FEEL IT

SENTIMENT ANALYSIS

SU INTERCETTAZIONI CRIMINALI

Scorzelli Lorenzo & Borrelli Giovanni

CONTENUTI DELLA PRESENTAZIONE

01 **INTRODUZIONE**

Introduzione e stato dell'arte

02 **OBIETTIVI**

Obiettivi e dati

03 **METODOLOGIA**

Spiegazione rigorosa della metodologia

04 **RISULTATI**

Analisi dei risultati e Sviluppi futuri

https://github.com/GiovanniBorrelli/RetiGeografiche

01
INTRODUZIONE

Introduzione al problema
e Stato dell'arte

INTRODUZIONE AL PROBLEMA

- Il nostro progetto si propone di utilizzare l'analisi del sentiment per estrarre informazioni cruciali dalle intercettazioni criminali in lingua italiana.
- Affrontiamo sfide come la varietà linguistica, la qualità dei dati e la complessità del dialetto napoletano. La comprensione accurata dello stato emotivo dei soggetti coinvolti, delle loro motivazioni e delle relazioni intercorrenti richiede un approccio multidisciplinare per massimizzare l'utilità delle informazioni estratte.

INTRODUZIONE AL PROBLEMA

LIMITAZIONI DELL'ANALISI DEL SENTIMENT

Sfide Attuali dell'Analisi del Sentiment:

- Comprensione di espressioni idiomatiche (essere a cavallo).
- Gestione delle sfumature culturali (ciò che è positivo per una cultura può essere negativo per un altra).
- Riconoscimento del contesto.

Obiettivi Futuri dell'Analisi del Sentiment:

- Utilizzo di modelli sempre più sofisticati.
- Integrazione di approcci di machine learning avanzati.
- Offrire una comprensione ancora più profonda del linguaggio e delle emozioni.

STATO DELL' ARTE

Modelli di Sentiment Analysis come:

- VADER
- Spacy
- Feel-It

Feel-It, a differenza di VADER e Spacy, riesce ad analizzare anche alcune emozioni di una frase.

02

OBIETTIVI

Obiettivi perseguiti e Dati

OBIETTIVI PERSEGUITI

Obiettivi del Progetto:

- Avendo a disposizione un documento di custodia cautelare, Analizzare le emozioni presenti nei testi delle intercettazioni, per comprendere lo status attuale dei criminali intercettati, le loro motivazioni, intenzioni e relazioni.
- Utilizzare un modello di sentiment già addestrato per estrarre informazioni utili dalle intercettazioni cri minali.
- Fasi del progetto: preparazione e pulizia del dataset, applicazione del modello di sentiment, analisi e visualizzazione dei risultati, discussione delle implicazioni e delle limitazioni nel contesto legale.

DATI DELLE INTERCETTAZIONI

Dati delle Intercettazioni Criminali:

- Provenienti da un documento di custodia cautelare anonimizzato.
- Contiene circa 30.000 frasi, sufficienti per valutare le capacità del modello.
- Principale sfida: l'impiego del dialetto napoletano, che potrebbe influenzare le risposte del modello.
- Controllo della percentuale di parole non italiane presenti nelle intercettazioni.

03

METODOLOGIA

Estrazione automatica di conversazioni e utilizzo di Feel-It

Estrazione Automatica delle Conversazioni:

- Utilizzo di un file di testo anonimizzato contenente tutte le conversazioni e altre informazioni non necessarie.
- Sfida nell'estrarre automaticamente le conversazioni senza impiegare settimane di lavoro manuale.
- Osservazione che ogni conversazione inizia con una o due lettere seguite da un trattino "-".
- Esempio di struttura: "P - Oggi è una bella giornata."
- Tecnologie Utilizzate: Python su Google Colab.
- Complicazioni dovute a invii a capo casuali e inizio di conversazioni a metà riga.

- Il codice identifica lettere seguite da un trattino "-" e salva la frase dell'interlocutore.
- Ogni "Frase" comprende l'inizio da quando qualcuno inizia a parlare e termina quando un altro inizia a parlare, anche se la conversazione è lunga o attraversa più righe.

```python
def extract_phrases_from_text(input_text):
    phrases = []
    lines = input_text.split("\n")
    for line in lines:
        # Verifica se la riga inizia con il modello di frase desiderato
        if line.strip().startswith(("A -", "B -", "C -", "D -", "E -", "F -", "G -", "H -", "I -",
            phrases.append(line.strip())
    return phrases
```

ESTRAZIONE DELLE CONVERSAZIONI

Interlocutore	Frasi
P	Natale io l' ho visto e quello all' improvvis
N	No .. i soldi ..
P	Dammelo ..
N	Tieni qua ..
P	Io tieni quell' altro ...
N	Prendili tutti e due ... inc ..
N	No ..
P	Voi li portate ..
N	Ne tengo due ..
N	Io non l' ho visto ancora ... già se lo guard
P	Natale ...
N	Lo può fare da un momento all' altro ..
P	Natale .. tutti i capelli bianchi tieni ... tutti
N	Eheèè .

- FRASI: Attributo che contiene la frase detta dall'interlocutore
- INTERLOCUTORE: Attributo che indica chi pronuncia la frase. In questo caso c'è solo 1 o 2 lettere perché è anonimizzato, ma volendo si possono anche fare analisi su un determinato interlocutore.
- Sono state estratte 30.000 frasi

UTILIZZO DI FEEL-IT

	A	B	C	D
1	**Emozione**	**Sentimento**	**Interlocutore**	**Frasi**
2	anger	negative	P	Natale io l' ho visto e quello all' improvvis
3	anger	negative	N	No .. i soldi ..
4	sadness	negative	P	Dammelo ..
5	sadness	negative	N	Tieni qua ..
6	sadness	negative	P	lo tieni quell' altro ...
7	fear	negative	N	Prendili tutti e due ... inc ..
8	sadness	negative	N	No ..
9	sadness	negative	P	Voi li portate ..
10	sadness	negative	N	Ne tengo due ..
11	fear	negative	N	Io non l' ho visto ancora ... già se lo guard
12	joy	positive	P	Natale ...
13	sadness	positive	N	Lo può fare da un momento all' altro ..
14	anger	negative	P	Natale .. tutti i capelli bianchi tieni ... tutti
15	joy	positive	N	Eheèè .

CONTEGGIO DI PAROLE ITALIANE

Utilizzo di un Dizionario Online:

- Utilizzo di un dizionario trovato online al link: https://github.com/sigmasaur/AnagramSolver.
- Il dizionario contiene tutte le parole italiane e tutte le coniugazioni dei verbi italiani.

Verifica delle Parole nel Dizionario:

- Utilizzo di un codice in Python per verificare se ogni parola delle frasi appartiene al dizionario.
- Utilizzo di un ulteriore dataset contenente tutti i nomi italiani, frequentemente utilizzati nelle conversazioni.

Analisi delle Parole Italiane:

- Per ogni frase, è riportato il numero di parole italiane e il numero di parole assenti nel dizionario.
- Le parole assenti potrebbero essere parole napoletane o particolari come nomi di marca di vestiti o bevande.

CONTEGGIO DI PAROLE ITALIANE

D	E	F
Frasi	Parole nel dizionario	Parole assenti nel dizionario
Natale io l' ho visto e quello all' improvvis	77	0
No .. i soldi ..	3	0
Dammelo ..	1	0
Tieni qua ..	2	0
lo tieni quell' altro ...	4	0
Prendili tutti e due ... inc ..	4	0
No ..	1	0
Voi li portate ..	3	0
Ne tengo due ..	3	0
Io non l' ho visto ancora ... già se lo guard	12	0
Natale ...	1	0
Lo può fare da un momento all' altro ..	8	0
Natale .. tutti i capelli bianchi tieni ... tutti	101	0
Eheèè .	0	1

PROBLEMATICHE

Problemi affrontati e relative soluzioni

PROBLEMATICHE

1) Estrarre le conversazioni dal file di testo, 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) 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. Abbiamo risolto con Feel-It;

3) Il secondo problema, riguarda la presenza del dialetto napoletano all'interno delle intercettazioni.
Però, bisogna tenere in considerazione che Feel-It è basato sulla rete neurale BERT, cerca quindi di capire anche il senso delle parole non conosciute utilizzando le parole vicine, il che rende il problema del dialetto meno ostacolante.

PROBLEMATICHE

3.1) Problema: Parole contate erroneamente come assenti.
Abbiamo usato un dizionario contenente tutte le parole italiane, ma oltre le parole è servito anche un insieme di tutti i nomi italiani.

Nel dizionario sono stati inseriti anche i nomi di alcune marche di auto.

Nonostante queste accortezze è comunque presente il nome di alcuni locali non presenti nei dataset; 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.

PROBLEMATICHE

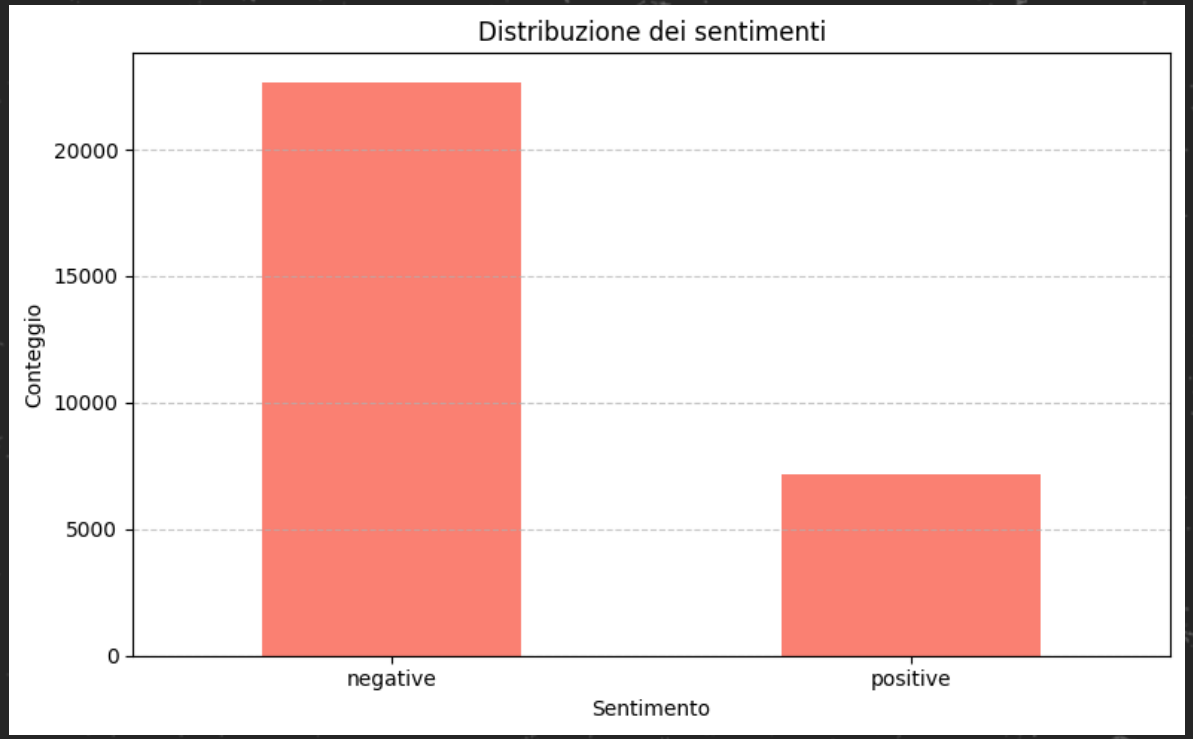
Alcune parole italiane non venivano considerate tali a causa degli accenti sbagliati, il caso più comune è la presenza della parola ”perchè” che nella forma corretta andrebbe scritta ”perché”, essendo questo un problema di scrittura si è fatto in modo che la parola ”perchè” così come altre parole con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

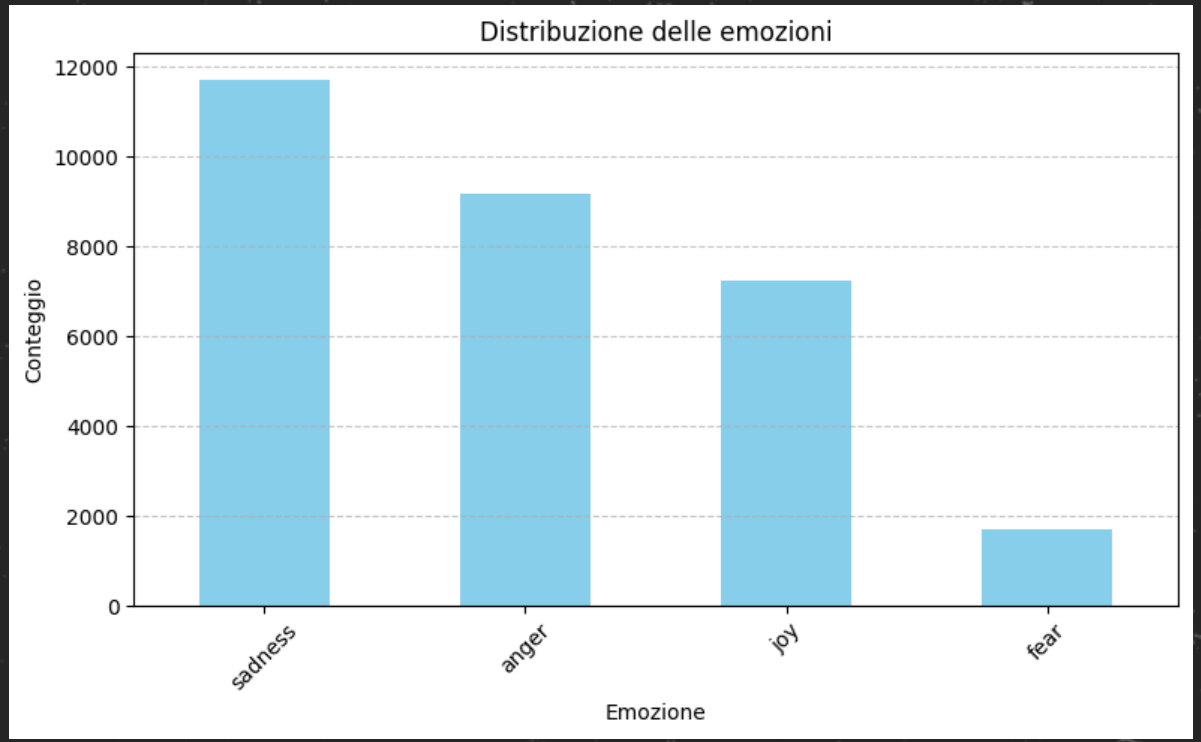
Abbiamo poi notato che in molti casi era presenta la parola ”inc” che serve per indicare una parte dell’intercettazione non compresa e quindi non trascritta, questa parola veniva contata quindi come parola non italiana, ma essendo che lo scopo è capire la presenza delle parole in dialetto si è deciso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

Avendo risolto tutti questi problemi, la percentuale di parole non italiane è di 1.41%.

04 RISULTATI

Risultati e Sviluppi Futuri


Distribuzione dei sentimenti
Conteggio
20000
15000
10000
5000
0
negative
positive
Sentimento


Distribuzione delle emozioni
Conteggio
12000
10000
8000
6000
4000
2000
0
sadness
anger
joy
fear
Emozione


Va bene
PERSONA
Since
adesso
ha detto
quello
io
Ciao Pronto
bene Ciao
la macchina
di PERSONA
per
chi
Ma tu
da
ci sta
non lo
casa
c è
perché
posto
chi è
già
noi
che non
sono
tutto
dove stai
la
l assegno
ma
ho capito
mi
ha
non ti
hai capito
Ma quello
non
là
dai
cosa
ha fatto
non ho più
incomp
fa
ci vediamo
niente
detto che
pure
io
Eheè
dopo
con la
allora
proprio
perchè
tu
un poco
di
se
il fatto
dice
te lo
ancora
euro
una cosa
lui
o
sempre
uno
questo
ho detto
che
però
soldi
comunque
PERSONA Si
qua
e non
sopra
quando
non mi
così
è
eh
gli ho
domani
Si Si
quella
poi
prima
Si ma
dove
come
sta
l ho
non ci
non è

ANALISI DEI RISULTATI

Analisi delle Emozioni nelle Intercettazioni Criminali:

Paura: Pochissime conversazioni suscitano questa emozione, ipotizzata essere legata alla spiccata sicurezza e all'ingiustificata convinzione dei criminali di non essere mai arrestati.

Gioia: Al penultimo posto, poiché la vita dei criminali è stressante e offre poche occasioni per emozioni positive, essendo il mondo della criminalità organizzata più dannoso che benefico per loro.

Rabbia: Al secondo posto, essendo il mondo dei criminali caratterizzato da regole d'onore, scontri di orgoglio, e un ricorso frequente alla forza e alla violenza, con scarsa propensione al dialogo e all'ascolto.

Tristezza: Al primo posto, poiché ipotizzata come emozione preponderante data la percezione che il mondo della malavita causi danni principalmente ai poveri malcapitati che rimangono intrappolati nella spirale del crimine e della depressione.

Aspettative rispetto alla rabbia: Nonostante la rabbia sia una presenza pervasiva nel campo criminale, ci si aspettava che fosse un'emozione preponderante, il che non si è verificato, forse dovuto al fatto che molte intercettazioni parlando di soldi persi, quindi la tristezza ha avuto la "meglio" come emozione preponderante.

LIMITAZIONI

Limitazioni di Feel-It e del documento:

Modello addestrato solo su testi in italiano: L'addestramento esclusivo su testi italiani limita l'efficacia del modello nell'interpretare il linguaggio in altre lingue, richiedendo l'addestramento su dataset specifici per lingue diverse per ottenere risultati accurati.

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, l'assenza di informazioni aggiuntive come la percentuale di emozione presente (50% rabbia, 25 tristezza, ecc.) avrebbe potuto aiutare nel fornire dati più accurati.

Presenza limitata di conversazioni e formattazione: La disponibilità ridotta di conversazioni e la formattazione poco chiara dei dati possono influenzare la precisione e l'affidabilità dell'analisi del sentiment. Un dataset più ampio e ben strutturato potrebbe migliorare la qualità dei risultati.

SVILUPPI FUTURI

Possibili Lavori Futuri:

Addestramento su frasi e/o parole napoletane: Migliorare la capacità del modello di interpretare il dialetto presente nel dataset delle intercettazioni criminali.

Aumento e miglioramento del dataset: Acquisire conversazioni più ampie e meglio strutturate per aumentare la capacità di generalizzazione e migliorare la precisione delle previsioni.

Sviluppo di strumenti per l'estrazione di informazioni aggiuntive: come la percentuale di ogni emozione espressa nelle frasi.

GRAZIE
PER L'ATTENZIONE

Lorenzo Scorzelli & Giovanni Borrelli