Sommario

[Abstract 1](#_Toc158475083)

[Dominio di applicazione 1](#_Toc158475084)

[Introduzione 1](#_Toc158475085)

[Stato dell’arte 2](#_Toc158475086)

[VADER 2](#_Toc158475087)

[Spacy 3](#_Toc158475088)

[Feel-It 4](#_Toc158475089)

[Differenze e scelta del modello 5](#_Toc158475090)

[Obiettivi perseguiti 5](#_Toc158475091)

[Obiettivi 5](#_Toc158475092)

[Dati 5](#_Toc158475093)

[Metodologia 6](#_Toc158475094)

[Estrazione di conversazioni 6](#_Toc158475095)

[Utilizzo di Feel It 6](#_Toc158475096)

[Conteggio di parole italiane e non 6](#_Toc158475097)

[Problematiche affrontate 6](#_Toc158475098)

[Risultati della sperimentazione 7](#_Toc158475099)

[Grafici dei risultati 7](#_Toc158475100)

[Analisi dei risultati 7](#_Toc158475101)

[Limitazioni e possibili lavori futuri 7](#_Toc158475102)

# Abstract

L'analisi del linguaggio naturale (NLP) rappresenta una risorsa cruciale per comprendere le complesse dinamiche delle intercettazioni criminali. Tuttavia, l'efficacia di tale analisi può essere compromessa da diverse sfide, tra cui la presenza del dialetto napoletano all'interno delle conversazioni e la limitata quantità di dati disponibili. In questo contesto, il presente studio si propone di utilizzare un modello di analisi del sentiment, Feel-It, addestrato su frasi in lingua italiana, per estrarre informazioni utili dalle intercettazioni criminali. Attraverso un'analisi accurata del testo delle conversazioni, si mira a comprendere lo stato emotivo dei soggetti coinvolti, le loro motivazioni e le relazioni intercorrenti. Tuttavia, le limitazioni legate alla varietà linguistica e alla qualità dei dati richiedono un approccio attento e multidisciplinare per massimizzare l'utilità delle informazioni ottenute. In conclusione, questo studio si propone di contribuire alla comprensione delle dinamiche criminali attraverso un'analisi avanzata del linguaggio naturale, offrendo insight preziosi per il lavoro investigativo e giuridico.

# Dominio di applicazione

## Introduzione E Stato dell’arte

L'analisi del sentiment rappresenta una disciplina fondamentale nell'ambito dell'elaborazione del linguaggio naturale (NLP), mirata a decifrare e interpretare le emozioni e le opinioni espresse nei testi scritti. Questo campo, in continua evoluzione, ha attraversato diverse fasi, dalla semplice assegnazione di polarità basata su dizionari di parole chiave alle sofisticate tecniche di machine learning che caratterizzano l'attuale panorama della sentiment analysis. Inizialmente, le prime metodologie di analisi del sentiment si basavano su regole linguistiche e liste di parole predefinite, assegnando a ciascuna parola una polarità predefinita.

Questi approcci, sebbene rudimentali, costituivano un primo tentativo di tradurre il linguaggio umano in un formato comprensibile per le macchine. Tuttavia, con l'avanzamento della capacità computazionale, il campo ha abbracciato con entusiasmo l'apprendimento automatico (machine learning), aprendo la strada a nuove prospettive. L'introduzione dei dizionari di sentimenti è stata una pietra miliare nell'evoluzione dell'analisi del sentiment. Modelli come il VADER Sentiment Analysis hanno implementato questo approccio, arricchendo i dizionari con informazioni sulla polarità e sull'intensità di ciascuna parola. Questa metodologia più sofisticata consente una valutazione più precisa, tenendo conto del contesto e delle sfumature linguistiche.

Il VADER, in particolare, ha dimostrato di essere un pilastro nell'analisi del sentiment, ispirando numerosi sviluppi e applicazioni in vari settori. Le applicazioni pratiche dell'analisi del sentiment sono ampie e trasversali. Nel marketing, le imprese sfruttano questa tecnologia per comprendere il feedback dei clienti, adattare le strategie di vendita e valutare l'accoglienza dei propri prodotti e servizi. Nei social media, l'analisi del sentiment è diventata uno strumento essenziale per monitorare l'opinione pubblica su questioni di rilievo, identificare trend emergenti e valutare la reputazione di brand e personaggi pubblici. Tuttavia, nonostante i progressi significativi, l'analisi del sentiment si trova ancora di fronte a diverse sfide. La comprensione di espressioni idiomatiche, la gestione delle sfumature culturali e il riconoscimento del contesto rimangono aree di miglioramento.

L'auspicio è che con l'impiego di modelli sempre più sofisticati e l'integrazione di approcci di machine learning avanzati, il campo possa superare queste sfide e offrire una comprensione ancora più profonda del linguaggio delle emozioni. In questa panoramica sull'analisi del sentiment, abbiamo delineato l'evoluzione di questa disciplina, dalle sue radici tradizionali fino alle attuali frontiere del machine learning. Nei prossimi paragrafi, ci immergeremo in due modelli specifici: il VADER Sentiment Analysis e il framework Spacy. Analizzeremo come questi modelli affrontano le sfide dell'analisi del sentiment, contribuendo a plasmare il modo in cui comprendiamo e interpretiamo il linguaggio delle emozioni nel contesto del NLP.

### VADER

Uno dei punti salienti nella storia dell'analisi del sentiment è l'avvento del VADER Sentiment Analysis, un modello concepito da Hutto e Gilbert nel 2014. Il nome VADER è un acronimo di Valence Aware Dictionary and sEntiment Reasoner, che sottolinea la sua duplice natura di dizionario di sentimenti e ragionatore di sentimenti.

Questo modello ha rivoluzionato l'analisi del sentiment attraverso l'implementazione di un approccio basato su dizionari e regole sofisticate. Il cuore del VADER è il suo vasto dizionario di sentimenti, nel quale ogni parola è associata a un punteggio sulla scala della polarità.

Questo dizionario, creato con attenzione per riflettere la complessità delle espressioni umane, è un elemento cruciale che consente al modello di valutare il sentiment complessivo di un testo.

A differenza di molti altri approcci, il VADER non si limita a classificare le parole come positive o negative; tiene anche conto dell'intensità del sentimento, catturando le sfumature del linguaggio in modo più accurato. Le regole di punteggio del VADER giocano un ruolo fondamentale nel determinare il sentiment complessivo di una frase. Ad esempio, il modello è in grado di rilevare negazioni e altre costruzioni linguistiche che possono invertire il significato di una parola.

Inoltre, le emoticon, spesso utilizzate nelle comunicazioni online, vengono integrate nel processo di analisi, aggiungendo ulteriori dettagli alla comprensione del sentiment complessivo. Le applicazioni pratiche del VADER Sentiment Analysis sono diversificate e ampiamente adottate in vari settori. Nel marketing, le aziende utilizzano il modello per valutare la reazione dei clienti ai loro prodotti e servizi. Nei social media, il VADER è impiegato per monitorare il sentiment pubblico su determinati argomenti, identificare tendenze emergenti e valutare la percezione di un brand o di un'organizzazione. L'impatto del VADER nella ricerca accademica e nell'industria è innegabile. Il modello ha ispirato lo sviluppo di nuovi approcci nell'analisi del sentiment, spingendo la comunità scientifica a esplorare ulteriormente le potenzialità dei modelli basati su dizionari e regole.

A livello pratico, il VADER è diventato uno strumento essenziale per molte organizzazioni che cercano di comprendere il sentiment del cliente e ottimizzare le loro strategie di comunicazione. Nonostante il suo successo, il VADER Sentiment Analysis non è immune da sfide. La gestione di contesti culturali diversi e la comprensione di espressioni idiomatiche rappresentano ancora ambiti di miglioramento. Tuttavia, gli sviluppatori stanno continuamente affinando il modello, integrando approcci di machine learning avanzati per affrontare queste sfide e migliorare ulteriormente le sue prestazioni. In conclusione, il VADER Sentiment Analysis ha aperto nuove prospettive nell'analisi del sentiment, fornendo un modello flessibile e preciso per decodificare le emozioni nel linguaggio umano. La sua combinazione di dizionari di sentimenti e regole di punteggio ha contribuito significativamente a una comprensione più approfondita del modo in cui le parole trasmettono il sentiment, gettando le basi per ulteriori innovazioni nel campo dell'elaborazione del linguaggio naturale.

### Spacy

Spacy rappresenta una pietra miliare nell'ambito dell'analisi del linguaggio naturale (NLP), offrendo un framework potente e flessibile che si estende ben oltre l'analisi del sentiment. Lanciato nel 2015, Spacy è diventato un punto di riferimento per la comunità di ricerca e per le applicazioni industriali, grazie alla sua efficienza, versatilità e alla capacità di gestire complessità linguistiche in modo accurato.

Il cuore di Spacy è il suo motore di analisi del linguaggio naturale, che utilizza algoritmi di machine learning per scomporre il testo in token, frasi e concetti grammaticali. A differenza di molti altri framework, Spacy è progettato per essere veloce ed efficiente, rendendolo ideale per l'analisi di grandi volumi di testo in tempo reale. La sua architettura modulare permette agli sviluppatori di selezionare le componenti specifiche di cui hanno bisogno, rendendo Spacy altamente personalizzabile per soddisfare le esigenze specifiche di ogni progetto.

Una delle caratteristiche distintive di Spacy è la sua capacità di integrare modelli pre-addestrati per svolgere una varietà di attività linguistiche, tra cui l'analisi del sentiment. I modelli di analisi del sentiment in Spacy possono essere adattati e personalizzati per contesti specifici, consentendo agli sviluppatori di ottenere risultati altamente specializzati. Questa flessibilità è un vantaggio significativo, specialmente quando si affrontano problemi complessi di analisi del testo.

Oltre all'analisi del sentiment, Spacy offre funzionalità avanzate come il riconoscimento di entità (NER), l'analisi grammaticale, e l'estrazione di concetti chiave. Queste capacità estese rendono Spacy una scelta preferita per una vasta gamma di applicazioni, dall'elaborazione di documenti legali alla creazione di chatbot intelligenti. La sua architettura modulare e la presenza di modelli pre-addestrati semplificano notevolmente il processo di sviluppo, riducendo il tempo e la complessità necessari per implementare soluzioni NLP avanzate.

L'ampia adozione di Spacy sia in ambito accademico che industriale testimonia della sua efficacia. Aziende di diverse dimensioni e settori integrano Spacy nelle loro applicazioni per ottenere una comprensione più profonda del linguaggio umano e migliorare le interazioni con gli utenti. L'ecosistema di Spacy continua a crescere, con la comunità di sviluppatori che contribuisce a nuovi modelli e miglioramenti costanti.

Nonostante i suoi successi, Spacy affronta costantemente nuove sfide. La comprensione di contesti altamente specifici e la gestione di lingue meno diffuse sono ambiti in cui il framework continua a evolversi. Gli sviluppatori di Spacy rispondono prontamente alle sfide emergenti, lavorando per migliorare continuamente la precisione e l'adattabilità del framework.

In conclusione, Spacy si posiziona come un pilastro fondamentale nell'analisi del linguaggio naturale, offrendo un approccio flessibile e potente per esplorare le complessità del linguaggio umano. La sua capacità di adattarsi a una vasta gamma di contesti, combinata con la sua efficienza e versatilità, continua a renderlo una scelta di spicco per coloro che cercano di sfruttare appieno le potenzialità dell'NLP nelle loro applicazioni e progetti.

### Feel-It

Feel-It è uno strumento innovativo che analizza i sentimenti espressi sui social media in lingua italiana. È stato sviluppato da due ricercatori, Federico Bianchi e Debora Nozza, e un professore associato, Dirk Hove, dell’Università Bocconi di Milano.

Questo strumento open source è in grado di riconoscere e analizzare i sentimenti e le emozioni trasmesse attraverso i social media in italiano. È stato progettato per essere facilmente installabile in qualsiasi pipeline e utilizzabile da chiunque.

Il funzionamento di Feel-It si basa sull’analisi di tweet in italiano. Gli sviluppatori hanno estratto le emozioni da alcuni tweet riguardanti vari argomenti di tendenza su Twitter. Hanno poi analizzato e classificato manualmente circa 2000 tweet associando a ciascuno un’emozione (gioia, rabbia, tristezza, paura). Questo dataset è stato utilizzato per addestrare un modello neurale chiamato “BERT” in italiano.

Feel-It è in grado di mappare le emozioni delle persone suddividendole in due cluster: positivo/negativo (“joy” per il positivo e le altre a negativo), riuscendo a replicare il modello BERT. I risultati hanno dimostrato che Feel-It è un modello che predice bene la sentiment analysis e le emozioni nel testo.

Un punto importante da sottolineare è che Feel-It non è limitato all’analisi dei soli testi dei social media. I risultati hanno dimostrato che questo strumento può essere utilizzato anche con testi non legati ai social media, ampliando così il suo campo di applicazione.

Feel-It può rivelarsi molto utile per aziende e PMI che hanno una presenza digitale, in quanto permette di capire il trend degli utenti e di analizzare come le persone rispondono a certi eventi sui social.

Il dataset e il modello di Feel-It sono disponibili in open source sul web e sono descritti in un paper accademico peer-reviewed che è stato presentato a WASSA 2021 durante l’undicesimo Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis della conferenza EACL.

Feel-It si dimostra quindi adatto all’obiettivo di questo progetto. Oltre a identificare se il sentimento espresso in un testo è positivo o negativo, Feel-It è in grado di distinguere tra diverse emozioni. Questa capacità fornisce informazioni più dettagliate che possono aiutare a comprendere lo stato emotivo del criminale analizzando la chiamata telefonica intercettata. Inoltre, l’analisi delle emozioni può contribuire a capire il tipo di legame che esiste tra i due interlocutori in chiamata.

### Differenze e scelta del modello

L’analisi del linguaggio naturale (NLP) è stata rivoluzionata da diversi modelli, tra i quali: il VADER Sentiment Analysis, il framework Spacy e Feel-It. Mentre tutti e tre si concentrano sull’esplorazione delle complessità del linguaggio umano, le loro architetture, metodologie e applicazioni variano significativamente.

Il VADER Sentiment Analysis si distingue per il suo approccio basato su dizionari e regole sofisticate. Questo modello è principalmente adatto all’inglese e si concentra sulla valutazione del sentiment complessivo di un testo, prendendo in considerazione sia la polarità delle parole che l’intensità del sentimento.

D’altra parte, Spacy adotta un approccio più modulare e flessibile basato su machine learning. Il framework offre un motore di analisi del linguaggio naturale che utilizza algoritmi di machine learning per scomporre il testo in token, frasi e concetti grammaticali. Spacy offre un’ampia suite di funzionalità NLP, tra cui l’analisi grammaticale, il riconoscimento di entità e l’analisi del sentiment.

Feel-It, sviluppato da ricercatori dell’Università Bocconi di Milano, è uno strumento open source che analizza i sentimenti espressi sui social media in lingua italiana. Utilizza un modello neurale chiamato “BERT” addestrato su un dataset di tweet in italiano. Feel-It è in grado di mappare le emozioni delle persone suddividendole in due cluster: positivo/negativo (“joy” per il positivo e le altre a negativo). A differenza di VADER e Spacy, Feel-It fornisce informazioni sulle emozioni e non solo sulla positività o negatività del sentimento.

Per lo scopo del progetto, Feel-It risulta essere il più adatto. Questo perché, oltre a fornire informazioni sulla positività o negatività del sentimento, Feel-It è in grado di distinguere tra diverse emozioni. Inoltre, Feel-It è stato addestrato su frasi italiane, il che lo rende particolarmente utile per l’analisi del linguaggio naturale in italiano.

# Obiettivi perseguiti

## Obiettivi

L’obiettivo di questo progetto è di usare un modello di sentiment già addestrato per estrarre informazioni utili per lo studio delle intercettazioni criminali. Il progetto si basa su un dataset contenente delle intercettazioni criminali in lingua italiana, che verranno sottoposte al modello di sentiment per rilevare le emozioni presenti nel testo delle intercettazioni. Questo può essere utile a capire lo status attuale dei criminali intercettati, le loro motivazioni, le loro intenzioni e le loro relazioni. Per raggiungere questo obiettivo, il progetto prevede le seguenti fasi: la preparazione e la pulizia del dataset di intercettazioni criminali; l’applicazione del modello di sentiment al dataset di intercettazioni criminali; l’analisi e la visualizzazione dei risultati del modello di sentiment; la discussione delle implicazioni e delle limitazioni del modello di sentiment per il campo della legge.

## Dati

I dati utilizzati per le intercettazioni criminali provengono da un documento di custodia cautelare anonimizzato, appositamente selezionato per testare il modello in modo legale ed efficace. Questo documento contiene circa 30.000 frasi, una quantità sufficiente per valutare le capacità del modello. La principale sfida legata ai dati riguarda l'impiego del dialetto napoletano, il quale potrebbe aumentare il rischio di imprecisioni nelle risposte del modello. Tuttavia, va notato che la presenza del dialetto non è così frequente da compromettere completamente l'obiettivo del progetto.

Per arrivare a questa conclusione si è effettuato un controllo per controllare la percentuale di parole non italiane presenti nelle intercettazioni, i quali dettagli saranno meglio spiegati

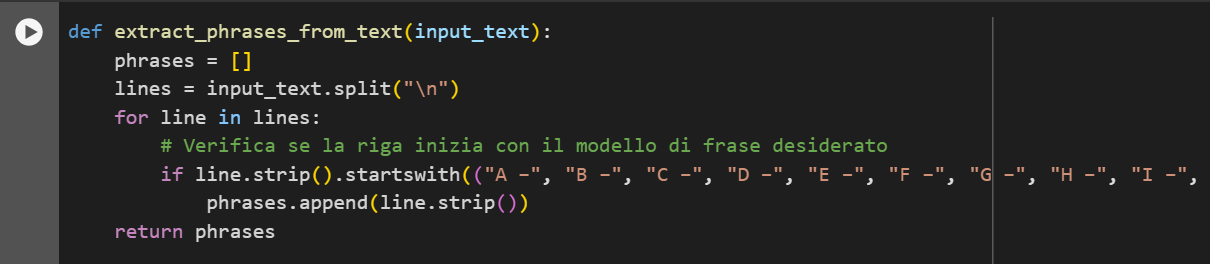
nella sezione "Problematiche e conseguenti trattamenti". Il risultato ha mostrato che solamente l'1.41% delle parole presenti nelle intercettazioni a disposizione non sono italiane, questo dato ci consente di ritenere applicabile l'utilizzo di un modello che riconosce la lingua italiana senza compromettere in maniera significativa i risultati.

# Metodologia

## Estrazione di conversazioni

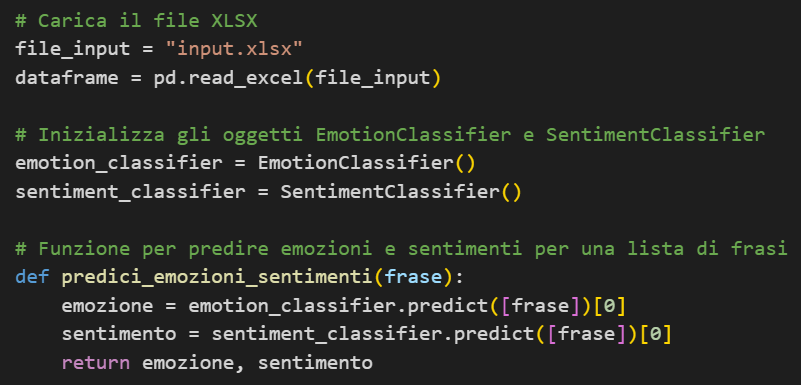
Siamo partiti da un file di testo anonimizzato contenente tutte le conversazioni, insieme ad altri informazioni di cui non avevamo bisogno, tra cui i dettagli di molte conversazioni o di alcuni processi. È stata quindi una sfida trovare un modo per rendere automatica l’estrazione di conversazioni. Infatti, estrarle manualmente avrebbe richiesto diverse settimane di lavoro. Abbiamo notato che ogni conversazione inizia con una lettera o due lettere, seguite da un trattino “-” e a seguire dalla frase. Ecco un esempio: P – Oggi è una bella giornata.

Di conseguenza, abbiamo scritto un codice in Python, su Colab (una piattaforma che permette di eseguire codice direttamente sul Cloud). È stato complicato anche perché nel documento c’erano diversi invii a capo casuali, e diverse conversazioni iniziavano a metà riga. Tale codice, quando in una riga trova una lettera o due, seguite da un trattino “-”, salva la frase detta da un determinato interlocutore. Da precisare che ogni “Frase” parte nel momento in cui qualcuno parla, fino a quando termina di parlare e inizia di parlare un altro. Quindi la frase “Che bella giornata! Oggi mi sento felice. E anche molto tranquillo per il bel tempo e il sole fuori.” viene contata come un’unica riga nell’attributo “Frasi”.



## Utilizzo di Feel It

A questo punto, abbiamo utilizzato il modello Feel It, sempre con Python su Colab. Il modello mette a disposizione la possibilità di analizzare sia il sentimento (positivo o negativo) che l’emozione (gioia, tristezza, paura, rabbia). Noi abbiamo utilizzato entrambe.



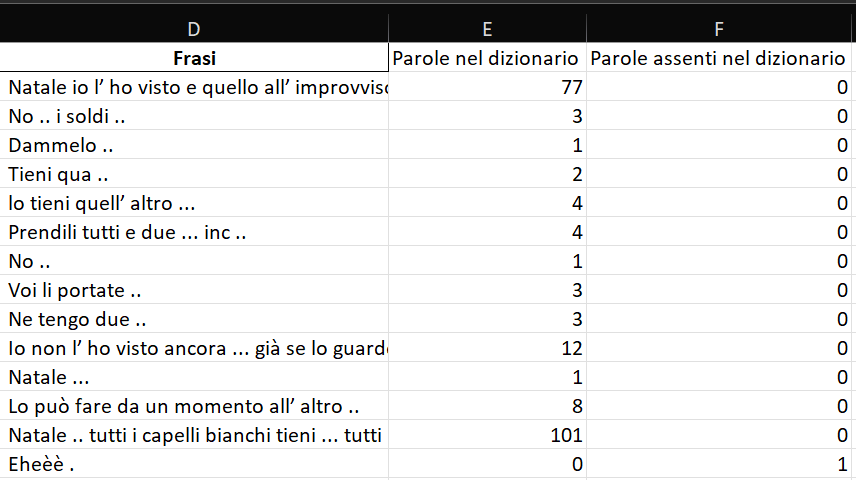
Abbiamo preso in input il file excel contenente le Frasi e l’Interlocutore (sempre anonimizzato), e in output abbiamo ricevuto lo stesso file con due colonne aggiuntive – Emozione e Sentimento – che vengono utilizzate in ogni riga.



## Conteggio di parole italiane e non

Abbiamo utilizzato un dizionario (file di testo) trovato online, al link: <https://github.com/sigmasaur/AnagramSolver>

Questo dizionario contiene tutte le parole italiane, e tutte le coniugazioni dei verbi italiani. Per ogni parola di ogni frase, abbiamo utilizzato un codice in Python (consultabile sulla nostra repository di github e non riportato qui poiché molto lungo: <https://github.com/GiovanniBorrelli/RetiGeografiche>) che verifica se quella parola appartiene o meno al dizionario. Abbiamo utilizzato inoltre un ulteriore dataset che contiene tutti i nomi italiano, dato che nelle conversazioni venivano usati frequentemente. Per ogni frase è riportato il numero di parole italiane, e parole assenti nel dizionario (che potrebbero essere sia parole napoletane, sia parole particolari come il nome di una marca di vestiti o di una bevanda).



# Problematiche affrontate

1) il problema iniziale è stato quello di estrarre le conversazioni dal file di testo anonimizzato, questo perché il file era formattato in modo inadatto all'estrazione automatica delle frasi, con molti invii a capo casuali e tante conversazioni che partivano a metà riga.

2) Il primo problema è stato individuare un modello adatto alla lingua italiana; infatti, i principali modelli di analisi del sentiment sono stati sviluppati interamente o principalmente per la lingua inglese, questa problematica è stata risolta dall'individuazione del modello Feel-It;

3) Il secondo problema, come già accennato, riguarda la presenza del dialetto napoletano all'interno delle intercettazioni.

Si è quindi deciso di controllare la percentuale di dialetto presente al loro interno. Per fortuna, i risultati sono stati positivi, evidenziando infatti una presenza esigua di parole dialettali.

Inoltre, dato che Feel-It è basato sulla rete neurale BERT, cerca di capire anche il senso delle parole non conosciute utilizzando le parole vicine, il che rende il problema del dialetto ancora meno ostacolante.

3.1) Riguardo la seconda problematica, non è stato semplice calcolare la giusta percentuale di parole dialettali contenute nelle intercettazioni, si è trovato innanzitutto un pool contenente tutte le parole italiane, ma oltre le parole è servito anche un pool di tutti i nomi italiani, questi due pool sono serviti per fare un confronto tra le parole delle intercettazioni e quindi constatare se le parole presenti sono italiane oppure no. Nel dizionario di parole italiane sono stati inseriti anche i nomi di alcune marche di auto per permetterne il riconoscimento durante il confronto.

Bisogna notare che nonostante queste accortezze è comunque presente il nome di alcuni locali non presenti nei pool; quindi, il loro nome verrà erroneamente contato come parola in dialetto.

Si è dovuto poi trovare il giusto metodo per non prendere in considerazione la punteggiatura, anche per gli apostrofi e le virgolette si è dovuto procedere in modo che venissero ignorati durante il confronto delle parole.

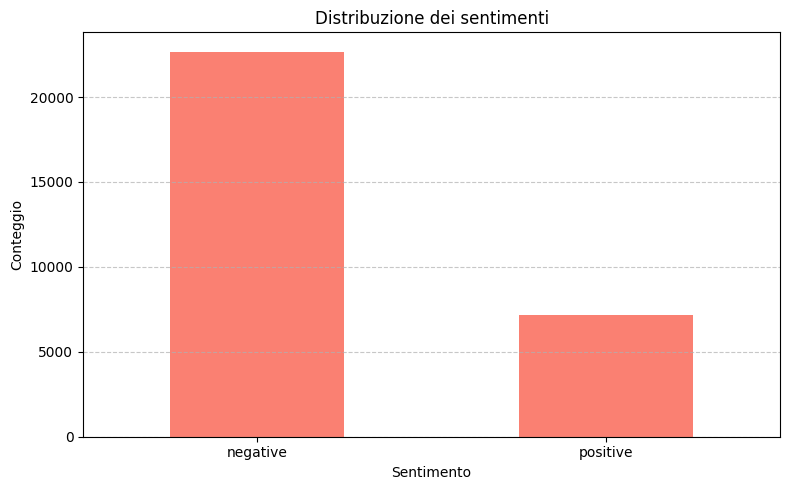
Infatti, nel dizionario delle parole italiane alcune parole avevano l'apostrofo, quindi per il confronto è stato ignorato, in maniera da confrontare le parole senza l'apostrofo.

In secondo momento ci si è resi conto che alcune parole italiane non venivano considerate tali a causa degli accenti sbagliati, il caso più comune è la presenza della parola "perchè" che nella forma corretta andrebbe scritta "perché", essendo questo un problema di scrittura si è fatto in modo che la parola "perchè" così come altre parole con la medesima problematica venissero giustamente contate come parole italiane, per fare ciò durante il confronto delle parole si è ignorata la presenza dell'accento.

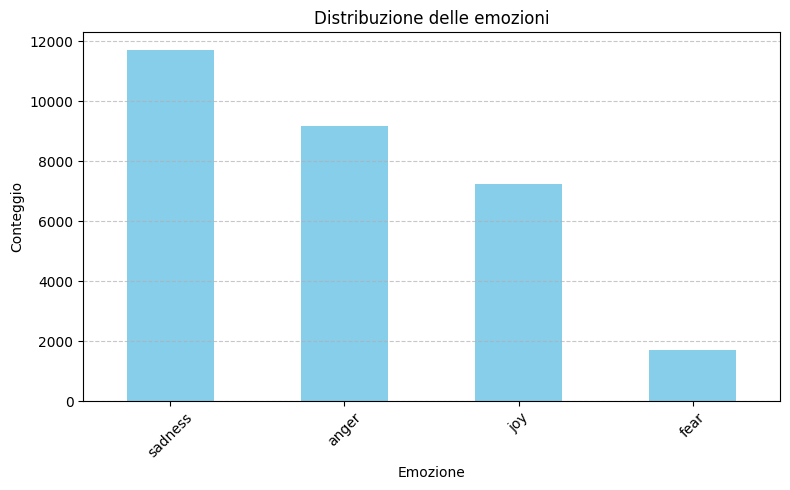
Abbiamo poi notato che in molti casi era presenta la parola "inc" che serve per indicare una parte dell'intercettazione non compresa e quindi non trascritta, questa parola veniva contata quindi come parola non italiana, ma essendo che lo scopo è capire la presenza delle parole in dialetto si è deciso di ignorare anche questa parola durante il conteggio delle parole italiane e dialettali.

# Risultati della sperimentazione

## Grafici dei risultati



Il grafico della distribuzione delle emozioni convalida ciò che ci aspettavamo, cioè che circa i 2/3 delle conversazioni sono di carattere negativo, andando a confermare ciò che l’ambiente criminale riserva, ovvero dei sentimenti generali di infelicità e insoddisfazione.



Innanzitutto, ci colpisce che siano pochissime le conversazioni che suscitano paura. Noi ipotizziamo che sia perché i criminali si contraddistinguono per una spiccata e spesso ingiustificata sicurezza di sé e di non essere mai arrestati. Successivamente, abbiamo la gioia, che è al penultimo posto. Questo perché la vita dei criminali è stressante e poche volte da emozioni positive. Infatti, il mondo nella criminalità organizzata fa molto più male che bene, e l’obiettivo di questo progetto è dare una mano ad estirpare questo grande male della società che ci affligge. Al secondo posto abbiamo la rabbia. Questo perché il mondo dei criminali è tempestato di regole d’onore e scontri di orgoglio, inoltre i criminali cercano di imporsi con la forza e la violenza, e raramente fanno discussioni e ascoltano l’altro. La tristezza è al primo posto perché noi ipotizziamo che il mondo della malavita arrechi soltanto dei danni, soprattutto ai poveri malcapitati che rimangono, spesso contro la loro volontà, bloccati in questa spirale di crimini e depressione. Ci aspettavamo che la rabbia fosse l’emozione prevaricante, poiché è una emozione pervasiva in questo campo.



Abbiamo generato un word cloud, che mostra le parole più frequenti in assoluto complessivamente. PERSONA rappresenta un nome anonimizzato, poiché abbiamo ricevuto un documento anonimizzato, quindi con i nomi censurati dalla parola ‘PERSONA’. Invece ‘inc’ significa incomprensibile, poiché spesso nelle conversazioni i criminali dicono qualcosa di incomprensibile.

# Limitazioni e possibili lavori futuri

Limitazioni:

1) Modello addestrato solo su testi in italiano: Una delle principali limitazioni di Feel-It è che il modello è stato addestrato esclusivamente su testi in italiano. Questo potrebbe limitare la sua efficacia nel comprendere il linguaggio in altre lingue. Per applicazioni che coinvolgono testi multilingue o lingue diverse dall'italiano, potrebbe essere necessario addestrare il modello su dataset specifici per ottenere risultati accurati.

2) Presenza limitata di conversazioni e formattazione: La disponibilità di un numero limitato di conversazioni e la formattazione poco chiara dei dati potrebbero influenzare la precisione e l'affidabilità dell'analisi del sentiment. Un dataset più ampio e meglio strutturato potrebbe migliorare la qualità dei risultati ottenuti.

3) Assenza di informazioni aggiuntive oltre all'emozione e al sentiment: Attualmente, Feel-It fornisce solo informazioni sull'emozione e sul sentiment presenti nel testo. Tuttavia, per ottenere una comprensione più approfondita del contenuto, potrebbe essere utile estrarre ulteriori informazioni come il tono della conversazione, il contesto sociale o la relazione tra gli interlocutori.

Possibili lavori futuri:

1) Addestramento su frasi e/o parole napoletane: Considerando la presenza del dialetto napoletano nel dataset delle intercettazioni criminali, un possibile sviluppo futuro potrebbe essere l'addestramento del modello su frasi e/o parole napoletane. Ciò potrebbe migliorare la capacità del modello di riconoscere e interpretare correttamente il linguaggio dialettale, aumentando così la precisione dell'analisi del sentiment.

2) Aumento del numero e miglioramento della qualità delle conversazioni: Acquisire un dataset più ampio e ben strutturato di conversazioni potrebbe migliorare significativamente le prestazioni del modello. Una maggiore varietà di conversazioni consentirebbe al modello di apprendere una gamma più ampia di espressioni linguistiche e contesti, aumentando la sua capacità di generalizzazione e migliorando la precisione delle previsioni.

3) Sviluppo di strumenti per l'estrazione di informazioni aggiuntive: Integrare strumenti o tecniche che consentono di estrarre informazioni aggiuntive dai testi, oltre all'emozione e al sentiment, potrebbe arricchire notevolmente l'analisi del contenuto. Ad esempio, l'identificazione del tono della conversazione o delle relazioni sociali potrebbe fornire una visione più completa e dettagliata del contesto delle intercettazioni criminali, contribuendo così a una migliore comprensione delle dinamiche in gioco.