NOTA METODOLOGICA

La violenza di genere raccontata dai social: *sentiment - emotion analysis e topic modelling.*

1. Introduzione

La violenza contro le donne è un problema persistente che colpisce la nostra società. L'Istat ha definito un sistema integrato volto a monitorare i dati sulla violenza contro le donne utilizzando un approccio multi-fonte. Al fine di aggiungere nuove e alternative fonti di informazione statistica, l'Istat ha iniziato a esplorare nuove frontiere di ricerca con lo scopo di migliorare la metodologia di analisi statistica utilizzando i *Big Data*. In particolare, è stato promosso un progetto di analisi che utilizza i social media, in linea con i recenti *framework* di *web-intelligence* del Sistema Statistico Europeo (ESS-net). È stata così prodotta un'analisi del *sentiment*, utilizzando i messaggi da *Twitter-X, Instagram e Facebook* (pagine pubbliche) e delle rassegne stampa online per capire come sono rappresentati la violenza di genere e gli stereotipi di genere. Inoltre, lo studio ha sperimentato anche metodi innovativi volti a catturare nuove forme di violenza di genere e le sue evoluzioni attraverso i *social media*, al fine di monitorarne le diverse forme digitali, come la cyber-violenza e il cyber-bullismo. Più in profondità, il metodo statistico adottato ha l'obiettivo di utilizzare *i Big Data* come fonti di dati affidabili e robuste, a complemento dell'indagine periodica sulla violenza contro le donne e sull'immagine sociale della violenza sessuale. Uno dei problemi principali, infatti, è che il processo che genera questi dati è sconosciuto e probabilmente selettivo rispetto alla popolazione target. I recenti lavori di Tut-Prats (2019), Gonzalez & Rodrıguez-Planas (2020) e Alesina et al. (2020), documentano il ruolo prominente delle norme culturali e sociali, che si riferiscono principalmente agli stereotipi sessuali e al ruolo della donna nella società. La forte correlazione tra stereotipi di genere e livello di accettazione della violenza di genere in Italia, studiata anche durante il periodo della pandemia, spinge a proseguire su questa strada di analisi. Prima di procedere nella illustrazione della metodologia adottata, occorre tuttavia fornire alcuni elementi di conoscenza che introducono il lettore alle tecniche di analisi del *sentiment* e all'utilizzo del *Natural Learning Process (NLP).*

2. La *sentiment analysis*

La *sentiment analysis (SA),* definita come lo studio delle opinioni e dei sentimenti espressi da dati testuali (Liu, 2012), è una tecnica in rapida crescita nell'ambito della ricerca del NLP, grazie anche all'ampia gamma di applicazioni effettuate in diversi campi di analisi. Le

tecniche di SA comprendono sia regole e metodi relativamente semplici sia procedure avanzate di *deep learning* (per una rassegna dettagliata si veda: Liu 2012; Medhat, Hassan e Korashy 2014).

Le diverse tecniche di SA possono essere classificate secondo due approcci principali:

- Le tecniche basate sulla conoscenza (o approccio basato sul lessico), che consistono nell'uso di alcune risorse lessicali, come dizionari ontologici o *corpus* annotati, per classificare le opinioni contenute nei documenti. Il principio di queste tecniche consiste nel determinare il *sentiment* di un documento in base alla somma delle polarità di ogni parola (indicata nel dizionario ontologico) contenuta nel testo. In base alla loro polarità, quindi, i termini sono contrassegnati con un peso negativo, positivo o neutro che, sommato alla polarità di tutti i termini del documento, fornisce l'orientamento generale del testo (Rudkowsky et al 2018).
- L'approccio *Machine Learning (ML),* che utilizza metodi di classificazione del testo basati su algoritmi di apprendimento automatico. La fase di apprendimento può essere di tipo “supervisionato” o “non supervisionato”. Nel primo caso, l'algoritmo viene addestrato attraverso un insieme di dati precedentemente classificati ed etichettati (*training set*). Nel secondo caso, l'algoritmo apprende dai dati senza informazioni aggiuntive (Yadav et al 2020; Alarifi et al 2018; Araque et al 2017; Bonadiman et al 2017; Su et al 2017; Tripathy et al 2015). Considerando i pro e i contro dei metodi di ML supervisionati e non supervisionati, il vantaggio dell'apprendimento supervisionato si basa principalmente sulla minore complessità del modello e sulla maggiore accuratezza dei risultati dovuta alla combinazione uomo-macchina, rispetto all'apprendimento non supervisionato che genera risultati affidabili ma complessivamente più poveri (Ceron et al 2014). Per questo motivo per la produzione del report sono state adottate tecniche di apprendimento supervisionato.

Negli ultimi anni, come conseguenza della crescente importanza e diffusione dei *social media*, il campo di ricerca sull'*opinion mining* ha registrato uno sviluppo rilevante. Infatti, l'aumento degli studi empirici e metodologici basati sulla *sentiment analysis* è chiaramente evidente (Maynard & Funk 2011; Agarwal et al 2011; Alicante et al 2016; Kumar & Jaiswal 2019; Mencarini et al 2019; Gagliardi et al 2020). Le ricerche, indipendentemente dall'approccio utilizzato, sottolineano direttamente o indirettamente la necessità di valutare la specificità dei contenuti pubblicati sulle piattaforme di *social network.* Infatti, per condurre una buona analisi è opportuno conoscere e considerare le peculiarità della specifica tipologia di testi utilizzati.

Il primo elemento da considerare nell'analisi di questi contenuti è che gli utenti di *Twitter-X* postano messaggi da molti dispositivi diversi (*smartphone*, PC, *tablet*, ecc.), quindi la frequenza di errori ortografici e l'uso di *slang* è molto più alta che in altri contesti comunicativi. In secondo luogo, i testi sono spesso accompagnati da altri elementi come *link*, menzioni (*tag* di altri utenti), *hashtag* ed *emoji*. I suddetti punti di criticità richiedono sia una pulizia e un pre-trattamento del testo particolarmente accurati, sia precise scelte metodologiche rispetto all'analisi o all'esclusione di alcuni elementi dal testo. A questo proposito, nel report realizzato, considerando le difficoltà nella gestione concettuale e operativa di questi elementi, i dati "extra-testuali" vengono esclusi dal testo per limitare il "rumore" da essi generato.

Bright e al. (2014) sostengono che è necessaria una certa cautela nell'interpretare i dati dei social media e che rimangono importanti interrogativi su come impiegare correttamente tali dati.
Non esiste un disegno di campionamento randomizzato che faciliti la generalizzazione delle conclusioni e dei risultati ottenuti con i dati disponibili a una popolazione target più ampia. Di conseguenza, estrarre informazioni statisticamente rilevanti da queste fonti è un compito impegnativo (Daas e Puts, 2014). È importante sottolineare che in questa sede ci si concentra sulla descrizione del metodo adottato per valorizzare i contenuti dei social media, ma non affronta la questione cruciale relativa al campione statistico rappresentativo di questi contenuti. Infatti, anche se non sono possibili collegamenti tra i *Big Data* e altre fonti, l'idea di questa statistica è quella di trovare un metodo che possa essere utilizzato a fini di comparazione. Se sono disponibili più istanze del set di dati nel tempo, esiste ancora la possibilità di combinare le fonti di dati attraverso un approccio basato sulle serie temporali. La correlazione temporale tra le serie può essere utilizzata per migliorare l'accuratezza del lancio o della previsione. Come infatti si evince dai dati, esiste una comparazione evidente tra il traffico registrato sui social attraverso questa analisi e le chiamate di aiuto delle vittime di violenza al numero telefonico 1522. Van den Brakel et al. (2017) migliorano l'accuratezza delle stime basate sulle indagini attraverso un approccio di modellazione strutturale delle serie temporali in cui una serie temporale di *Big Data* viene utilizzata come serie di covariate indipendenti. Le serie temporali di *Big Data* derivano dai messaggi dei social media e riflettono il *sentiment* nel testo dei messaggi. Sebbene i messaggi (come quelli di *X-Twitter*) non possano essere collegati agli individui e non possano essere aggregati a un livello sufficientemente dettagliato, questo approccio consente di sfruttare la correlazione temporale.

3. Il processo di *machine learning* di tipo supervisionato

In accordo con gli obiettivi della statistica, i ricercatori hanno condotto un processo di *machine learning*, al fine di verificare la solidità della metodologia di *tagging* e *testing* e la robustezza statistica dei primi risultati ottenuti. Utilizzando algoritmi di *Natural Language Processing*, attraverso la codifica del testo di *Twitter-X*, la metodologia si articola in tre fasi di lavoro:

1. Fase 1 - Definizione dei requisiti dello studio e dei criteri di estrazione, elaborazione e assemblaggio del corpus di annotazioni.
2. Fase 2 - Preparazione del *dataset.*
3. Fase 3 - Identificazione dei modelli di classificazione addestrati.

Al termine di tali fasi l'algoritmo è stato preparato per analizzare altri messaggi dei social media, come quelli previsti da *Facebook* e *Instagram* (pagine pubbliche), canale *YouTube*, forum pubblici (Web).

1. *Fase 1 - Definizione dei requisiti dello studio e dei criteri di estrazione, elaborazione e assemblaggio del corpus di annotazioni*

Si è proceduto con la definizione delle aree, dei temi, degli argomenti da raccogliere in termini di parole chiave e *hashtag*. I temi da monitorare riguardavano le conversazioni sul tema "Violenza di genere" e sul tema "Stereotipi di genere".

Una specifica piattaforma (IRIDE) ha consentito di captare le conversazioni pubbliche prodotte dagli utenti sui canali social e web sulla base di un set di parole chiave. Queste ultime hanno permesso di creare filtri specifici per l'estrazione dei contenuti di interesse dalle fonti identificate[1]. La scelta delle parole chiave è fondamentale per definire il perimetro della ricerca, definendo e guidando l'ambito tematico di monitoraggio e analisi.

Una volta elaborati, i testi sono stati sottoposti a un'ulteriore selezione volta a escludere i *tweet* troppo simili tra loro. La somiglianza è definita utilizzando come metrica il coseno di somiglianza, applicato a tutte le coppie di *tweet* acquisite ed elaborate a valle della loro proiezione in uno spazio vettoriale attraverso un approccio "*Bag-of-words*" (Qader Wisam A. et al., An Overview of Bag of Words; Importance, Implementation, Applications and Challenges, 2019).

Una volta raccolti i dati di interesse, il *dataset* è stato composto scegliendo le frasi che presentano il massimo grado di eterogeneità. In questo modo, il campione risulta rappresentativo della popolazione da cui è stato estratto. Il *dataset* di annotazioni consiste in un campione di *tweet* acquisiti dai connettori di *Twitter-X*. I *tweet* che sono stati campionati erano compresi tra quelli in lingua italiana pubblicati nel periodo 01/06/2020 - 30/09/2020.

2. *Fase 2 - Preparazione del dataset*

Il corpus di *tweet* estratto è stato suddiviso e ripetuto su più file, opportunamente formattati per facilitare il processo di annotazione: la ripetizione dei *tweet* ha permesso una lettura multipla dello stesso *tweet* da parte di diversi annotatori, in modo da consentire un meccanismo di preassegnazione a maggioranza.
Per ogni modello di classificazione sono stati fissati obiettivi specifici, che ne hanno determinato intrinsecamente la natura. La definizione di questi obiettivi è stata decisiva per l'organizzazione del processo di annotazione. In effetti, le caratteristiche fondamentali del modello di classificazione dei *tweet* sono state determinate sulla base dei seguenti elementi discriminanti:

1. **il contenuto dei *tweet*:** notizie o opinioni (info/no info). Non sempre il *tweet* contiene opinioni o esperienze personali, a volte il contenuto di un *tweet* rappresenta un'informazione oggettiva (riferimenti a notizie, contenuti promozionali o informativi, ecc.); in questo caso, è stato utilizzato un classificatore binario per identificare questo tipo di contenuto. In termini di disambiguazione, il contenuto contenente un sentimento o un'emozione dovrebbe essere incluso se si tratta, ad esempio, di una comunicazione giornalistica. Ad esempio: "Il 10 marzo durante un comizio il politico MARIO ROSSI ha detto di odiare tutte le donne" viene classificato come "INFO" perché si tratta di una notizia giornalistica anche se contiene un sentimento che viene

[1] Le parole chiave possono essere soggette a modifiche e aggiunte periodiche durante l'erogazione, al fine di creare un flusso di dati coerente e inerente all'oggetto dell'indagine.

riportato nella notizia stessa. Il giornalista (o chi ha postato il *tweet*), non ha espresso un sentimento o un'emozione ma ha solo registrato un evento. Alternativamente la frase che MARIO ROSSI riporta sui social è “odio le donne” è considerata no-info.

2. **Cardinalità dell'annotazione**: questo elemento ha riguardato quante classi potevano essere associate allo stesso *tweet*. Per classificare il *sentiment* le classi sono 3 (Positivo, Negativo e Neutro), per le emozioni la cardinalità è 7 (Amore, Gioia, Sorpresa, Rabbia, Tristezza, Paura, Neutro), il contenuto informativo dei *tweet* (Sì, No);
3. **Descrizione del *dataset***: è stato fatto un calcolo di rilevamento dell'oggettività in fase di *training* sull’argomento. Il *dataset* era composto da 376 osservazioni perfettamente uguali, distribuite tra le classi INFO (188) e NO INFO (188). Il *dataset* è stato suddiviso tra *training*, test e validazione nelle seguenti proporzioni 70%, 20%, 10%;
4. ***Sentiment analysis*:** Il *dataset* era composto da 1.025 osservazioni così distribuite tra le classi POS (346 casi), NEG (380) e NEU (299). Il *dataset* è stato suddiviso in *train* (70%), test (20%) e *validation* (10%), mediante campionamento stratificato al fine di mantenere la distribuzione delle osservazioni tra le varie classi di polarità;
5. **Rilevamento delle emozioni**: Il *dataset* si componeva di 2.727 osservazioni distribuite tra le classi come segue: Amore (107), Gioia (294), Neutro (749), Paura (394), Rabbia (518), Sorpresa (231), Tristezza (434). Il *dataset* è stato suddiviso tra addestramento, test e validazione nelle seguenti proporzioni 70%, 20%, 10%;

Valutazione della qualità dell'annotazione

Per valutare la qualità dell'annotazione prodotta dagli annotatori Istat, che hanno lavorato insieme per etichettare il *dataset* di *training* e di test, si è deciso di utilizzare *l'IRA e il Fleiss Kappa,* indici che indicano rispettivamente il grado di accordo tra un gruppo di annotatori e l'accordo raggiunto dal gruppo di annotatori rispetto a quello che si otterrebbe se il gruppo di annotatori annotasse casualmente le frasi che compongono il *dataset*.

In accordo con la letteratura sulla bontà dell'annotazione, il campione da annotare è stato etichettato nuovamente finché il valore dell'indice *Fleiss Kappa* non ha raggiunto il valore minimo di 0,8 e il valore dell'indice *IRA* ha raggiunto il valore minimo di 0,6 (*Interrater reliability, kappa statistics*).

Soddisfatti del grado di accordo tra gli annotatori, si è proceduto all'addestramento degli algoritmi descritti di seguito.

3. Fase 3 - Identificazione dei modelli di classificazione addestrati.

La base di ogni algoritmo di elaborazione del linguaggio naturale (NLP) è il *word embedding*, cioè la rappresentazione vettoriale delle parole. Esistono diversi algoritmi che permettono di calcolare i vettori che rappresenteranno le parole negli algoritmi NLP. Alcuni di essi sono *Word2Vec (Efficient Estimation of Word Representations in Vector Space*, Mikolov et al, 2013), *Glove (GloVe: Global Vectors for Word Representation*, Pennington et al, 2014), *FastText (Bag of Tricks for Efficient Text Classification*, Joulin et al, 2016).

Il problema di tutti gli algoritmi citati è che essi forniscono un incorporamento di parola unico per ogni parola e quindi non dipendente dal contesto in cui viene utilizzata.
Se utilizziamo un algoritmo NLP per analizzare le frasi "Sono una donna rosa" e "Il mio colore preferito è il rosa", la parola "rosa" sarà rappresentata dallo stesso vettore nonostante il significato in cui viene utilizzata sia molto diverso nelle due frasi.

Per risolvere questo problema, i modelli di classificazione implementati si basano su un algoritmo chiamato *Bidirectional Encoder Representations from Transformers (BERT*) (BERT: *Pre-training of Deep Bidirectional Transformers for Language Understanding* - Devlin et al. 2018) proposto dai ricercatori di Google. Grazie a questo algoritmo ogni volta che una parola ambigua come “rosa” figura in frasi diverse avrà una rappresentazione vettoriale che permetterà al modello di comprendere il significato di quella parola nel contesto.

Il meccanismo che rende potente questa categoria di algoritmi si chiama *Attention (Attention is All You Need* - Vaswani et al. 2017). Lo stesso motore di ricerca Google si basa su un algoritmo di questo tipo.

Sul modello linguistico pre-addestrato da BERT, è stato collegato uno strato neurale *feed forward*. Questa architettura ha permesso di sfruttare la potenza dell'algoritmo BERT pre-addestrato. Ogni modello, quindi, durante la fase di addestramento è stato messo a punto per assolvere al compito di analisi (*Sentiment*, *Emotion* e Info) nel modo più efficiente possibile.

Valutazione del modello

Per valutare la bontà dell'addestramento, i modelli sono stati valutati sul dataset di test osservando:

- osservazioni classificate positivamente ed effettivamente positive (TP);
- osservazioni classificate negativamente ed effettivamente positive (FN);
- osservazioni classificate positivamente ed effettivamente negative (FP);
- osservazioni classificate negativamente ed effettivamente negative (TN).

Sono stati quindi calcolati due indici di valutazione dei modelli:

- *l’Accuracy* che indica la percentuale delle previsioni corrette rispetto al totale; questo indice varia tra 0 e 1.
- *l’F1-Score* che costituisce la media del punteggio F1 calcolato sulle singole classi. Questo indice varia tra 0 e 1. Il punteggio F1 calcolato su una singola classe è uguale alla media armonica di *Precision e Recall* calcolati sulla classe stessa;

$$F1 = \frac{2}{\frac{1}{r} + \frac{1}{p}} = 2\,\frac{\mathrm{p}\,.\,\mathrm{r}}{\mathrm{p} + \mathrm{r}}$$

La precisione è pari al rapporto tra il numero di *True positives* e la somma dei *True positives* e dei *False positives*;

$$p = \frac{True\ positives}{True\ positives + False\ positives}$$

La *Recall* è pari al rapporto tra il numero di *True positives* e la somma di *True positives* e *False negatives.*

$$r = \frac{True\ positives}{True\ positives + False\ negatives}$$

Di seguito sono riportate le statistiche relative agli indicatori sopra descritti per i modelli di *sentiment analysis*, *emotion detection* e rilevamento del tipo di informazioni sul contenuto.

Sentiment analysis: **indici di qualità per la valutazione del modello**

Tavola 1 - Valori di Accuracy e F1 Score ottenuti per la valutazione del modello

Accuracy	0.71
F1-score	0.7

Tavola 2 - Indici di qualità per classe del modello di Sentiment Analysis

	Precision	***Recall***	***F1***
NEG	0.729	0.897	0.81
NEU	0.645	0.513	0.57
POS	0.711	0.692	0.7

Emotion detection: **indici di qualità per la valutazione del modello**

Tavola 3 - Indici di qualità della classificazione complessiva dell'Emotion Detection

Accuracy	0.82
F1-score	0.79

Tavola 4 - Indici di qualità per classe del rilevamento delle emozioni

	Precision	*Recall*	*F1*
Tristezza	0.717	0.874	0.788
Paura	0.925	0.785	0.85
Neutro	0.944	0.9	0.92
Rabbia	0.821	0.837	0.83
Gioia	0.688	0.746	0.72
Sorpresa	0.784	0.63	0.699
Amore	0.696	0.762	0.73

Tipo di informazioni sul contenuto rilevato (info/no info)

Tavola 5 - Indici di qualità della classificazione complessiva della rilevazione dell'"oggettività"

Accuracy	0.95
F1-score	0.95

Tavola 6 - Indici di qualità per classe del rilevamento dell'"oggettività"

	Precision	*Recall*	*F1-score*
info	0.95	0.95	0.95
no info	0.95	0.95	0.95

Il modello conferma l'ipotesi, espressa all'inizio, dell'importanza di tenere conto di diversi tipi di testo, quelli che si riferiscono a informazioni/notizie e quelli che trattano esclusivamente opinioni.

Il modello di rilevazione del "tipo di contenuto" è una parte fondamentale della *pipeline* di analisi del tema della violenza di genere, in quanto permette di selezionare contenuti testuali, non di natura puramente informativa, sui quali ha senso estrarre *sentiment* ed emozioni.

Una volta che i modelli sono stati adeguatamente addestrati, essi sono stati utilizzati per assegnare automaticamente una classe a ciascun contenuto testuale elaborato. Vale a dire che sono stati classificati i *post* e i commenti sui quali il modello non è stato addestrato nella fase di *training*. A ogni *record* valutato dai modelli è stato associato anche un indice di

entropia di classificazione, per un valore compreso tra 0 e 1, calcolato in funzione della distribuzione di probabilità assegnata alle singole classi. La formula utilizzata per calcolare l'entropia è:

$$-\sum_{k=1}^{n} p_k * \log_n(p_k)$$

dove n è il numero di classi e p_k i è la probabilità assegnata alla k-esima classe dal modello.

Se il punteggio supera 0,95, la classe assegnata è imprevedibile.

Questo *KPI* ha permesso di distinguere tra i *record* classificati con un livello di confidenza sufficiente e quelli non classificati e di valutarli in modo appropriato.

L'analisi basata sulle evidenze dei contenuti *social* raccolti, evidenzia l'importanza dei *social media* per comprendere meglio come l'opinione pubblica reagisce su alcuni argomenti e consente di rilevare alcune questioni emergenti. In particolare:

- permette di effettuare alcuni focus su temi rilevanti legati alla violenza di genere come il femminicidio, il *body shaming*, lo stupro e la violenza sessuale (su cui si proporranno in seguito specifiche note);

- consente di seguire le campagne di sensibilizzazione predisposte per il contrasto al fenomeno della violenza di genere e lo *stalking* (come, ad esempio, quella legata alla linea telefonica 1522).

Inoltre, oltre all'analisi di alcuni specifici "picchi" che guidano i *social*, la metodologia adottata consente di introdurre una nuova categoria di analisi: il linguaggio violento e la consapevolezza. In particolare, i contenuti *social* sono stati riclassificati in due polarità; un gruppo di parole ed espressioni complessivamente legate al sentimento di "odio" e un secondo composto da un insieme di parole ed espressioni legate al sentimento di "indignazione". Ciò aumenta la capacità interpretativa osservando se il canale *social* rafforzi maggiormente i contenuti violenti e il linguaggio di odio con cui si commentano fatti o giudizi, oppure se emerga la consapevolezza e la "indignazione" dell'opinione pubblica sugli argomenti che riguardano la violenza e lo stereotipo di genere.

Un'altra traccia importante per l'analisi è la complessità della discussione sulle questioni di genere, come già emerge dai primi risultati, che mostrano quali emozioni si associano ai contenuti dei post che esprimono indignazione.

Infine, l'impiego come base di algoritmi di classificazione di modelli linguistici pre-addestrati consente di sfruttare indirettamente risorse, in termini di potenza computazionale, che altrimenti non sarebbe stato possibile addestrare da soli a causa della scarsità di risorse e di dati. Tuttavia, gli algoritmi oggetto di studio, addestrati su *corpora* testuali come *Wikipedia*, sono probabilmente affetti da pregiudizi di genere, in quanto riflettono i pregiudizi presenti nella stessa raccolta di testi utilizzata per l'addestramento. Per superare questi pregiudizi

sistematici, può essere utile la proposta di costruire un dizionario italiano sensibile dedicato alla questione del genere. Un classico esempio di possibile *bias* è la maggiore predisposizione da parte dei modelli linguistici, addestrati su questi corpora, ad associare professioni come medico con una desinenza maschile e professioni come insegnante o infermiera con una desinenza femminile. È bene considerare, tuttavia, che questi limiti non dipendono dall'architettura dell'algoritmo, ma dalla qualità del testo in ingresso.

Una sfida per il futuro sarà quindi quella di mitigare queste distorsioni attraverso l'applicazione di strumenti in grado di misurare la correttezza dell'algoritmo una volta addestrato, nonché di selezionare per l'addestramento testi che non presentino distorsioni.

4. L'utilizzo della *Latent Dirichlet Allocation per la topic analisys*

La *Latent Dirichlet Allocation (LDA*) (Blei et al., 2003) è uno dei modelli bayesiani che viene utilizzato nel *natural language processing* per implementare il *topic modelling* (Vayansky et al., 2020; Kherwa et al., 2019; Blei et al., 2009). È un modello statistico utilizzato per l'analisi di temi (o argomenti) nascosti (o latenti) in un insieme di documenti. Si tratta di un algoritmo di apprendimento non supervisionato, che permette di scoprire in modo automatico i temi principali presenti in un corpus o collezione di testi. Nello specifico è un modello probabilistico generativo che tramite una distribuzione probabilistica considera ogni documento come una combinazione di temi (o *topic*), dove ogni tema è a sua volta una distribuzione di probabilità su parole. Il modello assume che ci siano K argomenti (il cui numero deve essere predefinito) e che ogni documento sia generato da una mistura di questi argomenti.
Le distribuzioni di parole che costituiscono i singoli *topic* e le distribuzioni di *topic* che costituiscono i singoli documenti sono modellizzate usando distribuzioni *prior* di *Dirichlet*. Una distribuzione di Dirichlet a priori K dimensionale ha una densità di probabilità data da:

$$f(p_1, \ldots, p_K \mid \alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_{i=1}^{K} p_i^{\alpha_i - 1}$$

Dove Γ rappresenta la funzione gamma. Nel caso della LDA viene solitamente utilizzata una distribuzione Dirichlet simmetrica, in cui il vettore di parametri α è uguale per tutti i K elementi (α=1/k), dove K sono il numero di *topic* latenti.

Le probabilità di appartenenza delle parole ai diversi *topic* si basano sulla co-occorrenza dei termini all'interno dei singoli documenti.

Il parametro della Dirichlet deve essere definito dall'utente sia per la distribuzione dei *topic* all'interno dei documenti che per la distribuzione delle parole all'interno dei *topic*.

La Document-Term Matrix (DTM) è una rappresentazione tabellare utilizzata per descrivere la frequenza delle parole in un insieme di documenti ed è l'input per l'LDA. È uno dei formati più comuni per preparare i dati di testo per analisi computazionali in cui le righe

rappresentano i documenti del corpus, mentre le colonne rappresentano i termini (parole uniche) presenti nel corpus. I valori della matrice indicano la frequenza di un termine specifico in un documento.

La DTM è generalmente ad alta dimensionalità e sparsa, quindi richiede degli step di pre-processing, come la rimozione di stopwords (es. "e", "di", "un"), lemmatizzazione (riduzione delle parole alle loro radici) e riduzione della dimensionalità, rimuovendo le rare occorrenze.

L'LDA utilizza la DTM per scoprire pattern latenti di co-occorrenza tra parole, modellando documenti e temi. La LDA trasforma la *DTM* in due sotto-matrici, ovvero la matrice *document-topic* e la matrice *topic-term.*

Come obiettivo, l'algoritmo cerca la rappresentazione ottimale delle due matrici, al fine di trovare le migliori distribuzioni di *topic* nei documenti e di parole nei *topic*. A partire dalle due assunzioni di cui sopra, tramite un processo di *backtracking*, la LDA riesce a identificare quali *topic* generano determinati documenti e quali parole generano i relativi *topic*.

Nella presente nota, dato che le collezioni di testi analizzate erano molto dissimili per struttura a seconda dei diversi canali social, al fine di ridurre la dimensionalità della DTM si è reso necessario analizzare ogni fonte separatamente. Infatti le caratteristiche dei corpus (o collezioni di testo) differiscono profondamente: X è caratterizzato da testi molto brevi, mentre Facebook (FB) e Instagram da testi molto più lunghi, tuttavia il secondo contiene moltissimi hashtag e testi in lingua inglese e quindi il dizionario (ovvero l'insieme delle parole) è molto dissimile.

Nel presente studio si è proceduto ad una lemmatizzazione che consentisse di mantenere il genere invariato, ovvero: ragazze diviene ragazza, ragazzi diventa ragazzo, violentata diviene violentare. Si è proceduto ad una rimozione di stop-words personalizzato a partire dalla liberia spacy di Python. Per quanto concerne la parte di text-cleaning si è provveduto a rimuovere i caratteri alfa-numerici ad eccezione degli #hashtags. Le parole con le minori occorrenze sono state eliminate (50 su X e FB, 25 su Instagram), lasciando un dizionario di circa 10mila parole per X e FB, circa 6000 per Instagram.

L'orizzonte temporale analizzato è 1/12/22 -31/8/24.

Per ogni analisi LDA, è necessario definire 3 parametri: il numero dei *topic*; il coefficiente di ciascuna delle due Dirichlet.

Esistono diverse metriche rispetto alle quali è possibile selezionare il numero ottimale dei *topic* (K). Nella presente analisi cono state analizzate le metriche; C_V, basato su co-occorrenze nei documenti e C_NPMI che utilizza l'entropia di mutua informazione (PMI) per valutare la correlazione tra parole. Tuttavia è l'interpretazione tematica quella che ha prevalso nella scelta del numero di *topic* definitivo. Per quanto riguarda la distribuzione delle

parole all'interno dei *topic* si è scelta la distribuzione uniforme (α= 1/k), mentre per la distribuzione dei *topic* all'interno del corpus per quanto riguarda X, Instagram e Facebook si è scelta una distribuzione (α del primo cluster = 1/K^1/2). Tale scelta è derivata per X dal fatto che la natura virale con pochi eventi rende i cluster sbilanciati tra di loro, mentre nel caso di FB ci si trova di fronte a lunghezze di testi molto variabili, infine su Instagram data la presenza di cluster sia in italiano che inglese li rende sbilanciati.

Le librerie *Python* utilizzate per effettuare l'analisi e la visualizzazione sono, rispettivamente, "*Gensim: Topic Modelling for Humans* (4.3.0)" (Řehůřek et al., 2011) e "pyLDAvis (3.4.1)" (Sievert et al., 2014). La libreria utilizzata pyLDAvis stima la dimensione di un cluster (tema) a partire da un sotto-insieme delle parole della DTM, pertanto ha un margine di errore di circa 0.2%.

Bibliografia

Abdulaziz A., Amr T., Zafer A.M., Wael S. (2018), *A big data approach to sentiment analysis using greedy feature selection with cat swarm optimization-based long short-term memory neural networks*, in "The Journal of Supercomputing", 76(6), pp. 4414-4429.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R. (2011), *Sentiment Analysis of Twitter Data, in: Proceedings of the Workshop on Language in Social Media* (LSM 2011), pp. 30–38, Association for Computational Linguistics, Portland.

Alicante, A., Corazza, A., Pironti, A. (2016), *Twitter Sentiment Polarity Classification using Barrier Features, in: Proceedings of the Third Italian Conference on Computational Linguistics* (CLiC-it 2016), pp. 34-39, Vol. 1749, Academia University Press, Torino, http://ceur-ws.org/Vol-1749/

Araque, O., Corcuera-Platas, I., Sánchez-Rada, F., Iglesias, C. (2017), *Enhancing deep learning sentiment analysis with ensemble techniques in social applications*, in "Expert Systems With Applications", 77, pp. 236–246, Elsevier Ltd, http://dx.doi.org/10.1016/j.eswa.2017.02.002.

Battisti N., Dolcetti F., *Emozioni e testo: costruzione di risorse per il tagging automatico*, in: "JADT 2012: 11es Journées internationales d'Analyse statistique des Données Textuelle", pp. 95-107.

Bergvall, V. (1999). *Toward a comprehensive theory of language and gender. Language in Society, 28*(2), 273-293. http://doi.org/10.1017/s0047404599002080
Retrieved from: https://digitalcommons.mtu.edu/michigantech-p/7752

Bing L. and Ian L., (2016). *Attention-based recurrent neural network models for joint intent detection and slot filling*. In Interspeech 2016, pages 685–689.

Blei, D. M., & Lafferty, J. D. (2009). *Topic models. Text mining: classification, clustering, and applications*, 10(71), 34.

Blei, D.M., A.Y. Ng, and M.I. Jordan. (2003). *Latent Dirichlet Allocation*, in Journal of Machine Learning Research, Volume 3 (Jan.): 993-1022

Bolasco S., (2013) *Analisi multidimensionale dei dati. Metodi, strategie e criteri d'interpretazione*, Carocci Editore.

Bonadiman D., Castellucci G., Favalli A., Romagnoli R. and Moschitti A. (2017), *Neural Sentiment Analysis for a Real-World Application, in: Proceedings of the Third Italian Conference on Computational Linguistics* (CLiC-it 2017), pp. 42-47, Academia University Press, Torino.

Bonadiman D., Castellucci G., Favalli A., Romagnoli R. and Moschitti A., (2017) *Neural Sentiment Analysis for a Real-World Application*, DOI: 10.4000/books.aaccademia.2357, Accademia University Press, Torino

Caliskan A., Parth Ajay P., Charlesworth T., Wolfe R., and R. Banaji M., (2022). *Gender Bias in Word Embeddings: A Comprehensive Analysis of Frequency, Syntax, and Semantics*. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society.

Daas P.J, Puts M.J., (2014), *Social Media Sentiment and Consumer confidence*. Series N.5. European Central Bank Statistics Paper, Frankfurt.

Daas P.J., Roos M., Van den Ven M., Neroni J., (2012), *Twitter as a potential Data Source for Statistics*. The Hague/ Herlen

De Waal T., van Delden A., Shoultus S., (2017), *Multisources Statistics: basic situations and methods*, The Hague /Helen.

Drakett J., Rickett, B., Day, K., & Milnes, K. (2018). *Old jokes, new media – Online sexism and constructions of gender in Internet memes*. Feminism & Psychology, 28(1), 109–127.

Fillmore C. J., 1985. Frames and the semantics of understanding. Quaderni di Semantica, 6(2):222–254.

Forest D. & Meunier, J. (2005). *NUMEXCO: A Text Mining Approach to Thematic Analysis of a Philosophical Corpus*. Digital Studies/Le champ numérique. 10.16995/dscn.247.

Gagliardi, G., Gregori, L., Suozi, A. (2020), *L'impatto emotivo della comunicazione istituzionale durante la pandemia di Covid-19: uno studio di Twitter Sentiment Analysis*, in: Proceedings of the Seventth Italian Conference on Computational Linguistics (CLiC-it 2020), pp. 205-210, Accademia University Press, Torino.

Hearst M. (1999). *Untangling Text Data Mining. Proc of ACL'99: The 37th Annual Meeting of the Association for Computational Linguistics*, University of Maryland, June 20–26 https://www.researchgate.net/publication/326928518_A_Survey_on_Sentiment_and_Emotion_Analysis_for_Computational_Literary_Studies

Ike Vayansky, Sathish A.P. Kumar (2020), *A review of topic modelling methods*, in "Information Systems", Volume 94, ISSN 0306-4379

Kherwa, P., & Bansal, P. (2019). *Topic modeling: a comprehensive review. EAI Endorsed transactions on scalable information systems*, 7(24).

Kim E. and Klinger R.,(2018), *A Survey on Sentiment and Emotion Analysis for Computational Literary Studies,*

Kiritchenko S., Zhu X. and S. M. Mohammad. (2014). *Sentiment analysis of short informal texts.* Journal of Artificial Intelligence Research, 50(1):723–762, May Kumar & Jaiswal 2019, Systematic literature review of sentiment analysis on Twitter using soft computing techniques, "Currency and Computation. Practice and Experience", 32:e5107, John Wiley & Sons, Ltd, DOI: 10.1002/cpe.5107.

Liu B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions.* Cambridge: Cambridge University Press

Liu J., Sarkar, K. M. & Chakraborty, G (2013). *Feature-based Sentiment Analysis on Android App Reviews Using SAS® Text Miner and SAS® Sentiment Analysis Studio*. Proceedings of the SAS® Global 2013 Conference. Available at http://support.sas.com/resources/papers/proceedings13/250-2013.pd

Liu, B. (2012), *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, https://doi.org/10.2200/S00416ED1V01Y201204HLT016.
McHugh M. (2012), *Interrater reliability: the kappa statistic*, Biochem Med (Zagreb). 2012;22(3):276-82.

Morrone A. (1993), *Alcuni criteri di valutazione della significatività dei segmenti ripetuti*, in Anastex S. J. (ed.), JADT93 - Actes des secondes Journées Internationales d'Analyse Statistique de Données Textuelles, ENS-Telecom, Paris, 445-453

Patone M., Zhang Li-Chun (2020), *On two existing approaches to statistical analysis of social media data*, International Statistical Review, doi 10.1111/insr.12404

Qader Wisam A. (2019), *An Overview of Bag of Words; Importance, Implementation, Applications and Challenges*, Fifth International Engineering Conference on Developments in Civil & Computer Engineering Applications 2019 - (IEC2019) - Erbil – IRAQ

Quanzeng You, (2016), *Sentiment and Emotion Analysis for Social Multimedia: Methodologies and Applications*, MM '16: Proceedings of the 24th ACM international conference on MultimediaOctober 2016 Pages 1445–1449 https://doi.org/10.1145/2964284.2971475

Řehůřek, Radim & Sojka, Petr. (2010). *Software Framework for Topic Modelling with Large Corpora*. 45-50. 10.13140/2.1.2393.1847

Rosenthal S., Nakov P., Kiritchenko S., Mohammad S., Ritter A., and Stoyanov V., (2015) *Semeval-2015 task 10: Sentiment analysis in twitter.* In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 451–463, Denver, Colorado, June. Association for Computational Linguistics.

S.Omrani Sabbaghi, A.Caliskan. (2022) *Measuring Gender Bias in Word Embeddings of Gendered Languages Requires Disentangling Grammatical Gender Signals*,

Sievert, C., & Shirley, K. (2014, June). *LDAvis: A method for visualizing and interpreting topics*. In Proceedings of the workshop on interactive language learning, visualization, and interfaces (pp. 63-70).

Sitikhu P., Pahi K., Thapa P., Shakya S., *A Comparison of Semantic Similarity Methods for Maximum Human Interpretability,* In: 2019 artificial intelligence for transforming business and society (AITB). IEEE, 2019. p. 1-4.

Stephen T. (2000). *Concept analysis of gender, feminist, and women's studies research in the communication literature*. Communication Monographs, 67(2), 193–214.

Vaswani A. et al. (2017), *Attention is all you need*. In Advances in Neural Information Processing Systems, pages 6000–6010.

Yadav, A., Vishwakarma, D. K. (2020) *A Deep Language-independent Network to analyze the impact of COVID-19 on the World via Sentiment Analysis*, in "Arxive", Cornell University, https://arxiv.org/.