Università di Pisa

Corso di Laurea Magistrale in Informatica Umanistica

Studio linguistico-computazionale sulla persuasione: il caso dei discorsi politici e delle recensioni online

Candidata:
Giulia Chiriatti

Relatore:
Dott. Felice Dell'Orletta

Anno Accademico 2020/2021

Indice

Capitolo 1

Introduzione

1.1 Background

La persuasione è stata per lungo tempo un tema di interesse nelle scienze umane. Tra i campi coinvolti si trova la retorica, che è stata definita essa stessa "fattrice di persuasione" nell'animo umano (Platone, *Gorgia*, 453a, 455a); la politica, per il ruolo chiave svolto dalla capacità di convincere l'altro all'interno di un dibattito in una democrazia funzionale (Partington e Taylor, 2017); il marketing, che integra strategie persuasive nella promozione di un servizio o un prodotto (Stanton, 1984). Ciò ha portato a un'ingente quantità di studi sui meccanismi propri della comunicazione persuasiva.

La ricerca, in particolare nella psicologia sociale, ha sottolineato l'importanza di attributi della fonte, come la sua credibilità o attrattiva (Chaiken, 1979; Chaiken, 1980), e del ricevente, come la predisposizione a analizzare il messaggio o le sue credenze pregresse (Petty e Cacioppo, 1986; Chaiken, 1980). Anche se è stato riconosciuto il ruolo di aspetti non verbali o paraverbali in quanto indicatori di persuasività, quali il tono di voce, le espressioni facciali e il contatto visivo (Segrin, 1993), il linguaggio rimane uno dei principali veicoli di persuasione (Chaiken e Eagly, 1976; Werner, 1978; Perelman e Olbrechts-Tyteca, 1973; Miller, 2013).

Più di recente, studi computazionali hanno esaminato le caratteristiche linguistiche che danno forma a un messaggio persuasivo. Alcuni di questi hanno riguardato il cambiamento di attitudine nei confronti di un tema controverso in un dibattito o un forum online (Habernal e Gurevych, 2016; Tan et al., 2016; Wei et al., 2016; Wang et al., 2017), con l'obiettivo di predire il successo di un'argomentazione. In altri, la persuasione è stata intesa in senso lato come l'adozione di uno specifico comportamento, per esempio la soddisfazione di una richiesta di beneficenza attraverso donazioni (Althoff et al., 2014; D. Yang et al., 2019), o l'esecuzione di un'azione, che può prendere forme diverse a seconda del dominio di riferimento (Guerini e Özbal, 2015).

1.2 Le domande di ricerca

Per questo lavoro, si è deciso di adottare una definizione generale di persuasione, in base alla quale si considera persuasivo ogni messaggio che miri a generare una risposta nel destinatario (Miller, 2013), per *modificarne* o *rafforzarne* attitudini e comportamenti. Lo scopo è valutare l'impatto di un ampio spettro di caratteristiche, che catturano diversi aspetti dello stile di scrittura, sulla modellazione di un uso persuasivo del linguaggio in due domini specifici (i discorsi politici e le recensioni online) e due lingue (l'inglese e l'italiano). A partire da corpora annotati con le reazioni del pubblico (nel primo caso) e degli altri utenti (nel secondo), sono state osservate le differenze nella distribuzione dei tratti linguistici estratti per verificare se esistano variazioni nello stile dei testi persuasivi (ossia in grado di provocare reazioni) e se i tratti più rilevanti per discriminarli rimangano gli stessi tra una tipologia testuale e l'altra.

In seguito, le stesse caratteristiche sono state testate in compiti di classificazione automatica per valutare il loro effetto sulla performance del sistema nel distinguere tra testi persuasivi e non persuasivi, in confronto a quello di predittori tipicamente impiegati per catturare informazione lessicale. Gli esperimenti

sono stati condotti sia in uno scenario in-domain che cross-domain, per esaminare in particolare l'impatto dell'informazione strutturale sull'accuratezza della classificazione nel caso in cui il test set sia costituito da esempi molto distanti da quelli di addestramento, perché appartengono a un'altra categoria di recensioni o, per esempio, perché sono stati pronunciati da un politico affiliato a un altro partito.

1.3 L'organizzazione della tesi

La tesi si apre con un capitolo dedicato a una rassegna dei lavori che hanno affrontato in passato il problema della predizione automatica delle reazioni dell'audience (come applausi e risate) nei discorsi politici e, più in generale, nei discorsi pubblici, per poi passare a discutere gli studi più rilevanti condotti nell'ambito della predizione automatica dell'utilità di una recensione (in termini dei voti di utilità ricevuti dagli utenti del portale su cui sono state pubblicate). Benché si tratti di fenomeni spesso trattati separatamente da quello della persuasione, come sarà discusso nel capitolo 2, sia le reazioni del pubblico che i voti di utilità possono essere considerati indizi di un tentativo di persuasione andato a buon fine, evidenziato da una reazione dei destinatari del messaggio. Il capitolo termina quindi con la descrizione dei corpora testuali impiegati come risorse per la successiva analisi sperimentale.

Nel capitolo 3, si introduce la metodologia del monitoraggio linguistico e le sue possibili applicazioni e si descrivono i risultati dell'analisi statistica effettuata sui tratti estratti dai discorsi e dalle recensioni. Si discutono poi le differenze più significative emerse a vari livelli di descrizione linguistica tra esempi persuasivi e non persuasivi sia nei discorsi politici che nelle recensioni, per passare dunque a un confronto tra i due casi di studio, in cui si sottolineano somiglianze e divergenze dal punto di vista stilistico nei modi di persuadere tipici dell'uno o dell'altro dominio.

Il capitolo 4 è invece dedicato agli esperimenti di classificazione automatica, condotti separatamente per i discorsi politici e per le recensioni, in cui gli esempi sono mappati nella classe persuasiva o non persuasiva utilizzando diverse configurazioni di caratteristiche per la creazione dei modelli, tra cui quelle estratte in fase di monitoraggio. Dopo aver descritto l'assetto sperimentale, si commentano i risultati e si discute sia delle configurazioni che permettono di ottenere le performance migliori che delle feature più rilevanti a questo scopo.
Nell'ultimo capitolo, si traggono le conclusioni su quanto emerso dall'analisi del monitoraggio linguistico e dagli esperimenti di classificazione per descrivere i tratti più rilevanti di variazione stilistica riscontrati nei testi associati a una risposta altrui e quindi efficaci nel loro intento persuasivo.

Capitolo 2

Studi e risorse sulla persuasione

Secondo Halmari e Virtanen (2005), ogni uso del linguaggio è in un certo senso persuasivo, perché si modella sempre ciò che viene detto in funzione del suo pubblico. Tuttavia, una definizione più specifica, considera persuasivo un messaggio che cerchi di rafforzare o modificare le risposte del destinatario (Miller, 2013), con l'obiettivo di cambiare il loro stato mentale o di far sì che compiano un'azione (Perelman e Olbrechts-Tyteca, 1973). Negli ultimi decenni, un grande sforzo di ricerca ha indagato da diverse prospettive gli aspetti che caratterizzano una comunicazione efficace o d'impatto che miri a modificare le attitudini e i comportamenti dei destinatari. Per esempio, Habernal e Gurevych (2016) hanno impiegato un ampio numero di caratteristiche linguisticamente motivate per predire l'argomentazione più convincente in coppie riguardanti lo stesso tema. Tan et al. (2016) hanno studiato i pattern lessicali e di interazione tra il persuasore e il persuaso in un forum su Reddit in cui gli utenti marcano esplicitamente le argomentazioni che hanno avuto successo nel cambiare la loro opinione. Wei et al. (2016), sempre a partire dallo stesso forum, hanno creato un ranking di commenti in base alla loro persuasività evidenziando anche l'impatto di carat-

teristiche basate sul numero di verbi modali, connettivi e frasi precedentemente classificate come argomentative. In altri contesti, Guerini e Özbal (2015) hanno studiato l'effetto di caratteristiche fonetiche (e.g. rime, allitterazioni) su vari tipi di persuasività, che influenzano per esempio la memorabilità di citazioni di film, le ricondivisioni di tweet e l'efficacia di slogan e discorsi politici. Nei prossimi paragrafi, ci si concentra principalmente sul dominio dei discorsi politici e delle recensioni e in particolare sugli studi che hanno studiato l'efficacia di diverse tipologie di caratteristiche dei testi per predire le reazioni del pubblico nel primo caso e l'utilità di una recensione nell'altro.

2.1 La persuasione nei discorsi politici

I discorsi politici sono considerati come un genere tipicamente persuasivo, perché mirano a rinforzare le posizioni politiche del proprio elettorato o a convincere altri a votarli. Nei raduni di massa, come quelli che si tengono durante una campagna elettorale, il discorso è organizzato in modo tale da trasformarsi in un gioco di coordinazione in cui il pubblico mostra il proprio assenso (e riconferma la propria affiliazione politica) tramite reazioni come applausi e risate. Anche se può capitare che gli applausi partano spontaneamente (Bull, 2006), in genere è l'oratore stesso a invitare l'audience a partecipare attraverso la modifica del tono di voce e della gestualità ma anche l'impiego di artifici retorici, quali la ripetizione (nella forma di liste in tre parti), l'antitesi o l'utilizzo di strutture del tipo *headline-punchline*, che gioca sul sovvertimento delle aspettative (Atkinson, 1984; Heritage e Greatbatch, 1986). Questo diventa evidente nei casi (alcuni dei quali ormai entrati nella cultura di massa) in cui il tentativo fallisce e l'oratore è costretto a richiamare l'applauso con una richiesta esplicita, come accaduto nel 2016 al candidato alle elezioni presidenziali statunitensi Jeb Bush che si ritrovò a dire "Per favore, applaudite" a un pubblico silenzioso.[1]

[1] `https://www.youtube.com/watch?v=OUXvrWeQU0g&ab_channel=CNN` (ultimo accesso: 07/04/2022).

Gli applausi sono stati considerati come *hot spot* di tentativi di persuasione andati a buon fine in studi precedenti, che si sono focalizzati su aspetti del lessico impiegato per ottenere una reazione dall'audience. Per esempio, Guerini et al. (2008) hanno introdotto una misura dell'impatto persuasivo di una parola (*pi*), basata sulla TF-IDF (Term Frequency-Inverse Document Frequency), che tiene conto anche della vicinanza di una parola ai tag di reazione in un corpus di trascrizioni di discorsi politici annotati con le reazioni del pubblico (v. par. 2.3.1). I punteggi di *pi* sono stati poi utilizzati per estrarre le parole con il più alto impatto e comprendere meglio il loro utilizzo (ad esempio, per valutare se l'impatto di termini come "guerra" sia soggetto a cambiamenti dopo eventi storici chiave). Strapparava et al. (2010) e Guerini e Özbal (2015) hanno anche studiato la possibilità di usare CORPS per la predizione automatica dei passaggi del discorso che generano reazioni, usando informazione lessicale ricavata da n-grammi e caratteristiche fonetiche relative all'eufonia o al "bel suono" del discorso.

Sempre per la predizione di applausi e risate, Gillick e Bamman (2018) hanno costruito un corpus a partire dalle trascrizioni di discorsi tenuti durante la campagna presidenziale statunitense del 2016 presenti in formato video sul sito web di C-SPAN, una tv americana.[2] Hanno poi adattato da lavori precedenti caratteristiche di vario genere per modellare applausi e risate, riscontrando in particolare l'impatto di informazione prosodica e lessicale sulla predizione delle reazioni. Lo stesso risultato è stato ottenuto sia all'interno di discorsi degli stessi autori che testando il sistema su autori diversi rispetto a quelli di addestramento. In un dominio diverso da quello dei discorsi politici ma in un compito simile, focalizzato solo sugli applausi, H. Liu et al. (2017) hanno osservato l'utilità di caratteristiche stilistiche, in particolare dei riferimenti alla propria persona o al pubblico tramite i pronomi personali. Hanno anche notato che, considerando un numero maggiore di frasi prima della reazione, la performance del sistema

[2] `https://www.c-span.org/` (ultimo accesso: 07/04/2022).

tende a diminuire.

2.2 La persuasione nelle recensioni online

Con l'affermarsi del commercio elettronico, il Web è diventato uno spazio digitale di compravendita. Di conseguenza, la possibilità di acquistare beni e servizi online ha favorito la proliferazione di una nuova tipologia di passaparola (*electronic Word-of-Mouth* o *eWOM*). Se il passaparola tradizionale si riferisce allo scambio di consigli sugli acquisti in conversazioni faccia a faccia con amici o parenti, l'eWOM è invece ogni dichiarazione fatta da un cliente, riguardo a un prodotto o una società, che sia resa disponibile a una moltitudine di persone e istituzioni tramite Internet (Hennig-Thurau et al., 2004). In questo caso, la comunicazione è mediata dal computer e dunque assume caratteristiche uniche: il messaggio permane nel tempo; gli individui che lo producono o lo ricevono non sono più uniti da legami forti (in certe circostanze, è possibile anche ricorrere all'anonimato); se il messaggio è di natura testuale, i contenuti e lo stile di scrittura acquisiscono particolare importanza per il consumatore nel valutarne la credibilità e utilità ai fini delle scelte di acquisto (King et al., 2014).

La facilità con cui i consumatori possono diffondere opinioni nel Web e la moltitudine di piattaforme che le ospitano (e.g. blog, siti di *e-commerce*, *Social Networking Services*) ha dato luogo a un volume di informazioni senza precedenti, di molto superiore alla capacità che hanno gli utenti del Web di processarle. Questo vale anche, nello specifico, per le recensioni generate dagli utenti (*user-generated reviews*), una delle forme assunte dall'eWOM, che consiste nei commenti, riguardo a prodotti, attività commerciali e servizi, pubblicati dai consumatori sui siti web di compagnie di *e-commerce* oppure su siti dedicati interamente alla raccolta di recensioni: per fare qualche esempio, Amazon.com, Google, Facebook, CNET (per prodotti e business); Citysearch, Trustpilot, Yelp! e Tripadvisor (per quanto riguarda viaggi, hotel e ristoranti). Secondo i dati ri-

levati in U.S.A. nel corso di un sondaggio, nel 2019 i consumatori statunitensi si aspettavano di trovare in media 112 recensioni relative a un prodotto di loro interesse su un qualsiasi rivenditore online (Statista, 2019). Tra le compagnie più popolari in questo ambito, a fine 2021 Yelp! ha dichiarato un numero cumulativo di recensioni pari a 224 milioni,[3] mentre il totale ammonta a 988 milioni per Tripadvisor.[4] In generale, sono le stesse piattaforme di *e-commerce* a fornire agli utenti l'opzione di recensire i prodotti ospitati sui propri siti web: da un lato, la condivisione di valutazioni da parte di clienti che hanno già fruito di un articolo influenza le decisioni degli altri potenziali acquirenti, con un conseguente impatto sulle vendite (Park et al., 2007); dal punto di vista del consumatore, invece, si riducono gli sforzi di valutazione, così come i rischi percepiti, e si trova rassicurazione nelle opinioni espresse dai propri pari (King et al., 2014). L'ingente quantità di materiale prodotto, che deriva da questa duplice opportunità per business e consumatori finali, ha evidenziato il bisogno di individuare le recensioni più utili per gli utenti, in modo tale da ridurre i costi dell'*information overload*.

Di solito, le piattaforme lasciano che siano i visitatori stessi a valutare manualmente l'utilità delle recensioni attraverso un'interfaccia (come quella in Figura 2.1), in cui l'utente può marcare la recensione come "utile" attraverso un pulsante apposito oppure scegliere di non esprimere alcun giudizio. Talvolta è possibile scegliere una terza opzione e giudicare la recensione "non utile" in modo esplicito, come accadeva nella versione originale del sistema implementata da Amazon.com: in tal caso, la dicitura in calce al testo (in grigio in fig. 2.1) riportava, oltre al numero di voti "utili", anche il totale dei voti ricevuti (nella forma *N out of M people found the following review helpful*). Diversamente

[3]Le statistiche relative al numero di recensioni pubblicate su Yelp!, web app che raccoglie recensioni su business di varia natura, sono pubblicate online all'indirizzo `https://www.yelp-press.com/company/fast-facts/default.aspx` (ultimo accesso: 26/12/2021).

[4]Il dato è reso disponibile da Tripadvisor all'indirizzo `https://tripadvisor.mediaroom.com/US-about-us` (ultimo accesso: 26/12/2021).

Figura 2.1: Esempio di recensione di un disco rigido tratta da Amazon.com. In questo caso, il rating del prodotto corrisponde a una stella; gli utenti che considerano utile la recensione sono invece 105.

dal rating, ossia il numero di stelle che l'utente assegna al prodotto su una scala di Likert da 1 a 5 per indicarne il proprio gradimento, i voti di utilità non costituiscono una valutazione di quanto è oggetto di recensione ma dell'opinione altrui sul prodotto recensito. Come sottolineano Danescu-Niculescu-Mizil et al. (2009), la domanda si sposta da "Cosa pensa Y di X?" a "Cosa pensa Z dell'opinione di Y su X?".

L'utilità delle recensioni in letteratura

Negli ultimi anni, l'utilità delle recensioni ha generato ampio interesse nella comunità del Natural Language Processing. La ricerca sul tema, che si inserisce nel campo più vasto dell'Opinion Mining (B. Liu e L. Zhang, 2012; Kim et al., 2006), si è focalizzata in particolar modo sulla modellazione computazionale e sulla predizione automatica dell'utilità di una recensione (Ocampo Diaz e Ng, 2018). In letteratura, si fa comunemente riferimento all'utilità delle recensioni come *helpfulness*, *usefulness* o *utility* ma alcune interpretazioni chiamano in causa anche la *persuasività*, considerando i voti di utilità come dei *proxy* del cambiamento di attitudine del consumatore nei confronti del prodotto, che può tradursi in acquisto e dunque influenzarne le vendite (Pentina et al., 2018; Li

e Zhan, 2011; W. Hong et al., 2020). I ricercatori si sono anche chiesti se sia possibile equiparare l'utilità di una recensione alla sua *qualità*, arrivando a un generale accordo sul fatto che le recensioni con punteggi più alti di utilità non corrispondano necessariamente a quelle migliori dal punto di vista qualitativo (Ocampo Diaz e Ng, 2018; Danescu-Niculescu-Mizil et al., 2009; Ghose e Ipeirotis, 2011; Tsur e Rappaport, 2009; Y. Yang et al., 2015), aspetto che sottolinea l'importanza di considerare l'informatività dei testi come soltanto uno dei molteplici fattori, interni e esterni alla componente testuale, in grado di influenzare le votazioni degli utenti. È poi interessante notare che studi sull'utilità delle recensioni sono stati condotti anche su domini diversi da quello delle recensioni online: per esempio, Xiong e Litman (2011) hanno impiegato nel dominio educativo le caratteristiche utilizzate da Kim et al. (2006) in uno dei primi studi sull'utilità delle recensioni di prodotti, focalizzandosi sui risultati di un processo di peer review relativo a temi realizzati da studenti universitari. Allo stesso modo, Ménard e Barrière (2016) hanno studiato l'utilità dei commenti realizzati dai ricercatori in ambito medico sui riassunti degli articoli di ricerca che la Canadian Medical Association inviava loro periodicamente.

Al di là del dominio di interesse, il problema della predizione automatica dell'utilità è stato generalmente inquadrato come un compito di apprendimento automatico supervisionato: di regressione (in cui l'obiettivo è predire il punteggio di utilità di una recensione in un intervallo da 0 a 1), classificazione (in cui l'utilità è definita in modo binario) o ranking (in cui l'obiettivo è ordinare un insieme di recensioni in base alla loro utilità), quest'ultimo talvolta ricavato dai risultati di una prima fase di classificazione o regressione. I dataset di pubblico accesso più comunemente impiegati comprendono Amazon Multi-Domain Sentiment Dataset (Blitzer et al., 2007) e Amazon Review Data (McAuley et al., 2015), entrambi costruiti per scopi diversi rispetto alla predizione dell'utilità. Come riportato in survey realizzati sul tema in varie aree di ricerca (Almutairi et al., 2019; Bilal et al., 2019; Ocampo Diaz e Ng, 2018), è stato studiato l'im-

patto sulla predizione dell'utilità di un ampio spettro di caratteristiche testuali, sia relative a aspetti strutturali delle recensioni che al loro contenuto.

In uno dei primi studi sul tema, dopo aver trovato una scarsa correlazione tra utilità e lunghezza delle recensioni di Amazon.com che evidenziava la necessità di costruire modelli non banali, Z. Zhang e Varadarajan (2006) hanno mostrato l'efficacia in un compito di regressione di caratteristiche basate sulla frequenza di alcune parti del discorso (e.g. nomi propri, verbi modali, aggettivi e avverbi comparativi e superlativi), considerate come in grado di catturare aspetti dello stile linguistico delle recensioni. In uno studio simile, condotto sempre sulle recensioni di Amazon.com, Kim et al. (2006) sono invece giunti a risultati opposti, ottenendo la migliore performance con una configurazione delle caratteristiche più comunemente usate per catturare aspetti lessicali dei testi (la frequenza di unigrammi di parole) in combinazione con la lunghezza della recensione e il rating. L'aggregazione di frequenze di parti del discorso, simili a quelle di Z. Zhang e Varadarajan (2006) ma meno granulari e con l'aggiunta del numero di verbi coniugati alla prima persona, non ha invece prodotto alcun miglioramento nei risultati. Allo stesso modo, in Xiong e Litman (2011), nonostante la forte correlazione trovata tra caratteristiche strutturali e morfo-sintattiche delle recensioni dei temi universitari e la loro utilità, il modello migliore è risultato ancora una volta quello che comprendeva la lunghezza delle recensioni, gli unigrammi e il rating.

Anche Mertz et al. (2014), che hanno invece investigato il ruolo di connettivi e dei bigrammi ottenuti da coppie di parole collegate da relazioni di dipendenza sintattica piuttosto che adiacenti, non hanno trovato alcun miglioramento rispetto agli unigrammi, nonostante il fatto che entrambi gli insiemi di caratteristiche avessero ottenuto una performance migliore della baseline e dunque fossero portatori di informazione utile nel discriminare le recensioni più votate. Y. Hong et al. (2012) hanno poi impiegato misure basate sul numero di verbi modali nelle recensioni e sul numero di verbi coniugati al passato per valutare l'affidabilità

delle recensioni, basandosi sull'ipotesi che l'uso dei verbi modali denoti maggiore incertezza da parte dell'autore e, al contrario, l'uso del passato sia indice di maggiore esperienza. Utilizzando queste caratteristiche in combinazione con la differenza del rating della recensione dal rating medio del prodotto e il numero di parole relative a caratteristiche del prodotto ritrovate anche nella recensione editoriale di Amazon, hanno ottenuto un incremento nella performance rispetto al modello migliore di Kim et al. (2006).

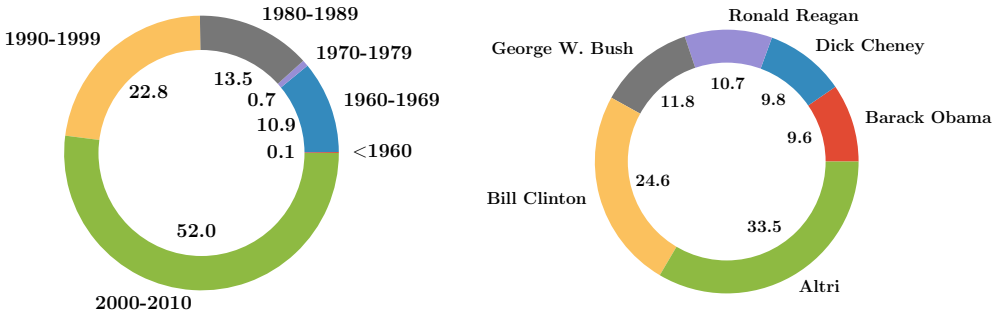
2.3 I corpora

Questa sezione è dedicata alla descrizione di tre corpora che sono stati utilizzati in questo studio per l'analisi linguistica e la predizione automatica della persuasività. I dati coprono vari domini e lingue, fornendo così esempi di testi che sono simili nell'intento persuasivo che sta dietro la loro produzione ma sono anche diversi per contenuto e scopo. Tutti i corpora sono arricchiti con annotazioni sulle reazioni delle persone che hanno risposto alle strategie persuasive incorporate nei testi. In alcuni casi, queste reazioni possono assumere la forma di applausi o risate in risposta al punto saliente di un discorso; mentre, quando si tratta di un ambiente online, sono codificate come metadati relativi ai voti di altri utenti.

2.3.1 CORPS

CORPS (CORpus of tagged Political Speeches) è un corpus di discorsi politici annotati con le reazioni del pubblico. È stato introdotto da Guerini et al. (2008) come risorsa adatta a una varietà di scopi di ricerca, che vanno dall'analisi qualitativa della comunicazione politica all'estrazione di espressioni persuasive. L'ipotesi di fondo è che gli applausi, le risate e le acclamazioni segnino i punti chiave di un discorso in cui il pubblico ha riconosciuto con successo un tentativo di persuasione da parte dell'oratore. La "persuasione" è intesa qui nel senso

più ampio: non riguarda strettamente un cambiamento di opinione o di atteggiamento, ma piuttosto l'efficacia di un passaggio nel provocare una risposta dal pubblico. Sfruttando i tag di reazione, è quindi possibile analizzare automaticamente le proprietà testuali dei passaggi cui sono associati per studiarne l'impatto persuasivo. Il corpus è stato costruito in modo semi-automatico, partendo da un nucleo di 900 discorsi in inglese, tutti pronunciati da madrelingua, che sono stati successivamente estesi a 3.600 in una nuova versione (Guerini et al., 2013), per un totale di circa 8 milioni di parole. Il processo ha comportato la raccolta dei discorsi dal Web e l'estrazione dei metadati descrittivi (titolo, evento, oratore, data e descrizione) dalle rispettive pagine HTML. I tag di reazione, già disponibili nelle fonti online (come gli aggregatori di notizie e i siti web dei politici), sono stati ridotti a quelli più comuni per mezzo di una conversione automatica e infine sottoposti a un controllo manuale di coerenza. La maggior parte dei discorsi sono stati pronunciati da oratori appartenenti alla scena politica statunitense tra il 1960 e il 2010, dalla presidenza di John F. Kennedy all'inizio del primo mandato di Barack Obama, ma alcuni di essi risalgono anche al 1917 (v. fig. 2.2).



(a) In base al periodo.

(b) In base agli oratori.

Figura 2.2: Distribuzione dei discorsi in CORPS (%).

Tutti i discorsi sono stati tenuti in un contesto monologico, in cui non sono previsti interventi del pubblico, e si rivolgono generalmente a spettatori ben disposti, per esempio quelli convenuti a un raduno di massa durante una campagna

elettorale. In tale contesto, si cerca di veicolare una serie di messaggi politici nel corso di un discorso lungo ed elaborato con l'obiettivo di catturare l'attenzione degli ascoltatori e generare una manifestazione di sostegno. Per quanto riguarda lo schema di annotazione, è da notare che i tag normalizzati (tra parentesi graffe in basso), che costituiscono il prodotto del processo di conversione automatica, sono stati ulteriormente raggruppati in tre categorie principali:

- **POSITIVE_FOCUS**, per i tag che indicano una manifestazione di sostegno da parte del pubblico (e.g. {APPLAUSE}, {CHEERS}, {STANDING OVATION});
- **IRONICAL_FOCUS**, per i tag che si riferiscono per lo più a risate in risposta a un'osservazione spiritosa dell'oratore (e.g. {LAUGHTER; APPLAUSE}, {LAUGHTER});[5]
- **NEGATIVE_FOCUS**, per i tag che indicano una reazione negativa a un passaggio del discorso (e.g. BOOING,{AUDIENCE}No!{AUDIENCE}).[6]

Si veda la Tabella 2.1 per un breve riassunto delle statistiche di base calcolate sul corpus.

La preparazione del corpus

Seguendo un approccio simile a quello di Strapparava et al. (2010), sono state estratte dal corpus delle finestre di testo *Persuasive* e *Non persuasive* (limitatamente ai discorsi tenuti da oratori statunitensi), a seconda che precedano o meno una reazione del pubblico. Le finestre etichettate come *Persuasive* precedono

[5]Guerini et al. (2008) hanno scelto il gruppo Ironical Focus per rietichettare più tag che si riferiscono sia alle risate che agli applausi, poiché questi ultimi si accompagnano alle prime, enfatizzandole.

[6]Questi commenti non dovrebbero essere intesi come negativi nei confronti dell'oratore ma piuttosto come risultato di un'antitesi, un espediente retorico che gli oratori adottano con l'intento di provocare l'applauso iniziando con un'affermazione negativa che prepara il terreno per l'enunciazione positiva finale (per informazioni più dettagliate sull'utilizzo dell'antitesi nei discorsi politici si veda, per esempio, Heritage e Greatbatch (1986)).

Numero di oratori	197
Numero di discorsi	3618
Numero di parole	7901893
Densità media dei tag	0.0084
Intervallo temporale	1917-2010
Numero di tag	66082
- Positive-Focus	49275
- Ironical-Focus	15660
- Negative-Focus	1147

Tabella 2.1: Statistiche principali calcolate su CORPS.

uno o più tag di qualunque natura (POSITIVE_FOCUS, IRONICAL_FOCUS o NEGATIVE_FOCUS); quelle *Non persuasive* sono invece formate da sequenze di frasi che non rientrano nelle finestre *Persuasive*. Sono state tuttavia escluse dagli esempi della classe persuasiva le finestre che presentavano ulteriori tag all'interno, basandosi sull'intuizione che queste ultime possano non rendere conto in modo corretto del range di frasi realmente necessario a catturare il fenomeno. Dato che si è scelto in fase di analisi di considerare anche caratteristiche linguistiche relative alla struttura sintattica della frase, le finestre non sono sono state tagliate immediatamente prima del tag nel caso in cui questo occorresse all'interno di una frase: i tag possono quindi occorrere anche all'interno dell'ultima frase della finestra ma non prima. Per quanto riguarda la dimensione delle finestre, sono state estratte inizialmente finestre lunghe 4 frasi[7] per poi provare anche un numero inferiore di frasi, in modo da verificare l'ipotesi che, avvicinandosi al tag, cresca anche la "persuasività" del testo.

In totale, sono state estratte 29266 finestre *Persuasive* e 47877 *Non persuasive* (v. fig. 2.3). Per ottenere le finestre di dimensione inferiore, sono state selezionate le ultime *n* frasi delle finestre *Persuasive* (le più vicine al tag) e le prime *n* frasi delle finestre *Non persuasive* (le più lontane dal tag), per *n* che va da 1 a 3. Si è scelto anche di dividere le finestre estratte in base al tipo di tag che vi è associato: anche se ci sono alcuni casi di sovrapposizione (è possibile,

[7]Il valore è ripreso da Strapparava et al. (2010).

per esempio, che una parte di discorso susciti sia risate che applausi, benché ciò non rappresenti la norma, come mostrato in fig. 2.3), si è deciso di associare alla categoria POSITIVE_FOCUS tutte le finestre che precedono almeno un tag di quella tipologia, utilizzando lo stesso criterio anche per le altre categorie.

Classe	Numero di finestre
Non persuasive	47877
Persuasive	29266
- Negative_Focus	435
- Ironical_Focus	5908
- Positive_Focus	23045

Tabella 2.2: Dati sulle finestre persuasive e non persuasive estratte da CORPS.

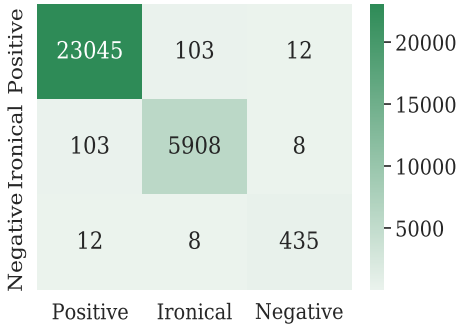


Figura 2.3: Matrice di co-occorrenza dei tag associati alle finestre estratte da CORPS.

Classe	Tag	Esempio
Persuasiva	Positive_Focus	We understand the threats before us, and we have the resources, the strength, and the moral courage to overcome them all. As our President has made clear to all, the terrorist enemies will fail because the direction of history is toward justice and human freedom. The terrorists will fail because the resolve of America and or allies wil not be shaken. And the terrorists will fail because men and women like you stand in their way.
Persuasiva	Ironical_Focus	And I think I should have done more to strengthen the bio-weapons convention. I think our industry people were a little wrong about that. The current administration is backed away from all of those four things because in every case, there is some worst-case nightmare they can come up with. Well if none of us joined anything unless we got our way all the time, nobody would ever get married.
Non persuasiva	-	The extraordinary growth we've seen in the clean energy sector is due first and foremost to the entrepreneurial drive of our businesses and our workers. But it's also due to the fact that we invested in them. One of these investments came in the form of clean energy manufacturing tax credits. What we said to clean energy firms was, if you're willing to put 70 percent of the capital for a worthy endeavor, we'll put up the other 30 percent.

Tabella 2.3: Esempi di finestre *Persuasive* e *Non persuasive*.

2.3.2 WIT³

WIT³ (Web Inventory of Transcribed and Translated Talks) è un corpus che ospita delle trascrizioni multilingue di TED talk.[8] I talk originali provengono dal sito di TED Conferences, una conosciuta organizzazione non-profit americana che organizza annualmente eventi in Nord America ma anche Europa, Asia e Africa a partire dal 1984, e dal 2007 li trasmette anche online.[9] Le conferenze TED sono nate con l'obiettivo di "diffondere idee" (secondo quanto espresso dallo slogan stesso dell'organizzazione, "ideas worth spreading") relative a temi disparati, nell'ambito della scienza e della cultura, attraverso brevi presentazioni della durata massima di 18 minuti, studiate perché siano il più possibile d'impatto: per questo motivo, l'organizzazione offre un percorso di formazione pre-conferenza per assicurare maggiori possibilità di successo agli speaker e invita i candidati che sono interessati a partecipare a fornire informazioni non solo sull'idea su cui sono basati i loro interventi ma anche sulla risposta che si aspettano da parte dell'audience, allegando video di precedenti interventi in pubblico. Le trascrizioni dei talk e le traduzioni sono realizzate da volontari reclutati online: i nuovi arrivati hanno la possibilità di aderire a programmi di *mentoring* per affinare le proprie abilità, mentre agli utenti più esperti sono assegnati ruoli di supervisione. Grazie a questo sistema sono attualmente disponibili traduzioni in 116 lingue dei talk, a partire dalle presentazioni originali in inglese, anche se non tutte riguardano la totalità degli interventi.

```
<transcription>
    <seekvideo id="1000">Good afternoon, good evening,
       whatever.</seekvideo>
    ...
```

---

[8] WIT³ può essere scaricato all'indirizzo `https://wit3.fbk.eu/` (ultimo accesso: 07/04/2022).

[9] La sezione del sito web di TED, grazie a cui è possibile accedere alla data odierna a registrazioni, trascrizioni e traduzioni di circa 3600 talk, pubblicati sotto licenza Creative Commons BY-NC-ND, è disponibile al link `https://www.ted.com/talks` (ultimo accesso: 07/04/2022).

```
    <seekvideo id="1399000">(Applause)</seekvideo>
```

riguarda le traduzioni presenti in WIT³, nella maggior parte delle lingue non è stato tradotto neanche il 10% dei TED talk tenuti entro il 2016, che corrisponde a circa 200 talk (v. fig. 2.4). Al contrario, in lingue come l'ebraico, lo spagnolo, il portoghese (parlato in Brasile) e il russo sono disponibili più del 98% delle presentazioni originali.

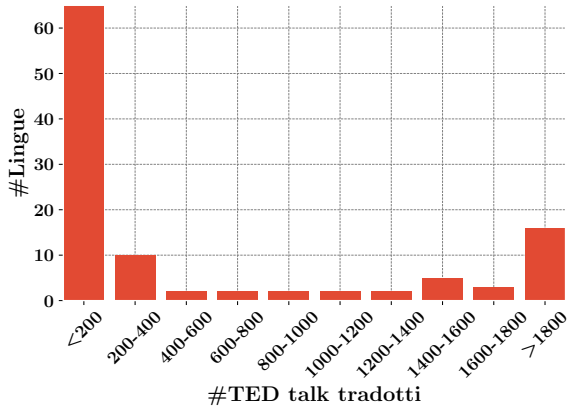


Figura 2.4: Distribuzione delle traduzioni dei TED talk tra le lingue presenti in WIT³ (versione del 2016).

Il corpus di TED talk è di interesse per la *Machine Translation*, in quanto presenta testi relativi a un vasto spettro di tematiche, tradotti grazie al lavoro di volontari, ma si configura anche come una risorsa utile per lo studio della persuasione. I sottotitoli, infatti, riportano, oltre alle parole degli speaker, le reazioni del pubblico, tra cui risate e applausi (non diversamente dai discorsi politici in CORPS).[11] I testi hanno, anche in questo caso, un chiaro intento persuasivo, considerata l'enfasi sulle risposte dell'audience e sulla creazione di una presentazione "d'impatto". Dai dati riportati in Tabella 2.4, ottenuti convertendo le reazioni del pubblico nei tag di CORPS secondo le modalità descritte in Guerini et al. (2008), emergono alcune differenze di superficie tra le tipologie di testi coinvolti: tra queste, la minore densità media dei tag per quanto riguarda i TED talk piuttosto che i discorsi politici, così come la prevalenza delle risate sugli applausi.

[11] v. il sottoparagrafo 2.3.1.

Numero di speaker	1767
Numero di talk	2085
Numero di parole	4372216
Densità media dei tag (*)	0.0028
Intervallo temporale	1984-2016
Numero di tag (*)	12430
- Positive-Focus	4215
- Ironical-Focus	8215

Tabella 2.4: Statistiche principali calcolate sui TED talk in inglese presenti nella versione di WIT³ rilasciata nel 2016. (*) Positive-Focus: (Applause), (Cheers); Ironical-Focus: (Laughter), (Laughter) (Applause).

Studi sul rilevamento degli applausi dei testi sono già stati condotti sui TED talk[12] ma le traduzioni offrono anche la possibilità di indagare i fenomeni persuasivi in contesti cross-linguistici, studiando la loro variazione in lingue differenti.

2.3.3 Dataset di Tripadvisor

Questo dataset è una raccolta di recensioni scaricate tramite *web crawling* dalla sezione italiana di Tripadvisor[13] (Chiriatti et al., 2019), l'applicazione web lanciata dalla società omonima nel 2000, quattro anni prima di Yelp!, uno dei suoi concorrenti più noti. Tripadvisor nasce inizialmente come aggregatore di guide turistiche e opinioni sulle mete turistiche provenienti da giornali e riviste ma nel 2001 si apre ai contenuti generati dagli utenti, dando loro la possibilità di postare le proprie recensioni di alloggi, ristoranti, musei, siti culturali. La mossa, sostenuta dal fatto che i visitatori risultavano attratti più dalle recensioni dei propri pari che da quelle ricavate da fonti ufficiali, ha contribuito alla crescita esponenziale dell'utenza, che è arrivata a una media mensile di 5 milioni nel 2004 ma a oggi si aggira sulle centinaia di milioni.[14]

[12]Il tema è discusso nella sezione 2.1.

[13]Il sito web di Tripadvisor è disponibile nella versione in italiano al link `https://www.tripadvisor.it/` (ultimo accesso: 07/04/2022).

[14]I dati riportati sono tratti dal sito web di Tripadvisor: `https://tripadvisor.mediaroom.com/US-about-us` (ultimo accesso: 07/04/2022).

Gli autori delle recensioni non ricevono alcun compenso per i propri contributi tranne punti e *badge* che possono guadagnare in base al numero di recensioni ed esporre poi sul proprio profilo. I contenuti sono sottoposti a moderazione per controllare che rispettino le linee guida stabilite e, una volta accettati e pubblicati, non possono essere modificati: in particolare, si richiede che le recensioni siano il resoconto imparziale di esperienze recenti, vissute in prima persona e rilevanti per i viaggiatori, e siano scritte in modo tale da essere accessibili a tutti, oltre che sintetiche a sufficienza da rientrare in un limite massimo di caratteri (di solito 100).[15]

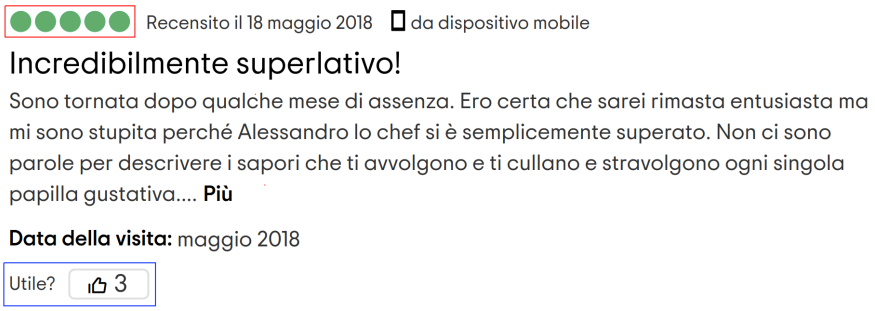


Figura 2.5: Esempio di recensione tratta da Tripadvisor. In *rosso* il *bubble rating* assegnato dall'autore al ristorante; in *blu* i *voti utili* assegnati dagli altri utenti alla recensione.

Gli *helpful votes* o *voti utili* sono anche soggetti a controlli volti ad assicurarsi che: non ci siano voti duplicati; gli utenti non votino per le proprie recensioni; i proprietari del business non votino per recensioni relative alla propria attività o a attività loro concorrenti. Il dataset, costruito con lo scopo di indagare quanto le caratteristiche lessicali (inerenti al contenuto) e strutturali (inerenti allo stile) delle recensioni influenzino la loro utilità percepita, comprende circa 42000 recensioni, associate al loro *bubble rating* e al numero di voti utili ricevuti. I dati sono stati scaricati dalle pagine di 1218 ristoranti e 383 attrazioni turistiche

[15]Fa eccezione, per esempio, la categoria degli hotel, per cui sono ammesse recensioni di 200 caratteri. Per maggiori informazioni sulle linee guida si rimanda alla pagina dedicata sul sito web di Tripadvisor: `https://www.tripadvisor.com/Trust-lvBd3L1aU38Y-Review_posting_guidelines.html` (ultimo accesso: 07/04/2022).

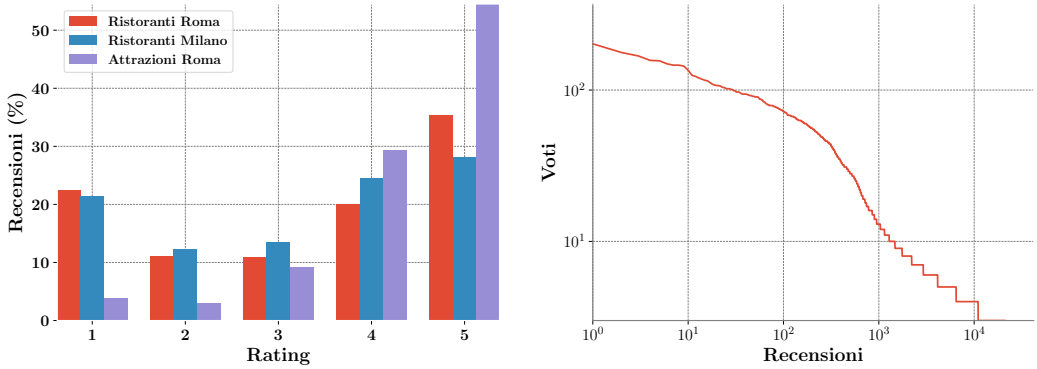
e presentano un'ulteriore suddivisione per città di appartenenza delle proprietà recensite nel caso dei ristoranti: le città prese in considerazione per i ristoranti sono Roma e Milano; mentre le attrazioni turistiche sono esclusivamente romane. Per motivazioni inerenti agli scopi di ricerca, i dati iniziali sono stati filtrati per isolare quelli in italiano, scartando anche le recensioni troppo corte (poco informative e scoraggiate da Tripadvisor stesso) e suddivisi in *utili* e *non utili* in base al numero di voti. In particolare, si considerano utili le recensioni che hanno ricevuto almeno 3 voti e non utili quelle che non ne hanno ricevuto nessuno. Poiché le recensioni utili costituivano solo il 5.1% del totale, il dataset è stato anche bilanciato per ottenere un numero comparabile di recensioni utili e non utili in ogni categoria.

Numero di utenti	30578
Numero di recensioni	42107
Numero di parole	4133312
Media dei voti per recensione ($\pm\sigma$)	2.89 ($\pm$7.32)
Intervallo temporale	2006-2019
Numero di recensioni per categoria:	
- Ristoranti di Roma	25039
- Ristoranti di Milano	12096
- Attrazioni di Roma	4972
Numero totale di voti utili	121822
- Recensioni utili (#voti ≥ 3)	21304
- Recensioni non utili (#voti $= 0$)	20803

Tabella 2.5: Statistiche principali calcolate sul dataset di Tripadvisor.

La Tabella 2.5 riporta alcune informazioni più dettagliate sulla composizione del dataset dopo il bilanciamento. In Figura 2.6 è invece possibile osservare la distribuzione dei metadati: il numero di voti utili segue una distribuzione a legge di potenza (è in generale molto vicino a zero, con poche eccezioni che corrispondono a recensioni in grado di raccogliere centinaia di preferenze); la distribuzione dei rating non presenta differenze significative nel caso dei ristoranti di Roma e Milano ma mostra un andamento differente nel caso delle attrazioni turistiche, che tendono a ricevere recensioni più positive rispetto ai ristoranti,

almeno per quanto riguarda il rating associato (da notare è, infatti, che più della metà delle recensioni raccolte assegna il punteggio massimo all'attrazione oggetto di valutazione).



(a) Distribuzione dei *rating* per categoria.

(b) Distribuzione dei *voti utili*.

Figura 2.6: Distribuzione dei *rating* e dei *voti utili* delle recensioni presenti nel dataset di Tripadvisor. La distribuzione dei voti utili è rappresentata in doppia scala logaritmica.

Classe	Categoria	Esempio
Utile	Ristoranti di Roma	La prima regola di un buon ristorante che fa pizza no stop è: Scegliere la pizza che preferisco. Qui non solo non si può scegliere la pizza ma capita spesso che escano le stesse pizze più volte così uno è costretto a mangiare sempre la stessa!! Per non parlare dell'ambiente poi, un vero casino, capisco che l'area bambini è la principale attrazione del ristorante, rivolto soprattutto alle famiglie, ma il casino che si crea non è cmq giustificabile. La pizza è di una qualità davvero scadente, praticamente era cruda!!! La pizza con la Lonza....una semplice focaccia con un pezzo di prosciutto preso molto probabilmente al discount! Ragazzi, carina l'idea di prendersi cura dei pargoli, ma non prendiamoci in giro però.
Non utile	Ristoranti di Milano	Devo dire che trovandomi per caso in quella zona con i miei amici abbiamo provato il posto è devo dire ché è molto accogliente e che la zona per mangiare nel cortile è proprio intima e carina...Per quanto riguarda il mangiare posso dire di essere soddisfatto perché le portate erano nelle mie corde ed avendo preso il pesce ero soddisfatto di quanto cucinato dal cuoco. Bravi mica male.

Tabella 2.6: Esempi di recensioni utili e non utili.

Capitolo 3

Il monitoraggio linguistico

Un approccio ormai consolidato per lo studio della variazione linguistica affonda le sue radici nella disponibilità di corpora testuali di grandi dimensioni e di tecnologie, basate sul Natural Language Processing, per la loro analisi automatica. La metodologia consiste nella ricostruzione del "profilo linguistico" di un testo (o di una collezione di testi) attraverso il computo di un ampio numero di tratti che spazino attraverso diversi livelli di descrizione linguistica (van Halteren, 2004; Montemagni, 2013). L'assunzione di base è che questi tratti contribuiscano a caratterizzare la variazione nella struttura linguistica dei testi, modellandone la forma o lo stile: dal confronto tra i profili di testi rappresentativi di varietà diverse si cerca dunque di valutare il modo in cui queste ultime differiscano rispetto ai parametri monitorati. Campi di applicazione comprendono l'analisi computazionale della variazione tra registri (Argamon, 2019), l'attribuzione di un testo al suo autore in base agli stilemi che ne caratterizzano le produzioni (Daelemans, 2013) oppure lo studio delle differenze nell'uso della lingua correlate a variabili sociolinguistiche (Nguyen et al., 2016). Le tecniche di monitoraggio si ispirano alla *Multi-Dimensional Analysis* di Biber, che nel suo lavoro sulla variazione di genere nello scritto e nel parlato sull'inglese (Biber, 1988) identifica cinque dimensioni di variazione, una delle quali riferita a "espressione esplicita di persuasione", cui sono associate positivamente

caratteristiche quali il maggior numero di infinitive e verbi modali e l'uso delle subordinate condizionali.

L'obiettivo di questo capitolo consiste nell'impiegare tale quadro di analisi per verificare se esista una differenza statisticamente significativa nella forma di testi "persuasivi" e "non persuasivi" appartenenti alle due tipologie prese in esame. Nel primo caso di studio, i dati linguistici che si considerano rappresentativi del fenomeno della persuasività sono finestre di frasi che precedono una reazione del pubblico all'interno del corpus di discorsi politici CORPS, in cui queste ultime sono individuate da tag posti in corrispondenza di risate, applausi o anche fischi ed espressioni di disprezzo. I discorsi politici sono generalmente pianificati prima di essere pronunciati ed è ragionevole pensare che gli autori (di solito non gli oratori stessi ma figure professionali preposte a questo scopo) adottino particolari strategie, codificate anche nello stile dei discorsi e che queste varino anche a seconda del tipo di reazione. Per questo motivo si è scelto di condurre l'analisi sui tag *Positive_Focus* e *Ironical_Focus*, considerati sia singolarmente che in combinazione, escludendo tuttavia il tag *Negative_Focus* per la sua scarsa presenza nel corpus.

Nel secondo caso di studio, che è invece inerente alle recensioni online, esempi di testi persuasivi sono le recensioni del portale di viaggi Tripadvisor che hanno ottenuto *voti utili* da parte degli altri utenti, da porre a confronto con quelle che non ne hanno ricevuto nessuno. Come anticipato nel capitolo precedente, i voti degli utenti sono considerati in letteratura al pari di giudizi umani sull'utilità delle recensioni, che può essere intesa anche come persuasività se si considera che la rilevanza per l'utente sta nel modo in cui queste influenzano il suo processo decisionale nell'acquistare un prodotto oppure nel frequentare un ristorante o un'altra attività (nel caso specifico di Tripadvisor). Benché una varietà di fattori esterni al testo influisca sulle votazioni di utilità, diversi studi, tra quelli basati sull'estrazione di caratteristiche definite manualmente, hanno coinvolto aspetti della struttura linguistica delle recensioni (v. par. 2.2), evidenziando anche la

necessità di indagare in modo più approfondito il loro profilo stilistico (W. Hong et al., 2020) attraverso "rappresentazioni testuali più sofisticate" (Ocampo Diaz e Ng, 2018): in questo caso può costituire un vantaggio analizzare un più ampio spettro di tratti.

Per poter procedere con il monitoraggio, i testi (in inglese nel caso dei discorsi politici e in italiano per le recensioni online) sono stati annotati in modo automatico. Per i discorsi politici è stata utilizzata UDPipe[1] (Straka e Straková, 2017), una pipeline per la tokenizzazione, il PoS Tagging, la lemmatizzazione e il parsing di quasi tutte le lingue presenti nel progetto Universal Dependencies[2] (Nivre, 2015), che ha l'obiettivo di sviluppare un'annotazione delle treebank coerente a livello cross-linguistico. L'annotazione condotta da UDPipe è basata sui modelli UD disponibili per la versione 2.5[3] e restituisce in output il risultato in formato CoNLL-U,[4] in cui le frasi sono separate da linee vuote e ciascuna parola, individuata da un numero intero che funge da indice, è seguita da campi dove sono riportati i risultati dell'analisi automatica in modo incrementale, separati da tabulazione. Per quanto riguarda le recensioni online, i testi sono stati annotati utilizzando gli strumenti realizzati presso l'Istituto di Linguistica Computazionale del CNR e l'Università di Pisa: il PoS-Tagger descritto in Dell'Orletta (2009) e il dependency parser DeSR (Attardi et al., 2009). A partire dai testi annotati, sono poi state estratte attraverso degli script in Python una serie di caratteristiche volte a modellare la forma linguistica dei testi, che si sono dimostrate efficaci per diversi compiti, tra i quali la classificazione di generi testuali (Cimino et al., 2017) e la predizione della complessità (Brunato et al., 2018). Le caratteristiche linguistiche considerate fanno riferimento ai differenti livelli

[1]UDPipe 1.2: `https://ufal.mff.cuni.cz/udpipe/1` (ultimo accesso: 07/04/2022).

[2]Universal Dependencies: `https://universaldependencies.org/` (ultimo accesso: 07/04/2022).

[3]Modelli UD 2.5 per UDPipe. `http://hdl.handle.net/11234/1-3131` (ultimo accesso: 07/04/2022).

[4]Descrizione del formato CoNLL-U dal sito di Universal Dependencies: `https://universaldependencies.org/format.html` (ultimo accesso: 07/04/2022).

di annotazione e possono essere raggruppate in:

- *Proprietà di base*, come la lunghezza del documento (calcolata in termini di token e frasi), delle frasi (in termini di token) e delle parole (equivalente al numero medio di caratteri nei token, esclusa la punteggiatura).

- Proprietà relative alla *ricchezza lessicale* del testo, come la Type/Token Ratio (TTR), ossia il rapporto tra il numero di parole tipo e parole token, calcolato sui primi 100 e 200 token. Sempre sotto il profilo lessicale, per l'italiano è stata anche calcolata la percentuale delle parole (in termini di forme e lemmi) appartenenti al Vocabolario di Base della lingua italiana di Tullio De Mauro (De Mauro, 2000) e singolarmente anche nei tre repertori in cui è ripartito (vocabolario fondamentale, di alto uso e di alta disponibilità).

- *Proprietà morfo-sintattiche*: la distribuzione percentuale delle categorie morfo-sintattiche definite nello Universal POS Tagset[5] (per i discorsi politici in inglese) e nel tagset utilizzato per l'annotazione della treebank ISST-TANL[6] (per le recensioni in italiano); la densità lessicale, intesa come il rapporto tra le parole lessicalmente piene (nomi, verbi, aggettivi, avverbi) e il numero di parole totali; le caratteristiche relative alla morfologia flessiva dei verbi (calcolate anche per gli ausiliari), quali le distribuzioni di modo, numero, tempo, forma, numero e persona.

- *Proprietà sintattiche*: *(i)* la percentuale di relazioni di dipendenza assegnate sulla base delle relazioni individuate nella versione 2 di Universal Dependencies[7] (per i discorsi) e dell'ISST-TANL dependency tagset[8] (per

[5]Universal POS Tagset, `https://universaldependencies.org/u/pos/index.html` (ultimo accesso: 07/04/2022).

[6]ISST-TANL PoS Tagset, `http://www.italianlp.it/docs/ISST-TANL-POStagset.pdf` (ultimo accesso: 07/04/2022).

[7]UD dependency tagset, `https://universaldependencies.org/u/dep/index.html` (ultimo accesso: 07/04/2022).

[8]ISST-TANL dependency tagset, `http://www.italianlp.it/docs/ISST-TANL-DEPtagset.pdf` (ultimo accesso: 07/04/2022).

le recensioni); *(ii)* caratteristiche relative alla struttura dei predicati, come l'arità verbale, intesa come il numero medio di dipendenti per testa verbale, e la distribuzione percentuale dei verbi per arità verbale; la percentuale di radici verbali e il numero di teste verbali per frase; *(iii)* le caratteristiche che descrivono gli alberi sintattici, come la loro profondità media (la media delle profondità massime di ogni frase), il numero medio di token per clausola (calcolato come il rapporto tra il numero di token e il numero di teste verbali o copule), la lunghezza media e massima dei link di dipendenza (calcolata linearmente considerando il numero di occorrenze tra la testa sintattica e il dipendente) e la media delle lunghezze massime; il numero e la profondità media delle catene di modificatori preposizionali; *(iv)* caratteristiche inerenti alla subordinazione, quali il numero di subordinate (in confronto a quello di proposizioni principali) e la lunghezza media delle catene di subordinazione; *(v)* le caratteristiche legate all'ordine delle parole (la distribuzione di soggetti e oggetti pre e post-verbali).[9]

Per valutare se esista una differenza statisticamente significativa nelle distribuzioni delle caratteristiche linguistiche tra i campioni di testi persuasivi e non persuasivi, è stato poi impiegato il Wilcoxon Rank-sum test (o test di Mann-Whitney-U), un test di verifica di ipotesi non parametrico, scelto per la sua robustezza alla differenza di dimensione dei campioni e perché i tratti studiati non seguono una distribuzione normale.[10] Ponendo la soglia di significatività a 0.05, quando il p-value risulta inferiore a 0.05 si rifiuta l'ipotesi nulla che le caratteristiche delle due categorie siano tratte dalla stessa distribuzione, ossia che per valori selezionati casualmente dal campione di testi persuasivi P e dal campione di testi non persuasivi $\neg P$, la probabilità di ottenere un valore di P maggiore di

[9] Questo tratto è stato preso in considerazione solo per l'italiano, dato che in inglese l'ordine SVO dei costituenti tende a essere fisso.

[10] Il fatto che i tratti non seguano una distribuzione normale è stato anche confermato dai risultati del test di normalità di Shapiro-Wilk: infatti per ciascun tratto è stato ottenuto un p-value inferiore a 0.05, che permette di rifiutare l'ipotesi di normalità.

un valore di $\neg P$ sia uguale alla probabilità di ottenere un valore di $\neg P$ maggiore di P.

Nel caso dei discorsi politici, un importante *caveat* per questo tipo di analisi sta nel fatto che CORPS contiene trascrizioni a posteriori di discorsi tenuti oralmente, quindi potrebbero esserci disallineamenti dovuti al modo in cui i testi sono stati trascritti rispetto a quello in cui sono stati prodotti (per esempio per le scelte nell'uso della punteggiatura). Un'altra problematica riguarda le modalità di produzione delle trascrizioni da parte delle fonti web originarie: non è possibile sapere se i discorsi siano stati trascritti manualmente o in modo automatico e come siano state gestite le interruzioni tipiche del parlato. Per quanto riguarda invece le recensioni, occorre tenere presente che i testi sono scritti in italiano non standard e che l'analisi automatica con modelli sviluppati per il linguaggio standard tende a essere soggetta a una significativa diminuzione dell'accuratezza. In ogni caso, sia per i discorsi politici che per le recensioni online, si può assumere che gli errori siano distribuiti in modo simile tra le due categorie, perché gli esempi persuasivi e non persuasivi appartengono allo stesso dominio: quindi, un'analisi delle differenze interne ai testi relativi ai due casi di studio non dovrebbe essere invalidata dagli errori dell'analisi automatica.

3.1 Il monitoraggio dei discorsi politici

Per quanto riguarda i discorsi politici, si è scelto di focalizzare l'analisi sulle finestre lunghe quattro frasi. Dall'osservazione delle caratteristiche estratte, si evidenziano una serie di differenze nella distribuzione dei tratti monitorati per le finestre che precedono una reazione del pubblico. Si tratta di differenze talvolta molto sottili, probabilmente perché in questo caso di studio non si cerca di trovare variazioni di stile da un documento all'altro ma piuttosto all'interno degli stessi testi. Queste emergono non solo tra esempi positivi e negativi ma anche internamente alle finestre persuasive, a seconda del tipo di tag che vi è

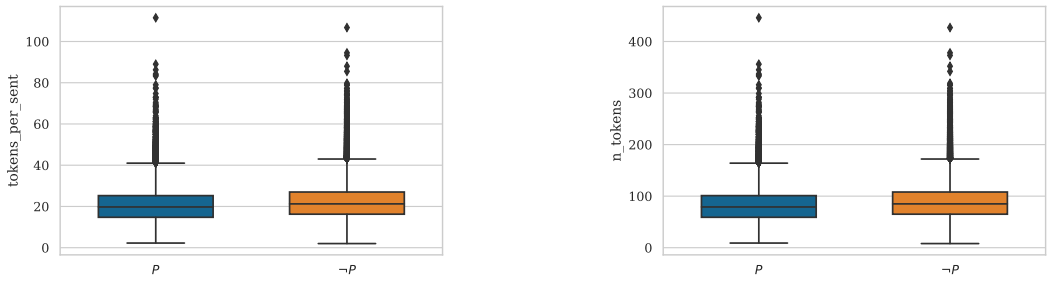
associato, offrendo evidenza del fatto che le strategie impiegate per suscitare applausi divergano da quelle impiegate per suscitare risate. I tratti più rilevanti ai fini dell'analisi sulla persuasività sono descritti nei sottoparagrafi successivi, organizzati in base al livello di annotazione linguistica cui fanno riferimento. Rispetto a un totale di 73 tratti considerati, il 71% varia in modo statisticamente significativo in base al Wilcoxon Rank-sum test.

3.1.1 Le caratteristiche di base

Un primo aspetto di interesse, ricavabile dal testo sottoposto solo a *sentence splitting* e tokenizzazione, riguarda la lunghezza degli estratti. Le finestre persuasive contengono frasi mediamente più corte di quelle non persuasive, in particolare nel caso in cui precedano una risata, benché la differenza tra i valori medi del numero di token per frase tra finestre persuasive e non persuasive sia piuttosto ridotta (inferiore a 2 token anche nel caso delle finestre *Ironical_Focus*, come mostrato in Tabella 3.1, in cui le statistiche descrittive relative alla lunghezza sono comparate tra le diverse categorie, oltre che ai valori medi per discorso in tutto il corpus). La stessa informazione può essere ricavata anche dal numero di token totale: in media, le finestra *Positive_Ironical* sono di circa 6 token più corte di quelle lontane dai tag, con una differenza maggiore (di circa 8 token) sempre nel caso delle frasi che precedono risate. Come si può vedere in Figura 3.1, le distribuzioni delle lunghezze presentano un'asimmetria a sinistra. Per esempio, per quanto riguarda la lunghezza media per frase, la maggior parte delle finestre è costituita da frasi mediamente inferiori a 40 token ma una serie di outlier con frasi più lunghe formano la coda della distribuzione. Osservando poi le statistiche relative alla lunghezza delle parole, è possibile notare che non sussistono grandi differenze nel numero medio di caratteri benché quest'ultimo sia di nuovo leggermente inferiore nel caso delle finestre persuasive e in particolare di quelle che precedono risate.

	CORPS		Non persuasive		Persuasive					
Caratteristiche	-		-		Positive_Ironical		Positive		Ironical	
n_tokens		-	89.34	(34.91)	83.42	(34.71)	83.95	(35.05)	81.47	(33.42)
tokens_per_sent	22.12	(4.17)	22.33	(8.73)	20.85	(8.68)	20.99	(8.76)	20.37	(8.35)
char_per_tok	4.35	(0.21)	4.37	(0.41)	4.27	(0.41)	4.31	(0.40)	4.13	(0.39)

Tabella 3.1: Statistiche descrittive (media e deviazione standard) relative alle caratteristiche di base di finestre persuasive e non persuasive nei discorsi politici in CORPS.



(a) Numero medio di token per frase

(b) Numero di token

Figura 3.1: Distribuzione della lunghezza nelle finestre persuasive e non persuasive.

3.1.2 Le caratteristiche morfo-sintattiche

Attingendo al livello di annotazione morfo-sintattico dei testi, si possono trarre alcune considerazioni a partire dalla distribuzione delle parti del discorso (v. tab. 3.2). In particolare, osservando la differenza nella distribuzione di nomi e verbi, le finestre persuasive appaiono contraddistinte da un minor numero di sostantivi rispetto a quelle non persuasive e da un maggior numero di verbi. Probabilmente perché si tratta di parametri correlati alla proporzione inferiore di sostantivi, le finestre persuasive presentano anche meno aggettivi, determinanti e preposizioni. Questi aspetti possono essere interpretati come una spia del fatto che nelle parti dei discorsi lontane dagli applausi venga impiegato uno stile nominale, tipicamente associato a maggiore densità informativa. A parte, si può invece considerare il caso dei nomi propri, che sono più frequenti nelle finestre persuasive. Non risulta difficile immaginare che la risata o l'applauso

scaturiscano in relazione alla citazione di persone, luoghi o istituzioni specifici: un'occorrenza comune del tag *Positive_Focus* nel corpus corrisponde infatti ai ringraziamenti in apertura o in chiusura del discorso. In relazione ai verbi, è poi interessante notare che questi si trovano in proporzione maggiore nelle finestre che precedono risate, accompagnati da pronomi e avverbi. Gli avverbi sono invece meno frequenti nelle finestre persuasive quando si considerano i due tag in combinazione.

	CORPS		Non persuasive		Persuasive					
Caratteristiche	-		-		Positive_Ironical		Positive		Ironical	
upos_dist_ADJ	5.88	(1.04)	6.08	(3.05)	5.59	(3.05)	5.75	(3.08)	4.96	(2.86)
upos_dist_ADP	9.27	(1.19)	9.13	(3.35)	8.60	(3.43)	8.74	(3.42)	8.04	(3.41)
upos_dist_ADV	5.07	(1.14)	5.24	(3.07)	5.05	(3.24)	4.80	(3.18)	6.04	(3.28)
upos_dist_DET	8.61	(1.10)	8.51	(3.10)	8.18	(3.21)	8.28	(3.24)	7.78	(3.08)
upos_dist_NOUN	16.04	(2.28)	16.53	(5.06)	15.20	(5.40)	15.66	(5.44)	13.40	(4.81)
upos_dist_PRON	11.29	(2.09)	11.30	(4.61)	12.21	(4.79)	11.94	(4.84)	13.29	(4.41)
upos_dist_PROPN	5.15	(2.38)	3.88	(4.34)	5.02	(5.06)	5.07	(5.09)	4.83	(4.91)
upos_dist_VERB	11.46	(1.34)	11.67	(3.52)	12.01	(3.74)	11.96	(3.78)	12.19	(3.59)

Tabella 3.2: Statistiche descrittive (media e deviazione standard) relative alla distribuzione percentuale delle parti del discorso in finestre persuasive e non persuasive nei discorsi politici in CORPS.

	CORPS		Non persuasive		Persuasive					
Caratteristiche	-		-		Positive_Ironical		Positive		Ironical	
verbs_form_dist_Fin	40.05	(6.41)	39.98	(20.46)	41.70	(21.47)	40.29	(21.32)	47.22	(21.11)
verbs_form_dist_Inf	32.59	(5.88)	31.26	(19.02)	32.01	(19.52)	33.22	(19.68)	27.29	(18.11)
verbs_num_pers_dist_Sing+3	7.18	(3.06)	7.51	(10.99)	7.19	(11.27)	7.28	(11.30)	6.84	(11.11)
verbs_tense_dist_Past	44.55	(12.02)	45.65	(31.22)	40.70	(31.12)	38.61	(30.96)	48.85	(30.39)
verbs_tense_dist_Pres	55.45	(12.02)	53.48	(31.32)	58.18	(31.44)	60.15	(31.39)	50.46	(30.41)

Tabella 3.3: Statistiche descrittive (media e deviazione standard) relative alle caratteristiche flessive dei verbi in finestre persuasive e non persuasive nei discorsi politici in CORPS.

Per quanto riguarda le caratteristiche flessive dei verbi, queste sono soggette a un alto grado di dispersione intorno alla media, come mostrano i valori di deviazione standard superiori a 3 per i parametri relativi all'intero corpus (v. tab. 3.3). La deviazione standard cresce ulteriormente se si esaminano separatamente le finestre delle diverse classi. Nonostante questo, è interessante evidenziare alcune differenze interne alle finestre persuasive che mostrano valori simili di

deviazione standard. Nelle frasi che precedono gli applausi prevalgono i verbi all'infinito, mentre la situazione è opposta nelle frasi che precedono risate, in cui si impiegano per lo più forme finite. Per quanto riguarda i tempi verbali, le finestre *Positive_Focus* presentano in media più occorrenze di verbi al presente; le risate, invece, tendono a seguire frasi con una prevalenza di verbi al passato, probabilmente dovuti al racconto di un aneddoto che termina con la battuta finale. La minore percentuale di verbi alla terza persona singolare (rispetto alla prima e alla seconda) per le finestre persuasive in confronto a quelle non persuasive indica uno stile più personale, adottato soprattutto nelle finestre *Ironical_Focus*.

3.1.3 Le caratteristiche sintattiche

I parametri estratti dall'annotazione a dipendenze in alcuni casi fungono da ulteriore conferma di alcune caratteristiche evidenziate nell'analisi della distribuzione delle parti del discorso, oltre a fornire nuove informazioni. In generale, le finestre persuasive tendono a contenere un maggior numero di argomenti sottocategorizzati dai verbi (come riportato in Tabella 3.4 per la distribuzione della relazione di soggetto *nsubj* e oggetto diretto *obj*) e lo stesso vale anche se si considera la maggiore presenza di subordinate argomentali (come *ccomp* e *xcomp*) nelle finestre persuasive.

Le frasi persuasive sono anche meno complesse: questo aspetto è evidente dal calcolo del numero medio di token per clausola, che tende a essere minore nel caso delle finestre che precedono un tag di reazione, ma anche ad altri aspetti associati alla complessità del testo, quali il numero e la lunghezza media delle catene preposizionali così come la profondità media degli alberi sintattici. Per quanto riguarda la subordinazione, nel caso degli applausi si riscontra un minor numero di subordinate; mentre nel caso delle finestre *Ironical_Focus* se ne ritrova un uso più ampio, anche se le subordinate tendono a occorrere in posizione canonica post-verbale. Nel caso delle finestre persuasive si evidenzia invece

un uso maggiore della coordinazione (in particolare per i tag *Positive_Focus*) e della paratassi (soprattutto per il tag *Ironical_Focus*).

	CORPS	Non persuasive	Persuasive		
Caratteristiche	-	-	Positive_Ironical	Positive	Ironical
dep_dist_amod	4.45 (1.03)	4.63 (2.77)	4.11 (2.72)	4.25 (2.76)	3.56 (2.48)
dep_dist_conj	3.80 (0.80)	3.44 (2.27)	3.45 (2.33)	3.57 (2.34)	2.97 (2.23)
dep_dist_nmod	4.04 (0.94)	3.89 (2.48)	3.59 (2.50)	3.73 (2.52)	3.06 (2.35)
dep_dist_nsubj	8.76 (1.32)	9.10 (3.04)	9.36 (3.12)	9.13 (3.08)	10.25 (3.13)
dep_dist_obj	5.65 (0.88)	5.60 (2.77)	6.07 (3.02)	6.20 (3.07)	5.55 (2.77)
dep_dist_parataxis	0.41 (0.24)	0.40 (0.76)	0.44 (0.83)	0.40 (0.79)	0.58 (0.96)
avg_token_per_clause	8.20 (0.93)	8.50 (2.64)	8.28 (2.64)	8.31 (2.68)	8.15 (2.51)
verbal_head_per_sent	2.71 (0.51)	2.77 (1.17)	2.66 (1.18)	2.67 (1.19)	2.62 (1.14)
avg_links_len	2.92 (0.26)	2.79 (0.52)	2.75 (0.54)	2.74 (0.54)	2.77 (0.54)
avg_max_depth	4.24 (0.49)	4.30 (1.22)	4.07 (1.21)	4.11 (1.22)	3.94 (1.19)
principal_proposition_dist	40.13 (7.03)	43.08 (19.32)	45.14 (20.19)	45.39 (20.25)	44.07 (19.97)
subordinate_proposition_dist	59.87 (7.03)	56.88 (19.34)	54.85 (20.20)	54.60 (20.25)	55.93 (19.97)

Tabella 3.4: Statistiche descrittive (media e deviazione standard) relative alle caratteristiche sintattiche in finestre persuasive e non persuasive nei discorsi politici in CORPS.

3.2 Il monitoraggio delle recensioni online

Come anticipato, l'analisi del monitoraggio delle recensioni online ha riguardato le caratteristiche dello stile di scrittura delle recensioni, con l'obiettivo di identificare dei tratti che permettano di distinguere le recensioni marcate come utili dagli altri utenti. Allo stesso modo dei discorsi politici, nei prossimi paragrafi si riporta una selezione di statistiche descrittive di base relative ai tratti che costituiscono il profilo linguistico delle due categorie prese in esame, con l'obiettivo di dare una prima rappresentazione delle principali dimensioni di variazione. Tutte quelle riportate nei paragrafi successivi fanno riferimento a tratti per cui, in base al Wilcoxon Rank-sum test, è possibile rifiutare con un p-value inferiore a 0.05 l'ipotesi nulla secondo la quale i valori dei due campioni sono tratti dalla stessa distribuzione, come accade per il 71% delle 212 caratteristiche monitorate.

3.2.1 Le caratteristiche di base

Una delle caratteristiche che distingue le recensioni utili più di tutte le altre è la loro lunghezza, sebbene ci sia un'alta deviazione standard, in particolare nel caso della categoria utile. Anche se le recensioni tendono a essere in linea di massima piuttosto corte, i testi votati dagli utenti sono in media più lunghi di una frase rispetto a quelli non votati e le frasi stesse contengono un numero molto maggiore di token (v. tab. 3.5). La correlazione tra l'utilità e la lunghezza non risulta inattesa, dato che testi più lunghi tipicamente veicolano anche più informazioni, che possono risultare utili nel decidere quale ristorante o quale luogo turistico frequentare. Per quanto riguarda invece la lunghezza delle parole, non si evidenziano forti differenze nel numero medio di caratteri per token.

Caratteristiche	Utili		Non utili	
N. di frasi	4.61	(4.23)	3.46	(2.49)
N. di token per frase	36.79	(38.39)	26.22	(24.86)
N. di caratteri per token	5.05	(0.45)	5.19	(0.56)

Tabella 3.5: Statistiche descrittive (media e deviazione standard) relative alle caratteristiche di base delle recensioni utili e non utili.

3.2.2 Le caratteristiche associate alla ricchezza lessicale

Un altro aspetto da sottolineare di divergenza tra recensioni utili e non utili è la Type/Token Ratio calcolata in rapporto ai primi 200 token del testo, sia quando al numeratore si considerano le forme di parole che i lemmi. Le recensioni utili sono caratterizzate da valori tipicamente più bassi di TTR, che corrispondono a una minore ricchezza lessicale. Per quanto riguarda invece la distribuzione delle parole nel Vocabolario di Base dell'italiano di De Mauro (De Mauro, 2000), la percentuale di lemmi o di forme che vi appartengono è pressoché la stessa, mentre è più interessante osservare come si distribuiscono le forme nei repertori d'uso. Infatti, anche se il lessico fondamentale è prevalente in entrambe le

categorie, le recensioni utili presentano una percentuale in media più bassa di forme appartenenti al Vocabolario Fondamentale (*FO* in tab. 3.6), mentre più alta nel caso di parole meno frequenti nell'italiano (classificate nel Vocabolario di De Mauro come ad alto uso -*AU* in tabella- e ad alta disponibilità -*AD*). In ogni caso, per tutti i tratti citati il Wilcoxon Rank-sum test permette di rifiutare l'ipotesi che i valori afferenti alle recensioni utili e alle recensioni non utili siano tratti dalla stessa distribuzione.

Caratteristiche	Utili		Non utili	
AD	10.27	(4.65)	9.65	(5.31)
AU	13.03	(4.91)	12.77	(5.75)
FO	76.70	(6.48)	77.58	(7.38)
Range 200.0 Type(forme)/token	0.78	(0.08)	0.82	(0.08)
Range 200.0 Type(lemmi)/token	0.70	(0.10)	0.75	(0.10)
Percentuale delle parole nel dizionario:	82.78	(6.63)	82.79	(7.40)
Percentuale dei lemmi nel dizionario	73.57	(7.44)	74.26	(8.02)

Tabella 3.6: Statistiche descrittive (media e deviazione standard) relative alla TTR e alla distribuzione nei lessici di frequenza di De Mauro delle recensioni utili e non utili.

3.2.3 Le caratteristiche morfo-sintattiche

Nel caso delle caratteristiche morfo-sintattiche, si osserva una differenza nella distribuzione dei sostantivi e dei verbi. Infatti, la categoria utile contiene recensioni con una percentuale di verbi in media maggiore rispetto a quella delle recensioni non utili e una minor percentuale di nomi. Il maggior numero di nomi in generale è correlato a una maggiore densità informativa, quindi il fatto che le recensioni utili contengano più verbi contraddice parzialmente quanto desunto dalle considerazioni sulla lunghezza esposte nel paragrafo precedente, dando invece un segnale di uno stile maggiormente focalizzato sul lettore. Ulteriori considerazioni che si possono trarre dalla distribuzione delle parti del discorso comprendono: il minor numero di aggettivi per quanto riguarda le recensioni utili; il maggior numero di avverbi, in particolare di negazione, e il maggior nu-

mero di pronomi (tutti tratti correlati al maggior numero di verbi). Per quanto riguarda invece le caratteristiche flessive dei verbi, tra i tratti che presentano una differenza più evidente per le due classi, si nota un maggior numero di verbi alla prima persona singolare e alla seconda persona plurale e un maggior numero di verbi all'imperfetto (v. tab. 3.8).

Caratteristiche	Utili		Non utili	
CPOS_S	23.50	(5.20)	24.50	(5.83)
CPOS_V	14.28	(5.17)	12.79	(5.69)
CPOS_A	8.32	(4.31)	10.73	(5.56)
CPOS_B	7.41	(3.63)	7.15	(4.11)
POS_BN	1.33	(1.45)	0.97	(1.45)
CPOS_P	4.99	(3.05)	4.14	(3.24)
CPOS_E	12.87	(4.11)	12.42	(4.80)
Densità Lessicale	0.58	(0.06)	0.61	(0.07)

Tabella 3.7: Statistiche descrittive (media e deviazione standard) relative alle caratteristiche morfo-sintattiche delle recensioni utili e non utili.

Caratteristiche	Utili		Non utili	
Verbi+Numero+Persona_V+s+1	9.23	(15.23)	8.15	(17.59)
Verbi+Numero+Persona_V+p+2	1.34	(6.57)	1.08	(6.08)
Verbi+Tempo_V+i	11.17	(16.66)	7.67	(16.24)

Tabella 3.8: Statistiche descrittive (media e deviazione standard) relative alle caratteristiche flessive dei verbi nelle recensioni utili e non utili.

3.2.4 Le caratteristiche sintattiche

Per quanto riguarda le caratteristiche estratte dall'analisi a dipendenze, si possono osservare alcuni tratti interessanti riguardo al tipo di relazioni, per esempio il fatto che i complementi e i modificatori di luogo e di tempo siano più frequenti nel caso delle recensioni utili, anche se entrambi i tratti sono caratterizzati da un'alta deviazione standard. Altre caratteristiche, correlate alla lunghezza delle recensioni utili, sono indice di una maggiore complessità sintattica per la categoria utile. Questa è evidenziata dalla maggiore lunghezza media delle relazioni

di dipendenza (calcolata linearmente in termini di numero di token interposti tra le testa della relazione e il dipendente) e una maggiore altezza media degli alberi sintattici delle frasi. Per quanto riguarda l'uso della subordinazione, si può notare che le recensioni utili presentano una maggiore proporzione di subordinate e, in più, catene di subordinate incassate mediamente più lunghe.

Caratteristiche	Utili		Non utili	
DIP_subj	3.94	(2.32)	3.46	(2.69)
DIP_obj	3.49	(2.23)	2.95	(2.44)
DIP_mod_temp	0.58	(0.97)	0.50	(1.12)
DIP_comp_loc	0.68	(1.11)	0.76	(1.36)
DIP_comp_temp	0.23	(0.54)	0.18	(0.60)
DIP_neg	1.31	(1.44)	0.96	(1.44)
DIP_conj	5.72	(2.96)	6.25	(3.70)
Numero di token per Clausola	10.00	(5.48)	11.65	(6.86)
Media della lunghezza dei Link	2.51	(0.78)	2.32	(0.69)
Media della lunghezza dei Link Massimi	13.67	(14.87)	10.33	(10.45)
Media delle Altezze Massime degli alberi	7.64	(5.41)	6.28	(4.06)
Frasi Principali	56.29	(16.74)	57.42	(14.17)
Frasi Subordinate	43.71	(16.74)	42.58	(14.17)
Lunghezza Media delle Catene Subordinanti	0.60	(0.68)	0.38	(0.60)
Lunghezza Media delle Catene preposizionali	1.21	(0.27)	1.19	(0.34)

Tabella 3.9: Statistiche descrittive (media e deviazione standard) relative alle caratteristiche sintattiche delle recensioni utili e non utili.

3.3 Discussione

Come anticipato, il Wilcoxon rank-sum test restituisce un p-value inferiore a 0.05 in circa il 70% dei tratti estratti da entrambe le tipologie di testi. Per individuare quelli che presentano maggiori differenze nella distribuzione, si è anche scelto di considerare una misura di *effect size* del test di verifica di ipotesi, calcolata a partire dalla statistica U risultante dal test insieme al p-value. In particolare, si è scelto di usare la *rank biserial correlation*, che corrisponde alla differenza tra la proporzione di coppie di istanze estratte dai due campioni (su tutte quelle possibili) per cui l'istanza della categoria persuasiva risulti

maggiore di quella della categoria non persuasiva e la proporzione di coppie per cui valga il contrario (e risulti dunque maggiore l'istanza della categoria non persuasiva). I valori possibili della *rank biserial correlation* spaziano da -1 (il caso in cui tutte le istanze della classe non persuasiva siano maggiori di tutte quelle della classe persuasiva) a 1 (che rappresenta invece il caso opposto, ossia quello in cui tutte le istanze della classe persuasiva siano maggiori di quella non persuasiva).[11]

CORPS			Tripadvisor		
Caratteristica	abs(r)	Segno	Caratteristica	abs(r)	Segno
xpos_dist_PRP	0.15	+	Totale Archi Entranti in Teste Verbali	0.38	+
xpos_dist_NNP	0.15	+	Totale Teste Verbali	0.37	+
upos_dist_PROPN	0.14	+	Numero di Token	0.37	+
upos_dist_NOUN	0.14	−	Totale catene preposizionali	0.34	+
xpos_dist_NNS	0.13	−	Range 200.0 Type(lemmi)/token	0.32	−
char_per_tok	0.14	−	Range 200.0 Type(forme)/token	0.29	−
avg_max_depth	0.11	−	POS_A	0.28	−
upos_dist_PRON	0.11	+	CPOS_A	0.28	−
dep_dist_amod	0.11	−	Media di Teste Verbali per frase	0.27	+
tokens_per_sent	0.11	−	Totale Radici Verbali	0.24	+
n_prepositional_chains	0.10	−	Numero di Token per Frase	0.23	+
xpos_dist_.	0.10	+	DIP_ROOT	0.21	−
upos_dist_ADJ	0.09	−	Densità Lessicale	0.20	−
xpos_dist_IN	0.09	−	Media della lunghezza dei Link Massimi	0.20	+
verbs_tense_dist_Past	0.09	−	Media delle Altezze Massime degli alberi	0.20	+
xpos_dist_JJ	0.09	−	Totale Strutture Subordinate	0.19	+
xpos_dist_WDT	0.09	−	POS_BN	0.19	+
verbs_tense_dist_Pres	0.09	+	DIP_neg	0.19	+

Tabella 3.10: Ranking basato sull'*effect size* del Wilcoxon Rank-sum test (in termini di *rank biserial correlation*) nei discorsi politici e nelle recensioni online.

Come mostrato in Tabella 3.10, la categoria persuasiva nei discorsi politici in CORPS si contraddistingue per uno stile ricco di pronomi, in particolare personali (PRP), aspetto che si riflette anche sul numero inferiore di caratteri per token. Altre due caratteristiche da evidenziare sono il minor numero di aggettivi e preposizioni (IN) nella classe persuasiva e le differenze nell'uso dei tempi verbali. Infatti, gli esempi persuasivi si caratterizzano per un uso più frequente del

[11]La formula per il calcolo della *rank biserial correlation* a partire dalle statistiche risultanti dal Wilcoxon Ranksum Test è descritta in Wendt (1972).

tempo presente piuttosto del passato. Risulta poi interessante notare che tra le caratteristiche con i valori più alti di *rank biserial correlation* emergano anche aspetti relativi alla complessità strutturale (e alla lunghezza delle frasi) come la media delle profondità degli alberi sintattici e il numero di catene preposizionali (a favore della classe non persuasiva). Un esempio di finestra persuasiva, rappresentativo in particolare dell'importanza dei pronomi personali nelle strategie impiegate per richiamare applausi nei discorsi politici, è un estratto di un discorso tenuto da Bill Clinton nel 1996 (con tag *Positive_Focus*):

> We can afford that and we can pay for it in our balanced budget plan. But you have to decide. Will you help us build that bridge? You have to decide.

Esempio 1: Bill Clinton, 24/10/96.

Come si può notare, il pronome personale passa dalla prima persona plurale alla seconda singolare, nel tentativo di chiamare in causa gli ascoltatori, spostando l'attenzione da quello che può fare l'amministrazione riguardo alle tasse al ruolo attivo degli elettori in questo tipo di decisioni. Lo stesso esempio 3.3 è anche caratterizzato da frasi brevi e incisive, semplici per quanto riguarda la struttura sintattica e caratterizzate da una bassa percentuale di nomi e modificatori nominali. Assimilabile a quello appena riportato è poi una finestra estratta da un discorso di Richard Nixon del 1960, anche se in questo caso il focus si sposta sull'audience attraverso un cambiamento nell'uso dei pronomi possessivi:

> But I ask you to think for a moment. Some people will say since he is going to spend more than you will spend, that means that his programs are better. But think for a moment. He isn't going to be spending his money, but your money, and that makes a big difference.

Esempio 2: Richard Nixon, 1/10/60.

Passando invece alle recensioni online, corrispondono a valori più alti di *rank biserial correlation* tratti relativi alla Type Token Ratio (che tende a essere più bassa negli esempi persuasivi) e alla densità lessicale, oltre a caratteristiche ine-

renti alla lunghezza e alla complessità, in particolare per quanto riguarda le strutture dei predicati, che sono più articolate nel caso delle recensioni persuasive. Alla complessità sono collegate anche caratteristiche come la media della lunghezza dei link massimi e la media delle altezze massime degli alberi sintattici, anche se in questo caso i valori maggiori si trovano in corrispondenza delle recensioni persuasive, come discusso nei paragrafi precedenti. Tratti inerenti alla distribuzione delle categorie morfosintattiche coinvolgono invece gli avverbi di negazione (più frequenti nelle recensioni persuasive) e il numero di aggettivi che, come nel caso dei discorsi politici, è inferiore nella categoria persuasiva. Per quanto riguarda il ruolo di tratti come il numero di relazioni di dipendenza per testa verbale, gli esempi 3 e 4 mostrano due recensioni in cui il numero totale di archi entranti nelle teste verbali è vicino ai valori medi della classe "non utile" nel primo caso e "utile" nel secondo.

> Sono stato qui con un amico la sera. Abbiamo preso degli antipasti di crostini misti e polpettine. Poi ho preso una classica ribollita, buona ma non eccezionale. Accompagnata da un mezzo vino toscano della casa di buona qualità. Infine una mousse al cioccolato, buona ma un po' secca accompagnata da un bicchierino di vino dolce. Buona cena, buona la compagnia. Il costo è decisamente elevato rispetto alla qualità.

Esempio 3: Recensione non utile.

> Prenotiamo il giorno prima, ci fanno presente che alle 21:40 avremmo dovuto liberare il tavolo, nessun problema se non che finiamo di consumare il secondo e alle 21:15 abbiamo chiesto il menu dei dolci che arriva solo dopo aver sollecitato. Ordiniamo i dolci e di lì a poco torna il cameriere con il conto in mano che ci chiede mortificato di lasciare il tavolo perché ormai erano le 21:40. Perché prendere ordini anche se il tempo è scaduto? L'atteggiamento del cameriere che ci ha gentilmente cacciati ci ha dato l'impressione di un servizio turistico e sciatto, molto male organizzato. Purtroppo la ristorazione è altra cosa, la qualità dei piatti è tutto sommato buona ma non basta, un filo di organizzazione e ospitalità in più costano fatica ma trasformano una pretenziosa (e poco ispirata) scontrinatrice in un'Osteria.

Esempio 4: Recensione utile.

L'esempio della classe votata è un resoconto dettagliato della cena, in forma narrativa e arricchito di riferimenti temporali specifici (e.g. *il giorno prima*, *alle 21:40*), che si conclude con alcune considerazioni sull'esperienza vissuta. Al contrario, la recensione che non ha ricevuto alcun voto è caratterizzata da uno stile scarno, quasi telegrafico. Talvolta il verbo principale è addirittura omesso, quando l'autore sceglie di usare frasi nominali (e.g. *Buona la cena, buona la compagnia.*). Un caso ancora più estremo in questo senso, sempre tra le recensioni non utili, è rappresentato dall'esempio 5, che è costituito per la maggior parte da frasi nominali:

> Locale in piena chinatown. Lista molto ampia con numerosi piatti inusuali e scelta di pietanza thailandesi e della cucina di Sichuan. Prese le tradizionali lingue d'anatra e un'ottima e morbidissima anatra con porri. Buoni i ravioli al vapore con carne, discreti quelli con verdure. Granchio allo zenzero con carne deliziosa anche se un poco troppo aromatizzata. Buono il maiale in agrodolce ed ottimi gli involtini vietnamiti (fritti con carne) serviti con le tradizionali verdure fresche. Delicate le cosce di rana alla piastra e le verdure cinesi spadellate. Soffice e leggero il pane al vapore. Birra cinese e acqua minerale con abbondanti porzioni alla fine più che onesti. 15€ a testa.

Esempio 5: Recensione non utile.

Capitolo 4

La predizione della persuasività

Questo capitolo è dedicato agli esperimenti che sono stati condotti su CORPS (CORpus of tagged Political Speeches) e su un dataset di recensioni di Tripadvisor per valutare l'impatto delle caratteristiche linguistiche, descritte nel capitolo 3, sull'individuazione automatica di varie forme di persuasività nei discorsi politici e nelle recensioni online. A tal fine, nel caso dei discorsi politici, le caratteristiche (o *feature*) linguistiche sono state comparate con quelle tradizionalmente impiegate per modellare aspetti lessicali e semantici dei testi, all'interno di task di classificazione binaria, in cui paragrafi di testo sono mappati nella classe persuasiva o non persuasiva da un algoritmo di apprendimento supervisionato addestrato su esempi della classe positiva (i paragrafi che precedono un tag di reazione) e negativa (i paragrafi lontani dai tag).

Tutti gli esperimenti hanno riguardato i tag presenti in CORPS (*Positive_Focus* e *Ironical_Focus*), escludendo tuttavia il tag *Negative_Focus* per la scarsa quantità di esempi, oltre che lunghezze variabili delle finestre di testo coinvolte (da 1 a 4 frasi). Le feature estratte sono state testate sul dataset estratto da CORPS (secondo le modalità descritte nel paragrafo 2.3.1), sia in uno scenario

in-domain (dividendo i discorsi politici in modo casuale in training e test set) che cross-domain (dividendo il corpus in base *(i)* agli oratori, *(ii)* alla loro appartenenza politica e *(iii)* al tipo di tag). Gli esperimenti sono stati anche condotti su un dataset diverso rispetto a quello di training per capire se le caratteristiche che permettono di individuare le parti persuasive dei discorsi politici siano rilevanti anche nel dominio dei TED talk. Nel caso delle recensioni online il problema è stato affrontato in modo simile, impostando un compito di classificazione binaria di una recensione come utile o non utile sia in uno scenario in-domain (addestrando il classificatore sulle recensioni romane e testandolo sempre sulla stessa tipologia) che cross-domain (testando i modelli addestrati nello scenario in-domain prima su recensioni di ristoranti milanesi e poi su quelle riguardanti le attrazioni turistiche).

4.1 Il set-up degli esperimenti

4.1.1 Il classificatore

Come algoritmo di classificazione, si è scelto di utilizzare una Support Vector Machine lineare (Vapnik, 1999), implementata tramite la libreria LIBLINEAR (Fan et al., 2008). Dato un dataset di addestramento di m istanze associate alla loro classe di appartenenza, in cui ogni istanza è individuata da n attributi (o *feature*), l'obiettivo di una SVM è trovare un iperpiano a $(n-1)$-dimensioni che separi correttamente le istanze appartenenti alle due classi, massimizzando anche la distanza (o margine) tra i *support vector*. I *support vector* sono gli esempi di entrambe le classi più vicini all'iperpiano e sono anche gli unici a determinarne la posizione. Quando i dati non sono perfettamente separabili, è possibile anche ammettere un certo livello di tolleranza nei confronti degli outlier che si trovano dalla parte sbagliata dell'iperpiano o all'interno dei confini individuati dal margine. Un parametro, solitamente chiamato C nelle librerie di Machine

Learning, controlla il trade-off tra la necessità di massimizzare il margine e al tempo stesso ridurre il numero di esempi classificati erroneamente: per valori alti di C, sarà selezionato anche un margine molto piccolo purché permetta di classificare correttamente più esempi, anche se questo può impattare in modo negativo la capacità di generalizzazione del modello su dati mai visti prima. Una caratteristica delle SVM è anche la capacità di trattare dati non linearmente separabili, attraverso il cosiddetto *kernel trick*. In questo caso, le istanze in training vengono implicitamente proiettate in uno spazio a dimensionalità più alta, in cui sia possibile individuare un confine di decisione lineare. Dato che in questo studio il focus degli esperimenti è sulle feature impiegate piuttosto che sulla performance, si è scelto di non ottimizzare la scelta del kernel e dei parametri, motivo per cui sono stati impiegati i valori di default di LIBLINEAR. Il vantaggio nell'impiegare LIBLINEAR piuttosto che un'altra libreria come LIBSVM (Chang e Lin, 2011) è invece nel minore tempo di training: infatti LIBLINEAR, pur non supportando il *kernel trick*, implementa un algoritmo ottimizzato per le SVM lineari che scala quasi linearmente (piuttosto che esponenzialmente come LIBSVM) con il numero di istanze in training, per quanto riguarda la complessità temporale. Le SVM sono state a lungo utilizzate per la classificazione di testi, perché sono in grado di gestire istanze rappresentate da vettori ad alta dimensionalità ma sparsi (Joachims, 1998), come quelli tradizionalmente impiegati nell'Information Retrieval per la rappresentazione di documenti in base alla frequenza dei termini che vi occorrono, ottenendo una buona performance anche quando il numero di feature è molto maggiore del numero di esempi in training.

4.1.2 Il ridimensionamento delle feature e il problema delle classi sbilanciate

Quando si utilizzano le Support Vector Machines, vi sono alcuni aspetti cui prestare attenzione. Per esempio, le SVM sono sensibili alla scala delle feature,

che può influenzare il modello risultante, perché le feature che cadono in un range numerico più ampio tendono a dominare sulle altre nel calcolo delle distanze necessario per trovare l'iperpiano ottimale, oltre ad aumentare il tempo di addestramento. Per questo motivo, negli esperimenti descritti nei prossimi paragrafi le feature sono state mappate nell'intervallo [0, 1] attraverso il ridimensionamento min-max, avendo cura di utilizzare per il test set gli stessi fattori di normalizzazione del training set. Inoltre, per ridurre il numero di feature utilizzate, si è deciso di filtrarle in base alla loro frequenza negli esempi di training, ponendo come soglia una frequenza minima di 5.

Un altro fattore che può compromettere la performance di una SVM riguarda poi il modo in cui gli esempi in training sono distribuiti nelle classi. Come riportato nel paragrafo 2.3.1, il numero di finestre negative estratte da CORPS è maggiore di quello delle finestre positive, conseguenza naturale del fatto che gli applausi e le risate occorrano raramente nei discorsi (la densità media dei tag in CORPS è infatti di 0.002, v. tab. 2.1). Il fatto che le due classi siano sbilanciate rispetto al numero di esempi in training costituisce una potenziale problematica per la classificazione, così come per la scelta delle metriche di valutazione. Un possibile approccio al problema prevede un intervento diretto sul modo in cui la SVM trova il confine di decisione: per esempio, si possono fornire al modello dei pesi (scelti in base alla conoscenza del dominio o sperimentalmente) che regolano il valore del parametro C in base alla categoria di appartenenza dell'istanza, in modo tale che eventuali violazioni del margine da parte degli esempi della classe minoritaria siano più penalizzati rispetto a quelle da parte di esempi della classe maggioritaria. Altri metodi invece agiscono sui dati di training, applicando tecniche di ricampionamento per ridurre il numero di esempi nella classe maggioritaria oppure aumentare sinteticamente il numero di esempi nella classe minoritaria. In questo caso, visto che la categoria di interesse è quella costituita da meno esempi, si è scelto di bilanciare al 50% tutti i dataset tramite *random undersampling* della classe negativa, ossia selezionando casualmente

gli esempi della classe negativa da eliminare. Anche per le recensioni, si è scelto di lavorare su un dataset bilanciato.

4.1.3 La valutazione del classificatore

Visto che si è scelto di trattare il problema in uno scenario bilanciato, è stata anche preferita l'accuratezza come misura della performance del classificatore. L'accuratezza corrisponde al numero di esempi classificati correttamente (sia negativi che positivi, *True Negative* e *True Positive*) rispetto al totale dei casi considerati (in cui si aggiungono ai *True Negative* e ai *True Positive* anche i *False Positive*, ossia gli esempi negativi erroneamente classificati come positivi e i *False Negative*, che viceversa corrispondono agli esempi positivi classificati come negativi). Quando le classi non sono bilanciate, l'accuratezza può infatti essere fuorviante: per esempio, se si considera un problema di classificazione a due classi in cui la classe A presenta 9990 esempi e la classe B solo 10, un classificatore che predice sempre la classe A avrebbe un'accuratezza del 99.9%, benché non sia in grado di classificare correttamente alcun esempio di B. Per ogni modello, i risultati sono stati poi comparati con una baseline rappresentata dall'accuratezza di un classificatore che predice sempre la classe più frequente, equivalente a 0.50 nello scenario bilanciato.

4.2 Gli esperimenti sui discorsi politici

4.2.1 I modelli

Gli esperimenti di classificazione sono stati condotti utilizzando diversi tipi di caratteristiche estratte dai testi:

- Feature lessicali basate su **n-grammi** (unigrammi e bigrammi, che comprendono:

  - *N-grammi di token*, ottenuti calcolando la frequenza di unigrammi e bigrammi di token, normalizzata in base al numero di token nel testo;
  - *N-grammi di lemmi*, ottenuti calcolando la presenza o assenza di unigrammi e bigrammi di lemmi all'interno del testo;
  - *N-grammi di caratteri*, calcolata come la frequenza di unigrammi e bigrammi di caratteri normalizzata in base al numero di caratteri nel testo. Gli n-grammi di caratteri sono considerati sia indipendentemente dalla loro posizione all'interno di una parola che come sequenze di caratteri che occorrono all'inizio, alla fine o nel mezzo di una parola.

- Feature distribuzionali basate su combinazioni di **word embeddings**, ottenute calcolando separatamente la media dei *word embeddings* di nomi, verbi e aggettivi. I *word embeddings* utilizzati per i discorrsi politici, di dimensione 32, sono stati addestrati tramite il toolkit word2vec (Mikolov et al., 2013) su UkWac (Ferraresi et al., 2008), un corpus di più di 2 miliardi di parole che raccoglie testi in inglese britannico trovati sul Web nel dominio .uk.

  Word2vec fornisce un'implementazione delle architetture CBOW (Continuous Bag-of-Words) e Skip-gram per il calcolo delle rappresentazioni vettoriali distribuite delle parole in un corpus di testi. Queste rappresentazioni sono dense, a bassa dimensionalità e in grado di catturare somiglianze e analogie nel significato delle parole. Nel modello CBOW si definisce un compito di predizione della parola *target* a partire dal suo contesto, ossia le parole che la precedono e la seguono all'interno di una finestra di lunghezza predefinita; al contrario, nell'architettura Skip-gram, speculare a CBOW, l'obiettivo è predire il contesto a partire da una parola. Il razionale risiede nell'ipotesi distribuzionale, secondo la quale il significato di una parola è determinato dal suo contesto, ossia dalle parole con cui co-occorre. Il compito è svolto in entrambi i casi da una rete neurale feed-

forward e gli embeddings vengono estratti dai pesi appresi durante la fase di addestramento.

- le **feature linguistiche** pensate per modellare lo stile di un testo descritte nel capitolo 3, che sono state calcolate a partire dai testi annotati automaticamente fino al livello sintattico.

Le caratteristiche sono state testate in tre diverse configurazioni, per studiare il ruolo svolto dalle feature che modellano il contenuto e di quelle che modellano la forma dei testi sia singolarmente che in combinazione le une con le altre:

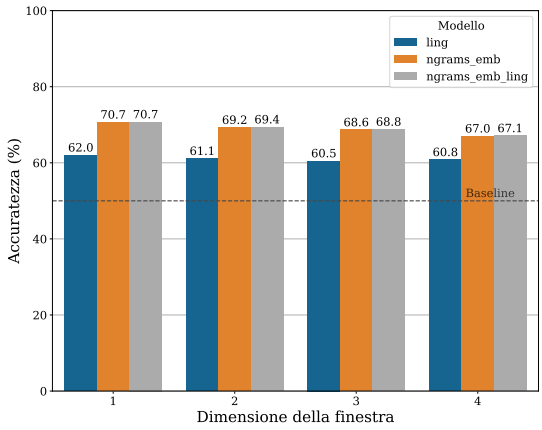
- *ngrams+emb* comprende gli n-grammi e le combinazioni di *word embeddings*;
- *ling* comprende le feature estratte nella fase di monitoraggio;
- *ngrams+emb+ling* include tutte le feature.

4.2.2 Gli esperimenti in-domain

Come prima fase sperimentale, si è scelto di testare il classificatore in uno scenario in-domain, dividendo in modo casuale il corpus in training set e test set in base ai discorsi di appartenenza degli esempi. A tale scopo, sono quindi stati selezionati casualmente 1714 discorsi per l'addestramento e 1713 per la fase di test, ottenendo due dataset di 14791 e 14040 esempi rispettivamente, composti al 50% da esempi della classe persuasiva (finestre di testo precedenti i tag *Positive_Focus* e *Ironical_Focus*) e per la restante metà da esempi della classe non persuasiva (lontani dai tag). Questi ultimi sono stati selezionati tramite ricampionamento casuale dal pool di esempi negativi perché fossero lo stesso numero di quelli positivi. Nella fase di training, il classificatore prende in input il training set in formato CoNLL-U e, per ogni esempio, estrae le caratteristiche selezionate in base al tipo di configurazione scelta, concatenandole poi tra loro per ottenere una rappresentazione del documento testuale originario. Le feature

vengono quindi normalizzate e indicizzate per poi essere utilizzate per la costruzione del modello. Nella fase di test, vengono estratte le feature dai nuovi esempi e la classe è predetta in base al modello.

In Figura 4.1 sono riportati i risultati degli esperimenti, condotti su diverse dimensioni delle finestre. I modelli che includono le feature relative al contenuto hanno ottenuto i punteggi di accuratezza migliori (quello più alto è di 70.07% per i modelli *ngrams+emb* e *ngrams+emb+ling*), come era ragionevole aspettarsi in uno scenario in-domain ma è interessante notare come, anche utilizzando solo le feature linguistiche, l'accuratezza del classificatore resti sopra il 60.05% (circa 10 punti sopra la baseline). In generale, le feature linguistiche non sembrano migliorare l'accuratezza quando aggiunte a quelle lessicali. Osservando l'andamento dei risultati in base alla dimensione della finestra, si può inoltre notare che le performance migliori sono state raggiunte utilizzando finestre lunghe una sola frase per tutti i modelli, anche se la differenza maggiore tra l'accuratezza dei modelli addestrati su finestre lunghe quattro frasi e su finestre lunghe una frase non supera i tre punti percentuali.



.

Figura 4.1: Risultati degli esperimenti di classificazione in-domain condotti sui discorsi politici.

4.2.3 Gli esperimenti cross-domain

In questa fase di esperimenti il corpus è stato suddiviso in modo tale da distinguere gli esempi in training e in test in base agli oratori che hanno tenuto i discorsi, per capire se i modelli siano applicabili anche a discorsi di oratori mai visti durante l'addestramento del sistema. Inizialmente gli oratori sono stati selezionati all'interno dello stesso partito (democratico o repubblicano), ottenendo per i democratici un training set di 7987 esempi e un test set di 5546 (con 18 oratori ciascuno), mentre per i repubblicani un training set di 8954 e un test set di 6016 (con 29 oratori ciascuno). Dai risultati si evidenzia una prima differenza rispetto allo scenario in-domain, ossia un generale calo della performance per quanto riguarda i modelli lessicali, che ha comportato una riduzione della distanza rispetto ai punteggi ottenuti dal modello delle feature linguistiche (v. tab. 4.2). Per quanto riguarda invece le differenze tra gli esperimenti sui discorsi tenuti da politici democratici e repubblicani, si possono osservare dei punteggi più bassi nel caso dei repubblicani, probabilmente dovuti a una maggiore varietà stilistica tra gli oratori nelle strategie retoriche impiegate per suscitare applausi e risate.

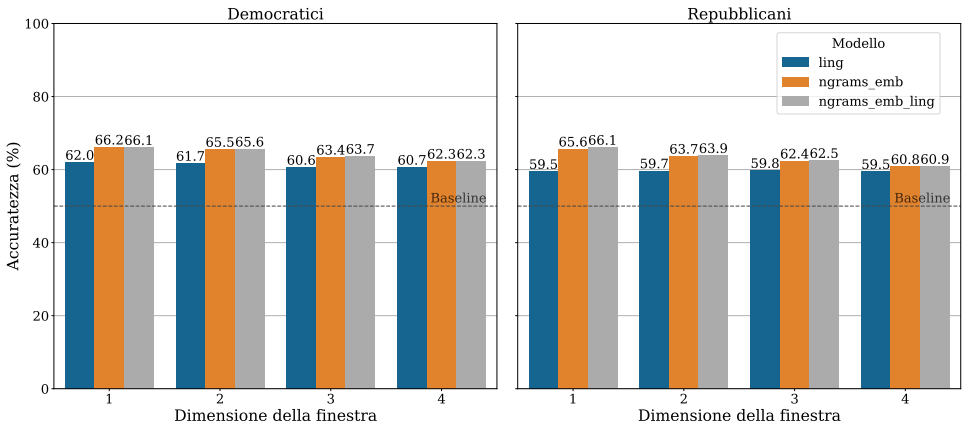


Figura 4.2: Risultati degli esperimenti di classificazione cross-domain condotti sui discorsi politici, distinguendo gli oratori in training e test per ciascuno dei due partiti di appartenenza.

Nel caso degli esperimenti su un partito diverso rispetto a quello in training,

come riportato in Figura 4.3, con tutti i modelli sono stati ottenuti risultati inferiori a quelli in-domain ma superiori alla baseline di almeno 7 punti percentuali. Ancora una volta le configurazioni che includono feature lessicali presentano performance migliori rispetto al modello che comprende solo le feature linguistiche. La differenza tra l'accuratezza ottenuta con le feature linguistiche e con quelle lessicali è più ridotta nel caso dei modelli addestrati sui discorsi dei repubblicani e testati sui discorsi dei democratici, mentre è più elevata nel caso inverso. Come in quasi tutti gli esperimenti precedenti (l'unica eccezione è rappresentata dal modello linguistico testato sui discorsi dei repubblicani), anche in questo caso sono stati ottenuti risultati migliori per finestre lunghe una frase, probabilmente perché è quella che corrisponde alla punchline finale per quanto riguarda le risate o che effettivamente suscita l'applauso.

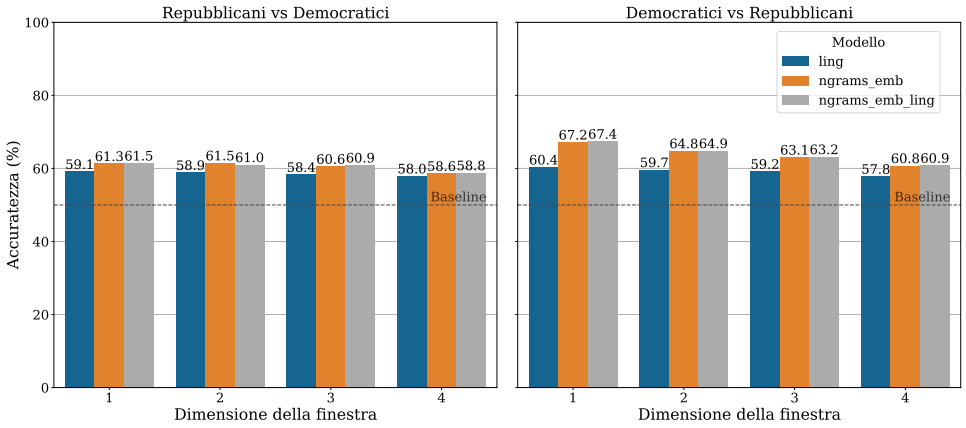


Figura 4.3: Risultati degli esperimenti di classificazione cross-domain condotti sui discorsi politici, distinguendo gli oratori in training e test in base al partito di appartenenza.

Si è anche deciso di provare ad addestrare i modelli su una combinazione di tag relativi alle reazioni del pubblico diversa rispetto a quella usata per la fase di test. I dataset utilizzati sono stati ottenuti a partire da quelli impiegati per gli esperimenti in-domain, riducendo gli esempi positivi a quelli associati solo ai tag *Positive_Focus* e poi solo ai tag *Ironical_Focus* (v. tab. 4.1).

Come mostrato in Figura 4.4, i risultati degli esperimenti nello scenario Positive vs Ironical e Ironical vs Positive sono in linea con la baseline, a evidenza del fatto che i fenomeni dell'applauso e della risata siano fondamentalmente diversi tra loro. Per questo motivo, potrebbe essere interessante condurre esperimenti interni allo stesso tag, anche se i dati di training sono ridotti, soprattutto per quanto riguarda il tag *Ironical_Focus*. Il fatto che siano stati raggiunti risultati migliori sui tag *Positive_Focus* piuttosto che sui tag *Ironical_Focus* con i modelli addestrati su entrambi è invece spiegabile per la composizione del training set, che presentava più finestre associate agli applausi piuttosto che alle risate. Per lo stesso motivo, la situazione è inversa nei risultati ottenuti addestrando il classificatore solo sui tag *Positive_Focus* o solo sui tag *Ironical_Focus* e poi testandolo su entrambi.

	Training set	Test set
#discorsi	1714	1713
	#esempi	
Positive_Ironical	14791	14040
Positive	11867	11166
Ironical	2980	2920

Tabella 4.1: Descrizione dei dataset ottenuti distinguendo i discorsi di origine delle finestre di testo ma non gli oratori.

Per quanto riguarda gli esperimenti cross-dataset, in Figura 4.5 sono riportati i risultati ottenuti testando i modelli (addestrati sul *training set* dello scenario in-domain) sulla porzione inglese del corpus multilingue WIT^3. Come anticipato nel capitolo 2, WIT^3, nella versione aggiornata al 2016, raccoglie le trascrizioni di 2085 TED talk tenuti da 1767 oratori, di cui 2044 contengono almeno un tag assimilabile alle reazioni del pubblico evidenziate in CORPS: nello specifico i tag *(Laughter)*, *(Applause)* e *(Cheering)*, con le loro varianti. In questo caso, si è scelto di focalizzarsi su un campione di 616 TED talk, che corrispondano a monologhi tenuti da un singolo speaker e privi di interventi da parte del

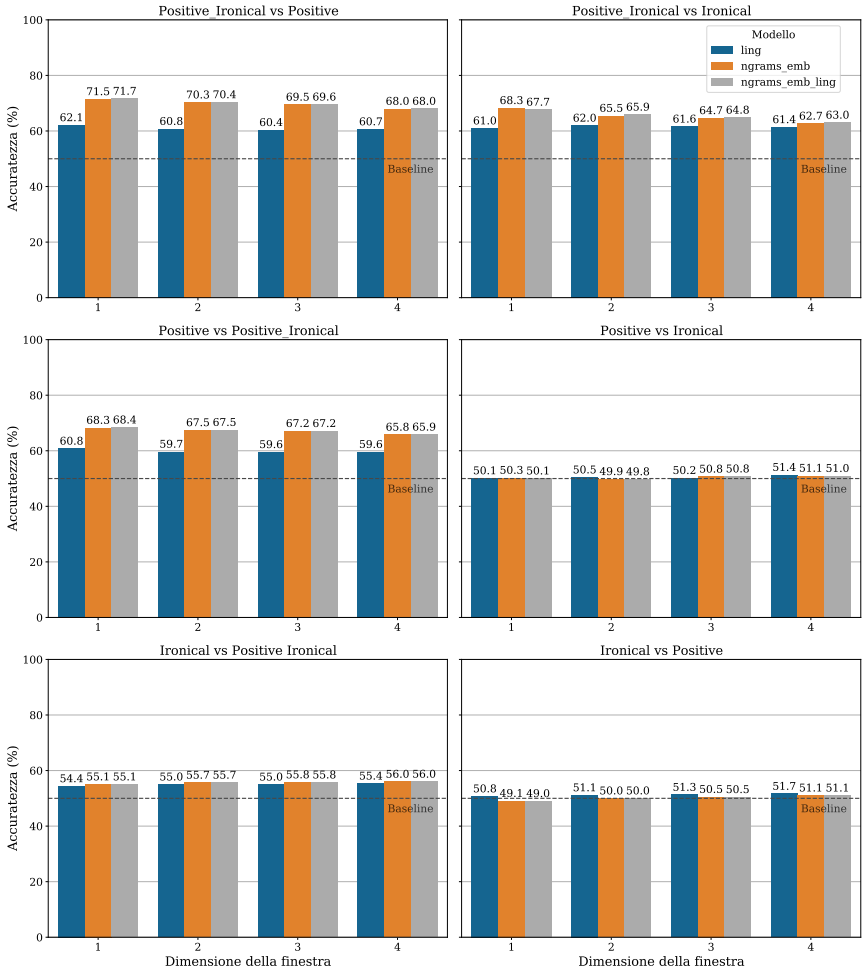


Figura 4.4: Risultati degli esperimenti di classificazione cross-domain condotti sui discorsi politici, distinguendo i tag in training e in test in base al partito di appartenenza.

pubblico (sul modello dei discorsi politici), da cui sono state estratte 5352 finestre, seguendo le stesse modalità illustrate nel capitolo 2 per i discorsi politici in CORPS. Nonostante il cambio di dataset, tutti i modelli hanno ottenuto accuratezze superiori al 56.3% per finestre lunghe una frase. Le performance più alte sono sempre quelle dei modelli *ngrams+emb* e *ngrams+emb+ling*, senza che ancora si evidenzi una differenza netta tra i due.

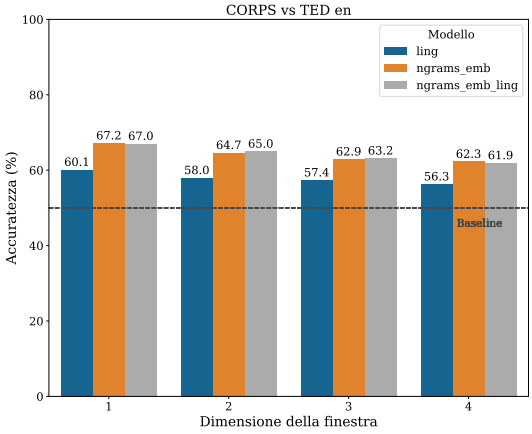


Figura 4.5: Risultati degli esperimenti di classificazione cross-dataset ottenuti addestrando il classificatore su CORPS e testandolo sui discorsi di WIT[3].

4.2.4 Discussione

Una conclusione generale che è possibile trarre dagli esperimenti sui discorsi politici riguarda il fatto che anche solo grazie all'informazione contenuta in un'unica frase è possibile predire una reazione da parte del pubblico (e con maggiore accuratezza rispetto a quando si considerano contesti più lunghi), sia in uno scenario in-domain che in uno cross-domain. Un risultato simile è stato riscontrato anche da Z. Liu et al. (2017) nel dominio dei TED talk. Come già evidenziato in altri lavori (Strapparava et al., 2010; H. Liu et al., 2017; Gillick e Bamman, 2018)), le caratteristiche lessicali di un testo svolgono un ruolo di primo piano nel suscitare una reazione da parte del pubblico. Questo tipo di informazione continua a essere rilevante anche spostandosi su domini più distanti da quello di addestramento. In effetti, in un ranking dei pesi del modello lessicale (per finestre di una sola frase) addestrato sull'intero CORPS, tra le prime trenta feature trova posto informazione relativa alla distribuzione della punteggiatura e dei pronomi (in particolare *I*, *you*, *me*), oltre alla distribuzione di n-grammi che includono termini piuttosto generici e non legati a un dominio specifico, quali *country*, *God*, *congratulations*. L'informazione linguistica è sufficiente a classificare una frase come persuasiva ma non consente di ottenere risultati pari a quelli del modello lessicale. Nel ranking delle feature linguistiche si possono

osservare dei punti di contatto con le feature lessicali. Infatti, trovano posto tra le prime venti posizioni le caratteristiche che riguardano la distribuzione delle categorie morfo-sintattiche relative a nomi (in particolare i nomi propri), pronomi (personali e interrogativi) e aggettivi, ma anche la punteggiatura (v. 4.2).

Caratteristica	Segno
1. char_per_tok	−
2. xpos_dist_PRP	+
3. xpos_dist_NNS	−
4. xpos_dist_"	+
5. upos_dist_PROPN	+
6. xpos_dist_WDT	−
7. xpos_dist_NNP	+
8. xpos_dist_WRB	−
9. upos_dist_NOUN	−

Tabella 4.2: Prime caratteristiche nel ranking dei pesi assegnati alle feature del modello linguistico negli esperimenti di classificazione sui discorsi politici nello scenario in-domain.

4.3 Gli esperimenti sulle recensioni online

4.3.1 I modelli

Nel caso delle recensioni online, sono state testate le stesse feature descritte in 4.2.1, con alcuni accorgimenti: le combinazioni di *word embeddings* sono state addestrate su ItWac (Baroni et al., 2009), un corpus costruito tramite *web crawling* a partire da pagine web raccolte nel dominio .it all'interno dello stesso progetto che ha portato alla creazione di UkWac, oltre che su una collezione di tweet italiani.[1] Come feature sono state aggiunte alle medie dei *word embeddings* corrispondenti a nomi, verbi e aggettivi per entrambi i set di embeddings pre-addestrati anche le medie delle distanze tra le due collezioni. Le feature

[1] `http://www.italianlp.it/resources/italian-word-embeddings.` (ultimo accesso: 07/04/2022)

linguistiche sono state invece estratte da testi annotati fino al livello sintattico secondo lo schema ISST-TANL (v. capitolo 3). Un'altra feature presa in considerazione e non presente per i discorsi politici, in quanto specifica del dominio delle recensioni online, è il rating, ossia il punteggio attribuito su una scala da 1 a 5 al luogo recensito da parte dell'autore della recensione. Anche per questo set di esperimenti, le feature sono state testate in diverse configurazioni. Nel caso dello scenario in-domain, si è scelto di impiegare configurazioni a livello più granulare, per testare l'impatto delle feature sui compiti di classificazione anche in isolamento:

- *ngrams* per gli n-grammi;
- *emb*, per le combinazioni di *word embeddings*;
- *ling*, per le feature linguistiche;
- *str*, per il rating della recensione.

Per quanto riguarda lo scenario cross-domain, invece si è scelto di testare solo le tre configurazioni principali, riproposte anche negli esperimenti sui discorsi politici:

- *ngrams+emb*, per le feature basate su n-grammi e *word embeddings*;
- *ling*, per le feature linguistiche;
- la loro combinazione *ngrams+emb+ling*.

In aggiunta, è stata anche testata la loro concatenazione con il rating (identificato sempre come *str* nei paragrafi successivi).

4.3.2 Gli esperimenti in-domain

Nello scenario sperimentale in-domain è stato ottenuto un generale miglioramento rispetto alla baseline con tutte le configurazioni di feature a parte quella

che utilizza solo il rating degli utenti (v. tab. 4.3). Nonostante questo, il rating migliora i punteggi di accuratezza di tutti i modelli di almeno un punto percentuale. I risultati evidenziano anche il ruolo svolto dall'informazione lessicale nella valutazione automatica dell'utilità di una recensione, benché questo sia in parte giustificato dallo scenario in-domain. In particolare, il modello che impiega sia le feature lessicali che il rating ha permesso di raggiungere il punteggio maggiore (71.14%), anche se con una differenza minima rispetto al modello delle sole feature lessicali (71.13%). Per quanto riguarda le caratteristiche linguistiche, l'accuratezza raggiunta dal classificatore è più bassa rispetto a quella ottenuta con le feature lessicali ma i risultati sono in linea con il modello *ngrams+emb*), con il modello in cui è stato aggiunto anche il rating della recensione, che ha permesso di ottenere un'accuratezza inferiore solo di un punto percentuale rispetto a quella del modello migliore.

Modello	Accuratezza (%)
str	49.6
ling	66
ling+str	70.81
ngrams	69.9
ngrams+str	71.13
emb	68.54
ngrams+emb	70.17
ngrams+emb+str	**71.14**
ngrams+emb+ling	70.04
ngrams+emb+ling+str	71.05
Baseline	50.46

Tabella 4.3: Risultati degli esperimenti di classificazione in-domain sulle recensioni online usando diversi modelli di feature.

4.3.3 Gli esperimenti cross-domain

Nello scenario cross-domain i modelli addestrati sulle recensioni romane sono stati testati prima sulle recensioni di ristoranti milanesi (cambiando solo l'area geografica del luogo recensito rispetto al dominio di training) e in seguito su

una categoria diversa, individuata dalle recensioni di attrazioni turistiche. Come riportato in Tabella 4.4, i risultati ottenuti sulle recensioni di Milano sono molto simili a quelli dello scenario in-domain, anche se leggermente più bassi. Questo risultato suggerisce il fatto che il classificatore sia in grado di determinare l'utilità di recensioni relative a località diverse sfruttando sia l'informazione lessicale (probabilmente legata solo marginalmente all'area geografica) che quella strutturale ma sarebbe necessario verificare questa ipotesi con altri esperimenti, che coinvolgano un più ampio numero di aree geografiche. Oltre a questo primo risultato, è interessante notare il fatto che la performance del modello addestrato sulle feature linguistiche e sul rating sia in linea con l'accuratezza del modello lessicale unito al rating. Entrambi i modelli hanno infatti permesso di raggiungere il risultato migliore, che corrisponde a un punteggio percentuale del 70.92%.

Modello	Recensioni di Milano	Attrazioni
ngrams+emb	69.38%	59.67%
ngrams+emb+str	**70.92%**	58.02%
ling	65.82%	**60.76%**
ling+str	**70.92%**	60.28%
ngrams+emb+ling	69.2%	59.9%
ngrams+emb+ling+str	70.78%	58.49%
Baseline	50.47%	51.56%

Tabella 4.4: Risultati in termini di accuratezza degli esperimenti di classificazione cross-domain sulle recensioni online usando diversi modelli di feature.

Un altro aspetto da sottolineare riguarda poi il fatto che il modello addestrato solo sulle feature linguistiche abbia un'accuratezza molto simile a quello dello scenario in-domain (65.82%), al contrario del modello lessicale, in cui si è verificato un calo del punteggio di 10.5 punti percentuali. L'accuratezza è diminuita invece in modo più consistente per tutti i modelli negli esperimenti sul dominio più distante, benché i risultati restino sopra la baseline. In questo caso è interessante notare che il modello migliore sia risultato quello addestrato solo sulle

feature linguistiche, che ha permesso di ottenere un'accuratezza pari al 60.76% (con una differenza di 5.24 punti percentuali rispetto ai risultati in-domain). Per i modelli lessicali la riduzione nel livello di accuratezza si è rivelata invece più consistente (sempre di almeno 10 punti percentuali). Per quanto riguarda il rating, nel caso della categoria più distante l'aggiunta dell'informazione su questo tipo di metadato ha provocato in generale un leggero calo nella performance del classificatore, probabilmente dovuto al fatto che i rating sono distribuiti in modo diverso rispetto a quelli delle recensioni dei ristoranti romani (v. par. 2.3.3).

4.3.4 Discussione

Caratteristica	Segno
1. Totale catene preposizionali	+
2. Totale Archi Entranti in Teste Verbali	+
3. Totale Teste Verbali	+
4. Numero di Token	+
5. POS_FB	+
6. Range 200.0 Type(lemmi)/token	−
7. Range 200.0 Type(forme)/token	−
8. POS_BN	+
9. AD	+
10. Totale Radici Verbali	+

Tabella 4.5: Prime caratteristiche nel ranking dei pesi assegnati alle feature del modello linguistico negli esperimenti sulle recensioni online.

Dagli esperimenti sulle recensioni è emerso che le caratteristiche linguistiche considerate sono in grado di catturare aspetti dello stile di recensioni persuasive, anche in modo trasversale rispetto al dominio di riferimento. Per capire quali di queste caratteristiche sono le più efficaci nel compito di classificazione automatica, si è deciso di ordinarle in base al valore assoluto del loro peso nel modello lineare. Tra le 50 migliori, oltre alle caratteristiche estratte dal testo grezzo (il cui ruolo nel predire l'utilità è già stato dimostrato in letteratura), sono presenti anche caratteristiche morfo-sintattiche e sintattiche. È il caso, ad esempio, delle

feature che riguardano i modificatori nominali, in particolare il numero di catene preposizionali (che occupano il 1° posto nella classifica) ma anche la distribuzione degli aggettivi e dei determinanti. Altre riguardano le strutture verbali, per esempio il numero di dipendenti delle teste verbali e la frequenza degli avverbi (specialmente quelli di negazione). Caratteristiche relative all'uso della subordinazione, come il numero di strutture subordinate e la profondità media degli alberi sintattici, occupano sempre le prime posizione. L'importanza delle caratteristiche linguistiche è ulteriormente confermata da una seconda verifica in cui lo stesso metodo di classificazione è stato applicato al modello che comprende tutte le feature, scoprendo che il 59,6% dell'intero set di 212 caratteristiche linguistiche cade nel 90° percentile dei pesi.

Capitolo 5

Conclusioni

In questo lavoro si è scelto di analizzare le caratteristiche inerenti a un uso del linguaggio persuasivo in due domini di interesse, quello dei discorsi politici e delle recensioni online, in cui gli esempi di testi persuasivi sono stati identificati a partire da una reazione da parte del pubblico di ascoltatori nel primo caso di studio (in forma di applausi e risate) e degli altri utenti del portale di recensioni nel secondo caso di studio. A questo scopo, sono stati impiegati due corpora annotati con le rispettive reazioni. Uno di questi è CORPS (CORpus of tagged Political Speeches), che comprende trascrizioni raccolte dal web di discorsi politici tenuti in inglese da politici americani in un periodo di tempo che si estende dal 1917 al 2010, in cui sono riportati anche i tag relativi a applausi, risate e fischi del pubblico nei punti del discorso in cui se ne è verificata un'occorrenza. L'altro è invece un dataset di recensioni scaricate dalla sezione italiana di Tripadvisor, associate a metadati quali il rating assegnato dall'autore al business recensito e il numero di *voti utili* o voti di utilità ricevuti dagli altri utenti. Entrambe le reazioni possono essere considerate come indizi di tentativi di persuasione andati a buon fine: in un caso rappresentano una manifestazione di consenso nei confronti del politico; nell'altro mostrano di aver preso in considerazione la recensione votata nel proprio processo decisionale finalizzato a scegliere se frequentare un luogo turistico o un ristorante il cui profilo sia

presente su Tripadvisor. La metodologia seguita ha previsto l'estrazione di un ampio numero di tratti volti a modellare lo stile di scrittura dei testi, che sono stati utilizzati in un primo momento per la ricostruzione del profilo linguistico delle categorie persuasive e non persuasive in entrambi i domini, così da valutare e quantificare le differenze nella loro distribuzione. In seguito le stesse caratteristiche sono state testate in un *framework* di apprendimento automatico per valutare il loro impatto sulla performance del sistema in un compito di classificazione binaria volto a discriminare esempi persuasivi e non persuasivi. Gli esperimenti sono stati condotti per entrambe le tipologie testuali sia in uno scenario in-domain che in uno cross-domain, sperimentando diverse configurazioni di caratteristiche che comprendono i tratti monitorati inerenti alla struttura linguistica e altri predittori che fanno riferimento a informazione lessicale.

Per i discorsi politici, è stato osservato un ruolo di primo piano per quanto riguarda l'informazione lessicale nella predizione automatica delle reazioni del pubblico, benché questa sia ridotta nello scenario cross-domain, come è stato notato in particolare negli esperimenti in cui i modelli sono stati testati su discorsi di oratori appartenenti a un partito diverso rispetto a quello di addestramento, in cui l'accuratezza del classificatore è in linea con quella ottenuta usando soltanto le caratteristiche linguistiche. Riguardo a queste ultime, si può notare che sono in grado di catturare informazione utili per la classificazione, come mostrato dai risultati sempre superiori alla baseline, benché non diano alcun apporto significativo quando sono utilizzate in combinazione con le caratteristiche che modellano il contenuto del testo. Tuttavia, sia gli esperimenti che l'analisi statistica del profilo linguistico, hanno evidenziato il fatto che risate e applausi corrispondano a due fenomeni essenzialmente diversi, ottenuti anche strutturando il discorso in modo diverso. Si può dunque presumere che l'analisi del fenomeno dell'applauso e della risata separatamente fornisca risultati migliori.

Nel dominio delle recensioni online, sono stati riscontrati risultati compara-

bili a quelli dei discorsi politici nello scenario in-domain (con le caratteristiche lessicali che danno forma al modello migliore) ma il discorso è differente nello scenario cross-domain. Infatti, gli esperimenti con i modelli addestrati sulle recensioni di ristoranti romani e testate sulle recensioni di un'altra città hanno ottenuto le accuratezze migliori per le caratteristiche lessicali in combinazione con il rating e di quelle linguistiche insieme al rating, oltre a mostrare un calo inferiore nell'accuratezza del modello linguistico rispetto a quello lessicale. In particolare, è interessante sottolineare il fatto che il modello che impiega solo le caratteristiche linguistiche abbia ottenuto l'accuratezza migliore sul dominio più distante da quello di addestramento (la categoria di recensioni relative alle attrazioni turistiche), sottolineando l'importanza dell'informazione strutturale per la modellazione del fenomeno della persuasività nel dominio delle recensioni.

Per quanto riguarda i tratti linguistici oggetto di variazione nei due domini tra esempi di testi persuasivi e non persuasivi, si può prima di tutto affermare che a seconda del dominio si evidenzia un tipo diverso di persuasività. Benché ci siano alcuni punti di somiglianza, come il maggior numero di verbi e pronomi nel caso degli esempi persuasivi, e il minor numero di nomi e aggettivi (che fanno riferimento a uno stile personale e focalizzato sul lettore), i due domini differiscono invece per aspetti collegati alla lunghezza dei testi e alla loro complessità strutturale. Se nel caso delle recensioni la lunghezza e la densità informativa delle recensioni risulta più rilevante, nel caso dei discorsi politici le parti a cui il pubblico risponde con un applauso o una risata tendono a essere più brevi e più semplici dal punto di vista della struttura sintattica.

Possibili sviluppi futuri di questo lavoro potrebbero riguardare variazioni dei testi persuasivi associate a caratteristiche dei destinatari o degli autori dei testi, per esempio per valutare come lo stile adottato allo scopo di persuadere cambi da un oratore all'altro oppure in che modo le differenze stilistiche influiscano sulla percezione della persuasività di una recensione da parte di utenti diversi.

Bibliografia

Almutairi, Yasamyian, Manal Abdullah e Dimah Alahmadi (2019). «Review Helpfulness Prediction: Survey». In: *Periodicals of Engineering and Natural Sciences* 7.1, pp. 420–432.

Althoff, Tim, Cristian Danescu-Niculescu-Mizil e Dan Jurafsky (2014). «How to Ask for a Favor: A Case Study on the Success of Altruistic Requests». In: *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.

Argamon, Shlomo Engelson (2019). «Register in Computational Language Research». In: *Register Studies* 1 (1), pp. 100–135.

Atkinson, Max (1984). *Our masters' voices: The language and body language of politics*. Psychology Press.

Attardi, Giuseppe, Felice Dell'Orletta, Maria Simi e Joseph Turian (2009). «Accurate Dependency Parsing with a Stacked Multilayer Perceptron». In: *Proceedings of EVALITA 2009 - Evaluation of NLP and Speech Tools for Italian*. Reggio Emilia, Italy.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi e Eros Zanchetta (2009). «The WaCky wide web: a collection of very large linguistically processed web-crawled corpora». In: *Language Resources and Evaluation* 43.3, pp. 209–226.

Biber, Douglas (1988). *Variation Across Speech and Writing*. Cambridge University Press.

Bilal, Muhammad, Mohsen Marjani, Ibrahim Abaker Targio Hashem, Akibu Mahmoud Abdullahi, Muhammad Tayyab e Abdullah Gani (14 dic. 2019). «Predicting Helpfulness of Crowd-Sourced Reviews: A Survey». In: 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS). Karachi, Pakistan.

Blitzer, John, Mark Dredze e Fernando Pereira (2007). «Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification.» In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 440–447.

Brunato, Dominique, Lorenzo De Mattei, Felice Dell'Orletta, Benedetta Iavarone e Giulia Venturi (2018). «Is this Sentence Difficult? Do you Agree?» In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 2690–2699.

Bull, Peter (2006). «Invited and uninvited applause in political speeches». In: *British Journal of Social Psychology* 45.3, pp. 563–578.

Cettolo, Mauro, Christian Girardi e Marcello Federico (mag. 2012). «WIT3: Web Inventory of Transcribed and Translated Talks». In: *Proceedings of the Conference of European Association for Machine Translation (EAMT)*. Trento, Italy, pp. 261–268.

Chaiken, Shelly (1979). «Communicator physical attractiveness and persuasion.» In: *Journal of Personality and social Psychology* 37.8, p. 1387.

– (1980). «Heuristic Versus Systematic Information Processing and the Use of Source Versus Message Cues in Persuasion.» In: *Journal of Personality and Social Psychology* 39.5, pp. 752–766.

Chaiken, Shelly e Alice H Eagly (1976). «Communication modality as a determinant of message persuasiveness and message comprehensibility.» In: *Journal of personality and social psychology* 34.4, p. 605.

Chang, Chih-Chung e Chih-Jen Lin (mag. 2011). «LIBSVM: A Library for Support Vector Machines». In: *ACM Trans. Intell. Syst. Technol.* 2.3.

Chiriatti, Giulia, Dominique Brunato, Felice Dell'Orletta e Giulia Venturi (nov. 2019). «What Makes a Review Helpful? Predicting the Helpfulness of Italian TripAdvisor Reviews». In: *Proceedings of the 6th Italian Conference on Computational Linguistics (CLiC-it)*. Bari, Italy.

Cimino, Andrea, Martijn Wieling, Felice Dell'Orletta, Simonetta Montemagni e Giulia Venturi (2017). «Identifying predictive features for textual genre classification: the key role of syntax». In: *Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it*, pp. 107–112.

Daelemans, Walter (2013). «Explanation in Computational Stylometry». In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pp. 451–462.

Danescu-Niculescu-Mizil, Cristian, Gueorgi Kossinets, Jon Kleinberg e Lillian Lee (2009). «How Opinions Are Received by Online Communities: A Case Study on Amazon.com Helpfulness Votes». In: *Proceedings of the 18th International Conference on World Wide Web (WWW '09)*. New York, NY, USA: Association for Computing Machinery, pp. 141–150.

De Mauro, Tullio (2000). *Grande dizionario italiano delluso*. Utet.

Dell'Orletta, Felice (2009). «Ensemble System for Part-of-Speech Tagging». In: *Proceedings of EVALITA 2009 - Evaluation of NLP and Speech Tools for Italian*. Reggio Emilia, Italy.

Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang e Chih-Jen Lin (2008). «LIBLINEAR: A library for large linear classification». In: *Journal of machine learning research* 9, pp. 1871–1874.

Ferraresi, Adriano, Eros Zanchetta, Marco Baroni e Silvia Bernardini (2008). «Introducing and evaluating ukWaC, a very large web-derived corpus of English». In: *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pp. 47–54.

Ghose, Anindya e Panagiotis G. Ipeirotis (2011). «Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics». In: *IEEE Transactions on Knowledge and Data Engineering* 23.10, pp. 1498–1512.

Gillick, Jon e David Bamman (1 giu. 2018). «Please Clap: Modeling Applause in Campaign Speeches». In: *Proceedings of NAACL-HLT 2018*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 92–102.

Guerini, Marco, Danilo Giampiccolo, Giovanni Moretti, Rachele Sprugnoli e Carlo Strapparava (2013). «The New Release of CORPS: A Corpus of Political Speeches Annotated with Audience Reactions». In: *Multimodal Communication in Political Speech. Shaping Minds and Social Action*. Springer Berlin Heidelberg, pp. 86–98.

Guerini, Marco e Gözde Özbal (giu. 2015). «Echoes of Persuasion: The Effect of Euphony in Persuasive Communication». In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics.

Guerini, Marco, Carlo Strapparava e Oliviero Stock (2008). «CORPS: A Corpus of Tagged Political Speeches for Persuasive Communication Processing». In: *Journal of Information Technology & Politics* 5.1, pp. 19–32.

Habernal, Ivan e Iryna Gurevych (ago. 2016). «Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM». In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 1589–1599.

Halmari, Helena e Tuija Virtanen (2005). *Persuasion across genres: A linguistic approach*. John Benjamins Publishing.

Hennig-Thurau, Thorsten, Kevin P. Gwinner, Gianfranco Walsh e Dwayne D. Gremler (2004). «Electronic Word-of-Mouth via Consumer-Opinion Plat-

forms: What Motivates Consumers to Articulate Themselves on the Internet?» In: *Journal of Interactive Marketing* 18.1, pp. 38–52.

Heritage, John e David Greatbatch (1986). «Generating Applause: A Study of Rhetoric and Response at Party Political Conferences». In: *American Journal of Sociology* 92.1, pp. 110–157.

Hong, Wei, Zemin Yu, Linhai Wu e Xujin Pu (2020). «Influencing Factors of the Persuasiveness of Online Reviews Considering Persuasion Methods». In: *Electronic Commerce Research and Applications* 39.

Hong, Yu, Jun Lu, Jianmin Yao, Qiaoming Zhu e Guodong Zhou (2012). «What Reviews Are Satisfactory: Novel Features for Automatic Helpfulness Voting». In: *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '12. New York, NY, USA: Association for Computing Machinery, pp. 495–504.

Joachims, Thorsten (1998). «Text Categorization with Support Vector Machines: Learning with Many Relevant Features». In: *European Conference on Machine Learning*. Springer, pp. 137–142.

Kim, Soo-Min, Patrick Pantel, Tim Chklovski e Marco Pennacchiotti (lug. 2006). «Automatically Assessing Review Helpfulness». In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*. Conference on Empirical Methods in Natural Language Processing (EMNLP 2006). Sydney, pp. 423–430.

King, Robert Allen, Pradeep Racherla e Victoria D. Bush (2014). «What We Know and Don't Know About Online Word-of-Mouth: A Review and Synthesis of the Literature». In: *Journal of Interactive Marketing* 28.3, pp. 167–183.

Li, Jin e Lingjing Zhan (2011). «Online Persuasion: How the Written Word Drives WOM. Evidence from Consumer-Generated Product Reviews». In: *the Journal of Advertising Research* 51.1, pp. 239–257.

Liu, Bing e Lei Zhang (2012). «A Survey of Opinion Mining and Sentiment Analysis». In: *Mining Text Data*. Springer, pp. 412–463.

Liu, Haijing, Gao Yang, Lv Pin, Li Mengxue, Geng Shiqiang, Li Minglan e Wang Hao (set. 2017). «Using Argument-based Features to Predict and Analyse Review Helpfulness». In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 1358–1363.

Liu, Zhe, Anbang Xu, Mengdi Zhang, Jalal Mahmud e Vibha Sinha (2017). «Fostering user engagement: Rhetorical devices for applause generation learnt from ted talks». In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 11. 1.

McAuley, Julian, Christopher Targett, Qinfeng Shi e Anton Van den Hengel (2015). «Image-based recommendations on styles and substitutes». In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information*, pp. 43–52.

Ménard, Pierre André e Caroline Barrière (giu. 2016). «Classification of comment helpfulness to improve knowledge sharing among medical practitioners.» In: *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. San Diego, California: Association for Computational Linguistics, pp. 72–81.

Mertz, Matthias, Nikolaos Korfiatis e Roberto V. Zicari (2014). «Using Dependency Bigrams and Discourse Connectives for Predicting the Helpfulness of Online Reviews». In: *International Conference on Electronic Commerce and Web Technologies*. Springer, pp. 146–152.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado e Jeff Dean (2013). «Distributed Representations of Words and Phrases and their Compositionality». In: *Advances in Neural Information Processing Systems 26*. A cura di

C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani e K. Q. Weinberger. Curran Associates, Inc., pp. 3111–3119.

Miller, Gerald R (2013). «On being persuaded: Some basic distinctions.» In: *The SAGE Handbook of Persuasion: Developments in Theory and Practice*.

Montemagni, Simonetta (2013). «Tecnologie linguistico-computazionali e monitoraggio della lingua italiana». In: *Studi Italiani di Linguistica Teorica e Applicata (SILTA)* 42.1, pp. 145–172.

Nguyen, Dong, A. Seza Doğruöz, Carolyn P. Rosé e Franciska de Jong (set. 2016). «Computational Sociolinguistics: A Survey». In: *Computational Linguistics* 42.3, pp. 537–593.

Nivre, Joakim (2015). «Towards a Universal Grammar for Natural Language Processing». In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pp. 3–16.

Ocampo Diaz, Gerardo e Vincent Ng (lug. 2018). «Modeling and Prediction of Online Product Review Helpfulness: A Survey». In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2018. Melbourne, Australia: Association for Computational Linguistics, pp. 698–708.

Park, Do-Hyung, Jumin Lee e Ingoo Han (2007). «The Effect of On-Line Consumer Reviews on Consumer Purchasing Intention: The Moderating Role of Involvement». In: *International Journal of Electronic Commerce* 11.4, pp. 125–148.

Partington, Alan e Charlotte Taylor (2017). *The Language of Persuasion in Politics: An Introduction*. 1ª ed. London: Routledge.

Pentina, Iryna, Ainsworth Anthony Bailey e Lixuan Zhang (17 feb. 2018). «Exploring effects of source similarity, message valence, and receiver regulatory focus on yelp review persuasiveness and purchase intentions». In: *Journal of Marketing Communications* 24.2. Publisher: Routledge, pp. 125–145. ISSN: 1352-7266.

Perelman, C. e L. Olbrechts-Tyteca (1973). *The New Rhetoric: A Treatise on Argumentation*. University of Notre Dame Press. ISBN: 9780268175092.

Petty, Richard E. e John T. Cacioppo (1986). «The Elaboration Likelihood Model of Persuasion». In: *Communication and Persuasion*. Springer Series in Social Psychology. New York, NY: Springer, pp. 1–24.

Platone (2003). *Gorgia*. Trad. da Francesco Adorno. Vol. Platone, Opere complete Vol. V (Eutidemo, Protagora, Gorgia, Menone, et altri). Laterza.

Segrin, Chris (1993). «The effects of nonverbal behavior on outcomes of compliance gaining attempts». In: *Communication Studies* 44.3-4, pp. 169–187.

Stanton, William J. (1984). *Fundamentals of Marketing*. Marketing Series. Fundamentals of Marketing.

Statista (2019). *Number of online product reviews expected by U.S. digital shoppers 2019, by age.* URL: https://www.statista.com/ (visitato il 07/04/2022).

Straka, Milan e Jana Straková (ago. 2017). «Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe». In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada: Association for Computational Linguistics, pp. 88–99.

Strapparava, Carlo, Marco Guerini e Oliviero Stock (mag. 2010). «Predicting Persuasiveness in Political Discourses». In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA), pp. 1342–1345.

Tan, Chenhao, Vlad Niculae, Cristian Danescu-Niculescu-Mizil e Lillian Lee (apr. 2016). «Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions». In: *Proceedings of the 25th international conference on world wide web*, pp. 613–624.

Tsur, Oren e Ari Rappaport (2009). «Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews». In: *Proceedings of the Inter-*

national AAAI Conference on Web and Social Media. Vol. 3. 1, pp. 154–161.

van Halteren, Hans (2004). «Linguistic Profiling for Authorship Recognition and Verification». In: *Proceedings ACL 2004*. East Stroudsburg: Association for Computational Linguistics, pp. 199–206.

Vapnik, Vladimir (1999). *The Nature of Statistical Learning Theory*. Information Science and Statistics. Springer, New York.

Wang, Lu, Nick Beauchamp, Sarah Shugars e Kechen Qin (2017). «Winning on the Merits: The Joint Effects of Content and Style on Debate Outcomes». In: *Transactions of the Association for Computational Linguistics* 5, pp. 219–232.

Wei, Zhongyu, Yang Liu e Yi Li (2016). «Is This Post Persuasive? Ranking Argumentative Comments in the Online Forum». In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany: Association for Computational Linguistics, pp. 195–200.

Wendt, Hans W (1972). «Dealing with a common problem in social science: A simplified rank-biserial coefficient of correlation based on the statistic.» In: *European Journal of Social Psychology*.

Werner, Carol (1978). «Intrusiveness and Persuasive Impact of Three Communication Media». In: *Journal of Applied Social Psychology* 8.2, pp. 145–162.

Xiong, Wenting e Diane Litman (2011). «Automatically predicting peer-review helpfulness». In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 502–507.

Yang, Diyi, Jiaao Chen, Zichao Yang, Dan Jurafsky e Eduard Hovy (2019). «Let's Make Your Request More Persuasive: Modeling Persuasive Strategies Via Semi-Supervised Neural Nets on Crowdfunding Platforms». In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational LInguistics: Human Language Technologies. Vol. 1 (Short and Long Papers), pp. 3620–3630.

Yang, Yinfei, Yaowei Yan, Minghui Qiu e Forrest Bao (lug. 2015). «Semantic Analysis and Helpfulness Prediction of Text for Online Product Reviews». In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Beijing, China, pp. 38–44.

Zhang, Zhu e Balaji Varadarajan (2006). «Utility scoring of product reviews». In: *Proceedings of the 15th ACM international conference on Information and knowledge management (CIKM '06)*. New York, NY, USA: Association for Computing Machinery, pp. 51–57.