Conoscere la comunità con la sentiment analysis: un progetto di CSBNO in tempo di crisi

ANNA MARIA TAMMARO
annamaria.tammaro@unipr.it

MICHELE TOMAIUOLO
michele.tomaiuolo@unipr.it

MONICA MORDONINI
monica.mordonini@unipr.it

LUCA CALDERINI
luca.calderini@studenti.unipr.it

PAOLO D'ALESSANDRO
paolo.dalessandro@studenti.unipr.it

Tutti gli autori afferiscono al Dipartimento
di Architettura e Ingegneria, Università di Parma

DOI: 10.3302/0392-8586-202206-027-1

Introduzione

Il CSBNO è un consorzio di circa 60 biblioteche nel Nord Ovest di Milano. Il Consorzio partecipa al Progetto europeo NEWCOMER[1] che intende realizzare la visione della biblioteca innovativa che mette al centro le comunità. Conoscere le comunità, non solo i bisogni ma anche le aspettative e i desideri, è considerata l'attività prioritaria della biblioteca di comunità che sceglie un approccio partecipativo per il disegno di programmi e servizi.
Nell'ambito del Progetto NEWCOMER le biblioteche del Consorzio CSBNO hanno condotto un'intervista telefonica durante il primo lockdown da marzo a maggio 2020 cercando di comprendere i desideri e le aspettative della comunità verso la biblioteca digitale. La più grande difficoltà per il CSBNO in tempo di crisi pandemica è stata quella di cambiare il modello di servizio, da servizi in presenza a servizi remoti. Il CSBNO ha cercato di rimanere connesso alla comunità, informandola che le biblioteche, anche se chiuse, continuavano il servizio a distanza.

1. Scopo ed obiettivi

Lo scopo dell'indagine del Consorzio CSBNO era quello di conoscere l'opinione del servizio della biblioteca digitale presso coloro che non la usano. Per la raccolta dei dati sono state effettuate oltre 30.000 telefonate da parte dei bibliotecari del Consorzio CSBNO ai membri classificati come membri attivi e inattivi.[2]

Sono stati raccolti due set di dati:
1. risposte telefoniche ricevute da membri inattivi;
2. risposte telefoniche ricevute dai membri attivi.

Il progetto di collaborazione tra il CSBNO e l'Università di Parma[3] è nato dal desiderio del CSBNO di sperimentare come metodo di indagine la *sentiment analysis* e dalla opportunità di coinvolgere i ricercatori e gli studenti del Dipartimento di Architettura e Ingegneria su dati e utenti reali. Il progetto collaborativo ha avuto l'obiettivo di introdurre il metodo di ricerca della sentiment analysis nelle biblioteche per conoscere le esigenze e aspettative della comunità durante un periodo di crisi. Il progetto, inoltre, si propone anche come un modo di condurre indagini sulla comunità, utilizzando gli algoritmi per l'analisi dei sentimenti che sono stati sperimentati e che potranno essere riutilizzati anche in futuro in circostanze diverse.

2. Cos'è la sentiment analysis?

Sentiment analysis (o Opinion mining o Emotion analysis) è un metodo per analizzare e classificare automaticamente sentimenti, opinioni e percezioni espresse dai loro autori in base a documenti testuali. L'analisi del sentimento funziona fornendo al computer un cosiddetto *bag of words*, che gli consente di riconoscere la sintassi e le parole che gli esseri umani usano per esprimere opinioni positive e negative. La sentiment analysis impiega programmi informatici di Natural Language Processing (NLP), data mining e linguistica computazionale per estrarre significato da un corpus ampio di testo. Come suggerisce il nome, l'analisi del sentimento implica l'identificazione e l'analisi di elementi soggettivi all'interno di un determinato testo come opinioni, commenti, punti di vista, preferenze, emozioni positive e negative (Wilson et al., 2005).
Per eseguire l'analisi del sentimento del testo è possibile utilizzare una serie di tecniche NLP. Alcune tecniche usano un lessico, altre usano un set di apprendimento e altre ancora usano entrambi.

- *Lessico*: richiede un corpus (*sentiment lexicon*) delle parole e dei sentimenti in cui ogni parola e frase è connotata, dall'orientamento positivo o negativo. Per determinare i sentimenti evidenziabili in un testo, si identificano le parole-opinioni nel testo. Il corpus è quindi usato per determinare la polarità di ogni parola-opinione ed un orientamento semantico +1 o -1 viene assegnato rispettivamente a parole con orientamento positivo o negativo. Se il *total polarity score* è positivo, il testo è considerato positivo o viceversa.
- *Apprendimento*: in questo metodo, il rilevamento del sentimento è un'attività di classificazione binaria, ovvero positiva o negativa. Il processo della sentiment analysis è una forma di apprendimento automatico (*machine learning*) supervisionato. Dei set di dati pre-etichettati vengono utilizzati come esempi di formazione per "insegnare" al computer e creare le basi per classificare le informazioni future senza etichetta (Mohri et al., 2012). Fornendo parole "positive" (come buono, educato, eccellente ecc.) e "negative" pre-etichettate (come terribile, scadente, scortese ecc.), il software di *data mining* può stabilire un modello che verrà applicato ai commenti futuri per decifrare la loro polarità, cioè se hanno un sentimento positivo o negativo. Un dato testo è classificato sulla base di esempi etichettati. Un set di dati di apprendimento viene utilizzato per estrarre le funzionalità rilevanti e quindi utilizzato per addestrare l'algoritmo (come ad esempio Naive Bayes, Maximum Entropy, SVM ecc.).

Altri metodi sono basati sull'individuazione di parole chiave e concetti.

- *Parole chiave*: questo metodo include lo sviluppo di un elenco di parole "emotive", cioè parole chiave che si riferiscono a un determinato sentimento. Questi sono solitamente aggettivi positivi/negativi usati come forti indicatori di sentimento.
- *Concetti*: questo metodo si basa su ampie basi di conoscenza semantica e utilizza reti semantiche e ontologie web per l'analisi semantica del testo. A differenza dei metodi che utilizzano parole chiave e conteggi di co-occorrenza di parole, utilizzano il significato implicito associato a determinati concetti del linguaggio naturale.

2.1 Sentiment analysis nelle biblioteche

Ricercatori e bibliotecari hanno cercato di usare la sentiment analysis come metodo di ricerca per capire le opinioni (a favore o contro), gli atteggiamenti (negativi o positivi), le preferenze (piace o non piace) e le emozioni (buono o cattivo) percepite dal pubblico in relazione ai vari servizi delle biblioteche. È importante notare che questo metodo di ricerca non vuole studiare fatti oggettivi, come statistiche ed evidenza concreta, e neppure indagare la soddisfazione per i servizi ma solo i sentimenti positivi o negativi del pubblico verso la biblioteca, analizzati in un determinato testo di solito di grandi dimensioni.
La comunità bibliotecaria internazionale ha utilizzato diversi testi per l'analisi del sentimento: i social media,

il testo a risposta libera dei questionari e altri *corpora*.
Un'esperienza importante di sentiment analysis è stata realizzata dalle biblioteche accademiche canadesi (Moore, 2017). I bibliotecari hanno analizzato i sentimenti dell'utenza per due servizi fondamentali della biblioteca: accesso alla collezione e allo spazio fisico. Hanno aggiunto inoltre la percezione emotiva che la biblioteca in generale suscita negli utenti. Per costruire un modello di sentiment analysis specifico della biblioteca sono stati utilizzati i commenti qualitativi alle domande aperte nei questionari LibQUAL+ (Kyrillidou and Cook, 2008) e sono stati creati vettori di parole con associati sentimenti positivi o negativi relativi a biblioteche universitarie di medie dimensioni. Il set di apprendimento è stato costruito selezionando casualmente 105 commenti da ciascuno dei risultati del sondaggio LibQUAL+ nel 2008 (614 commenti), nel 2010 (755 commenti), nel 2013 (528 commenti), nel 2015 (343 commenti) e nel 2016 (326 commenti), con 2.566 commenti totali. Per ottenere un apprezzamento comparativo delle opinioni degli intervistati nel tempo, i commenti aperti sono stati tutti analizzati tenendo traccia sia del loro sentimento che degli argomenti di riferimento. Il modello di analisi del sentimento di questa ricerca è stato creato utilizzando RapidMiner,[4] una piattaforma software open source per l'analisi predittiva e il data mining.
Numerose applicazioni della sentiment analysis alle biblioteche si possono trovare all'interno della letteratura sul marketing delle biblioteche, con lo scopo di applicare metodi predittivi per capire gli orientamenti del pubblico e successivamente prendere decisioni informate. Alcune di queste applicazioni si basano sui post dei social media (Palmer, 2014; Lund, 2020). Lund (2020) discute l'uso del linguaggio di programmazione "R" per eseguire un'analisi del sentimento raccogliendo i tweet relativi alla biblioteca. In questa ricerca viene presentato un codice riutilizzabile, completo di istruzioni su come gli utenti possono eseguirlo per analizzare i tweet. Sono anche proposti due esempi, creati utilizzando il codice (analisi dei tweet delle biblioteche delle migliori università e analisi dei tweet sui principali editori di libri) e che vengono utilizzati per dimostrare la funzionalità del codice. Al-Daihani (2016, 2018) ha analizzato un set di dati riguardanti biblioteche accademiche su Twitter e Facebook. Il set di dati per la prima ricerca è stato raccolto dalle linee temporali complete di Twitter di dieci biblioteche accademiche. Il set di dati totale comprendeva 23.707 tweet con 17.848 menzioni, 7.625 hashtag e 5.974 retweet. Le biblioteche accademiche analizzate hanno in genere pubblicato meno di 50 tweet al mese, sebbene il volume dei tweet sia cresciuto rapidamente tra la fine del 2013 e il 2014. I risultati mostrano la differenza tra le biblioteche accademiche nella distribuzione dei tweet nel tempo. La seconda ricerca ha applicato un approccio di analisi di un set di dati raccolto dai post di Facebook delle biblioteche accademiche nelle prime 100 università di lingua inglese, come valutate dal Shanghai World University Rankings nel 2014. Il set di dati proviene da una cronologia di pubblicazione di due anni di 18.333 post unici, 113.621 Mi piace e 3.401 Commenti. Meno di un quarto delle biblioteche esaminate ha avuto più di 2000 Mi piace e solo sette hanno ricevuto più di 100 commenti sui loro post. L'analisi del contenuto ha identificato la singola parola (unigram), sequenze di due parole (bigram) e sequenze di tre parole (trigram) prevalenti nel contenuto, con coinvolgimento alto e basso della comunità. L'analisi semantica ha identificato le categorie semantiche per i post con coinvolgimento positivo e negativo.
Lamba (2018) descrive uno studio sperimentale per introdurre un nuovo servizio per gli utenti delle biblioteche. In una prima fase, i dati sono stati recuperati da Twitter su 20 diverse query relative alla produttività della biblioteca. È stata utilizzata la piattaforma RapidMiner e successivamente è stata eseguita la sentiment analysis utilizzando il software di analisi del testo Aylien.[5] Un totale di 6.416 tweet sono stati estratti da Twitter per un periodo di 13 giorni. Inoltre, sono stati identificati 676 hashtag importanti, in cui 83 hashtag sono stati trovati associati a una posizione geografica nei tweet.
Un'altra base di riferimento che è stata usata per l'applicazione della sentiment analysis alle biblioteche è stata la chat. Logan (2019) ha cercato di identificare fattori e comportamenti associati ai sentimenti negativi degli utenti analizzando l'interazione nella chat del servizio di reference, con lo scopo di fornire agli operatori della chat suggerimenti sui comportamenti da evitare. I ricercatori hanno esaminato 473 trascrizioni della chat del servizio di reference raccolte da un consorzio di biblioteche accademiche da giugno a dicembre 2016. Le trascrizioni sono state codificate per 13 comportamenti che sono stati successivamente analizzati statisticamente confrontandoli con le valutazioni dei sondaggi.

3. Metodologia

Nel progetto del CSBNO la collaborazione tra ricercatori e bibliotecari ha permesso di utilizzare le tecniche più rappresentative e consolidate di sentiment analysis e *topic detection*. Mentre la sentiment analysis rileva i sentimenti positivi o negativi presenti in un testo, topic detection è una tecnica di machine learning che consente di identificare gli argomenti semantici presenti all'interno di un testo.
In particolare, sono stati presi in considerazione i migliori algoritmi di diverse famiglie, inclusi quelli basati su al-

cune nozioni di distanza geometrica tra campioni (come ad esempio knn, svm) (Huq et al., 2017), alberi decisionali (come rf, xgboost) (Hama et al., 2021), probabilità e statistica (Laksono et al., 2019), percettrone piccole reti neurali (Akhtar et al., 2017). In effetti, tali algoritmi, o la loro composizione (Angiani et al., 2016), hanno dimostrato la loro buona accuratezza su molti dataset diversi, di piccole e medie dimensioni (Tomaiuolo et al., 2020). Inoltre, alcuni algoritmi più recenti hanno migliorato l'accuratezza su insiemi di dati più grandi, sfruttando le cosiddette architetture di rete neurale profonda, insieme a tecniche più avanzate per raccogliere i vettori delle caratteristiche dell'insieme di apprendimento. La vettorizzazione tradizionale, basata sull'algoritmo del bag of words, crea un dataset con un numero molto elevato di caratteristiche e richiede una fase di selezione delle caratteristiche accurata e sensata, per ottenere i migliori risultati. Invece, le tecniche di incorporamento di parole e rappresentazioni dense (Yadav et al., 2020) sono in grado di mappare ogni parola in uno spazio multidimensionale, dove le parole semanticamente correlate sono rappresentate come punti a breve distanza. Il vettore che rappresenta ciascun campione è calcolato sulla base delle posizioni delle parole in questo spazio multidimensionale. Inoltre, le reti neurali profonde hanno mostrato risultati efficienti in molte applicazioni, inclusa l'analisi del sentimento (BERT) (Sun et al., 2019). Uno svantaggio è che queste reti sono caratterizzate da un numero elevatissimo di parametri che devono essere appresi, richiedendo l'utilizzo di campioni dell'ordine di grandezza dei big data. Nel presente lavoro collaborativo tra CSBN e Dipartimento di Architettura e ingegneria dell'Università di Parma sono state utilizzate tecniche consolidate insieme a nuove tecniche, sfruttando modelli pre-addestrati e fasi aggiuntive di machine learning e apprendimento, per adattare i modelli al particolare scopo e obiettivi che il CSBNO si proponeva. I passaggi aggiuntivi di solito richiedono set di dati molto più piccoli rispetto a quelli che sono stati utilizzati dal progetto collaborativo per addestrare l'intero modello.

3.1 Fasi

Il modello di sentiment analysis utilizzato dal progetto prevede essenzialmente tre passaggi: preelaborazione, analisi del testo, creazione di un modello.

3.1.1 Pre-elaborazione

Il testo utilizzato è stato creato dai bibliotecari del CSBNO che hanno effettuato le interviste e trascritto le risposte. Il testo preso in considerazione dal Consorzio CSBNO è in italiano. Per poter eseguire un modello di sentiment analysis è necessaria una prima fase di preelaborazione del testo. Questa fase è molto importante in quanto rende il testo utilizzabile. Per creare un modello, è necessario selezionare sia gli elementi di testo da analizzare che le assegnazioni dei sentimenti corrispondenti a ciascuna parola e frase. Nel progetto la prima fase di preelaborazione è consistita di due attività.
La prima ha avuto lo scopo di rimuovere dall'elenco le risposte non analizzabili, come risposte lasciate vuote, risposte rappresentate da un numero o da dei caratteri privi di significato come sequenze di punti, virgole o altri simboli. Successivamente, una volta ottenuta la lista delle risposte analizzabili, la preelaborazione del testo ha eliminato le differenze linguistiche minori, come lettere minuscole e maiuscole, la pluralizzazione e i tempi verbali, utilizzando tecniche comuni di *stemming* e *stopword*. All'interno delle risposte alcune lettere, soprattutto accentate, erano sostituite da sequenze di caratteri, e, non sostituendole, parti delle frasi non sarebbero state analizzate in quanto prive di significato. Un altro problema presente all'interno del dataset utilizzato sono stati gli errori di battitura: purtroppo non è stato possibile sistemare tutti questi errori, in quanto, a causa della loro natura, non seguono una sequenza o delle regole, e per questo motivo sarebbe stato necessario individuare e risolvere manualmente ogni errore, cosa non fattibile con un dataset di svariate migliaia di righe e il tempo limitato a disposizione. Questo, tuttavia, ha portato a perdita d'informazione all'interno delle frasi, poiché queste parole errate sono prive di significato. Tuttavia, poiché alcuni modelli utilizzano la struttura grammaticale del testo, anche il testo originale è stato mantenuto nel dataset, per un possibile utilizzo nelle fasi di analisi successive.
La seconda attività è consistita nella traduzione delle frasi ottenute precedentemente. Per eseguire una sentiment analysis è necessario eseguire una prima fase di apprendimento di un modello precedentemente addestrato. Tuttavia, per far questo, è necessario avere a disposizione un lessico di grandi dimensioni, opportunamente etichettato con la polarità negativa o positiva per ogni frase contenuta in esso. Questo dataset viene poi suddiviso in due parti, una per effettuare l'apprendimento e una invece per la fase di testing. In questo modo il modello viene "specializzato" sulle frasi, e quindi sugli argomenti, contenuti all'interno del dataset, in modo da ottenere un rendimento migliore una volta applicato per eseguire la sentiment analysis sulle opinioni della comunità. Nel nostro caso questo non è stato possibile, in quanto il dataset a disposizione non era precedentemente etichettato, e realizzare questa fase avrebbe richiesto vario tempo e le polarità sarebbero state

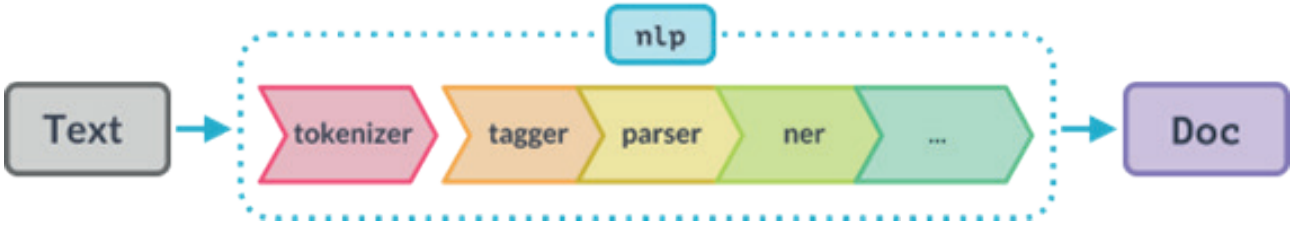


Figura 1 - *Pipeline* di SpaCy

definite in base alla nostra soggettività. Durante lo svolgimento di questo progetto collaborativo, è stato scelto di utilizzare dei lessici in lingua inglese. Questa scelta porta sia dei vantaggi che degli svantaggi: dal punto di vista dei vantaggi, è molto più ampia la scelta di classificatori in lingua anglofona, e, molti di questi, sono più evoluti e di maggiori dimensioni rispetto a quelli disponibili in lingua italiana. Ciò ha reso necessaria la traduzione del testo prima di essere sottoposto all'analisi: uno svantaggio è che questo può creare delle frasi con un significato meno preciso, rispetto alla frase originale in italiano. È tuttavia stata scelta la traduzione, in quanto si è pensato che l'uso dei classificatori in inglese fosse migliore dell'utilizzo di un classificatore che supportasse l'italiano con il vantaggio di poter utilizzare un modello pre-addestrato su una grande quantità di parole. Questo è stato rafforzato dal fatto che il dataset analizzato fosse incentrato sull'argomento particolare dei libri e delle biblioteche, considerato di nicchia. Si è deciso quindi che fosse necessario un modello con una base di apprendimento di grandi dimensioni.

3.1.2 Elaborazione del testo

Per realizzare la sentiment analysis è necessario elaborare il testo e trasformare le parole e le frasi in vettori, in modo da posizionarle all'interno di uno spazio vettoriale in base al loro significato e area semantica: due parole semanticamente simili si trovano più vicine all'interno dello spazio, mentre due parole diverse saranno posizionate distanti l'una dall'altra. Questo processo è detto *word embedding* (o rappresentazione distribuita delle parole), ed è uno dei processi chiave utilizzati all'interno del NLP. La tecnica più utilizzata per raggiungere il word embedding è *word2vec*: questa tecnica consiste in una elaborazione del testo letterale che restituisce una rappresentazione delle frasi e delle parole sotto forma di vettori.
Ai dataset del progetto collaborativo sono stati applicati sia l'algoritmo di sentiment analysis (Spacy) che l'algoritmo di *topic detection* (BERT). I risultati ottenuti sono stati analizzati, confrontando i risultati ottenuti dal primo tipo di algoritmo con quelli prodotti dal secondo.

SpaCY

SpaCy è una libreria open source Python che consente di eseguire la trasformazione di parole in vettori. Spacy utilizza una "pipeline" di operatori (Figura 1) i quali vengono applicati al testo analizzato, e restituisce un oggetto "Doc", contenente varie informazioni e attributi.
Come osservabile nella Figura 1, SpaCy applica vari operatori sul testo da analizzare. Tra queste il *tokenizer* è l'unico obbligatorio, in quanto si occupa della trasformazione delle parole e frasi in *token*, e quindi del word embedding. Il tagger, parser e ner sono presenti nella pipeline di default, ma possono essere rimossi se non necessari, essendo opzionali.
Viene così restituito l'oggetto Doc: questo contiene i *tokenized data*, cioè i token prodotti dalla *pipeline*, e in ognuno di questi token creati è presente un vettore posizionato nello spazio rappresentante una parola. Il Doc possiede inoltre un vettore che identifica la posizione della frase, risultato dei vettori delle varie parole di cui la frase è composta.
Utilizzando i token così creati, il progetto ha poi utilizzato *SpaCyTextBlob*, che rappresenta uno dei componenti aggiuntivi che si possono aggiungere alla pipeline SpaCy. Questo componente permette di effettuare una sentiment analysis sul testo elaborato e vettorizzato da SpaCy, restituendo una polarità sotto forma di valore compreso tra [-1, 1]: un valore molto negativo rappresenta una polarità negativa, mentre di conseguenza, un valore vicino all'1 rappresenta una buona positività: SpaCyTextBlob consente di utilizzare token e vettori generati da SpaCy per eseguire una sentiment analysis su di essi.
Questo componente si basa sulla libreria TextBlob di Python, che permette l'applicazione della sentiment analysis sul testo, oltre a contenere altre funzionalità di elaborazione del testo. TextBlob utilizza a sua volta attivamente il Natural Language ToolKit (NLTK) per eseguire i propri task. NLTK è una libreria che fornisce accesso a molte risorse lessicali, e consente di eseguire varie attività sul testo.
Seguendo questo procedimento sono stati ricavati i vari valori e risultati della sentiment analysis delle frasi qualitative contenute all'interno dei dataset del progetto.

BERT

Una volta effettuata la sentiment analysis attraverso il metodo precedente, è stato deciso di utilizzare la *topic detection*, in modo da confrontare i due set di risultati, per vedere se ci fossero differenze sostanziali. La topic detection è una tecnica di machine learning che permette di individuare i vari argomenti (topic) presenti all'interno di una grande quantità di testo. Per fare questo, utilizza tecniche di NLP per scomporre le frasi sotto forma di linguaggio umano, in modo da poter individuare schemi o strutture semantiche all'interno di essi. Per realizzare la topic detection il progetto ha utilizzato BERT (Bidirectional Encoder Representations from Transformers) per eseguire sentiment analysis su dati trasformati in token, restituendo il *sentiment* identificato sotto forma di cinque valori: molto positivo, positivo, neutro, negativo e molto negativo. In particolare, è stato usato il modello *bert-base-multilingual-uncased-sentiment*, ossia un modello ottenuto attraverso un processo di fine tuning del modello *bert-base-multilingual-uncased*. Il fine tuning permette un nuovo allenamento di alcuni dei livelli di un modello che era stato preaddestrato per uno scopo più generale, utilizzando un *training set* adeguato e ottenendo un nuovo modello specializzato per un particolare task. Per utilizzare questo modello è stato prima creato un tokenizer, con lo scopo di trasformare il testo in token. Successivamente questi token sono stati passati al modello vero e proprio, che ha restituito il sentimento analizzato sotto forma di questi cinque valori. In particolare, è stato utilizzato BERTopic che rappresenta una tecnica di modellazione degli argomenti che sfrutta gli incorporamenti BERT, basato su classi, per creare cluster densi che consentono di individuare argomenti facilmente interpretabili.

3.1.3 Creazione del modello

Sono stati applicati i due algoritmi di sentiment analysis e di topic detection sulle risposte di tipo qualitativo presenti all'interno dei dataset da elaborare, sono stati quindi analizzati i risultati ottenuti, confrontando quelli risultanti dal primo tipo di algoritmo con quelli prodotti dal secondo.
Le risposte su cui è stato possibile applicare questi algoritmi sono poche, in quanto la maggior parte delle risposte erano di tipo chiuso e quindi non adatte all'essere analizzate. In particolare, sono state trovate solo cinque colonne di risposte su cui ha avuto senso applicare una sentiment analysis, in quanto contenenti commenti liberi. È stata effettuata la sentiment analysis su questa colonna, utilizzando prima SpaCyTextBlob e poi BERT. Per verificare l'accuratezza dei modelli, questi vengono testati su alcuni dati di test pre-taggati, per misurare la precisione, il richiamo e l'accuratezza della classificazione. Il modello del progetto è stato salvato per un ri-uso futuro.
Per creare il modello, come compito ortogonale, i commenti selezionati sono anche stati classificati in base al loro argomento. Questa ulteriore classificazione fornisce una comprensione più profonda e completa delle opinioni raccolte. Con lo stesso approccio, è anche possibile classificare un testo in base al suo argomento specifico. In alternativa, è possibile utilizzare algoritmi di clustering per raggruppare testi con caratteristiche simili, in uno scenario non supervisionato.

4. Risultati

Sono analizzati di seguito i tre macro argomenti trattati all'interno delle interviste telefoniche effettuate dai bibliotecari del CSBNO, in modo da capire le percezioni delle persone su ognuno di questi servizi offerti: lettura digitale, MLOL, OPAC. Infine, sono stati considerati i commenti sulla biblioteca digitale in generale.

4.1 Lettura digitale

In Figura 2 sono evidenziati i commenti negativi, positivi e neutri (1.695 in totale) espressi dagli utenti rispetto alla lettura di tipo digitale. Per la sentiment analysis si è dapprima usato SpaCYTextBlob. Le risposte positive sono 125, quelle negative 1.056, quelle neutre 512. Le soglie utilizzate per la definizione della polarità per questa colonna e per le successive sono state: positivo > 0.05, negativo < -0.05 e neutro compreso nei due valori.
Successivamente si è usato l'algoritmo BERT. Le risposte molto positive sono 163, quelle positive 91, quelle neutre 412, mentre negative sono 761, infine molto negative 268.
I risultati ottenuti mostrano una grande maggioranza di polarità negativa rispetto a quella positiva, con entrambi i modelli utilizzati. La maggior parte dei commenti negativi individuati provengono da risposte del tipo "preferisco il cartaceo". Per questo motivo, la negatività ottenuta non è strettamente relativa a un servizio non buono, ma piuttosto indica una preferenza della comunità per la lettura di libri fisici, essendo abituata maggiormente a questa tipologia.
Una volta ottenute le valutazioni dei commenti, è stata effettuata un'ulteriore fase di analisi statistica in base all'età degli utenti.
L'età media degli individui rispondenti è stata di circa 48.0 anni.

spaCyTextBlob

Risposte totali	Positive	Negative	Neutre
1695	125	1056	512

BERT

Risposte totali	Molto positive	Positive	Neutre	Negative	Molto negative
1695	163	91	412	761	268

Età media: 48.0 anni

Polarità	Positiva	Negativa	Neutra
Età media	45.8	48.3	48.1

Figura 2 - Lettura digitale: sentiment analysis

Topic individuati: 38

Topic	0	1	2	3	4	5	6	7
Occorrenze	179	51	50	45	43	40	40	40



Figura 3 - Lettura digitale: topic detection

Le risposte ottenute erano così suddivise: 77 rispondenti con meno di 30 anni, 1 solo con più di 65 anni, e 1102 compresi tra i 30 e 65. È stata poi ottenuta l'età media degli intervistati che hanno rilasciato commenti con le tre polarità:

- età media delle risposte positive: 45.8;
- età media delle risposte negative: 48.3;
- età media delle risposte neutre: 48.1.

Si può osservare che le risposte positive hanno un'età media più bassa rispetto alle altre. Questo può essere legato all'abitudine maggiore che hanno le persone di età più giovane nell'utilizzo di dispositivi e servizi digitali. Tuttavia, essendo i commenti positivi di numero poco elevato, questa statistica non è tanto significativa (Figura 3).

Nell'analisi della topic detection il primo topic individuato è presente nella stragrande maggioranza delle risposte, ed è quindi il più importante tra quelli contenuti in questa colonna. Questo topic riguarda la preferenza da parte degli utenti della lettura cartacea. Questi risultati della topic detection combaciano perfettamente quindi con i risultati della sentiment analysis. Oltretutto, possono essere confrontati anche con i topic individuati precedentemente, rappresentati da quelle che sono le cose più importanti per gli utenti. Infatti, il motivo primario che porta le persone in biblioteca sono proprio i libri cartacei. Gli altri topic rappresentano avverbi di tempo, che mostrano un raro uso della lettura digitale, la preferenza di libri o audio, ovviamente l'aspetto digitale, e il possesso o meno di dispositivi per la lettura digitale da parte degli utenti.

spaCyTextBlob

Risposte totali	Positive	Negative	Neutre
1473	341	267	865

BERT

Risposte totali	Molto positive	Positive	Neutre	Negative	Molto negative
1473	214	154	354	414	227

Età media: 44.7 anni

Polarità	Positiva	Negativa	Neutra
Età media	44.0	43.9	45.2

Figura 4 - MLOL: sentiment analysis

4.2 Media Library Online (MLOL)

La sentiment analysis è stata effettuata sui 1.473 commenti degli utenti relativi a MLOL (Media Library OnLine), il primo servizio italiano di biblioteche digitali (Figura 4).

Anche per le opinioni su MLOL è stata effettuata la sentiment analysis con entrambi i modelli. Usando SpaCyTextBlob: le risposte positive sono 341, quelle negative sono 267 e quelle neutre sono 865. Usando BERT le risposte molto positive sono 214, quelle positive 154, quelle neutre sono 354, infine quelle negative sono 414 e quelle molto negative sono 227.

È subito possibile notare una differenza tra i risultati delle due analisi. Infatti, nel primo caso le polarità positive e negative sono molto minori rispetto a quelle neutre, mentre utilizzando BERT, le risposte neutre diminuiscono altamente, aumentando di conseguenza quelle negative.

Questo può essere dovuto al fatto che nella colonna dei commenti, molti di questi sono rappresentati da frasi del tipo "non lo utilizza". Queste frasi vengono considerate neutre nei risultati ottenuti attraverso SpaCyTextBlob, mentre esse diventano negative nel caso di utilizzo di BERT. Il non utilizzo non è di per sé un commento negativo, e per questo motivo la prima classificazione potrebbe essere più accurata; tuttavia, se è interessante determinare il successo di utilizzo di quel servizio, possono essere corrette anche le classificazioni di BERT.

L'analisi statistica sull'età degli utenti che hanno lasciato una risposta a questo campo ha rivelato un'età media di 44.7. L'età media degli intervistati che hanno rilasciato commenti con le tre polarità:

- età media delle risposte positive: 44.0;
- età media delle risposte negative: 43.9;
- età media delle risposte neutre: 45.2.

Topic individuati: 42

Topic	0	1	2	3	4	5	6	7
Occorrenze	187	56	45	42	41	40	39	37



Figura 5 - MLOL: topic detection

In questo caso le medie delle età ottenute sono più o meno costanti, e non è effettuabile nessuna considerazione.
È stata poi effettuata la topic detection su questa colonna, che ha prodotto i seguenti risultati evidenziati in Figura 5.
I topic individuati sono stati 42, con le occorrenze evidenziate in Figura 5 per i primi sette topic. Anche in questo caso vi è il primo topic (il numero 0) prevalente su tutti gli altri individuati. L'argomento di questo topic è appunto MLOL e le sue offerte, ossia l'oggetto centrale di questo set di risposte. Gli altri sette topic principali individuati che sono stati riportati, rappresentano la preferenza degli utenti rispetto al cartaceo, la lettura digitale e le biblioteche chiuse. È ovvia la presenza del primo e secondo topic, mentre, per quanto riguarda gli altri, essi possono essere usati per degli spunti di ricerca. Infatti, una presenza sostanziale della preferenza dei libri cartacei combacia con i risultati di sentiment analysis e topic detection anche precedenti, in cui è stata evidenziata la preferenza da parte degli utenti dei libri fisici. Per quanto riguarda invece l'ultimo topic evidenziato, da esso si può dedurre che il servizio MLOL viene utilizzato in caso di biblioteche chiuse.

4.3 OPAC

Successivamente è stata effettuata la sentiment analysis degli utenti rispetto all'OPAC (Online Public Access Catalog), il catalogo informatizzato reso disponibile dalle biblioteche. Sono state considerate 605 risposte totali ed anche in questo caso sono stati applicati entrambi i modelli sulla colonna contenente i commenti degli utenti su cosa ne pensassero del servizio offerto. I risultati sono evidenziati in Figura 6.
Utilizzando SpaCyTextBlob le risposte positive sono 344, quelle negative sono 42, quelle neutre sono 219.
Utilizzando l'algoritmo di BERT le risposte molto positive sono 176, le risposte positive sono 179, quelle neutre

spaCyTextBlob

Risposte totali	Positive	Negative	Neutre
605	344	42	219

BERT

Risposte totali	Molto positive	Positive	Neutre	Negative	Molto negative
605	176	179	118	76	56

Età media: 44.0 anni

Polarità	Positiva	Negativa	Neutra
Età media	43.5	44.5	44.6

Figura 6 - OPAC: sentiment analysis

sono 118, quelle negative sono 76 e quelle molto negative sono 56.
Anche in questo caso, a causa della diversa natura dei due modelli, nel primo set sono di più le risposte neutre ottenute rispetto al secondo. Parte di questo è probabilmente dato dal fatto che le risposte presenti simili a "non lo uso spesso" vengono classificate come negative da BERT. In entrambi i casi, comunque, le opinioni positive sono maggiori rispetto a quelle negative, denotando una buona performance del servizio offerto agli utenti, che in generale si ritengono soddisfatti.
Anche per OPAC è stata effettuata una analisi statistica dell'età ottenendo un valore medio di 44.0 anni. L'età media degli intervistati che hanno rilasciato commenti con le tre polarità:

- età media delle risposte positive: 43.5;
- età media delle risposte negative: 44.5;
- età media delle risposte neutre: 44.6.

In questo caso si può vedere che gli utenti con commenti positivi hanno una media di anni leggermente inferiore. Questo può avvalorare l'ipotesi precedente, che vede una popolazione più giovane, anche se di poco, più abituata ai servizi digitali. La differenza rimane comunque molto sottile in questo caso per farne una statistica rilevante.
È stata poi effettuata la topic detection sulla colonna relativa a come viene usato l'OPAC dagli utenti. I topic individuati sono stati 23, con le seguenti occorrenze per i primi sette (Figura 7).
In questo caso non vi è un topic prevalente sugli altri, ma tutti questi argomenti sono presenti in misura simile. Dal primo topic si nota subito la ragione della presenza dei commenti negativi: anche in questo caso alcuni utenti non sono soddisfatti del servizio a causa della loro preferenza nei confronti della biblioteca fisica, intesa come luogo. A parte questo, gli argomenti dei commenti vertono sia sulle funzionalità per le quali OPAC viene utilizzato prevalentemente, come rinnovi, prestiti o prenotazioni, con presenza all'interno delle frasi di commenti sulle credenziali e password necessari per l'utilizzo, sia sul livello del servizio stesso, con

Topic individuati: 42

Topic	0	1	2	3	4	5	6	7
Occorrenze	187	56	45	42	41	40	39	37



Figura 7 - OPAC: topic detection

spaCyTextBlob

Risposte totali	Positive	Negative	Neutre
7178	5354	767	1057

BERT

Risposte totali	Molto positive	Positive	Neutre	Negative	Molto negative
7178	2610	1103	1899	175	1391

Età media: 47.5 anni

Polarità	Positiva	Negativa	Neutra
Età media	47.4	47.9	47.6

Figura 8 - Commenti finali: sentiment analysis

parole e commenti molto positivi. Confrontando quindi proprio questi ultimi due topic, in cui sono presenti parole molto buone nei confronti di OPAC, con i risultati della sentiment analysis, abbiamo la prova dell'opinione positiva della maggior parte degli utenti che usano questo servizio.

4.4 Commenti finali

L'ultima analisi approfondita è stata effettuata sull'ultima colonna del dataset dell'intervista telefonica. Questa colonna rappresenta 7.178 commenti finali rilasciati dagli utenti riguardanti il servizio di biblioteca digitale in generale. I commenti, di tipo libero, danno una valutazione complessiva, e quindi si prestano benissimo per effettuare una sentiment analysis (Figura 8).
Utilizzando i due diversi algoritmi, sono stati evidenziati i seguenti risultati:

- con SpaCyTextBlob i commenti positivi sono 5.354, quelli negativi sono 767 e quelli neutri sono 1.057;
- con BERT le risposte molto positive sono 2.610, quelle positive sono 1.103, quelle neutre sono 1.899 e quelle negative e molto negative sono rispettivamente 175 e 1.391.

Analizzando i dati ottenuti si può subito osservare la maggioranza dei commenti positivi rispetto a quelli negativi, con entrambi gli algoritmi utilizzati. Le differenze tra questi due modelli sono costituite dal fatto che attraverso BERT, i numeri di neutri e negativi sono un po' superiori, a discapito dei positivi. Questo può essere dipeso sia dalla differenza dei due modelli in sé, sia dalle soglie utilizzate, nel caso di SpaCyTextBlob, per differenziare le tre possibili polarità.
Essendo questo il dataset più grande tra quelli analizzati, ed essendo i commenti contenuti relativi al servizio nella sua totalità, i dati ottenuti sono molto significativi, e denotano un buon grado di curiosità o interesse, entrambi sentimenti positivi trattandosi di un servizio relativamente nuovo e in crescita.
A seguito dell'analisi statistica effettuata sulle età degli utenti intervistati, la media ottenuta è di 47.5 anni. Entrando nei particolari, le età medie degli intervistati che hanno rilasciato commenti con le tre polarità sono le seguenti:

- età media delle risposte positive: 47.4;
- età media delle risposte negative: 47.9;
- età media delle risposte neutre: 47.6.

Topic individuati: 190

Topic	0	1	2	3	4	5	6	7
Occorrenze	791	606	530	342	222	132	121	96



Figura 9 - Commenti finali: topic detection

Inoltre, gli intervistati sono suddivisi in tre fasce di età nel seguente modo: 491 con età minore di 30, 12 con età superiore a 65 e 6.675, quindi la stragrande maggioranza, con anni compresi tra 30 e 65. A seguito di questa statistica, in questo caso, non si evidenziano differenze sostanziali nella suddivisione delle età nelle tre polarità, avendo solamente una differenza di sei mesi tra i commenti positivi e negativi, con un'età minore nel caso dei positivi.
Anche per l'analisi dei commenti finali è stato applicato l'algoritmo di topic detection BERTopic. I topic individuati sono stati 190, con le seguenti occorrenze per i primi 7 (Figura 9).
Si può subito notare che i primi topic rappresentano l'interesse degli utenti rispetto al servizio di biblioteca digitale. Era stato ottenuto un quarto topic rappresentante le persone interessate, simile al primo, ma con parole differenti. Essendo simile a quello già riportato è stato deciso di menzionare solamente il primo.

A seguito della sentiment analysis, il topic più presente è quello derivante dai commenti di utenti abbastanza interessati, seguito da quello degli utenti un po' interessati e non interessati. Ricordiamo la presenza del quarto topic di persone interessate, che andrebbe a rafforzare ancora di più l'interesse dei primi due. Quindi questi topic consolidano a loro volta le polarità precedentemente ottenute, confermando una positività generale da parte dell'utenza nel confronto dei vari servizi digitali offerti. Altri topic presenti in questi commenti sono relativi a MLOL, considerato servizio centrale della biblioteca digitale, e a vari aggettivi qualificativi usati per MLOL, anche questi molto positivi, confermando a loro volta i buoni risultati della sentiment analysis.

Conclusioni

La sentiment analysis della comunità durante la crisi pandemica evidenzia un approccio verso le biblioteche digitali con commenti sia positivi che negativi, che dimostra comunque un interesse prevalentemente positivo da parte della comunità.
Per quanto riguarda invece la lettura digitale, la maggior parte degli intervistati hanno mostrato una opinione negativa nei suoi confronti, ribadendo il piacere nel leggere un libro cartaceo. Tuttavia, alcuni di questi commenti negativi possono appartenere a quegli utenti che si sono detti in ogni caso interessati a scoprire i servizi digitali in futuro.
Analizzando i singoli servizi, nel caso di MLOL la differenza tra commenti positivi e negativi è molto bassa, in favore dei primi, considerando come neutri i commenti di non utilizzo, mentre nel caso dell'OPAC la polarità generale tende invece al positivo.
I principali vantaggi dell'utilizzo dell'analisi del sentimento nel settore delle biblioteche riguardano la possibilità di conoscere opinioni e percezioni della comunità su grandi quantità di dati e in tempi relativamente brevi. Questa caratteristica è stata considerata particolarmente importante durante la pandemia, quando il CSBNO ha dovuto prendere decisioni in tempi brevi. Di solito le indagini qualitative vengono contrapposte ad indagini con metodi quantitativi e rappresentano sempre casi di studio limitati. La sentiment analysis consente di raccogliere invece dati qualitativi utilizzando grandi quantità di dati su opinioni e percezioni soggettive. I risultati ottenuti vanno comunque approfonditi, utilizzando altri metodi di raccolta e di analisi dei dati.
Il Consorzio CSBNO potrà prevedere ad esempio di usare il modello di sentiment analysis come prima fase di indagini sull'utenza, integrando il metodo prevalente nelle biblioteche attualmente, che è quello dell'osservazione non strutturata o del questionario. Il vantaggio della sentiment analysis è che può fornire i primi dati di orientamento per evidenziare argomenti specifici, come un'analisi più approfondita delle opinioni e aspettative delle comunità e delle sotto-comunità per specifici servizi e funzionalità. Così, ad esempio, il modello realizzato dal progetto CSBNO che applica la sentiment analysis e la topic detection potrà essere riutilizzato per il servizio MLOL analizzato nel tempo e consentendo così di presentare le opinioni della comunità delle biblioteche in modo dinamico.

NOTE

[1] Il Progetto NEWCOMER è stato finanziato dal programma ERASMUS+ (Reference 2020-1-NL01-KA204-064611). Lo scopo è quello di creare una rete globale di biblioteche che si riconoscono nella biblioteconomia della comunità. È coordinato da PL2030 e dalla Biblioteca nazionale olandese. I partner del Progetto sono otto e comprendono: CSBNO, Cubiss (Paesi Bassi), Biblioteca di Colonia, Biblioteca di Kranj, Biblioteca nazionale della Repubblica Ceca, Biblioteche di Aarhus/DOKK1. Maggiori informazioni: https://webopac.csbno.net/azienda-speciale-csbno/newcomer.

[2] Per membri attivi si intendono gli utenti iscritti alle biblioteche CSBNO a partire dal 31/12/2017 e ancora attivi nell'utilizzo del servizio di prestito, selezionando quelli di età compresa tra i 25 e i 65 anni. Per membri inattivi si intendono gli utenti iscritti alle biblioteche CSBNO a partire dal 31/12/2017, ma non più attivi nell'utilizzo del servizio di prestito.

[3] Il progetto di collaborazione tra CSBNO e Dipartimento di Ingegneria e architettura dell'Università di Parma è stato coordinato da Anna Maria Tammaro e ha coinvolto Michele Tomaiuolo e Monica Mordonini insieme agli studenti di dottorato Luca Calderini e Paolo D'Alessandro

[4] Sviluppato presso l'Università Tecnica di Dortmund, RapidMiner fornisce un'interfaccia grafica e una serie di estensioni gratuite, come un modulo di elaborazione del testo, che è stato usato per l'analisi dei commenti di LibQUAL+.

[5] https://docs.aylien.com/textapi/#getting-started.

RIFERIMENTI BIBLIOGRAFICI

Akhtar, M.S., Kumar, A., Ghosal, D., Ekbal, A., & Bhattacharyya, P. (2017) *A multilayer perceptron based ensemble technique for fine-grained financial sentiment analysis*, in *Proceedings of the 2017 conference on empirical methods in natural language processing*, p. 540-546.

Al-Daihani S.M., Abrahams A. (2016) *A text mining analysis of academic libraries' Tweets*. "Journal of Academic Librarianship", 42 (2), p. 135-143, https://doi.org/10.1016/j.acalib.2015.12.014

Al-Daihani S.M., Abrahams A. (2018) *Analysis of academic libraries' Facebook posts: text and data analytics*. "Journal Academic Librarianship", 44, p. 216-225, https://doi.org/10.1016/j.acalib.2018.02.004.

Angiani, G., Cagnoni, S., Chuzhikova, N., Fornacciari, P., Mordonini, M., & Tomaiuolo, M., (2016) *Flat and hierarchical classifiers for detecting emotion in tweets*, in *Conference of the Italian Association for Artificial Intelligence*, p. 51-64.

Hama Aziz, R.H., & Dimililer, N. (2021) *SentiXGboost: enhanced sentiment analysis in social media posts with ensemble XGBoost classifier*, "Journal of the Chinese Institute of Engineers", 44 (6), p. 562-572.

Huq, M. R., Ali, A., & Rahman, A. (2017) *Sentiment analysis on Twitter data using KNN and SVM*, "International Journal of Advanced Computer Science and Applications", 8 (6), p. 19-25.

Kyrillidou, M. and Cook, C. (2008), *The evolution of measurement and evaluation of libraries: a perspective from the Association of Research Libraries*, "Library Trends", 56 (4), p. 888-909.

Laksono, R. A., Sungkono, K. R., Sarno, R., & Wahyuni, C. S. (2019) *Sentiment analysis of restaurant customer reviews on tripadvisor using naïve bayes*, in *12th International Conference on Information & Communication Technology and System, IEEE*, p. 49-54.

Lamba, M., Madhusudhan, M. (2018) *Application of sentiment analysis in libraries to provide temporal information service: a case study on various facets of productivity*. "Social Network Analysis and Mining", 8 (63), https://doi.org/10.1007/s13278-018-0541-y.

Lund, B. D. (2020) *Assessing library topics using sentiment analysis* in *R: a discussion and code sample*. "Public Services Quarterly", 16 (2), p. 112-123, https://doi.org/10.1080/15228959.2020. 1731402.

Logan, J., Barrett, K. and Pagotto, S. (2019) *Dissatisfaction* in *Chat Reference Users: A Transcript Analysis Study*. "College & Research Libraries", 80 (7), p. 925, https://crl.acrl.org/index.php/crl/article/view/23611.

Moore, M.T. (2017), *Constructing a sentiment analysis model for LibQUAL+ comments*, "Performance Measurement and Metrics", 18 (1), p. 78-87, https://doi.org/10.1108/PMM-07-2016-0031.

Palmer, S. (2014) *Characterizing university library use of social media a case study of Twitter and Facebook from Australia*. "Journal Academic Librarianship", 40 (6), p. 611–619. https://doi.org/10.1016/j.acalib.2014.08.007.

Sun, C., Huang, L., & Qiu, X., (2019) *Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentences*, arXiv preprint, arXiv:1903.09588.

Tomaiuolo, M., Lombardo, G., Mordonini, M., Cagnoni, S., & Poggi, A. (2020) *A survey on troll detection*, "Future Internet", 12 (2), p. 31.

Wilson, T., Wiebe, J. and Hoffmann, P. (2005) *Recognizing contextual polarity in phrase-level sentiment analysis*, in *Proceedings of the conference on human language technology and empirical methods in natural language processing*, Association for Computational Linguistics, p. 347-354.

Yadav, Ashima, and Dinesh Kumar Vishwakarma (2020) *Sentiment analysis using deep learning architectures: a review*. "Artificial Intelligence Review", 53 (6), p. 4335-4385.

ABSTRACT

The CSBNO project aimed to understand the needs and expectations of the community for digital services during a period of crisis. Two sentiment analysis and topic detection algorithms were applied to the datasets collected by librarians with the transcription of phone calls to library users. The results showed an interest in the digital library even though the print book is preferred, with positive sentiments for the services of MLOL and OPAC. The project is proposed as an example of conducting community surveys, using algorithms for the analysis of the sentiments and that can be reused in the future in different circumstances.