEGORIZZAZIONE DEL TESTO

Per categorizzazione (o classificazione) testuale, in ambito di intelligenza artificiale si intende il processo automatico di attribuzione di un documento testuale, espresso in una lingua naturale, a una certa categoria o classe. Per realizzare la categorizzazione, si utilizzano solitamente approcci di apprendimento automatico di tipo supervisionato: si addestrano dei cosiddetti classificatori attraverso esempi da cui si genera un modello per la classificazione automatica.

Per capire come funziona questo approccio l'immagine seguente mostra come il classificatore viene addestrato, attraverso un training corpus in modo che, a partire da nuovi documenti, esso sia in grado di classificarli, in base all'esperienza acquisita.

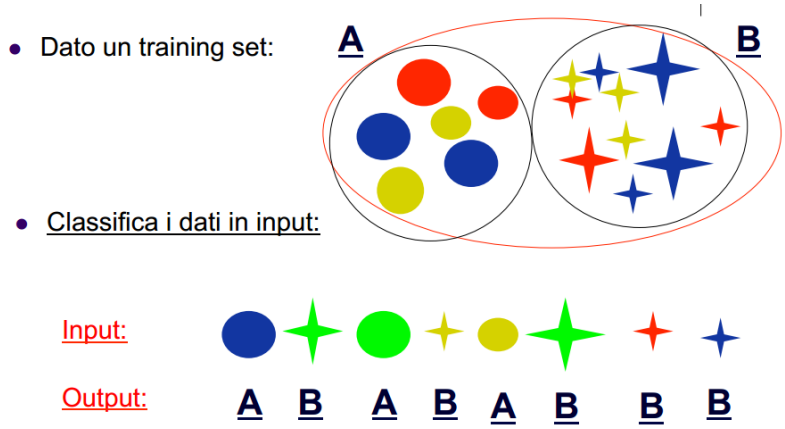

Immagine 1: classificazione[12]

Specificatamente per questo contributo, il classificatore utilizzato è stato addestrato sugli n-grammi (sottosequenze di n elementi di una data sequenza; gli elementi possono essere fonemi, sillabe, lettere, parole) in particolare qui, unigrammi, bigrammi e trigrammi di lemmi unstemmed (lo stemming è il processo di riduzione dalla forma flessa di una parola alla sua forma radice, detta "tema"). Dunque il training set costruito viene segmentato negli n-grammi appena specificati; i nuovi testi da classificare dovrebbero presentare una distribuzione di n-grammi simile ai testi dell'addestramento.

3.3.2 SECONDO APPROCCIO: RILEVAMENTO PSICOLINGUISTICO INGANNO

Il software *Linguistic Inquiry and Word Count* (LIWC) (Pennebaker et al., 2007) è uno strumento di analisi del testo automatizzato, ampiamente utilizzato nelle scienze sociali. È stato usato per rilevare i tratti della personalità, per studiare la dinamica di tutoraggio, e, più pertinentemente, per analizzare l'inganno. LIWC, specificamente, calcola il grado in cui le persone utilizzano diverse categorie di parole in una vasta gamma di testi, tra cui e-mail, discorsi, poesie, o trascrizioni del parlato quotidiano. Lo fa calcolando la percentuale di parole che corrispondono ciascuna a varie dimensioni psicologiche, come emozioni positive o negative, auto-riferimenti, parole causali, fino a 80 dimensioni psicologicamente significative. In particolare, questo software conta e raggruppa le istanze di circa 4500 parole chiave. Mentre LIWC non include un classificatore di testo, lo studio dei ricercatori della Cornell University ne ha creato uno con le caratteristiche derivate dalle produzione LIWC. In particolare, è stata costruita una caratteristica per ciascuna delle 80 dimensioni LIWC, che si possono riassumere sostanzialmente nelle seguenti quattro categorie:

[12] L'immagine è stata presa dal materiale del corso di Linguistica Computazionale. Professore Alessandro Lenci. (Corso di studi di Informatica Umanistica triennale. Uuniversità di Pisa).

1. I processi linguistici: aspetti funzionali del testo (ad esempio, il numero medio di parole per frase, il tasso di errore ortografico, imprecazioni, etc.)
2. I processi psicologici: include tutti i processi sociali, emotivi, cognitivi, percettivi e biologici, così come qualsiasi cosa che riguarda il tempo o lo spazio.
3. Le preoccupazioni personali: tutti i riferimenti a lavoro, tempo libero, soldi, religione, etc.
4. Le categorie del parlato: principalmente parole di riempimento e accordo.

Sebbene altre caratteristiche sono state prese in considerazione nei precedenti lavori di rilevazione dell'inganno, i recenti esperimenti trovano le caratteristiche LIWC le migliori. Infatti, il software LIWC2007 utilizzato dai ricercatori della Cornell University assume maggior parte delle funzioni introdotte in altri lavori.

3.3.3 TERZO APPROCCIO: IDENTIFICAZIONE DEL GENERE

La linguistica computazionale ha dimostrato che la distribuzione di frequenza di tag part-of-speech (POS) in un testo spesso dipende dal genere del testo (Biber et al, 1999;. Rayson et al., 2001).

Nell'approccio di identificazione del genere per rilevare lo spam d'opinione ingannevole, lo studio che stiamo affrontando in questa relazione si è concentrato sulla rilevazione dell'esistenza di un rapporto tra la distribuzione di part-of-speech e il genere testuale delle recensioni.

In particolare, lo spunto di ricerca è costituito da un contributo prodotto dai ricercatori dello University Centre for Computer Corpus Research on Language, di Lancaster che ha esaminato la relazione tra la frequenza di part-of-speech e la tipologia del testo nel British National Corpus Sampler. Sono stati fatti quattro confronti a coppie di frequenze di part-of-speech:

- lingua parlata vs. lingua scritta;
- scrittura informativa vs. scrittura creativa;
- discorsi di conversazione vs. discorsi 'task-oriented';
- scrittura creativa vs discorsi 'task-oriented'.

Qui riportiamo i risultati solo relativi alla scrittura creativa e alla scrittura informativa che riportano forti differenze tra i due generi. In particolare, il genere informativo è costituito da più sostantivi, aggettivi, preposizioni, determinanti e congiunzioni coordinative, mentre quello creativo è caratterizzato dalla presenza maggiore di verbi, avverbi, pronomi e pre-determinanti.

Nello specifico dei verbi, nel confronto tra scrittura creativa e informativa, i verbi modali, e la maggior parte delle forme di verbi lessicali, erano più comuni nella scrittura creativa. L'eccezione nel caso di verbi lessicali era il participio passato, che era più comune negli scritti informativi. Questo risultato riflette quasi certamente un maggior ricorso al passivo nei generi informativi (cfr

Biber et al 1999. 477; Svartvik 1966). Altre forme verbali preferite dal genere informativo sono stati: essendo, essere (come infinito), stato (participio passato), sono, e ha.

In effetti, lo studio della Cornell University ha ritrovato che i pesi appresi dal classificatore basato sulle POS (per peso si intende la presenza o meno di una parola nel testo calcolata mettendo in rapporto il numero di occorrenze della parola con il numero totale di parole nel documento) sono in gran parte d'accordo con questi risultati, ad eccezione di avverbi e aggettivi superlativi. Tuttavia, che le opinioni ingannevoli contengono più superlativi non è inaspettato, in quanto la scrittura ingannevole (ma non necessariamente scrittura creativa in generale) contiene spesso il linguaggio esagerato.

I risultati sono riportati in tabella 2.

TRUTHFUL/INFORMATIVE			DECEPTIVE/IMAGINATIVE		
Category	**Variant**	**Weight**	**Category**	**Variant**	**Weight**
NOUNS	Singular	0.008	VERBS	Base	-0.057
	Plural	0.002		Past tense	**0.041**
	Proper, singular	**-0.041**		Present participle	-0.089
	Proper, plural	0.091		Singular, present	-0.031
ADJECTIVES	General	0.002		Third person singular, present	**0.026**
	Comparative	0.058		Modal	-0.063
	Superlative	**-0.164**	ADVERS	General	**0.001**
PREPOSITION	General	0.064		Comparative	-0.035
DETERMINERS	General	0.009	PRONOUNS	Personal	-0.098
COORD. CONJ.	General	0.094		Possessive	-0.303
VERBS	Past Participle	0.053	PRE-DETERMINERS	General	**0.017**
ADVERBS	Superlative	**-0.094**			

Tabella 2: La media dei pesi assegnati da POS_SVM. Basandosi sul lavoro di Rayson et al. (2001). Ci si aspetta che i pesi sulla sinistra siano positivi (predittivo di opinioni veritiere), e che i pesi sulla destra siano negativi (predittivi di opinioni ingannevoli). I valori in grassetto sono in contrasto con le aspettative.

4. RISULTATI

Sia la classificazione testuale basata sugli n-grammi che la classificazione sulle caratteristiche del rilevamento psicologico dell'inganno superano in prestazione la semplice classificazione di genere. Specificamente, l'approccio psicolinguistico (LIWC_SVM) è più accurato del 3,8% e l'approccio di categorizzazione del testo standard si attesta tra 14,6% e il 16,6% in accuratezza. Tuttavia, nel complesso, le migliori prestazioni si ottengono combinando le caratteristiche di questi due approcci. In particolare, il modello combinato LIWC + bigrammi funziona all'89,8% ad individuare lo spam di opinione ingannevole.

Per capire meglio i modelli appresi da questi approcci automatizzati, riportiamo nella tabella 3 le prime 15 occorrenze più utilizzate per ogni classe (veritiera e ingannevole), catalogate secondo il modello combinato LIWC + bigrammi e da LIWC. In accordo con le teorie di monitoraggio della realtà (Johnson e Raye, 1981), si osserva che le opinioni veritiere tendono ad includere un linguaggio più sensoriale e concreto delle opinioni ingannevoli; in particolare, le opinioni veritiere specificano le configurazioni spaziali (ad esempio, piccolo, bagno, su, posizione). Questo risultato è supportato anche da recenti lavori che suggeriscono che i bugiardi hanno una notevole difficoltà nel codificare l'informazione spaziale nelle loro bugie. Di conseguenza, nei pareri ingannevoli si osserva una maggiore attenzione sugli aspetti esterni all'hotel recensito (per esempio, il marito, gli affari, le vacanze).

Sono stati riconosciuti anche diversi risultati che sono in contrasto con i precedenti studi di inganno psicolinguistico. Per esempio, sebbene l'inganno sia spesso associato a termini di emozioni negative, le recensioni ingannevoli analizzate hanno termini più positivi e meno emozioni negative. Questo modello ha senso se si considera l'obiettivo degli ingannatori, e cioè di creare una recensione positiva. All'inganno è, inoltre, già stato associato ad una diminuzione dell'utilizzo della prima persona singolare, un effetto attribuito al distanziamento psicologico. Al contrario, è stato notato che l'aumento della prima persona singolare risulta essere tra i maggiori indicatori di inganno, probabilmente dovuto al fatto che gli ingannatori tentano di rafforzare la credibilità delle recensioni enfatizzando la propria presenza nella revisione.

[LIWC + BIGRAMS]svm		LIWC_svm	
TRUTHFUL	DECEPTIVE	TRUTHFUL	DECEPTIVE
-	Chicago	Hear	I
…	My	Number	Family
On	Hotel	Allpunct	Perspron
Location	,_and	Negemo	See

)	Luxury	Dash	Pronoun
Allpunct(LIWC)	experience	Exclusive	Leisure
Floor	Hilton	We	Exclampunct
(	Business	Sexual	Sixletters
The_hotel	Vacation	Period	Posemo
Bathroom	I	Otherpunct	Comma
Small	Spa	Space	Cause
Helpful	Looking	Human	Auxverb
$	While	Past	Future
Hotel_.	husband	Inhibition	Perceptual
other	my_husband	assent	feel

Tabella 3: le prime 15 occorrenze più utilizzate per ogni classe (veritiera e ingannevole), catalogate secondo il modellocombinato LIWC + bigrammi e da LIWC.

5. CONCLUSIONI

I lavoro di ricerca degli studiosi della Cornell University ha portato ai seguenti risultati: lo sviluppo del primo set di dati standard di grandi dimensioni contenti spam di opinione ingannevole. Con esso, è stato dimostrato che l'individuazione di spam di opinione ingannevole è ben oltre le capacità dei giudici umani; lo sviluppo di tre approcci automatici per il rilevamento dello spam di opinione ingannevole, sulla base di conoscenze provenienti dalla ricerca in linguistica computazionale e in psicologia. È stato scoperto che un approccio combinato di caratteristiche psicolinguistiche e funzioni n-grammi è in grado di eseguire meglio il compito. Infine, è stato scoperto che esiste un rapporto plausibile fra l'opinione ingannevole e la scrittura creativa, basata su somiglianze tra distribuzioni di Part-of-speech.

6. LAVORI CORRELATI

Altri contributi hanno portato alla luce nuovi obiettivi ottenibili con queste tecniche linguistico-computazionale. In particolare, all'istituto di Linguistica Computazionale del Consorzio Nazionale delle Ricerche di Pisa, queste nuove metodologie, che esplorano la struttura linguistica dei documenti testuali, sono state impiegate per analizzare la leggibilità di un testo. I risultati sono

stati presentati durante uno dei seminari di cultura digitale del corso di laurea magistrale di informatica umanistica, da cui questa relazione prende spunto.

L'ILC non si è occupato solo dell'analisi della leggibilità, ma anche di altri task tra cui il riconoscimento del genere testuale, in particolare della prosa narrativa italiana all'interno del macro genere Letteratura. Il lavoro dei ricercatori si è concentrato, specificamente, sull'analisi dei testi letterari destinati a un pubblico di bambini e ragazzi e dei testi letterari per adulti che mostrano diverse differenze semantiche al fine di dimostrare che, nonostante queste diversità rilevate, le peculiarità del genere letteratura sono ancora chiare e visibili.

7. BIBLIOGRAFIA

D. Biber, S. Johansson, G. Leech, S. Conrad, E. Finegan, andR.Quirk. 1999. *Longman grammar of spoken and written English*, volume 2. MIT Press.

M.L. Newman, J.W. Pennebaker, D.S. Berry, and J.M. Richards. 2003. *Lying words: Predicting deception from linguistic styles. Personality and Social Psychology Bulletin*, 29(5):665.

J.W. Pennebaker, C.K. Chung, M. Ireland, A. Gonzales, and R.J. Booth. 2007. *The development and psychometric properties of LIWC2007*. Austin, TX, LIWC. Net.

P.Rayson, A.Wilson, and G.Leech. 2001. *Grammatical word class variation within the British National Corpus sampler. Language and Computers*, 36(1):295– 306.

A. Vrij, S. Leal, P.A. Granhag, S. Mann, R.P. Fisher, J. Hillman, and K. Sperry. 2009. *Outsmarting the liars: The benefit of asking unanticipated questions. Law and human behavior*, 33(2):159–166.

8. WEBGRAFIA

www.ilfattoquotidiano.it

www.lemond.fr

www.repubblica.it

www.wikipedia.it

http://labcd.humnet.unipi.it/seminario/cultura_digitale6732/wp/uploads/sites/6/2013/12/11dic2013.pdf (SEMINARIO DI CULTURA DIGITALE)