Università Politecnica delle Marche

Facoltà di Ingegneria

Dipartimento di Ingegneria dell'Informazione

Corso di Laurea in Ingegneria Informatica e dell'Automazione

Tesi di Laurea

Esperienze con la Sentiment Analysis con particolare riferimento al Marketing e all'Healthcare

Experiences with Sentiment Analysis with a specific focus on Marketing and Healthcare

Relatore

Prof. Domenico Ursino

Candidato

Walter Di Sabatino

Anno Accademico 2022-2023

Che colpa abbiamo, io e voi, se le parole, per sé, sono vuote?
Vuote, caro mio. E voi le riempite del senso vostro, nel dirmele;
e io nell'accoglierle, inevitabilmente, le riempio del senso mio.
Abbiamo creduto d'intenderci, non ci siamo intesi affatto.

Luigi Pirandello, "Uno, nessuno e centomila"

Sommario

Negli ultimi anni la Sentiment Analysis ha acquisito crescente rilevanza grazie alla sua capacità di interpretare le emozioni e i sentimenti espressi dalle persone. Questo processo ha dimostrato di semplificare e ottimizzare vari ambiti. In questa tesi, è stato condotto un approfondimento tecnico e storico della Sentiment Analysis, descrivendo anche le sue potenziali applicazioni in diversi settori, come healtchare, bussiness e government intelligence. Successivamente, sono stati analizzati i principali servizi che offrono strumenti per effettuare quest'importante tecnica nell'ambito dell'elaborazione del linguaggio naturale, al fine di confrontarli, evidenziarne le differenze e sottolineare il livello tecnologico attuale di questo ambito. In conclusione, è stata effettuata un'analisi etica sull'impiego della Sentiment Analysis, esplorando le sue possibili problematiche e delineando le azioni necessarie, sia già intraprese che da intraprendere, per garantirne un suo utilizzo responsabile e efficace.

Keyword: Sentiment Analysis, Opinion Mining, Natural Language Processing, opinioni, sentimenti, sentimento mirato, emozioni, Intelligenza Artificiale, insight, entità

Indice

Elenco delle figure

Introduzione

I sentimenti sono una componente fondamentale dell'esperienza umana. Sono le emozioni, le risposte istintive che ci guidano attraverso la vita, influenzando le nostre decisioni, le nostre azioni e, persino, la nostra salute mentale. Pertanto, la capacità di comprendere e analizzare in modo accurato questi sentimenti diventa cruciale per acquisire una prospettiva più profonda su noi stessi e sulla società in cui viviamo.

In questo contesto, quindi, la Sentiment Analysis, campo dell'elaborazione del linguaggio naturale, ha un ruolo di grande rilevanza. Questa tecnologia, infatti, sfruttando sofisticati algoritmi di apprendimento automatico, consente di catturare e misurare i sentimenti e le emozioni che sono presenti all'interno del linguaggio scritto, fornendo, quindi, la possibilità di avere una più agevole e profonda comprensione, dal punto di vista emotivo, delle parole, aiutandoci anche ad approfondire le sfumature dell'animo umano.

La Sentiment Analysis, dunque, si è dimostrata essere una risorsa essenziale all'interno di un grande numero di ambiti. Infatti, essa, si è rivelata estremamente versatile e utile in ambiti come l'healthcare, il marketing e la politica. Con l'avvento dell'era digitale, il web è divenuto la più grande fonte di accesso di dati in formato testuale, tra cui feedback e commenti. Questi ultimi sono i dati ideali per la Sentiment Analysis, poiché da essi si possono estrapolare importanti informazioni che, poi, possono essere usate a vantaggio di organizzazioni ed aziende nei settori sopra menzionati.

Nel contesto aziendale, la Sentiment Analysis, ad esempio, può essere sfruttata per analizzare feedback e commenti al fine di valutare l'accoglienza di un prodotto tra il pubblico. Nell'ambito dell'healthcare può essere impiegata per analizzare le esperienze dei pazienti riguardo ad un determinato farmaco, valutandone l'efficacia e la presenza di effetti collaterali indesiderati. In politica, invece, può essere uno strumento prezioso per comprendere le opinioni dei cittadini riguardo a scelte politiche specifiche.

Perciò, anche soltanto da questo breve numero di esempi, si può intuire l'estrema versatilità della Sentiment Analysis, potendo anche dedurre tutti i potenziali utilizzi che potranno contribuire in modo significativo alla comprensione dei nostri sentimenti e al miglioramento delle diverse sfaccettature della vita moderna.

Avendo dato un'idea di tutte le sue potenzialità, in questa tesi verrà effettuata un'analisi approfondita della Sentiment Analysis. Saranno spiegati il suo funzionamento, la sua evoluzione storica e tutte le varie tecniche di questo tipo di analisi, cioè la document-level, sentence-level e aspect-level Sentiment Analysis. Inoltre, saranno anche presentati e analizzati, in maniera più approfondita, tutti i contesti in cui si può fare uso di questa tecnica, cioè healthcare, Bussiness Intelligence, Government Intelligence e Reccomender

Systems, illustrando, anche, un ampio numero di esempi di possibile utilizzo in questi contesti.

Successivamente, sarà svolta un'analisi dei principali servizi di Sentiment Analysis, sia per il marketing che per l'healthcare, di tre importanti provider: AWS, Google e Azure. Per AWS verranno esaminati i servizi Amazon Comprehend e Amazon Comprehend Medical. Per Google verranno analizzati i servizi API Natural Language e API Healthcare Natural Language. Infine, per Azure verrà analizzato il servizio Azure AI Language, che include sia funzioni per il marketing che per l'healthcare.

Per ogni servizio verranno esaminati gli insight che fornisce, verrà spiegato il suo funzionamento generale, e verranno forniti degli esempi di utilizzo, accompagnati, anche, dalle rispettive risposte in formato JSON.

Al termine delle singole analisi saranno confrontati i tre provider. Il confronto sarà svolto sui tipi di analisi offerti dai servizi e sulla loro facilità d'uso. Per valutare quest'ultimo aspetto, saranno presi in considerazione, per ciascun fornitore, la completezza della documentazione, l'intuitività dell'interfaccia utente e l'integrazione con gli strumenti esistenti.

In conclusione, sarà svolta un'analisi etica della Sentiment Analysis, esaminando importanti aspetti, quali la privacy, la sicurezza, le possibili implicazioni socio-culturali del suo utilizzo e l'importanza della trasparenza e delle regolamentazioni in questo campo.

Quindi, in definitiva, si può dire che questo documento ha l'obbiettivo di fornire una visione completa su un ambito importante come la Sentiment Analysis, tecnologia che, con il passare del tempo, diventerà sempre più influente all'interno delle nostre vite.

La presente tesi è, quindi, composta da sette capitoli strutturati come di seguito specificato:

- Nel Capitolo 1 sarà presentata la Sentiment Analysis insieme ad una sua delineazione storica affiancata dai tipi diversi di analisi; successivamente, verranno anche presentati gli ambiti di utilizzo di questa tecnica.
- Nel Capitolo 2 si spiegherà il funzionamento, con degli esempi, dei servizi di AWS adatti alla Sentiment Analysis, ovvero Amazon Comprehend e Amazon Comprehend Medical.
- Nel Capitolo 3 si spiegherà il funzionamento, con degli esempi, dei servizi di Google adatti alla Sentiment Analysis, ovvero API Natural Language e API Healthcare Natural Language.
- Nel Capitolo 4 si spiegherà il funzionamento, con degli esempi, del servizio di Azure utile alla Sentiment Analysis, ovvero Azure Ai Language.
- Nel Capitolo 5 verranno confrontati tra loro i tre provider.
- Nel Capitolo 6 verrà effettuata un'analisi etica della Sentiment Analysis.
- Nel Capitolo 7 verranno tratte le conclusioni e verranno delineati alcuni possibili sviluppi futuri.

CAPITOLO 1

La Sentiment Analysis

In questo primo capitolo si propone una visione generale della Sentiment Analysis. Si spiegherà in cosa consiste e le difficoltà che si presentano nello sviluppo di sistemi per la sua esecuzione. Verranno mostrati i principali livelli di complessità di questo task i quali saranno approfonditi ed analizzati uno ad uno. Sarà, anche, riportata la storia della Sentiment Analysis, con riferimenti a documenti importanti che hanno segnato la sua evoluzione, e anche ad esempi di suo interesse risalenti fin dall'antica Grecia.

Infine, verranno presentati anche i principali campi di utilizzo della Sentiment Analysis quali: bussiness intelligence, reccomender systems, e-governance e healthcare ciascuno dei quali sarà successivamente approfondito.

1.1 Cos'è la sentiment analysis

La Sentiment Analysis (SA), nota anche come Opinion Mining (OM) è un campo dell'elaborazione del linguaggio naturale (Natural Language Processing, NLP) che si occupa di costruire sistemi per l'identificazione, estrazione, quantificazione e studio di sentimenti ed opinioni contenuti all'interno del linguaggio. Questo tipo di elaborazione fa grande uso del Data Mining, del Machine Learning (ML), dell'Intelligenza Artificiale e della linguistica computazionale al fine di poter estrarre il grado di sentimento dai testi e comprendere se esso è positivo, negativo o neutro, come evidenziato nella figura 1.1.

Attualmente questa area di ricerca è una di quelle in più rapida crescita all'interno della computer science, soprattuto grazie alle enormi mole di testi e opinioni ricavabili dal web. Questo processo può sembrare semplice, ma, in realtà, non lo è, infatti i processi moderni di Sentiment Analysis si devono interfacciare con testi che, la maggior parte delle volte, non sono del tutto puliti o corretti grammaticalmente. Infatti, i testi online spesso sono soggetti ad errori grammaticali, espressioni idiomatiche, utilizzo di emoji, abbreviazioni, punteggiatura impropria e una struttura prevalentemente informale, rendendo molto più difficile il compito di comprensione da parte di tecniche statistiche o di Natural Language Processing. Un altro aspetto che rende estremamente difficile il lavoro di analisi dei sistemi odierni riguarda il sarcasmo o l'ironia, che risultano essere estremamente problematici da comprendere a causa della loro implicità e situazionalità. Proprio per questi motivi, per poter fare un analisi adeguata, bisogna sviluppare dei sistemi opportunamente allenati nell'analisi di testi.

Un altro aspetto che rende la Sentiment Analysis un task complesso da svolgere è che esso si deve applicare anche in un ambiente che è multilingua e multidisciplinare. Quindi, per poterla mettere in atto con successo, bisogna sviluppare sistemi che riescano ad interfacciarsi con semantiche, grammatiche, lessici, slang, alfabeti e, addirittura, tonalità, come nel caso del Mandarino, completamente diversi. Come abbiamo detto, la Sentiment Analysis deve essere anche multidisciplinare, nel senso che si deve saper interfacciare anche con terminologie di discipline diverse, ampie e in costante evoluzione. Un esempio è proprio quello della medicina, campo estremamente vasto che racchiude decine di migliaia di terminologie che, di anno in anno, vanno sempre ad aumentare.

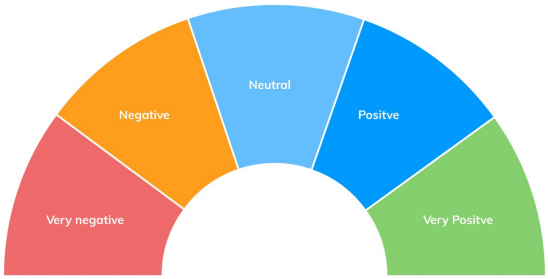


Figura 1.1: I gradi di sentimento

1.2 Breve storia della Sentiment Analysis

L'interesse per l'interpretazione dei sentimenti non è solo attuale, bensì si può far risalire anche a secoli fa. Esempi di tentativi di individuazione del dissenso interno, e quindi del sentimento, si trovano già ai tempi dell'Antica Grecia. In particolare, un esempio è Atene nel V secolo A.C. in cui, al fine di misurare l'opinione pubblica riguardo determinate politiche, venivano utilizzati opportuni sistemi di votazione. Un altro ancora, invece, sempre nell'ambito della cultura greca, può essere quello di Agamennone nell'Iliade, quando quest'ultimo valuta opinioni e disposizioni del suo consiglio di guerra cercando di spronarli ad attaccare i Troiani.

Sforzi per comprendere l'opinione pubblica, quantificandola e misurandola attraverso questionari, sono apparsi solo nei primi decenni del ventesimo secolo, andandosi a concentrare, però, solo su opinioni pubbliche di esperti, piuttosto che su opinioni di utenti o clienti. Alcuni esempi riguardano studi pubblicati dopo la seconda guerra mondiale per studiare l'opinione pubblica all'interno di paesi che avevano sofferto durante la guerra (*Italia*, *Giappone* e *Cecoslovacchia*), pubblicati tutti all'interno del giornale Public Opinion Quarterly.

Solo dalla metà degli anni 90 i sistemi di tipo informatico iniziarono ad diventare sempre più comuni, e, quindi, ad influenzare direttamente il campo della Sentiment Analysis. Il lavoro che ha gettato le fondamenta per creare l'idea moderna di Sentiment Analysis è stato svolto dalla *Association for Computational Linguistics*, associazione fondata nel 1962 e che, con i suoi studi sul Natural Language Processing e Computational Linguistics, è stata precursore della computer based Sentiment Analysis.

L'idea corrente di Sentiment Analysis, comunque, si è sviluppata solo all'inizio del ventunesimo secolo. Un inizio potrebbe essere identificato con l'anno 2002 in corrispondenza della pubblicazione della ricerca *"Thumbs up? Sentiment Classification using Machine*

Learning Techniques" pubblicata da Bo Pang e Lillian Lee in cui, utilizzando come dati le recensioni di film reperite da *IMDb*, è stato scoperto che le tecniche di classificazione standard di Machine Learning basate sul sentiment superano in maniera assoluta quelle di base utilizzate dall'uomo. Un'altra ricerca rilevante, svolta sempre nello stesso anno e con nome simile, è *"Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews"*, scritta da Peter D. Turney, in cui è stato sviluppato un algoritmo automatico di classificazione basato sul sentimento per stabilire quali recensioni possano essere o meno raccomandate. Questo studio è stato condotto su vari domini (recensioni di automobili, banche, film e destinazioni di viaggio, tutte ricavate da *Epinions*) ed è riuscito a raggiungere un'accuratezza del ben 74%.

Da quegli anni, quindi, l'interesse per la Sentiment Analysis è sempre aumentato, ed il numero di ricerche e studi a riguardo è arrivato anche alle migliaia diventando, quindi, una delle aree di ricerca più interessanti e in crescita della computer science. A conferma di questa crescente rilevanza, basti guardare alle statistiche fornite da servizi autorevoli come Google Trends. La figura 1.2 mostra, infatti, chiaramente come l'interesse per la Sentiment Analysis sia costantemente in aumento, con un andamento ascendente nel tempo.

Figura 1.2: Popolarità della Sentiment Analysis negli anni secondo Google Trends

1.3 I livelli della Sentiment Analysis

Si identificano principalmente tre livelli di studio e di complessità della Sentiment Analysis: la document-level Sentiment Analysis, la sentence-level Sentiment Analysis e la aspect level Sentiment Analysis.

1.3.1 La document-level Sentiment Analysis

La document-level Sentiment Analysis o analisi del sentimento basata sui documenti ha come obbiettivo quello di classificare se un intero documento esprime un sentimento o un'opinione positiva o negativa. Ogni documento è classificato in base al sentimento complessivo del detentore dell'opinione su una singola entità. Per ottenere un'analisi opportuna, il documento deve essere scritto da una singola persona e non deve trattare di più argomenti. Nonostante ciò, il contenuto del testo può includere sentimenti opposti che andranno ad influenzare il sentimento complessivo del documento.

1.3.2 La sentence-level Sentiment Analysis

La sentece-level Sentiment Analysis, o analisi del sentimento basata sulla frase, ha come scopo quello di comprendere l'opinione generale che viene espressa da una singola frase al fine, anche qui, di classificarla in maniera opportuna. Però, con questo tipo

di analisi, prima di poter fare una opportuno esame della frase, vi è la necessità di distinguere se quest'ultima sia soggettiva o oggettiva, quindi se esprime dati fattuali o opinioni.

1.3.3 La aspect-level Sentiment Analysis

La aspect-level Sentiment Analysis, o analisi del sentimento basata sugli "aspetti", è un tipo di analisi a grana fine che approfondisce le analisi delle frasi considerando anche gli "aspetti" specifici delle entità di cui si tratta all'interno del testo. Ad esempio, in un feedback, si riuscirebbe a comprendere in maniera specifica ciò che è stato apprezzato e ciò che non lo è stato. Pertanto, questo tipo di analisi, riesce ad afferrare a pieno, e in maniera precisa, cosa piace e non piace alle persone, comprendendo le opinioni sui singoli aspetti delle entità che vengono trattate nel testo. Per esempio, considerando la frase: "Mi piacciono molto i colori della mia nuova TV Samsung", tramite la aspect-level Sentiment Analysis il sistema riuscirebbe a capire che la recensione è positivia e riguarda "i colori" che sono un aspetto dell'entità "TV Samsung".

1.4 I domini applicativi della Sentiment Analysis

La Sentiment Analysis, quindi, può essere uno strumento estremamente prezioso nell'analisi dei dati; infatti, grazie alla proliferazione di Internet e del web, sono stati generati enormi volumi di informazioni non strutturate, non solo documenti web, ma anche email, blog e feedbacks. Perciò i dati generati attraverso la comunicazione online possono fungere da vere e proprie miniere d'oro per l'accumulo di conoscenza. Proprio per questi motivi, il tema della Sentiment Analysis è più attuale che mai, e sistemi per svolgere questo tipo di task vengono utilizzati in un numero sempre maggiore di domini applicativi.

1.4.1 Business Intelligence

Un dominio in cui la Sentiment Analysis può essere estremamente vantaggiosa è la Business Intelligence. Le aziende, infatti, avendo a disposizione grandi quantità di feedback dei loro clienti, possono sfruttare sistemi dediti alla Sentiment Analysis a loro favore. Infatti, analizzando questi feedback, le aziende possono sviluppare degli insight più chiari inerenti ai sentimenti e alla fidelizzazione delle persone riguardo i loro prodotti o servizi, aiutandoli, quindi, a capire ciò che è implementato in maniera corretta e ciò che, invece, ha bisogno di miglioramenti.

Questo task, ovviamente, può essere svolto non solo con prodotti già esistenti, ma anche con delle nuove idee al fine di poterle testare più facilmente all'interno del mercato. Questo processo si chiama concept testing e fornisce ulteriori libertà e vantaggi alle aziende.

Le singole imprese possono sfruttare la Sentiment Analysis non solo per analizzare le percezioni sui loro prodotti o servizi, ma anche per poter approfondire quelle della concorrenza. Quindi, mettendo a confronto i dati, le aziende possono prendere delle decisioni che permettano loro di avere dei vantaggi competitivi sul mercato, migliorando le loro performance e anticipando i customer trend per poter capitalizzare su di essi.

Bisogna ricordare, però, che le aziende non sono definite solo dai prodotti e dai servizi che offrono, ma anche, e soprattutto, dai loro brand.

Il nome e la fama che circondano un grande brand dipendono principalmente dagli aspetti che costruiscono la sua *brand equity* (figura 1.4), cioè: la brand loyalty, che esprime quanto è forte la fedeltà verso il brand, la brand association, che rappresenta ciò che viene associato al brand, la brand quality, che esprime la qualità globale del brand che viene percepita, e la Brand awareness, cioè la notorietà e il riconoscimento del nome del brand da parte dei clienti. Quindi, i dati di Sentiment Analysis riguardanti la brand equity possono aiutare a comprendere le percezioni attuali e potenziali dei clienti riguardo il brand. Perciò i dati aquisiti possono essere sfruttati per trasformare un brand insignificante in uno di successo.

Figura 1.3: Aspetti della brand equity

1.4.2 Recommender system

Un sistema di raccomandazione, o motore di raccomandazione, è un algoritmo che crea dei suggerimenti personalizzati specifici per l'utente al fine di aiutarlo nelle sue scelte. Questo tipo di algoritmo viene sfruttato in molti ambiti e per molti tipi di prodotti, come libri, musica e film come viene anche mostrato nella figura 1.4.

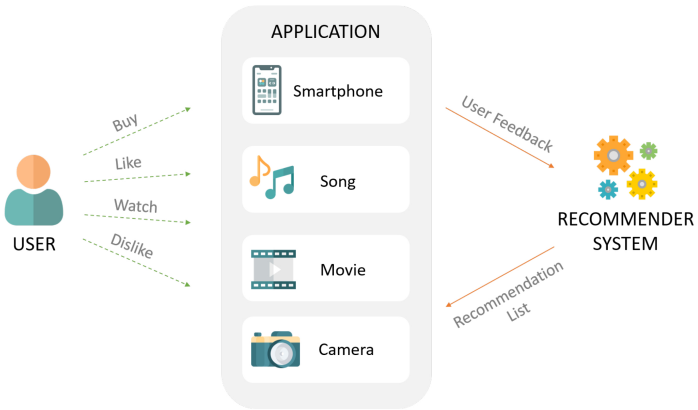


Figura 1.4: Funzionamento dei reccomender system

Un utilizzo estremamente intelligente della sentiment analysis è quello di utilizzarla per migliorare i reccomender systems. Infatti, considerando, ad esempio, delle recensioni di prodotti all'interno di un sito di e-commerce, si può sfruttare il testo della recensione,

oltre che il rating, per poter consigliare all'utente dei prodotti che possano generare in lui un sentimento positivo.

Un esempio può essere quello dei siti di viaggio. Infatti, potendo sfruttate le recensioni di siti come TripAdvisor, si possono consigliare agli utenti ristoranti oppure hotel che rispecchino maggiormente i loro gusti.

1.4.3 Government intelligence

Un altro ambito in cui la Sentiment Analysis si è rivelata essere estremamente preziosa è quello della government intelligence; infatti le persone online non esprimono pareri soltanto riguardanti prodotti o servizi ma anche riguardo alla politica, alla religione o questioni sociali.

Difatti, utilizzando social network di micro-blogging come Twitter, è possibile comprendere i sentimenti dei cittadini riguardo determinate politiche o questioni. Questo rappresenta un prezioso aiuto nel migliorare i processi di e-governance, che si riflettono nella gestione elettronica dei servizi governativi, nello scambio di informazioni, nelle comunicazioni e nell'integrazione di sistemi autonomi all'interno delle istituzioni governative. I vantaggi di tale approccio, come mostrato nella figura 1.5, sono molteplici, poiché permettono di comprenderne appieno le preferenze e le opinioni dei cittadini riguardo a scelte politiche e sociali, contribuendo così a un'efficace e partecipativa governance.

Quindi, grazie a queste tecnologie, si riesce a migliorare l'interazione diretta tra cittadino ed ente pubblico, permettendo a quest'ultimo di migliorare l'efficienza dei servizi e di rendersi maggiormente trasparente.

Figura 1.5: Aspetti della e-governance

La Sentiment Analysis è stata utilizzata anche per tentare di predire quali potessero essere i risultati di alcune elezioni governative.

Un esempio è quello del 2018 in cui uno *studio* di alcuni ricercatori dell'università degli studi di Milano ha tentato di predire, utilizzando i post di Twitter, quali potessero essere i risultati delle elezioni del Parlamento Italiano.

Le analisi hanno mostrato che il sentimento positivo era più frequente per i partiti di centro-sinistra, come il movimento 5 Stelle, e che quest'ultimo era quello anche aveva il più alto sentimento positivo tra tutti. Infatti, questo è riuscito anche a prendere più seggi al Senato e alla Camera.

1.4.4 Healthcare

Al centro di ogni sistema di healthcare c'è il rapporto tra il paziente e il servizio sanitario; per questo motivo strumenti di Sentiment Analysis possono essere, anche in questo ambito, molto preziosi al fine di poter comprendere dove il sistema stesso stia fallendo.

Attraverso il feedback dei pazienti, con l'utilizzo della sentiment analysis, si possono anche semplificare le diagnosi di questi ultimi. Infatti, grazie a determinati sistemi opportunamente allenati, si possono anche estrarre dati medici da testi non strutturati, semplificando, quindi il lavoro di diagnosi degli operatori sanitari, aiutandoli anche a comprendere l'efficacia di trattamenti a cui i pazienti sono soggetti.

Anche i social network possono essere fonte di informazioni riguardanti vari aspetti dell'healthcare. Infatti, studiosi e operatori sanitari, con l'utilizzo della sentiment analysis applicata a post e tweet, possono avere una più chiara visione riguardante argomenti come malattie, epidemie ed efficacia o reazioni avverse di nuovi farmaci.

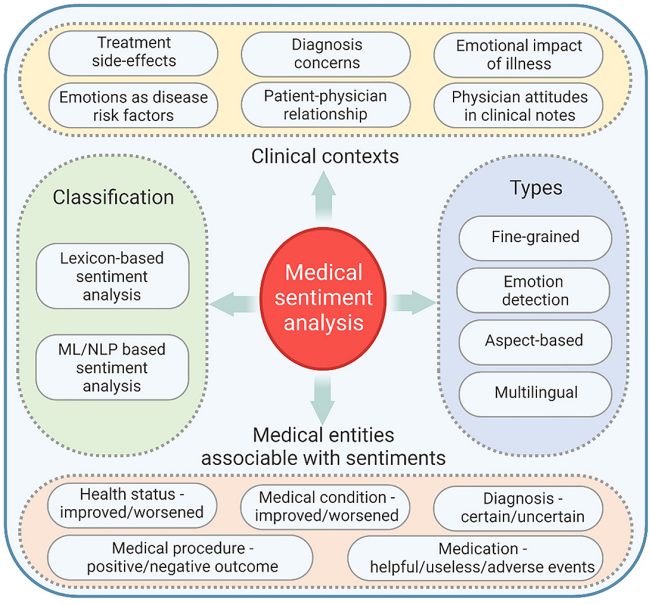


Figura 1.6: La sentiment analysis nell'ambito medico

CAPITOLO 2

La Sentiment Analysis con AWS

Il presente capitolo delinea un'analisi approfondita dei servizi di AWS (Amazon Web Services) che si rivelano utili per la Sentiment Analysis. In particolare, verrà esaminato in dettaglio il servizio Amazon Comprehend, focalizzandosi sul suo impiego nel campo del Marketing. Sarà fornita una chiara spiegazione del suo funzionamento e saranno presentati esempi illustrativi.

successivamente, sarà dedicata un'ulteriore sezione all'analisi di Amazon Comprehend Medical, un servizio specificamente progettato per il settore medico. In particolare, verranno esplorati il suo funzionamento e i contesti in cui è utilizzato, corredati da esempi esplicativi.

2.1 Cos'è Amazon Comprehend

Amazon Comprehend è un servizio di Natural Language Processing (NLP) sviluppata da Amazon.com, Inc. che usa il machine learning per trovare insight e relazioni all'interno di testi. Questo servizio elabora gli insight riconoscendo entità, frasi chiave, linguaggio, sentimenti e altri elementi comuni in un documento. Amazon Comprehend può essere utilizzato per creare nuovi servizi e prodotti basati sulla comprensione della struttura di documenti. Ad esempio, utilizzando il servizio, è possibile cercare nei feed dei social network le menzioni di prodotti oppure scansionare interi archivi di documenti alla ricerca di frasi chiave, come viene anche mostrato nella Figura 2.1.

Le funzionalità di analisi dei documenti di Amazon Comprehend possono essere adoperate utilizzando la console oppure le API fornite dal servizio. L'analisi dei testi può essere eseguita in tempo reale per piccoli carichi di lavoro oppure, nel caso di grandi set di documenti, possono essere avviati dei processi di analisi asincrona. Oltre ad utilizzare i modelli pre-addestrati forniti dal servizio si possono addestrare i propri modelli personalizzati per la classificazione e il riconoscimento delle entità in base alle proprie necessità.

Tutte le funzionalità di Amazon Comprehend accettano documenti di testo UTF-8 come input. Inoltre, per la classificazione personalizzata e il riconoscimento personalizzato delle entità, è possibile utilizzare come input anche file di immagine, file PDF e file Word.

2.1.1 Gli insight di Amazon Comprehend

Amazon Comprehend, analizzando un documento, offre vari insight quali:

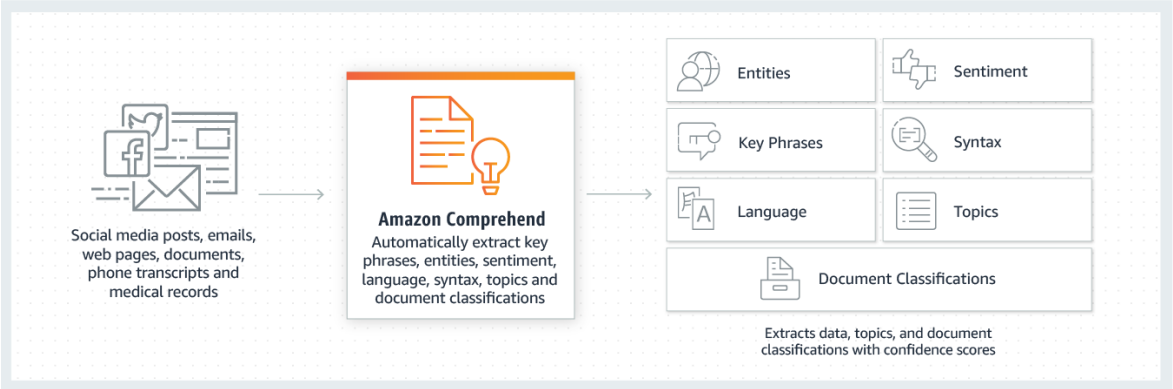


Figura 2.1: Il funzionamento di Amazon Comprehend

- Entità;
- Eventi;
- Frasi chiave;
- Informazioni di identificazione personale (PII);
- Lingua dominante;
- Sentimento;
- Sentimento mirato;
- Analisi della sintassi;

Analisi delle entità

Amazon Comprehend è in grado di svolgere l'analisi delle *entità*. Con questo termine si intendano riferimenti testuali ai nomi effettivi di soggetti del mondo reale, come persone, luoghi e articoli commerciali, nonché riferimenti precisi a misure, come date e quantità. Ad esempio, nel testo 'Giorgio nel 2012 si è trasferito al 156 di Viale Della Vittoria', 'Giorgio' potrebbe essere riconosciuto come una PERSON, quindi persona, '2012' come una DATE, quindi data, e '156 di Viale Della Vittoria' come una LOCATION, quindi luogo.

Per ogni entità, Amazon Comprehend assegna anche un punteggio, chiamato 'Score', che indica il livello di fiducia che il servizio ha riguardo alla correttezza del riconoscimento del tipo di entità. È possibile filtrare le entità con punteggi inferiori per ridurre il rischio di utilizzare rilevamenti errati.

Nelle API, per ottenere l'analisi delle entità, si utilizza il metodo `DetectEntities`, che fornisce un JSON di ritorno strutturato nel seguente modo:

```
1       {
2           "Entities": [
3               {
4                   "Text": "today",
5                   "Score": 0.97,
6                   "Type": "DATE",
7                   "BeginOffset": 14,
8                   "EndOffset": 19
9               },
```

```
                {
                    "Text": "Seattle",
                    "Score": 0.95,
                    "Type": "LOCATION",
                    "BeginOffset": 23,
                    "EndOffset": 30
                }
            ],
            "LanguageCode": "en"
        }
```

Il campo "BeginOffset" indica il punto nel testo in cui inizia la menzione dell'entità, mentre il campo "EndOffset" indica dove termina. Inoltre, il campo "LanguageCode" indica la lingua in cui è stata identificata l'entità. Queste informazioni saranno presenti, anche, in alcune analisi degli insight che vedremo a seguire.

Analisi degli eventi

Amazon Comprehend offre anche la capacità di individuare *eventi*, che rappresentano avvenimenti correlati ad entità specifiche. Ciascun evento restituisce una lista di dettagli; questi sono:

- *Type*: Indica il tipo di evento rilevato.
- *Arguments*: È un elenco di argomenti correlati all'evento rilevato, in cui ogni argomento rappresenta un'entità collegata all'evento. Ogni argomento contiene i seguenti dettagli:
  - *EntityIndex*: È un valore di indice che fa riferimento a un'entità dalla lista di entità ottenuta durante l'analisi.
  - *Role*: Indica il ruolo dell'entità rispetto all'evento.
  - *Score*: Rappresenta il livello di fiducia di Amazon Comprehend nella correttezza del rilevamento dei ruoli.
- *Triggers*: È un elenco di trigger per l'evento rilevato. Un *trigger* è una singola parola o frase che indica il verificarsi dell'evento.
- *BeginOffset*: È un offset di carattere che indica il punto in cui inizia il trigger (il primo carattere è in posizione 0).
- *EndOffset*: È un offset di carattere che indica il punto in cui termina il trigger.
- *Score*: Indica il livello di fiducia di Amazon Comprehend nella precisione del rilevamento del trigger.
- *Text*: Denota il testo del trigger.
- *GroupScore*: Rappresenta il livello di fiducia di Amazon Comprehend nel raggruppamento corretto di questo trigger con altri trigger per lo stesso evento.
- *Type*: È il tipo di evento indicato da questo trigger.

Utilizzando la funzione di *event detection*, possiamo ottenere una lista di entità correlate all'evento stesso, che viene fornita in formato JSON come segue:

```
{
"Entities": [
     {
       "Mentions": [
         {
           "BeginOffset": number,
           "EndOffset": number,
           "Score": number,
           "GroupScore": number,
           "Text": "string",
           "Type": "string"
         }, ...
       ]
     }, ...
   ],
   "Events": [
     {
       "Type": "string",
       "Arguments": [
         {
           "EntityIndex": number,
           "Role": "string",
           "Score": number
         }, ...
       ],
       "Triggers": [
         {
           "BeginOffset": number,
           "EndOffset": number,
           "Score": number,
           "Text": "string",
           "GroupScore": number,
           "Type": "string"
         }, ...
       ]
     }, ...
   ]
 }
```

Analisi delle frasi chiave

Un altro insight analizzabile da Amazon Comprehend è quello delle *frasi chiave*. Con questo termine si indica una stringa contenente una frase sostantiva che descrive una cosa in particolare e rilevante nel documento. Generalmente è composta da un sostantivo e da dei modificatori che lo contraddistinguono. Ad esempio, "giorno" è un sostantivo; "una bella giornata" è una frase sostantiva che include un articolo ("una") e un aggettivo ("bella"). Ogni frase chiave include anche un punteggio che indica il livello di fiducia di Amazon Comprehend nel fatto che la stringa sia una frase chiave. Nelle API, per ottenere l'analisi delle frasi chiave, si utilizza il metodo `DetectKeyPhrases` che fornisce un JSON di ritorno che ha questa struttura:

```
{
    "LanguageCode": "en",
    "KeyPhrases": [
        {
            "Text": "today",
```

```
                    "Score": 0.89,
                    "BeginOffset": 14,
                    "EndOffset": 19
                },
                {
                    "Text": "Seattle",
                    "Score": 0.91,
                    "BeginOffset": 23,
                    "EndOffset": 30
                }
            ]
        }
```

Analisi dei PII

Amazon Comprehend può essere utilizzato anche per individuare entità *PII* (Personally Identifiable Information), cioè informazioni di identificazione personale. Ad esempio, nel testo "Salve Giorgio Colombo. L'ultimo estratto conto della sua carta di credito 1111-0000-1111-0000 è stato spedito al 123 di Via Carlo Goldoni", l'output della analisi sarebbe: "Giorgio Colombo" come tipo NAME, quindi nome, "1111-0000-1111-0000" come tipo CREDIT_NUMBER, quindi numero di carta, e "123 di Via Carlo Goldoni" come tipo ADDRESS, quindi indirizzo. Quindi la risposta in formato JSON riguardante il testo descritto precedentemente sarebbe:

```
        {
            "Entities": [
                {
                    "Score": 0.9999669790267944,
                    "Type": "NAME",
                    "BeginOffset": 6,
                    "EndOffset": 18
                },
                {
                    "Score": 0.8905550241470337,
                    "Type": "CREDIT_DEBIT_NUMBER",
                    "BeginOffset": 69,
                    "EndOffset": 88
                },
                {
                    "Score": 0.9999889731407166,
                    "Type": "ADDRESS",
                    "BeginOffset": 103,
                    "EndOffset": 138
                }
            ]
        }
```

Bisogna considerare, però, che l'esempio riportato sopra è solo rappresentativo poichè, in realtà, l'analisi dei dati *PII* è supportata soltanto per la lingua inglese.

Analisi della lingua dominante

Come si è potuto osservare dagli insight analizzati in precedenza, grazie all'utilizzo delle analisi fornite da Amazon Comprehend, è possibile ottenere informazioni sulla lingua di singole parole. Inoltre, quindi, è altrettanto possibile identificare la *lingua dominante* di

un'intera frase. Per effettuare questa operazione, si utilizza il metodo `DetectSentiment` dell'API, il quale fornisce un risultato in formato JSON che è strutturato come segue:

```
{
    "Languages": [
        {
            "LanguageCode": "en",
            "Score": 0.9793661236763
        }
    ]
}
```

Analisi del sentimento

Un altro insight che viene fornito dall'analisi di Amazon Comprehend è quello del *sentimento*, cioè le emozioni che vengono espresse da una determinata frase. Questa analisi, che potrebbe essere considerata come una *document-level Sentiment Analysis*, fornisce vari valori indicativi presenti all'interno del "SetimentScore":

- *Positive*: che indica con un numero quanto sentimento positivo esprime la frase.
- *Negative*: che indica con un numero quanto sentimento negativo esprime la frase.
- *Mixed*: che indica con un numero quanto sentimento contrastante esprime la frase.
- *Neutral*: che indica con un numero quanto sentimento neutro esprime la frase.

Oltre a questi valori viene anche riportato quale tra loro ha prevalenza maggiore.

Per ottenere questa analisi è necessario utilizzare il metodo dell'API `DetectSentiment` che restituirà un risultato in formato JSON strutturato come segue:

```
{
"SentimentScore": {
        "Mixed": 0.030585512690246105,
        "Positive": 0.94992071056365967,
        "Neutral": 0.0141543131828308,
        "Negative": 0.00893945890665054
    },
    "Sentiment": "POSITIVE",
    "LanguageCode": "en"
}
```

Analisi del sentimento mirato

Un'evoluzione dell'insight precedente è quello del *sentimento mirato*, o *targeted sentiment*, che fornisce una comprensione dettagliata delle emozioni associate alle singole entità presenti nei documenti di input, potendo essere considerato una vera e propria *aspect-level Sentiment Analysis*. Questo tipo di analisi fornisce varie informazioni, ovvero:

- Le identità delle entità presenti nei documenti.
- La classificazione del tipo di entità per ogni sua menzione.
- Il sentimento principale e un punteggio riguardante i singoli tipi di sentimento per ogni menzione di entità.

- I gruppi di menzioni (gruppi di co-riferimento), che rappresentano un insieme di parole che si riferiscono tutte ad una singola entità.

È da tenere presente che questo tipo di analisi è disponibile solo in lingua inglese e che, in formato JSON, è presentata come segue:


```
{"Entities":[
  {
    "DescriptiveMentionIndex": [0],
    "Mentions": [
      {
       "BeginOffset": 0,
       "EndOffset": 1,
       "Score": 0.999997,
       "GroupScore": 1,
       "Text": "I",
       "Type": "PERSON",
       "MentionSentiment": {
         "Sentiment": "NEUTRAL",
         "SentimentScore": {
           "Mixed": 0,
           "Negative": 0,
           "Neutral": 1,
           "Positive": 0
         }
       }
     }
   ]
  }
 ],
 "File": "Input.txt",
 "Line": 0
}
```


Analisi della sintassi

Un ultimo insight fornito da Amazon Comprehend è quello dell'*analisi della sintassi*. Attraverso questa funzione è possibile identificare, per ogni parola del documento, i sostantivi, i verbi, gli aggettivi e tutti gli elementi della sintassi del testo. Ad esempio, nel testo "mia nonna ha spostato il divano", l'analisi di Amazon Comprehend ci mostrerà che i termini "nonna" e "divano" sono sostantivi, mentre "ha spostato" è un verbo. L'analisi di ogni singola parola è restituita in formato JSON nel modo seguente:


```
{
    "SyntaxTokens": [
      {
        "TokenId": 2,
        "Text": "nonna",
        "BeginOffset": 4,
        "EndOffset": 9,
        "PartOfSpeech": {
            "Tag": "NOUN",
            "Score": 0.9898086190223694
        }
      }
   ]
}
```

2.1.2 Come funziona Amazon Comprehend

Con Amazon Comprehend, l'analisi dei testi può essere effettuata attraverso vari strumenti. Quello più semplice ed immediato, il quale viene messo a disposizione direttamente all'interno del sito di Amazon Comprehend, è la console dedicata per analizzare i documenti in tempo reale o eseguire processi di analisi asincrona.

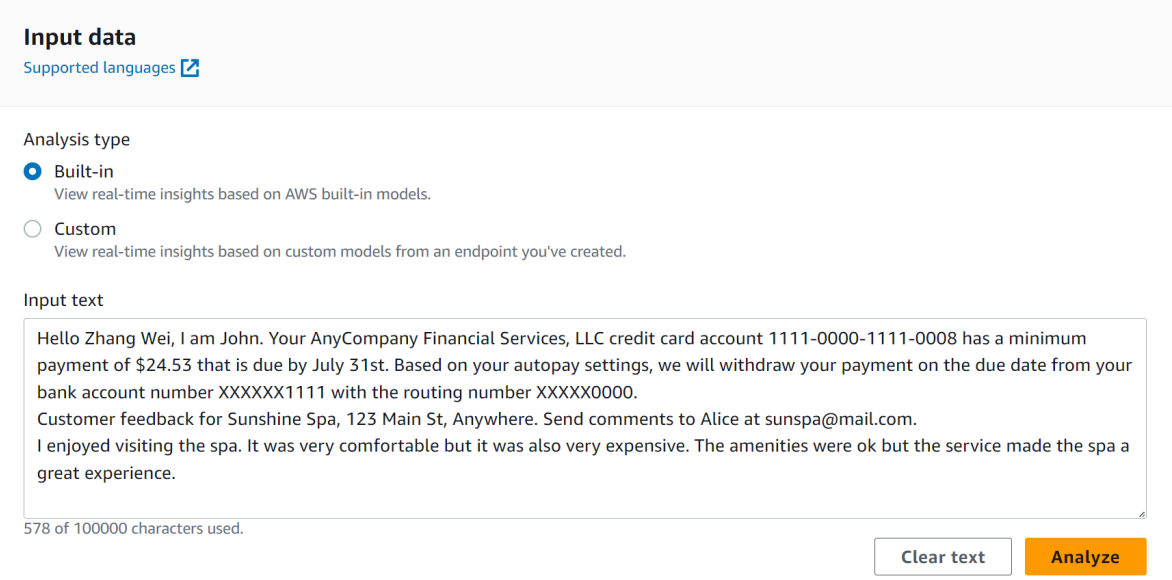


Figura 2.2: La console di Amazon Comprehend

Come si può vedere nella Figura 2.2, che rappresenta la console con il testo di default fornito da AWS, per l'analisi si può scegliere quale tipo di modello utilizzare, se quello fornito direttamente da Amazon, "Built-in", oppure se usare uno progettato dall'utente, "Custom". Il testo da analizzare deve essere inserito all'interno del campo di testo "input text", per eseguire l'analisi bisogna premere il pulsante "Analyze". I risultati dell'analisi vengono riportati in una sezione sottostante il campo di inserimento.

Nella Figura 2.3 si può notare come è strutturata la sezione dedicata ai risultati della console. Nella parte superiore si trovano i pulsanti per passare dall'analisi di un tipo insight ad un altro; nella parte subito sotto, invece, nella sezione "Analyzed text", viene riportato il testo dell'analisi in cui vengono messe in evidenza le parole che sono relative a quale insight si sta analizzando. Queste parole vengono analizzate singolarmente nella sezione "Results", in cui, quindi, si riporta un elenco di esse con tutti i dati.

Come si può vedere dalla figura, al di sotto della sezione "Results", se ne trova un'altra, ovvero la sezione "Application Integration" visualizzata nella Figura 2.4. In questa sezione, vengono fornite informazioni riguardanti il metodo dell'API utilizzato per eseguire l'analisi basata sull'insight. Inoltre, vengono presentate, in formato JSON, sia la chiamata effettuata all'API sia la relativa risposta fornita dall'API stessa.

Un altro modo per poter sfruttare l'analisi di Amazon Comprehend consiste nell'utilizzare direttamente l'API all'interno di vari ambienti quali: l'AWS command line, Java, Python e la piattaforma .NET. In ognuno di questi, quindi, vi sono delle funzioni opportune che semplificano le chiamate all'API stessa.

Un esempio è Python, in cui, sfruttando la libreria `boto3` e le chiavi fornite da AWS, si può effettuare un collegamento diretto all'API per utilizzare i metodi di Amazon

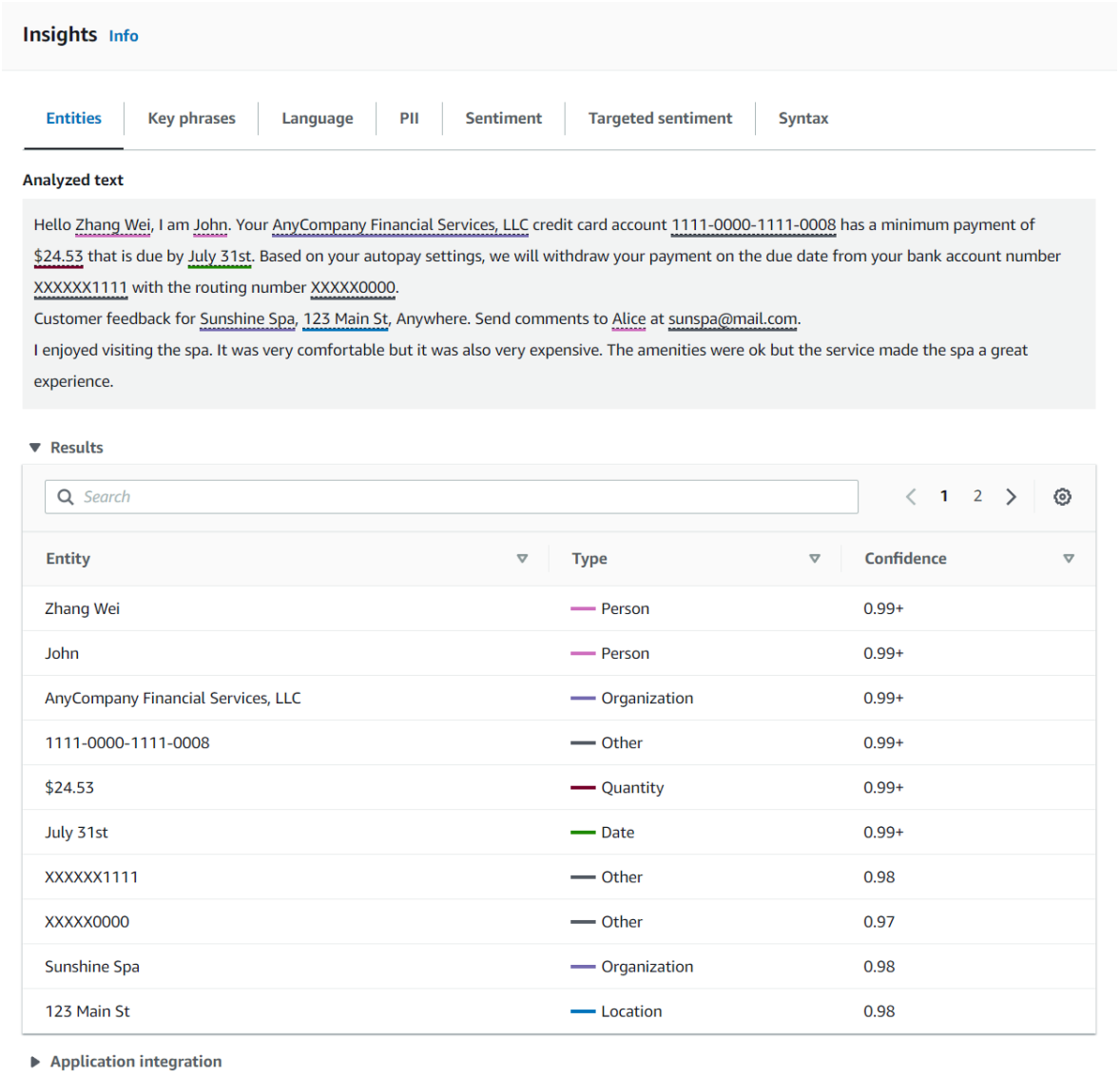


Figura 2.3: La sezione della console di Amazon Comprehend dedicata ai risultati dell'analisi

Comprehend all'interno del codice. Un esempio di utilizzo è quello riportato nel repository al seguente indirizzo:

`https://github.com/Walter-Di-Sabatino/Amazon-Comprehend-Example.git`

Nell'esempio, creando un oggetto `BaseClient` con l'utilizzo di `boto3`, è possibile accedere a tutti i metodi di analisi di Amazon Comprehend che vengono sfruttati per effettuare il `print` dei risultati. Questi metodi in Python sono:

- `detect_dominant_language`
- `detect_entities`
- `detect_key_phrases`
- `detect_sentiment`
- `detect_targeted_sentiment`

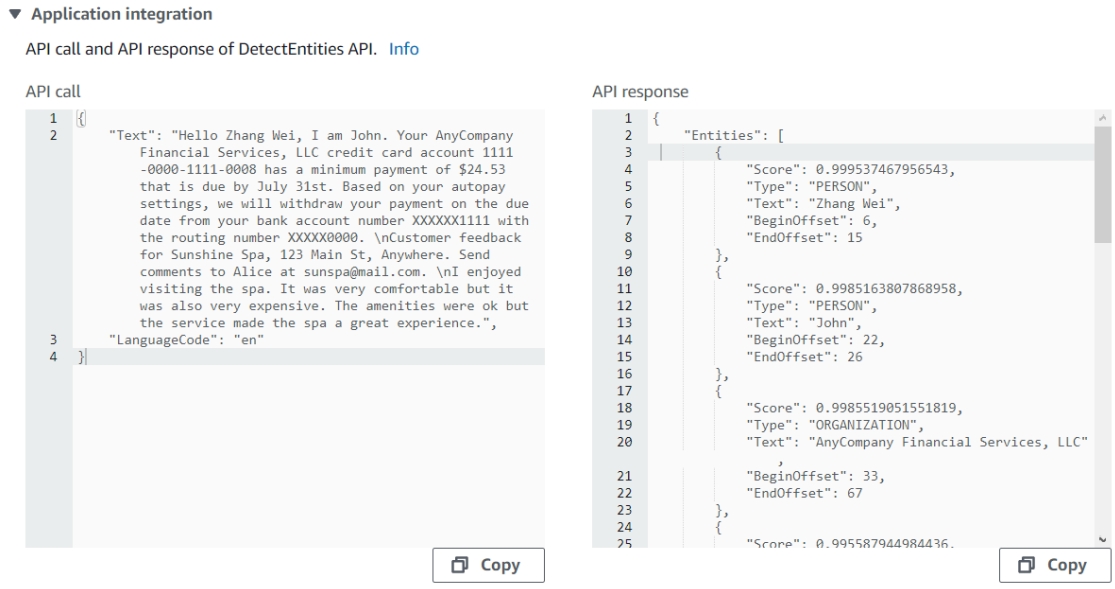


Figura 2.4: La sezione della console di Amazon Comprehend dedicata all'analisi in formato JSON

- `detect_syntax`

Tali metodi, come si può intuire, servono ognuno per analizzare un insight differente.

2.1.3 Esempi con Amazon Comprehend

Al fine di illustrare il funzionamento di Amazon Comprehend è stato scelto il testo seguente in inglese:

> John Smith is a software engineer who lives happly in New York City. He enjoys playing videogames and loves listening to rock music. His favorite book series is 'The Lord of the Rings', and he has a pet Labrador named Max. If you you would like to contact him his email is: fictionalEmail@gmail.com.

È stata selezionata la lingua inglese in modo tale da sfruttare appieno le capacità di analisi di Amazon Comprehend. Di seguito, è riportata la traduzione letterale del testo:

> John Smith è un ingegnere informatico che vive felicemente a New York. Si diverte a giocare ai videogiochi e ama ascoltare la musica rock. La sua serie di libri preferita è "Il Signore degli Anelli" e ha un Labrador di nome Max. Se volete contattarlo, il suo indirizzo e-mail è: fictionalEmail@gmail.com.

Al fine di semplificare la presentazione dei risultati, verranno mostrate soltanto le immagini della sezione "Results" per ogni insight e il file JSON di risposta.

Risultati dell'analisi delle entità

Il risultato dell'analisi delle *entità* all'interno della console è quello riportato all'interno della Figura 2.5; invece l'analisi in formato JSON appare così:

```
{
    "Entities": [
        {
            "Score": 0.9994526505470276,
            "Type": "PERSON",
            "Text": "John Smith",
            "BeginOffset": 0,
            "EndOffset": 10
        },
        {
            "Score": 0.9980171918869019,
            "Type": "LOCATION",
            "Text": "New York City",
            "BeginOffset": 55,
            "EndOffset": 68
        },
        {
            "Score": 0.9946677684783936,
            "Type": "TITLE",
            "Text": "Lord of the Rings",
            "BeginOffset": 164,
            "EndOffset": 181
        },
        {
            "Score": 0.9769556522369385,
            "Type": "PERSON",
            "Text": "Max.",
            "BeginOffset": 212,
            "EndOffset": 216
        },
        {
            "Score": 0.9945738911628723,
            "Type": "OTHER",
            "Text": "fictionalEmail@gmail.com",
            "BeginOffset": 274,
            "EndOffset": 298
        }
    ]
}
```

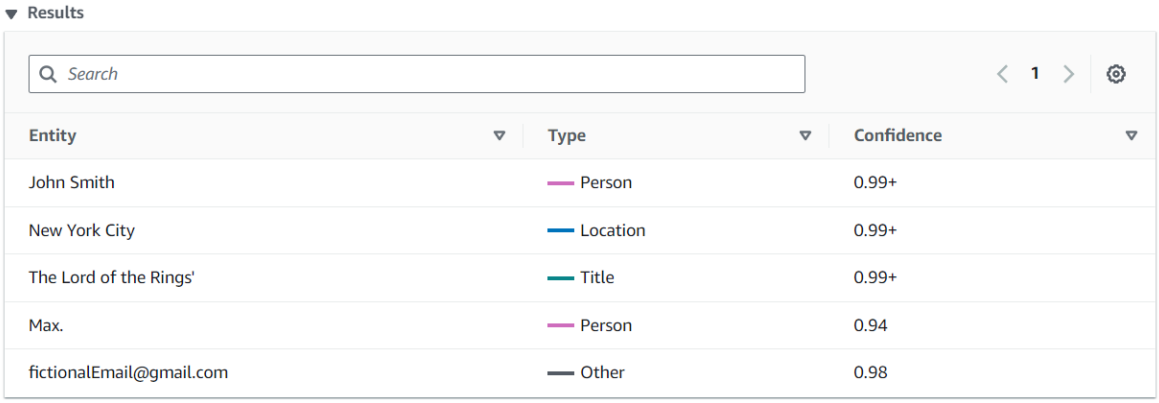



Figura 2.5: I risultati dell'analisi delle entità dell'esempio

Dall'analisi dell'entità è evidente come i risultati siano generalmente accurati, con il servizio in grado di distinguere chiaramente le diverse entità presenti nella frase fornendo

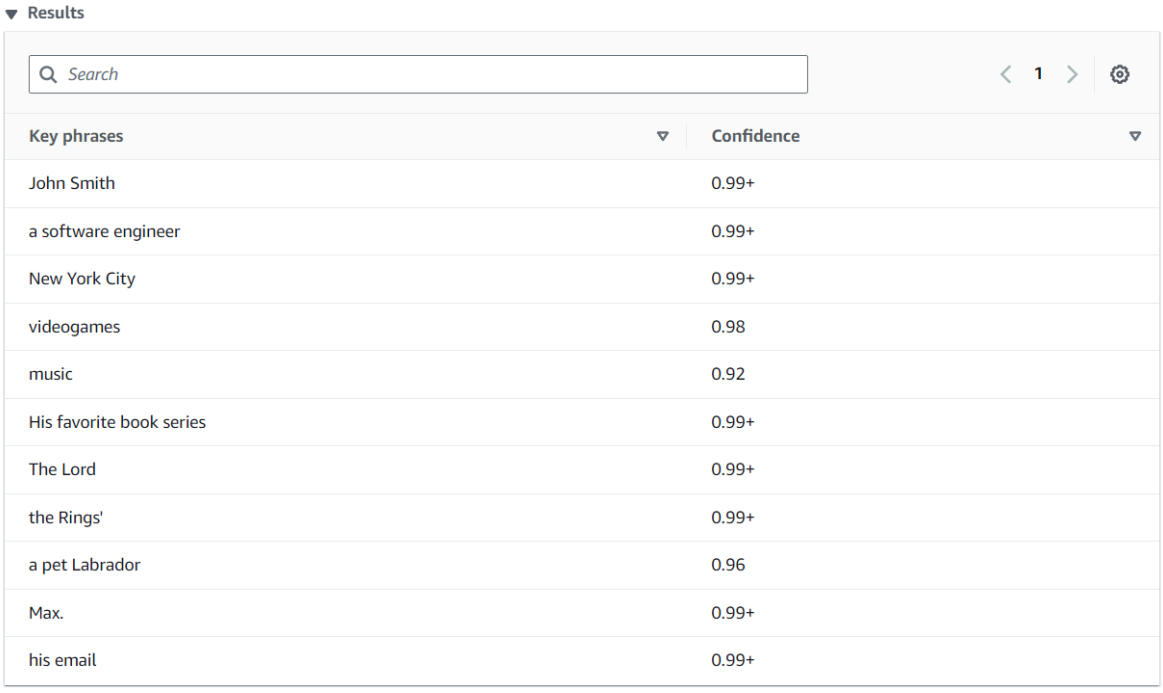
valutazioni di sicurezza molto elevate (SCORE). L'unico risultato che potrebbe essere oggetto di discussione riguarda la parola "Max", identificata come entità di tipo PERSON con uno SCORE di 0,94 su 1, quando, in realtà, potrebbe anche essere classificata come OTHER, poiché rappresenta il nome del cane del protagonista. Tuttavia, anche questa interpretazione potrebbe essere considerata opinabile.

Risultati dell'analisi delle frasi chiave

Il risultato dell'analisi delle *frasi chiave* all'interno della console è quello riportato all'interno della Figura 2.6; invece l'analisi in formato JSON appare in questo modo:


```
{
    "KeyPhrases": [
        {
            "Score": 0.9999722838401794,
            "Text": "John Smith",
            "BeginOffset": 0,
            "EndOffset": 10
        },
        {
            "Score": 0.9999691843986511,
            "Text": "a computer engineer",
            "BeginOffset": 14,
            "EndOffset": 33
        },
        {
            "Score": 0.9999877214431763,
            "Text": "New York City",
            "BeginOffset": 55,
            "EndOffset": 68
        },
        {
            "Score": 0.9990238547325134,
            "Text": "video games",
            "BeginOffset": 88,
            "EndOffset": 99
        },
        {
            "Score": 0.8688954710960388,
            "Text": "music",
            "BeginOffset": 128,
            "EndOffset": 133
        },
        {
            "Score": 0.9988552331924438,
            "Text": "His favorite book series",
            "BeginOffset": 135,
            "EndOffset": 159
        },
        {
            "Score": 0.9928750991821289,
            "Text": "Lord",
            "BeginOffset": 164,
            "EndOffset": 168
        },
        {
            "Score": 0.9996851682662964,
            "Text": "the Rings",
```

```
                "BeginOffset": 172,
                "EndOffset": 181
            },
            {
                "Score": 0.999833881855011,
                "Text": "a Labrador",
                "BeginOffset": 195,
                "EndOffset": 205
            },
            {
                "Score": 0.9981802701950073,
                "Text": "Max.",
                "BeginOffset": 212,
                "EndOffset": 216
            },
            {
                "Score": 0.9996006488800049,
                "Text": "his e-mail address",
                "BeginOffset": 251,
                "EndOffset": 269
            }
        ]
    }
```

Results

Search 1

Key phrases	Confidence
John Smith	0.99+
a software engineer	0.99+
New York City	0.99+
videogames	0.98
music	0.92
His favorite book series	0.99+
The Lord	0.99+
the Rings'	0.99+
a pet Labrador	0.96
Max.	0.99+
his email	0.99+

Figura 2.6: I risultati dell'analisi delle frasi chiave dell'esempio

In questo caso, l'analisi risulta complessivamente accurata, ma vi è un punto discutibile nel risultato riguardante le frasi "The Lord" e "the Rings'". In modo errato, queste due frasi vengono separate come se non facessero parte di un unica frase chiave, e, nell'ultima frase, viene aggiunto erroneamente un apostrofo alla parola "Rings".

Risultati dell'analisi della lingua

Il risultato dell'analisi della *lingua* all'interno della console è quello riportato nella figura 2.7; invece, l'analisi in formato JSON è così rappresentata:

```
{
    "Languages": {
        "LanguageCode": "en",
        "Score": 0.9912170767784119
    }
}
```

Figura 2.7: I risultati dell'analisi della lingua dell'esempio

Qui, con molta facilità, il servizio riesce ad individuare in maniera sicura il linguaggio del testo senza commettere alcun errore e con uno SCORE pressoché pari ad uno.

Risultati dell'analisi dei PII

Il risultato dell'analisi delle *PII* (Informazioni di identificazione personale) all'interno della console è quello riportato all'interno della Figura 2.8; invece l'analisi in formato JSON appare come segue:

```
{
    "Entities": [
        {
            "Score": 0.9999579191207886,
            "Type": "NAME",
            "BeginOffset": 0,
            "EndOffset": 10
        },
        {
            "Score": 0.9999725222587585,
            "Type": "ADDRESS",
            "BeginOffset": 55,
            "EndOffset": 68
        },
        {
            "Score": 0.9999785423278809,
            "Type": "EMAIL",
            "BeginOffset": 274,
            "EndOffset": 298
        }
    ]
}
```

Anche in questo caso il servizio svolge il lavoro in maniera eccellente, riuscendo, inoltre, a distinguere chiaramente l'entità "fictionalEmail@gmail.com" come tipo EMAIL e, quindi, come informazione personale.

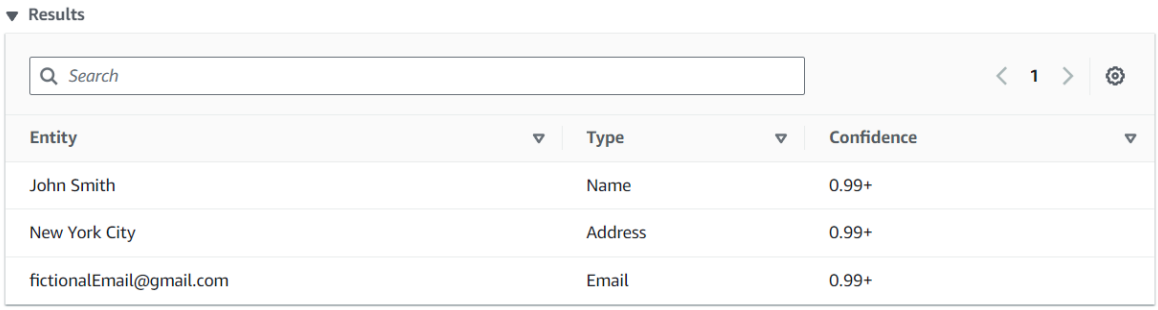


Figura 2.8: I risultati dell'analisi delle PII dell'esempio

Risultati dell'analisi del sentimento

Il risultato dell'analisi del *sentimento* all'interno della console è quello riportato all'interno della Figura 2.9; invece l'analisi in formato JSON appare in maniera seguente:


```
{
    "Sentiment": {
        "Sentiment": "POSITIVE",
        "SentimentScore": {
            "Positive": 0.773431658744812,
            "Negative": 0.0003041046147700399,
            "Neutral": 0.2261437475681305,
            "Mixed": 0.000120512158900965
        }
    }
}
```


Figura 2.9: I risultati dell'analisi del sentimento dell'esempio

Il sentimento predominante nel testo risulta essere positivo, con uno SCORE pari a 0.77 su 1. Al secondo posto si colloca il sentimento neutro, con uno SCORE pari a 0.22. Questo risultato potrebbe essere attribuito alle parti descrittive del testo in cui non emerge chiaramente né un'accentuata positività né negatività. Invece, il sentimento negativo e il sentimento misto hanno uno SCORE che è pari allo zero, visto che il testo riporta una descrizione che emana un sentimento quasi interamente positivo dovuto, principalmente, alla descrizione delle passioni del protagonista.

Risultati dell'analisi del sentimento mirato

Il risultato dell'analisi del *sentimento mirato* all'interno della console è quello riportato all'interno delle Figure 2.10 e 2.11, in cui vengono mostrate anche le menzioni; invece, l'analisi in formato JSON, che non verrà mostrata tutta per brevità, appare nel modo seguente:


```
{
    "Entities": [
        {
```

```
                    "DescriptiveMentionIndex": [
                        2
                    ],
                    "Mentions": [
                        {
                            "Score": 0.9999899864196777,
                            "GroupScore": 0.5184260010719299,
                            "Text": "His",
                            "Type": "PERSON",
                            "MentionSentiment": {
                                "Sentiment": "NEUTRAL",
                                "SentimentScore": {
                                    "Positive": 0,
                                    "Negative": 0,
                                    "Neutral": 1,
                                    "Mixed": 0
                                }
                            },
                            "BeginOffset": 135,
                            "EndOffset": 138
                        },
                        {
                            "Score": 0.9999939799308777,
                            "GroupScore": 0.21274800598621368,
                            "Text": "his",
                            "Type": "PERSON",
                            "MentionSentiment": {
                                "Sentiment": "NEUTRAL",
                                "SentimentScore": {
                                    "Positive": 0,
                                    "Negative": 0,
                                    "Neutral": 1,
                                    "Mixed": 0
                                }
                            },
                            "BeginOffset": 251,
                            "EndOffset": 254
                        },
                        {
                            "Score": 0.9999380111694336,
                            "GroupScore": 1,
                            "Text": "John Smith",
                            "Type": "PERSON",
                            "MentionSentiment": {
                                "Sentiment": "NEUTRAL",
                                "SentimentScore": {
                                    "Positive": 0.0027379998937249184,
                                    "Negative": 0.00002400000848900527,
                                    "Neutral": 0.9972190260887146,
                                    "Mixed": 0.00018999999156221747
                                }
                            },
                            "BeginOffset": 0,
                            "EndOffset": 10
                        },
                        ...
                    ]
                },
                {
```

```
                    "DescriptiveMentionIndex": [
                        0
                    ],
                    "Mentions": [
                        {
                            "Score": 0.9517210125923157,
                            "GroupScore": 1,
                            "Text": "engineer",
                            "Type": "PERSON",
                            "MentionSentiment": {
                                "Sentiment": "NEUTRAL",
                                "SentimentScore": {
                                    "Positive": 0.00000900000318337698,
                                    "Negative": 0.00000300000106112566,
                                    "Neutral": 0.9999120235443115,
                                    "Mixed": 0.0000759999662488699
                                }
                            },
                            "BeginOffset": 25,
                            "EndOffset": 33
                        }
                    ]
                },
                {
                    "DescriptiveMentionIndex": [
                        0
                    ],
                    "Mentions": [
                        {
                            "Score": 0.9997259974479675,
                            "GroupScore": 1,
                            "Text": "New York City",
                            "Type": "LOCATION",
                            "MentionSentiment": {
                                "Sentiment": "NEUTRAL",
                                "SentimentScore": {
                                    "Positive": 0.000006000000212225132,
                                    "Negative": 0.000006000000212225132,
                                    "Neutral": 0.9999859929084778,
                                    "Mixed": 9.999999974752427e-7
                                }
                            },
                            "BeginOffset": 55,
                            "EndOffset": 68
                        }
                    ]
                },
                ...
            ]
        }
```


Questa analisi rappresenta un passo ulteriore rispetto a quella precedente, in cui abbiamo puntualmente evidenziato i sentimenti espressi dalle principali entità presenti nel testo, comprese le menzioni specifiche delle singole entità.

Entity	Entity type	Entity score	Primary sentiment	Positive score	Negative scc
fictionalEmail@gmail.com (2)	OTHER	-	NEUTRAL	-	-
John Smith (6)	PERSON	-	NEUTRAL	-	-
Labrador	OTHER	0.99+	NEUTRAL	0.00	0
Lord of the Rings	MOVIE	0.71	NEUTRAL	0.00	0.00
Max.	PERSON	0.99+	POSITIVE	0.79	0.00
music	OTHER	0.93	NEUTRAL	0.00	0.00
New York City	LOCATION	0.99+	NEUTRAL	0.00	0.00
video games	OTHER	0.66	NEUTRAL	0.00	0.00
you	PERSON	0.99+	NEUTRAL	0	0

Figura 2.10: I risultati dell'analisi del sentimento mirato con menzioni dell'esempio

Entity	Entity type	Entity score	Primary sentiment	Positive score	Negative scc
fictionalEmail@gmail.com (2)	OTHER	-	NEUTRAL	-	-
e-mail address	OTHER	0.99+	NEUTRAL	0.00	0.00
fictionalEmail@gmail.com	OTHER	0.73	NEUTRAL	0.00	0.00
John Smith (6)	PERSON	-	NEUTRAL	-	-
he	PERSON	0.99+	NEUTRAL	0	0
He	PERSON	0.99+	NEUTRAL	0.00	0
him	PERSON	0.99+	NEUTRAL	0	0
his	PERSON	0.99+	NEUTRAL	0	0
His	PERSON	0.99+	NEUTRAL	0	0
John Smith	PERSON	0.99+	NEUTRAL	0.00	0.00

Figura 2.11: I risultati dell'analisi del sentimento mirato con menzioni dell'esempio

Risultati dell'analisi della sintassi

Ultima analisi fornita da Amazon Comprehend, che non verrà mostrata tutta per brevità, è quella della *sintassi* all'interno della console. Essa è riportata nella Figura 2.10; invece l'analisi in formato JSON appare in maniera seguente:


```
{
    "SyntaxTokens": [
        {
            "TokenId": 1,
            "Text": "John",
            "BeginOffset": 0,
            "EndOffset": 4,
            "PartOfSpeech": {
                "Tag": "PROPN",
                "Score": 0.9999986290931702
            }
        },
```

```
            {
                "TokenId": 2,
                "Text": "Smith",
                "BeginOffset": 5,
                "EndOffset": 10,
                "PartOfSpeech": {
                    "Tag": "PROPN",
                    "Score": 0.999944806098938
                }
            },
            {
                "TokenId": 3,
                "Text": "is",
                "BeginOffset": 11,
                "EndOffset": 13,
                "PartOfSpeech": {
                    "Tag": "VERB",
                    "Score": 0.9995461702346802
                }
            },
            {
                "TokenId": 4,
                "Text": "a",
                "BeginOffset": 14,
                "EndOffset": 15,
                "PartOfSpeech": {
                    "Tag": "DET",
                    "Score": 0.9999964237213135
                }
            },
            {
                "TokenId": 5,
                "Text": "computer",
                "BeginOffset": 16,
                "EndOffset": 24,
                "PartOfSpeech": {
                    "Tag": "NOUN",
                    "Score": 0.9968807101249695
                }
            },
            {
                "TokenId": 6,
                "Text": "engineer",
                "BeginOffset": 25,
                "EndOffset": 33,
                "PartOfSpeech": {
                    "Tag": "NOUN",
                    "Score": 0.9998804330825806
                }
            },
            ...
        ]
    }
```


In questo caso, quindi, Amazon Comprehend svolge l'analisi della sintassi, esaminando ogni singola parola in maniera puntuale e senza commettere errori.

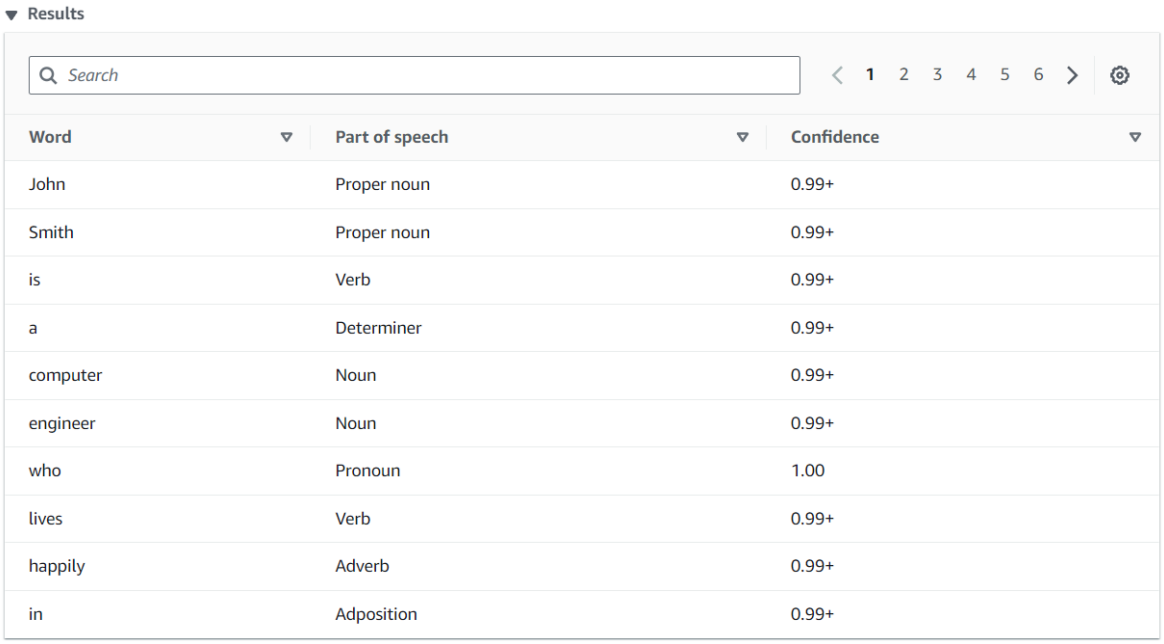


Figura 2.12: I risultati dell'analisi della sintassi dell'esempio

2.2 Cos'è Amazon Comprehend Medical

Amazon Comprehend Medical è un servizio avanzato di natural language processing (NLP) sviluppato da Amazon.com, Inc. che sfrutta l'apprendimento automatico per estrarre informazioni sanitarie da testi medici non strutturati.

Questo servizio offre la possibilità di individuare e comprendere dati rilevanti all'interno dei testi clinici, consentendo, ad esempio, di migliorare notevolmente la farmacovigilanza. Quest'ultima consiste nella capacità di monitorare gli effetti collaterali dei farmaci e valutarne l'efficacia terapeutica dopo il loro rilascio sul mercato, analizzando test clinici o recensioni.

Inoltre, Amazon Comprehend Medical semplifica e rende più preciso il monitoraggio della risposta dei pazienti a specifiche terapie, permettendo di analizzare le loro esperienze direttamente dai testi clinici. Questo approccio fornisce ai ricercatori una prospettiva più approfondita delle reazioni dei pazienti, migliorando così la comprensione dei trattamenti e aumentando le possibilità di adattare le terapie in modo personalizzato.

Il servizio consente inoltre a medici e operatori sanitari di gestire e accedere agevolmente alle informazioni mediche che non si adattano ai moduli tradizionali. Grazie alla possibilità per i pazienti di riportare i propri problemi di salute fornendo una maggiore quantità di informazioni rispetto ai moduli standard, le organizzazioni possono individuare i candidati che necessitano di uno screening precoce delle proprie condizioni mediche, evitando che queste si complichino e diventino più costose da trattare.

Amazon Comprehend Medical, sfortunatamente supporta solamente la lingua inglese.

2.2.1 Gli insights di Amazon Comprehend Medical

Con l'analisi di testi clinici Amazon Comprehend Medical offre vari insights quali:

1. Entità
2. RxNorm concepts

3. ICD-10-CM concepts
4. SNOMED CT concepts

Analisi delle entità

Quindi, come nel servizio base di Amazon Comprehend, anche in Amazon Comprehend Medical vi è la possibilità di analizzare *entità*, intese come riferimenti testuali a soggetti e oggetti rilevanti, attraverso la funzione `DetectEntities`. Con questa operazione si individuano quindi varie informazioni riguardanti le entità, quali:

1. *Category*: che indica il raggruppamento generalizzato a cui appartiene un'entità rilevata. Ad esempio, "Ibuprofene" fa parte della categoria MEDICATION.
2. *Type*: che indica il tipo di entità rilevata, che ha come ambito una categoria. Ad esempio, "Ibuprofene" rientra nel tipo GENERIC_NAME della categoria MEDICATION.
3. *Attribute*: che indica informazioni relative a un'entità rilevata, come il dosaggio di un farmaco. Ad esempio, "200 mg" è un attributo dell'entità "Ibuprofene"
4. *Trait*: che indica qualcosa che Amazon Comprehend Medical capisce di un'entità, in base al contesto. Ad esempio, un farmaco ha il tratto NEGATION se il paziente non lo assume.

In questa analisi, ogni entità e attributo è accompagnato da un punteggio che riflette il livello di fiducia attribuito da Amazon Comprehend Medical all'accuratezza del rilevamento. Inoltre, ogni attributo è dotato di un punteggio di relazione, che indica il grado di certezza di Amazon Comprehend Medical riguardo alla corretta connessione tra l'attributo e la sua entità madre.

Inoltre, per ciascuna entità e attributo, sono forniti i dati di offset, che consentono di individuare la posizione esatta delle singole parole nel testo analizzato. Questi dati di offset risultano essenziali per comprendere la contestualizzazione e l'organizzazione delle informazioni estratte dal testo.

Il risultato di questa analisi in formato JSON per la frase "Severe Rash on face, slightly itchy." è del tipo:


```
{
    "Entities": [
        {
            "Id": 3,
            "BeginOffset": 8,
            "EndOffset": 12,
            "Score": 0.8499777317047119,
            "Text": "Rash",
            "Category": "MEDICAL_CONDITION",
            "Type": "DX_NAME",
            "Traits": [],
            "Attributes": [
                {
                    "Type": "QUALITY",
                    "Score": 0.8698788285255432,
                    "RelationshipScore": 1,
                    "RelationshipType": "QUALITY",
```

```
                            "Id": 2,
                            "BeginOffset": 1,
                            "EndOffset": 7,
                            "Text": "Severe",
                            "Category": "MEDICAL_CONDITION",
                            "Traits": []
                        },
                        {
                            "Type": "SYSTEM_ORGAN_SITE",
                            "Score": 0.9949906468391418,
                            "RelationshipScore": 0.9941522479057312,
                            "RelationshipType": "SYSTEM_ORGAN_SITE",
                            "Id": 1,
                            "BeginOffset": 17,
                            "EndOffset": 21,
                            "Text": "face",
                            "Category": "ANATOMY",
                            "Traits": []
                        }
                    ]
                },
                {
                    "Id": 1,
                    "BeginOffset": 17,
                    "EndOffset": 21,
                    "Score": 0.9949906468391418,
                    "Text": "face",
                    "Category": "ANATOMY",
                    "Type": "SYSTEM_ORGAN_SITE",
                    "Traits": []
                }
            ],
            "UnmappedAttributes": [
                {
                    "Type": "MEDICAL_CONDITION",
                    "Attribute": {
                        "Type": "QUALITY",
                        "Score": 0.5131389498710632,
                        "Id": 4,
                        "BeginOffset": 23,
                        "EndOffset": 37,
                        "Text": "slightly itchy",
                        "Category": "MEDICAL_CONDITION",
                        "Traits": []
                    }
                }
            ],
            "ModelVersion": "2.4.0"
        }
```


Analisi secondo il servizio InferRxNorm

Amazon Comprehend Medical offre anche la possibilità di analizzare i testi clinici tramite il servizio *InferRxNorm*. Questo servizio è progettato per identificare i farmaci menzionati nel testo clinico come entità e associarli agli identificatori concettuali normalizzati (RxCUI) presenti nel database RxNorm della US National Library of Medicine.

Utilizzando InferRxNorm, è possibile, quindi, ottenere una comprensione più accurata e strutturata delle informazioni riguardanti i farmaci presenti nei testi clinici, semplificando così la ricerca, l'analisi e l'elaborazione dei dati medici.

La risposta in formato JSON per la frase "patient is not on warfarin" in questo tipo di analisi è la seguente:


```
{
    "Entities": [
        {
            "Id": 1,
            "Text": "warfarin",
            "Category": "MEDICATION",
            "Type": "GENERIC_NAME",
            "Score": 0.9970192909240723,
            "BeginOffset": 18,
            "EndOffset": 26,
            "Attributes": [],
            "Traits": [
                {
                    "Name": "NEGATION",
                    "Score": 0.8079015016555786
                }
            ],
            "RxNormConcepts": [
                {
                    "Description": "warfarin",
                    "Code": "11289",
                    "Score": 0.9439865350723267
                },
                {
                    "Description": "warfarin sodium 2 MG Oral Tablet",
                    "Code": "855302",
                    "Score": 0.5045595169067383
                },
                {
                    "Description": "warfarin sodium 10 MG Oral Tablet",
                    "Code": "855296",
                    "Score": 0.40246912837028503
                },
                {
                    "Description": "warfarin sodium 2 MG Oral Tablet [Coumadin]",
                    "Code": "855304",
                    "Score": 0.22325271368026733
                },
                {
                    "Description": "warfarin sodium 10 MG Oral Tablet [Jantoven]",
                    "Code": "855300",
                    "Score": 0.13163453340530396
                }
            ]
        }
    ],
    "ModelVersion": "2.2.0.20221003"
}
```

Analisi secondo il servizio InferICD10CM

Un altro insight fornito da Amazon Comprehend Medical è rappresentato dal servizio *InferICD10CM*. Questo strumento ha la capacità di identificare potenziali condizioni mediche all'interno del testo, trattandole come entità e collegandole a codici univoci presenti nella versione 2019 della International Classification of Diseases, 10th Revision, Clinical Modification (ICD-10-CM) (una classificazione medica sviluppata dall'Organizzazione Mondiale della Sanità). Il principale obiettivo di questo processo è garantire la massima precisione nelle diagnosi, facilitando così il lavoro dei professionisti della salute e migliorando la qualità delle cure mediche fornite.

La struttura della risposta in formato JSON utilizzando questo servizio è la seguente:


```
{
    "Entities": [
        {
            "Id": 1,
            "Text": "abdominal pain",
            "Category": "MEDICAL_CONDITION",
            "Type": "DX_NAME",
            "Score": 0.9606665968894958,
            "BeginOffset": 153,
            "EndOffset": 167,
            "Attributes": [
                {
                    "Type": "ACUITY",
                    "Score": 0.764342725276947,
                    "RelationshipScore": 0.9999940395355225,
                    "Id": 2,
                    "BeginOffset": 183,
                    "EndOffset": 193,
                    "Text": "persistent",
                    "Traits": []
                }
            ],
            "Traits": [
                {
                    "Name": "SYMPTOM",
                    "Score": 0.7559975981712341
                }
            ],
            "ICD10CMConcepts": [
                {
                    "Description": "Unspecified abdominal pain",
                    "Code": "R10.9",
                    "Score": 0.7775180339813232
                },
                {
                    "Description": "Epigastric pain",
                    "Code": "R10.13",
                    "Score": 0.6876822710037231
                },
                {
                    "Description": "Lower abdominal pain, unspecified",
                    "Code": "R10.30",
                    "Score": 0.6758853197097778
                },
                {
                    "Description": "Generalized abdominal pain",
```

```
                            "Code": "R10.84",
                            "Score": 0.6746202707290649
                        },
                        {
                            "Description": "Upper abdominal pain, unspecified",
                            "Code": "R10.10",
                            "Score": 0.6702126860618591
                        }
                    ]
                }
        ...
            "ModelVersion": "2.5.0.20220401"
        }
```


Analisi secondo il servizio InferSNOMEDCT

Un ulteriore insight fornito da Amazon Comprehend Medical è il servizio *InferSNOMEDCT*. Questo strumento è in grado di individuare concetti medici, come condizioni mediche, anatomia, test medici o trattamenti e procedure, trattandoli come entità e collegandoli ai codici del Systematized Nomenclature of Medicine, Clinical Terms (SNOMED CT). Si tratta di un vocabolario standardizzato e multilingue di terminologia clinica utilizzato da medici e altri operatori sanitari per lo scambio elettronico di informazioni cliniche sulla salute.

L'uso del servizio InferSNOMEDCT permette, quindi, una migliore comprensione e organizzazione delle informazioni cliniche, facilitando la comunicazione tra i professionisti della salute e contribuendo a un'analisi più accurata dei dati medici. Questo aiuta a migliorare l'efficienza delle cure mediche e a fornire un servizio più completo e personalizzato ai pazienti.

La struttura della risposta in formato JSON utilizzando questo servizio è la seguente:


```
{
    "Entities": [
        {
            "Category": "ANATOMY",
            "BeginOffset": 0,
            "EndOffset": 5,
            "Text": "HEENT",
            "Traits": [],
            "SNOMEDCTConcepts": [
                {
                    "Code": "69536005",
                    "Score": 0.8183674812316895,
                    "Description": "Head structure (body structure)"
                },
                {
                    "Code": "429031000124106",
                    "Score": 0.8062137961387634,
                    "Description": "Review of systems, head, ear, eyes, nose and
                        throat (procedure)"
                },
                {
                    "Code": "385383008",
                    "Score": 0.7023276090621948,
                    "Description": "Ear, nose and throat structure (body structure
                        )"
                },
```

```
                        {
                            "Code": "64237003",
                            "Score": 0.6886451840400696,
                            "Description": "Structure of left half of head (body structure
                                )"
                        },
                        {
                            "Code": "113028003",
                            "Score": 0.6595167517662048,
                            "Description": "Ear, nose and throat examination (procedure)"
                        }
                    ],
                    "Score": 0.9941003918647766,
                    "Attributes": [],
                    "Type": "SYSTEM_ORGAN_SITE",
                    "Id": 0
                }
            ],
            "SNOMEDCTDetails": {
                "Edition": "US",
                "VersionDate": "20200901",
                "Language": "en"
            },
            "Characters": {
                "OriginalTextCharacters": 59
            },
            "ModelVersion": "2.6.0.20220301"
        }
```


2.2.2 Come funziona Amazon Comprehend Medical

Amazon Comprehend Medical mette a disposizioni vari strumenti al fine di effettuare l'analisi dei testi medici non strutturati. Quello più immediato è la console che viene fornita direttamente all'interno del sito di Amazon Comprehend Medical.

Nella Figura 2.13, possiamo osservare la sezione della console dedicata all'inserimento del testo da analizzare, dove è incluso di default un esempio fornito direttamente da AWS. Per ottenere i risultati desiderati, è sufficiente premere il pulsante "Analyze", il quale avvierà l'analisi in tempo reale.

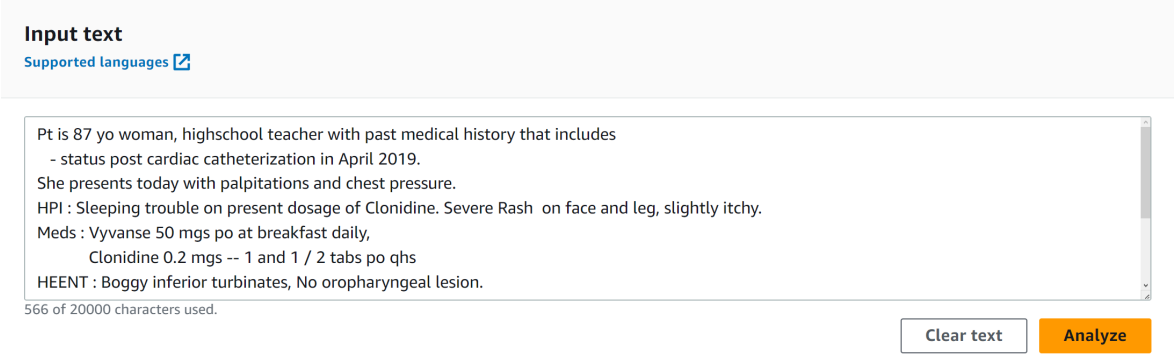


Figura 2.13: La console di Amazon Comprehend Medical

Subito sotto la sezione illustrata nella Figura 2.13, troviamo la sezione corrispondente alla Figura 2.14. Qui sono presenti i pulsanti che consentono di passare dall'analisi di un

insight all'altro; viene anche visualizzato il testo analizzato, dove tutte le parole rilevanti rispetto all'analisi sono evidenziate e collegate.

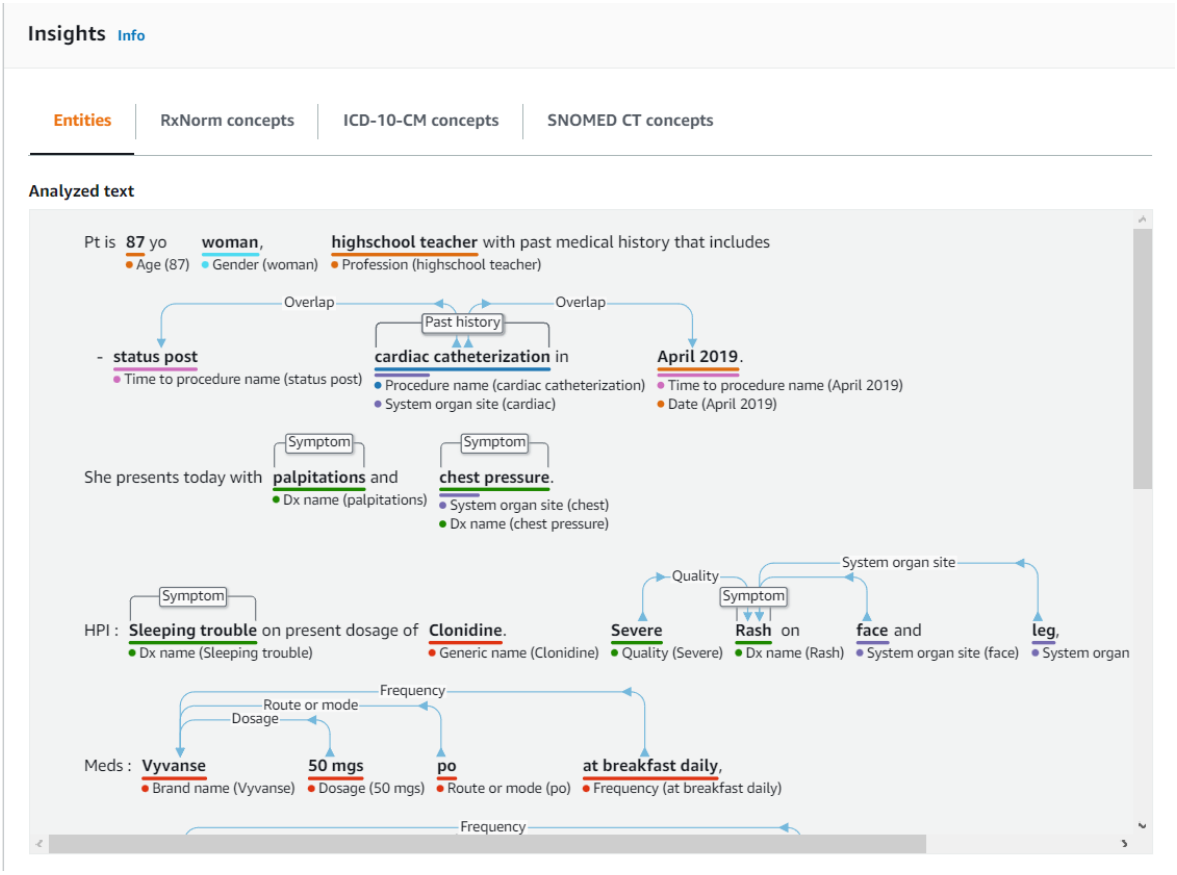


Figura 2.14: Il testo analizzato di Amazon Comprehend Medical

Nella Figura 2.15 viene, invece, riportata la sezione della console in cui sono presenti tutte le parole rilevanti rispetto all'analisi e in cui, quindi, sono riportati i valori dell'*entity*, dello *score*, della *category*, del *type* e del *traits*.

Al di sotto della sezione appena descritta, è presente un'area specifica, riportata in Figura 2.16, dedicata all'analisi in formato JSON. Qui troviamo sia la chiamata all'API in formato JSON che la risposta dell'API nello stesso formato.

Un altro modo per interagire con Amazon Comprehend Medical è attraverso l'uso diretto dell'API, utilizzando funzioni disponibili in AWS command line, Java e Python. Questi strumenti sono utili per lo sviluppo di applicazioni che sfruttino le funzionalità di Amazon Comprehend Medical.

2.2.3 Esempi con Amazon Comprehend Medical

Al fine di illustrare il funzionamento di Amazon Comprehend Medical è stato scelto il testo in inglese:

> Patient John Smith, 45-year-old male, presented with severe headaches and dizziness at the ER. Noted family history of migraines. Mild hypertension observed during the physical exam. Further lab tests requested to determine the underlying cause of symptoms. Administered acetaminophen for immediate pain relief. Goal: identify the underlying cause and establish an appropriate

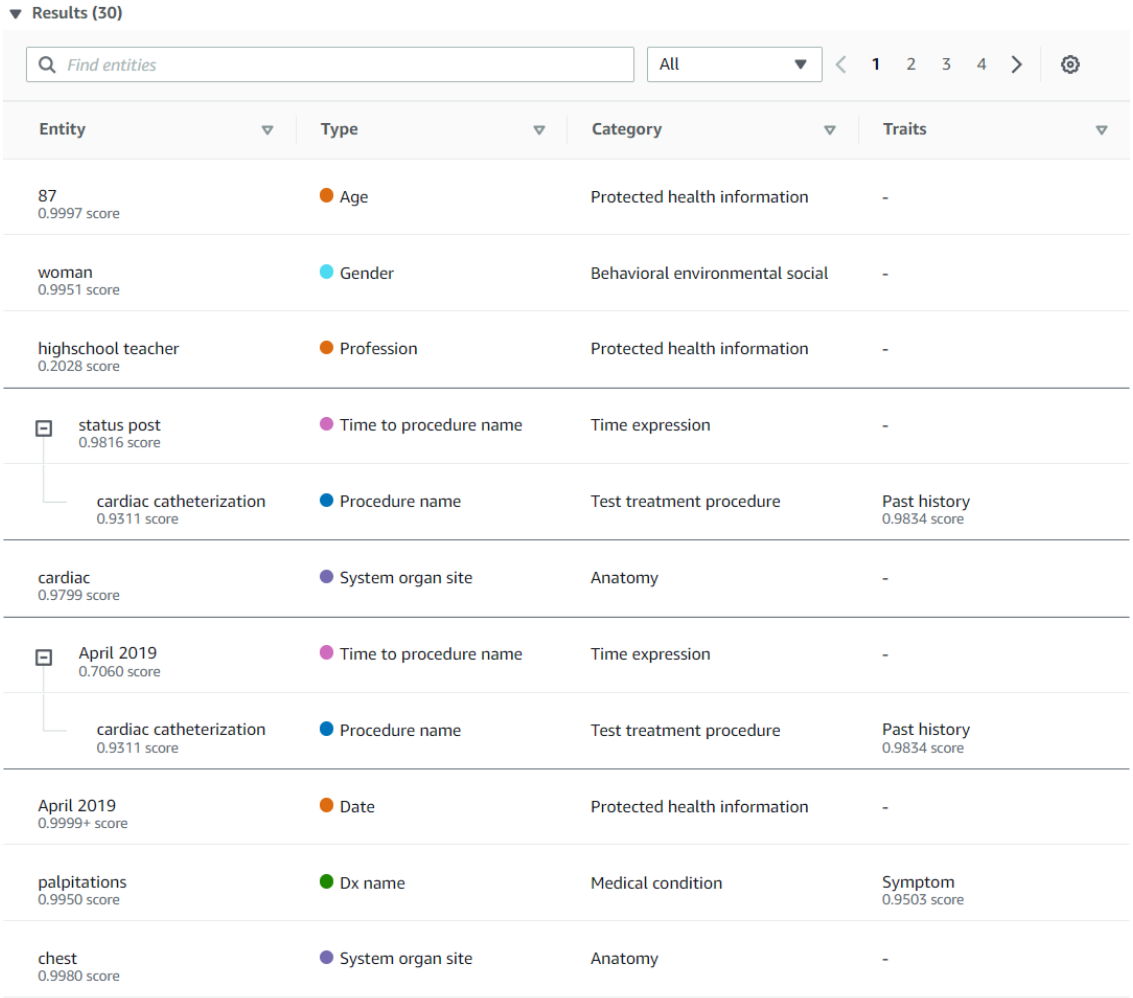


Figura 2.15: I risultati dell'analisi di Amazon Comprehend Medical

> treatment plan, which may involve prescription medications such as triptans or analgesics.

È stata selezionata la lingua inglese poiché il servizio non supporta altre lingue. Di seguito, è riportata la traduzione letterale del testo:

> Il paziente John Smith, uomo di 45 anni, si è presentato al Pronto Soccorso con forti mal di testa e vertigini. Anamnesi familiare di emicrania. Durante l'esame fisico è stata osservata una lieve ipertensione. Sono stati richiesti ulteriori esami di laboratorio per determinare la causa dei sintomi. Somministrazione di acetaminofene per alleviare immediatamente il dolore. Obiettivo: identificare la causa sottostante e stabilire un piano di trattamento appropriato, che può comportare la prescrizione di farmaci come triptani o analgesici.

Al fine di semplificare la presentazione dei risultati, verranno mostrate soltanto le immagini della sezione "Results" per ogni insight e il file JSON di risposta.

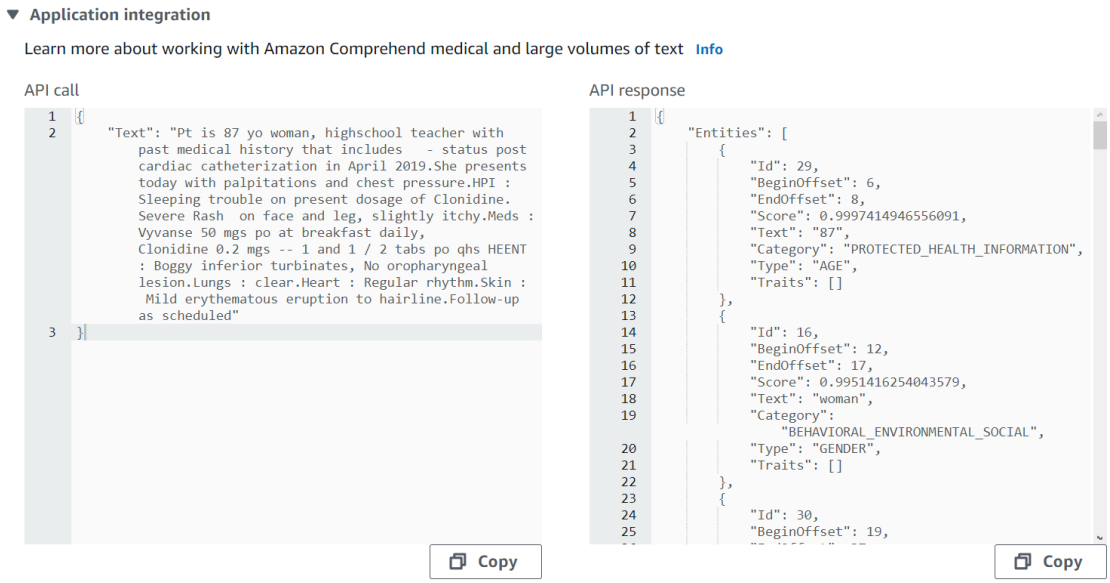


Figura 2.16: I risultati in JSON di Amazon Comprehend Medical

Risultati dell'analisi delle entità

Il risultato dell'analisi delle *entità* all'interno della console è quello riportato nelle Figure, 2.17 e 2.18; invece l'analisi in formato JSON, che, per brevità, non verrà mostrata tutta, appare così:


```
{
    "Entities": [
        {
            "Id": 4,
            "BeginOffset": 8,
            "EndOffset": 18,
            "Score": 0.9985851049423218,
            "Text": "John Smith",
            "Category": "PROTECTED_HEALTH_INFORMATION",
            "Type": "NAME",
            "Traits": []
        },
        {
            "Id": 5,
            "BeginOffset": 22,
            "EndOffset": 24,
            "Score": 0.9999154806137085,
            "Text": "45",
            "Category": "PROTECTED_HEALTH_INFORMATION",
            "Type": "AGE",
            "Traits": []
        },
        {
            "Id": 7,
            "BeginOffset": 34,
            "EndOffset": 38,
            "Score": 0.9995986819267273,
            "Text": "male",
            "Category": "BEHAVIORAL_ENVIRONMENTAL_SOCIAL",
```

```
                    "Type": "GENDER",
                    "Traits": []
                },
                {
                    "Id": 11,
                    "BeginOffset": 62,
                    "EndOffset": 71,
                    "Score": 0.9928064346313477,
                    "Text": "headaches",
                    "Category": "MEDICAL_CONDITION",
                    "Type": "DX_NAME",
                    "Traits": [
                        {
                            "Name": "SYMPTOM",
                            "Score": 0.9771478176116943
                        }
                    ],
                    "Attributes": [
                        {
                            "Type": "QUALITY",
                            "Score": 0.9884397983551025,
                            "RelationshipScore": 1,
                            "RelationshipType": "QUALITY",
                            "Id": 10,
                            "BeginOffset": 55,
                            "EndOffset": 61,
                            "Text": "severe",
                            "Category": "MEDICAL_CONDITION",
                            "Traits": []
                        }
                    ]
                },
                ...
            "ModelVersion": "2.4.0"
        }
```


Amazon Comprehend Medical fornisce risultati accurati, riuscendo a distinguere tra condizioni mediche e trattamenti. In questo caso, è stata individuata un'unica parola errata, "ER.", a cui erroneamente è stato attribuito un punto finale e che è stata identificata come "ADDRESS", invece di essere classificata correttamente come "FACILITY".

Risultati dell'analisi secondo il servizio InferRxNorm

Il risultato dell'analisi rispetto al servizio *InferRxNorm* all'interno della console è quello riportato in Figura 2.19; invece l'analisi in formato JSON appare così:


```
        {
            "Entities": [
                {
                    "Id": 1,
                    "Text": "acetaminophen",
                    "Category": "MEDICATION",
                    "Type": "GENERIC_NAME",
                    "Score": 0.9921550154685974,
                    "BeginOffset": 274,
                    "EndOffset": 287,
                    "Attributes": [],
                    "Traits": [],
```

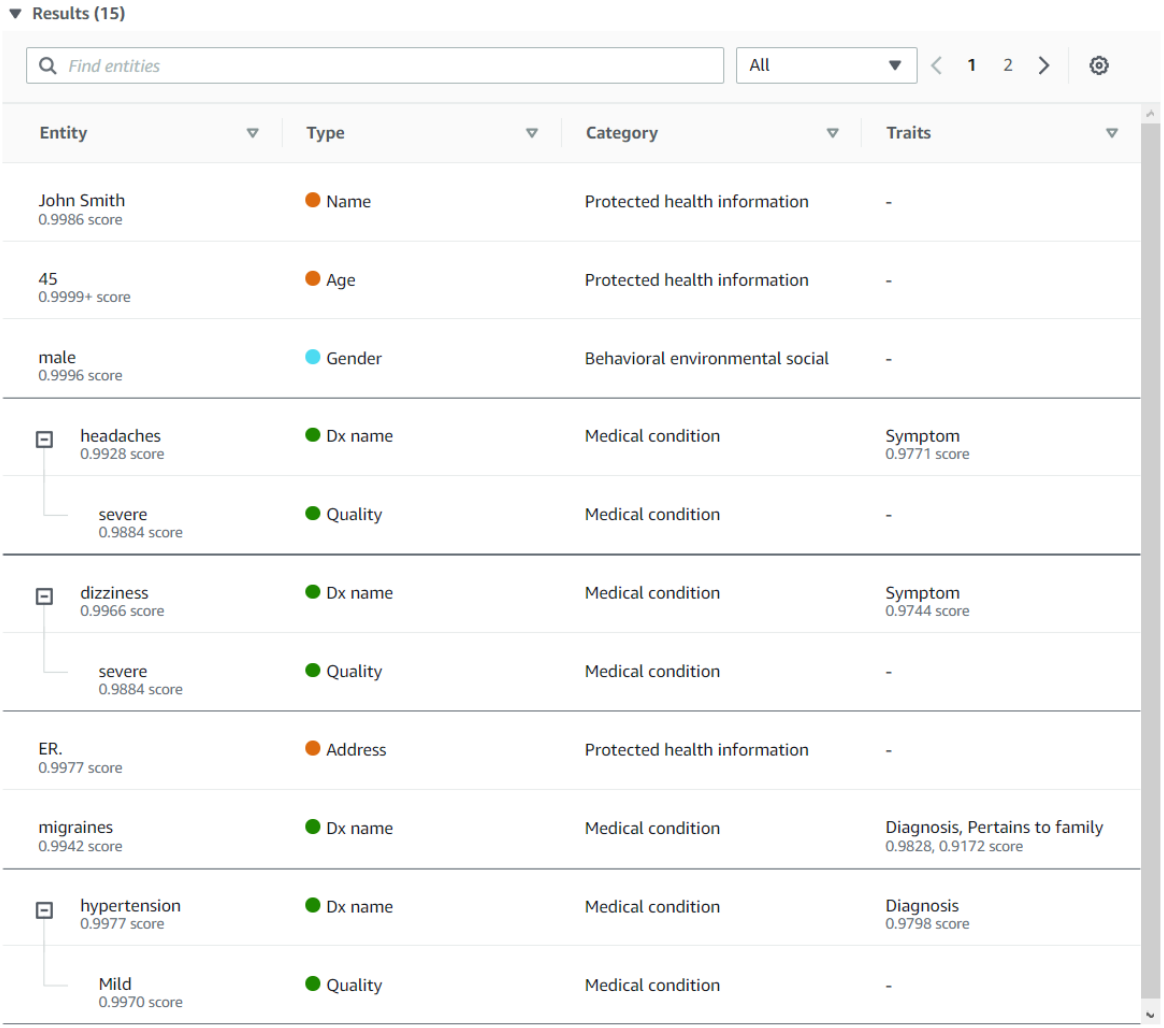


Figura 2.17: La pagina 1 dei risultati dell'analisi delle entità dell'esempio


```
                "RxNormConcepts": [
                    {
                        "Description": "acetaminophen",
                        "Code": "161",
                        "Score": 0.8187639713287354
                    },
                    {
                        "Description": "acetaminophen 300 mg oral tablet",
                        "Code": "348978",
                        "Score": 0.10138837993144989
                    },
                    {
                        "Description": "acephen",
                        "Code": "225064",
                        "Score": 0.09056191146373749
                    },
                    {
                        "Description": "acetaminophen 1000 mg oral tablet",
                        "Code": "430837",
                        "Score": 0.06560156494379044
                    },
                    {
```

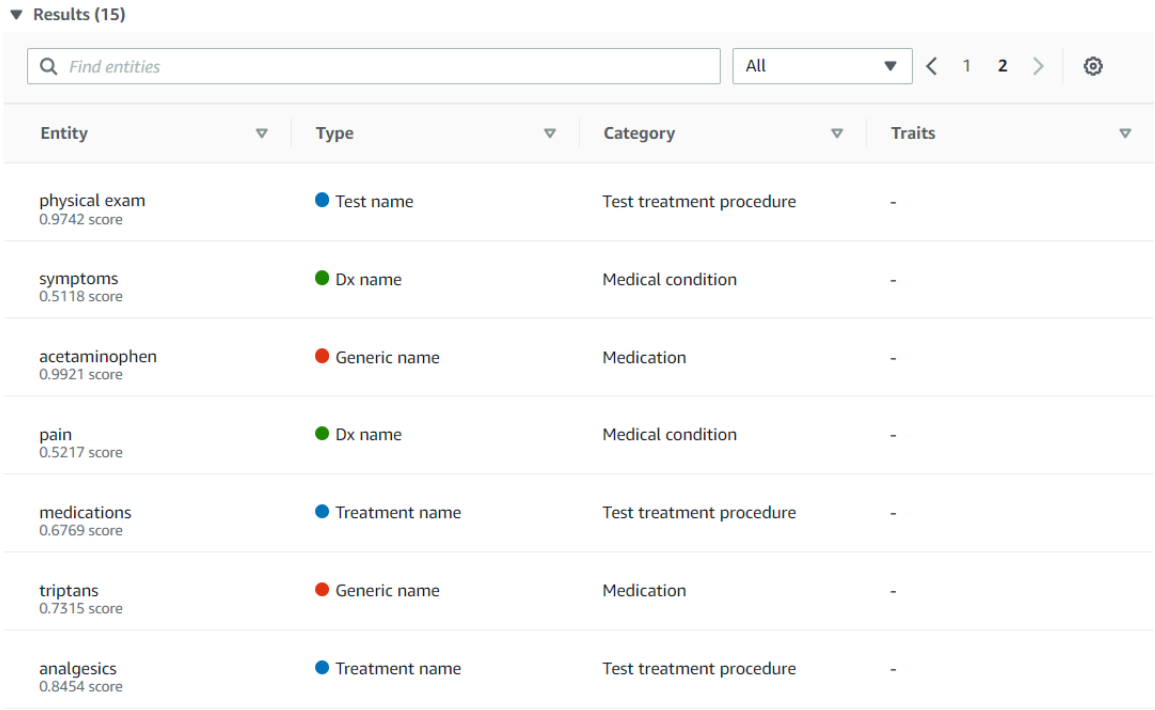


Figura 2.18: La pagina 2 dei risultati dell'analisi delle entità dell'esempio


```
                        "Description": "acetaminophen 650 mg oral tablet",
                        "Code": "198444",
                        "Score": 0.06405171006917953
                    }
                ]
            },
            {
                "Id": 2,
                "Text": "triptans",
                "Category": "MEDICATION",
                "Type": "GENERIC_NAME",
                "Score": 0.7121685743331909,
                "BeginOffset": 447,
                "EndOffset": 455,
                "Attributes": [],
                "Traits": [],
                "RxNormConcepts": [
                    {
                        "Description": "triptone",
                        "Code": "220486",
                        "Score": 0.25794175267219543
                    },
                    {
                        "Description": "triptodur",
                        "Code": "1944385",
                        "Score": 0.1797599345445633
                    },
                    {
                        "Description": "naratriptan",
                        "Code": "141366",
                        "Score": 0.1596847027540207
                    },
```

```
                        {
                            "Description": "naratriptan / sumatriptan",
                            "Code": "1007757",
                            "Score": 0.10175459086894989
                        },
                        {
                            "Description": "isopentane",
                            "Code": "1368703",
                            "Score": 0.05336438864469528
                        }
                    ]
                }
            ],
            "ModelVersion": "2.2.0.20221003"
        }
```

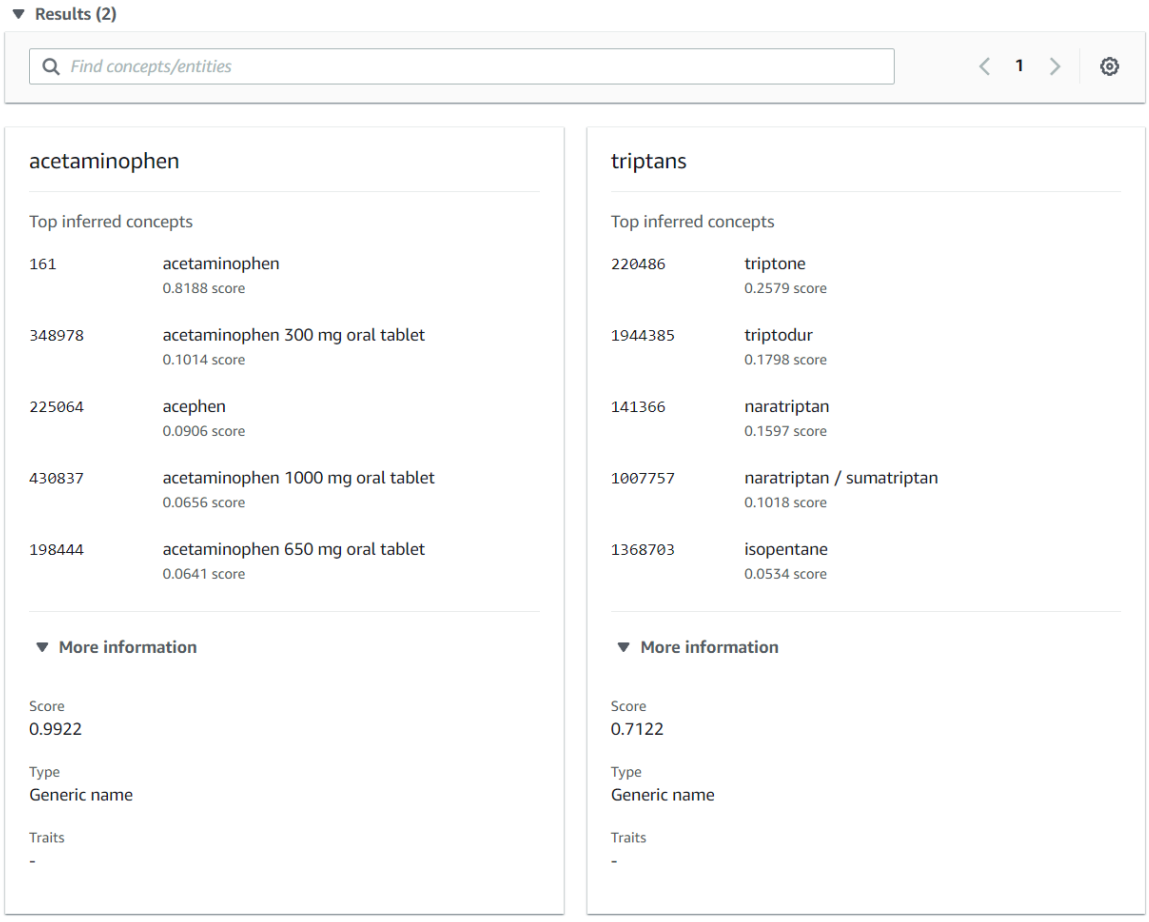



Figura 2.19: I risultati dell'analisi rispetto al servizio InferRxNorm dell'esempio

I risultati di questa analisi sono corretti; l'unico appunto che bisogna fare riguarda il termine "analgesics" che non viene rilevato da Amazon Comprehend Medical nell'analisi dei medicinali rispetto ai concetti RxNorm.

Risultati dell'analisi secondo il servizio InferICD10CM

Il risultato dell'analisi rispetto al servizio *InferICD10CM* all'interno della console è quello riportato nelle Figure 2.20, 2.21 e 2.22; invece l'analisi in formato JSON,che non

verrà mostrata tutta per brevità, appare così:


```
{
    "Entities": [
        {
            "Id": 2,
            "Text": "headaches",
            "Category": "MEDICAL_CONDITION",
            "Type": "DX_NAME",
            "Score": 0.9928064346313477,
            "BeginOffset": 62,
            "EndOffset": 71,
            "Attributes": [
                {
                    "Type": "QUALITY",
                    "Score": 0.9884397983551025,
                    "RelationshipScore": 1,
                    "Id": 1,
                    "BeginOffset": 55,
                    "EndOffset": 61,
                    "Text": "severe",
                    "Traits": []
                }
            ],
            "Traits": [
                {
                    "Name": "SYMPTOM",
                    "Score": 0.9771478176116943
                }
            ],
            "ICD10CMConcepts": [
                {
                    "Description": "Headache",
                    "Code": "R51",
                    "Score": 0.7334675192832947
                },
                {
                    "Description": "Migraine, unspecified, not intractable,
                        without status migrainosus",
                    "Code": "G43.909",
                    "Score": 0.1043735072016716
                },
                {
                    "Description": "Headache, unspecified",
                    "Code": "R51.9",
                    "Score": 0.10375749319791794
                },
                {
                    "Description": "Migraine, unspecified, intractable, without
                        status migrainosus",
                    "Code": "G43.919",
                    "Score": 0.08496489375829697
                },
                {
                    "Description": "Migraine without aura, not intractable,
                        without status migrainosus",
                    "Code": "G43.009",
                    "Score": 0.0765332281589508
                }
            ]
```

```
                }
                ...
            ],
            "ModelVersion": "2.5.0.20220401"
        }
```

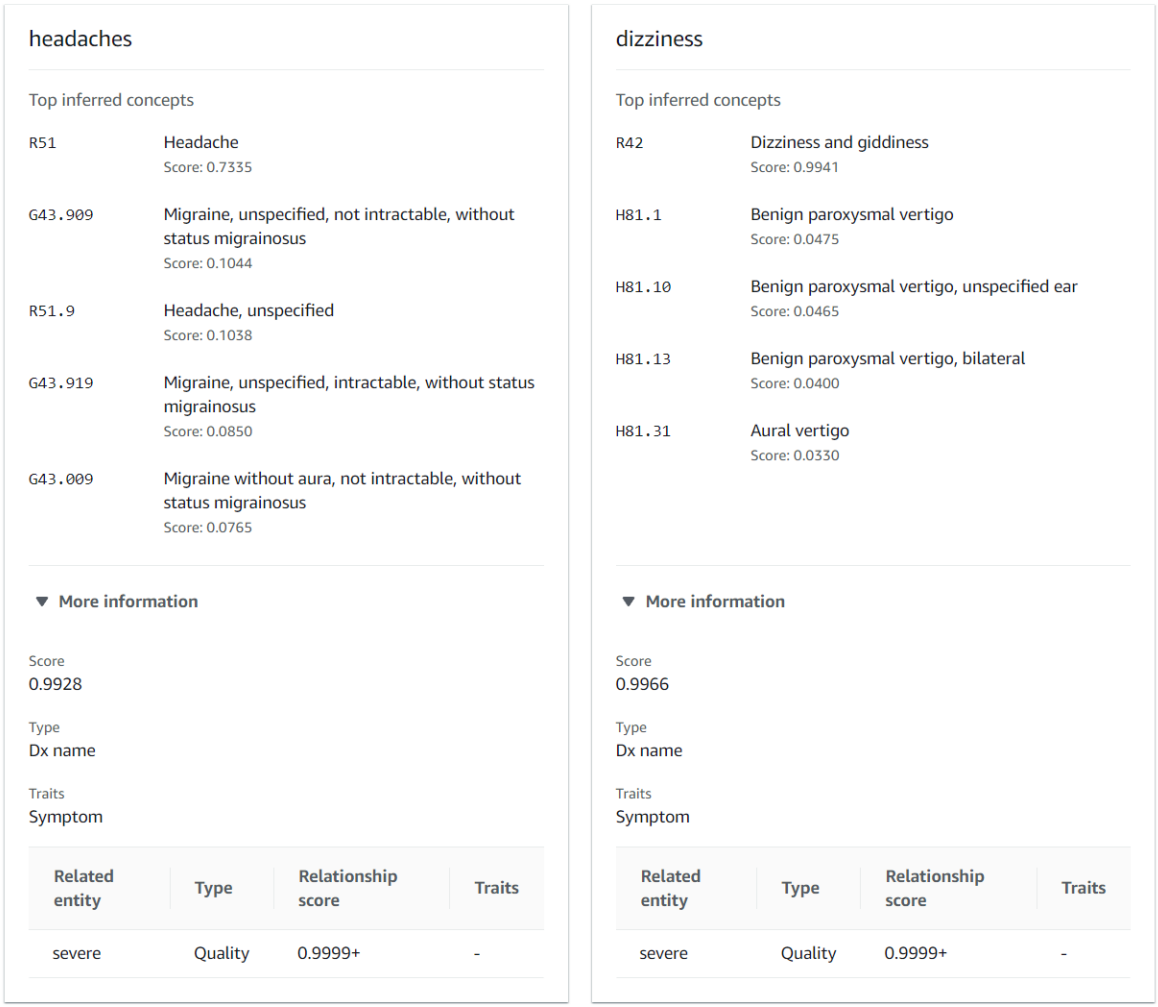


Figura 2.20: La pagina 1 dei risultati dell'analisi rispetto al servizio InferICD10CM dell'esempio

In questo caso l'analisi Amazon Comprehend non contiene errori fornendo dati corretti affiancati ai concetti forniti dalla International Classification of Diseases.

Risultati dell'analisi secondo il servizio InferSNO-MEDCT

Il risultato dell'analisi rispetto al servizio *InferSNO-MEDCT* all'interno della console è quello riportato nelle Figure 2.23 - 2.28; invece l'analisi in formato JSON, che non verrà mostrata tutta per brevità, appare così:

```
{
    "Entities": [
        {
            "Id": 2,
            "Text": "headaches",
            "Category": "MEDICAL_CONDITION",
            "Type": "DX_NAME",
```

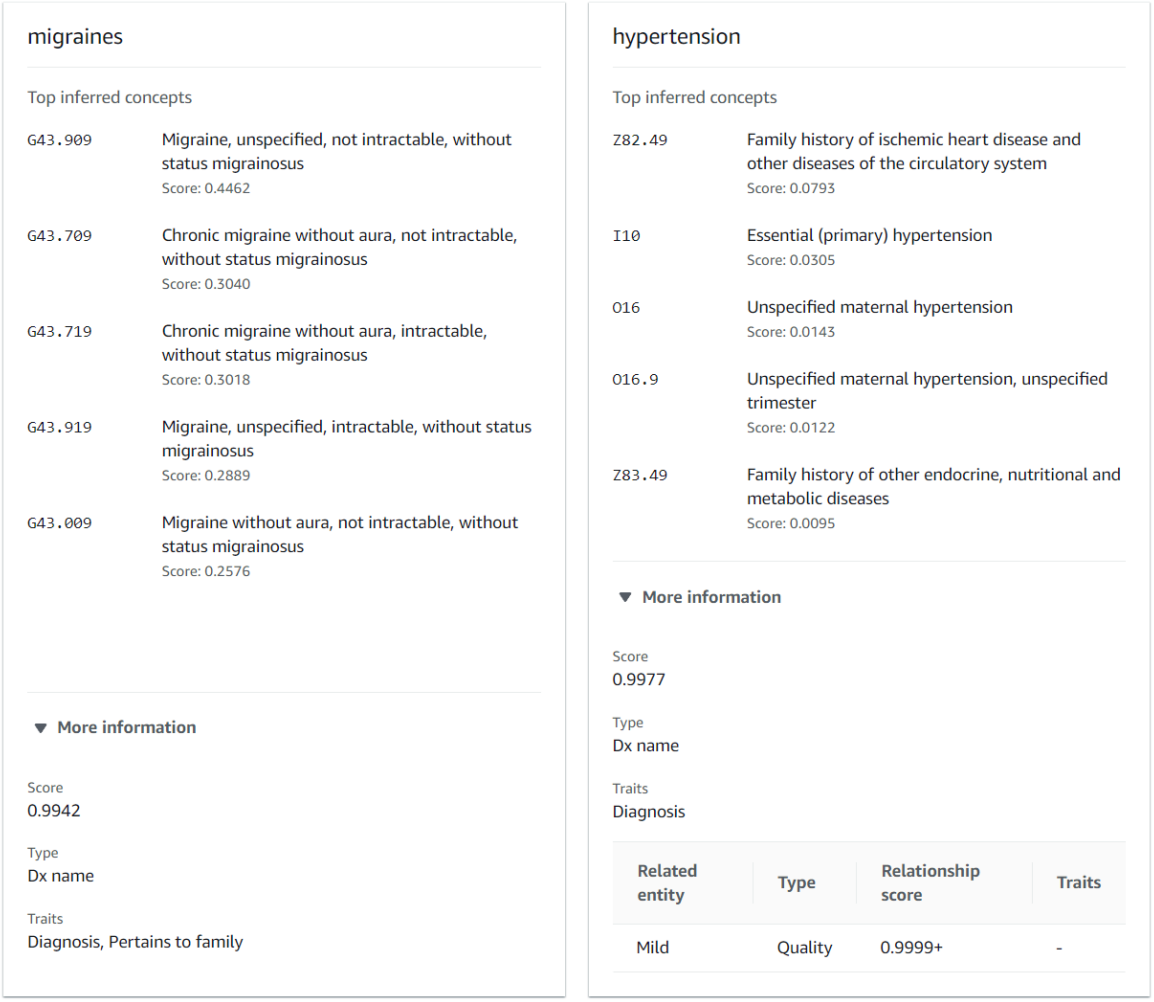


Figura 2.21: La pagina 2 dei risultati dell'analisi rispetto al servizio InferICD10CM dell'esempio


```
                "Score": 0.9928064346313477,
                "BeginOffset": 62,
                "EndOffset": 71,
                "Attributes": [
                    {
                        "Category": "MEDICAL_CONDITION",
                        "Type": "QUALITY",
                        "Score": 0.9884397983551025,
                        "RelationshipScore": 1,
                        "RelationshipType": "QUALITY",
                        "Id": 1,
                        "BeginOffset": 55,
                        "EndOffset": 61,
                        "Text": "severe",
                        "Traits": [],
                        "SNOMEDCTConcepts": [
                            {
                                "Description": "Severe (severity modifier) (qualifier
                                    value)",
                                "Code": "24484000",
                                "Score": 0.2926347851753235
                            },
                            {
```

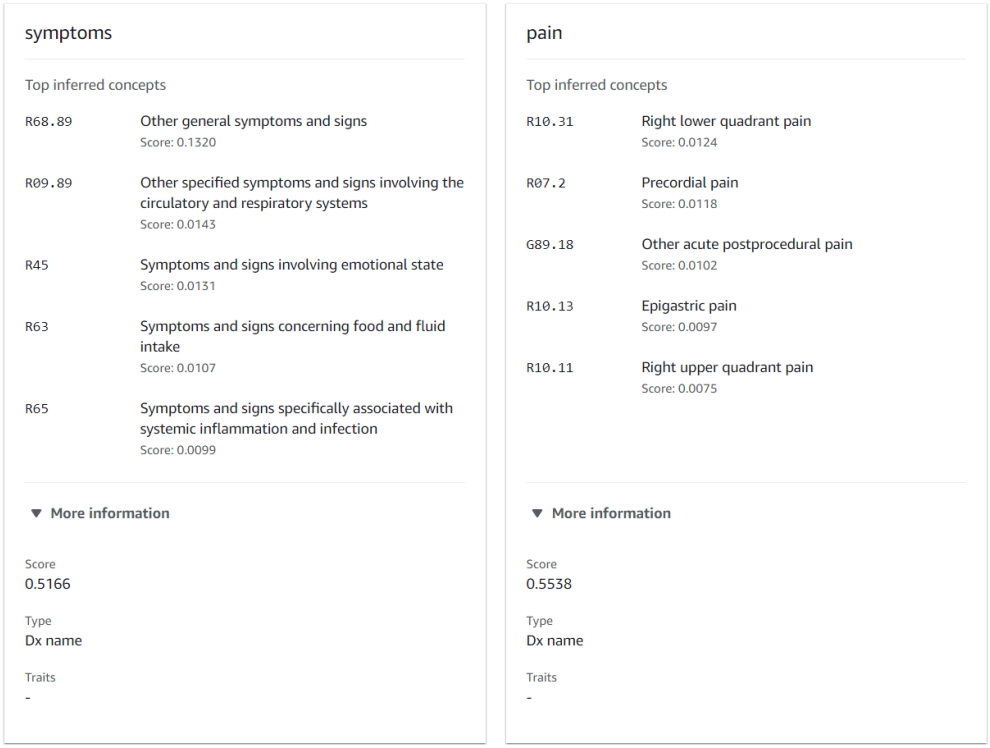


Figura 2.22: La pagina 3 dei risultati dell'analisi ispetto al servizio InferICD10CM dell'esempio


```
                                "Description": "Severe myopia (disorder)",
                                "Code": "34187009",
                                "Score": 0.014972607605159283
                            },
                            {
                                "Description": "Symptom severe (finding)",
                                "Code": "162470006",
                                "Score": 0.012160452082753181
                            },
                            {
                                "Description": "Severe pain (finding)",
                                "Code": "76948002",
                                "Score": 0.010815835557878017
                            },
                            {
                                "Description": "Disability - severe (finding)",
                                "Code": "161045001",
                                "Score": 0.005536999553442001
                            }
                        ]
                    }
                ],
                "Traits": [
                    {
                        "Name": "SYMPTOM",
                        "Score": 0.9771478176116943
                    }
                ],
                "SNOMEDCTConcepts": [
                    {
                        "Description": "Headache (finding)",
```

```
                            "Code": "25064002",
                            "Score": 0.16066789627075195
                        },
                        {
                            "Description": "Headache caused by drug (finding)",
                            "Code": "294091000119104",
                            "Score": 0.05955628678202629
                        },
                        {
                            "Description": "Headache disorder (disorder)",
                            "Code": "230461009",
                            "Score": 0.029438398778438568
                        },
                        {
                            "Description": "Headache caused by high altitude (disorder)",
                            "Code": "425422007",
                            "Score": 0.021283376961946487
                        },
                        {
                            "Description": "Pain in face (finding)",
                            "Code": "95668009",
                            "Score": 0.01750575192272663
                        }
                    ]
                },
                ...
            ],
            "ModelVersion": "2.6.0.20220301",
            "SNOMEDCTDetails": {
                "Edition": "US",
                "Language": "en",
                "VersionDate": "20220301"
            },
            "Characters": {
                "OriginalTextCharacters": 471
            }
        }
```


L'analisi fornita da Amazon Comprehend Medical è accurata e completa. Il servizio è in grado di analizzare e fornire informazioni dettagliate su tutte le parole connesse al Systematized Nomenclature of Medicine. L'unico aspetto da notare riguarda la mancata identificazione del termine "triptans" durante il processo di analisi, che sarebbe, invece, importante approfondire essendo il nome di un trattamento.

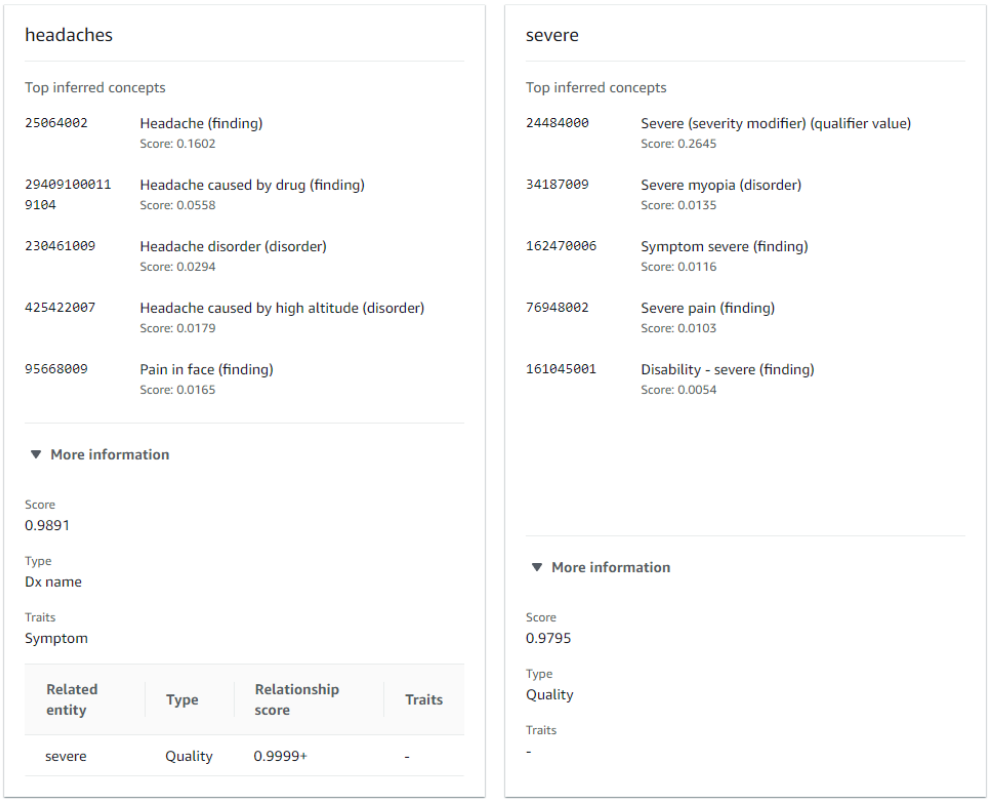


Figura 2.23: La pagina 1 dei risultati dell'analisi rispetto al servizio InferSNOMEDCT dell'esempio

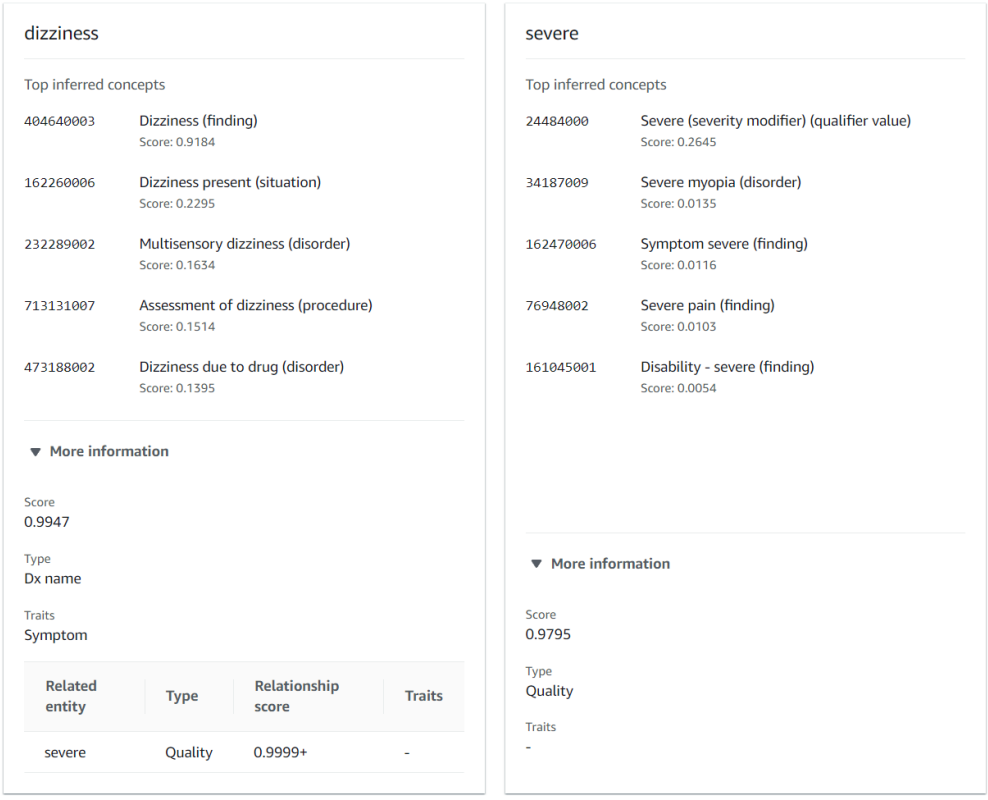


Figura 2.24: La pagina 2 dei risultati dell'analisi rispetto al servizio InferSNOMEDCT dell'esempio

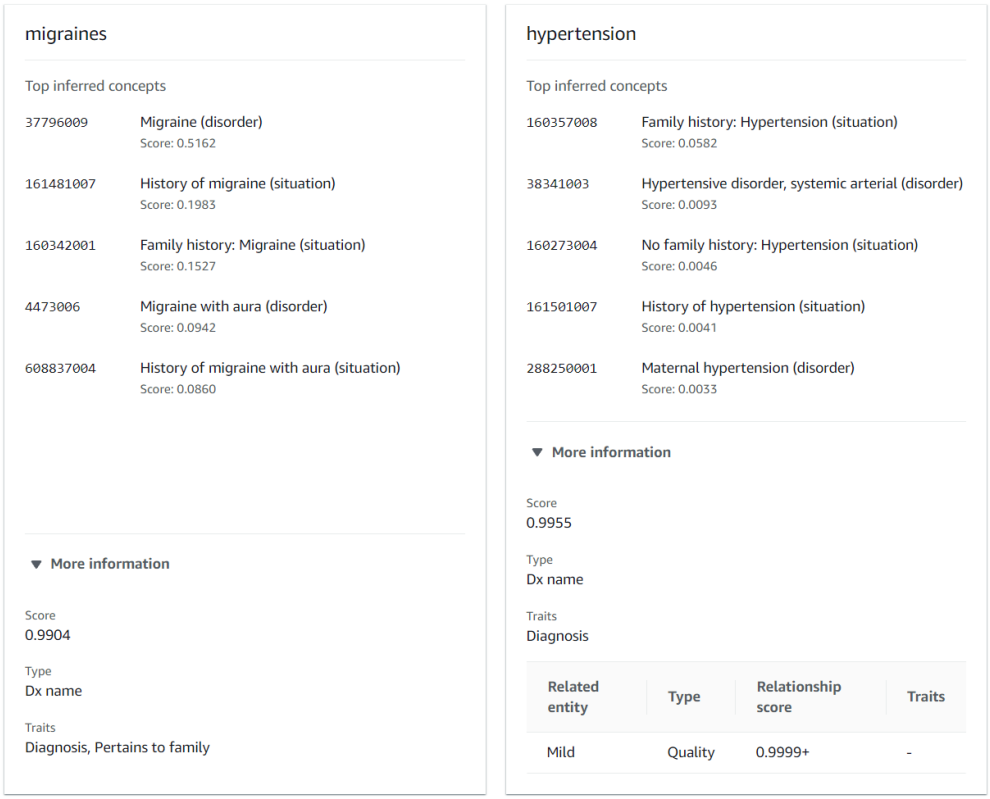


Figura 2.25: La pagina 3 dei risultati dell'analisi rispetto al servizio InferSNOMEDCT dell'esempio

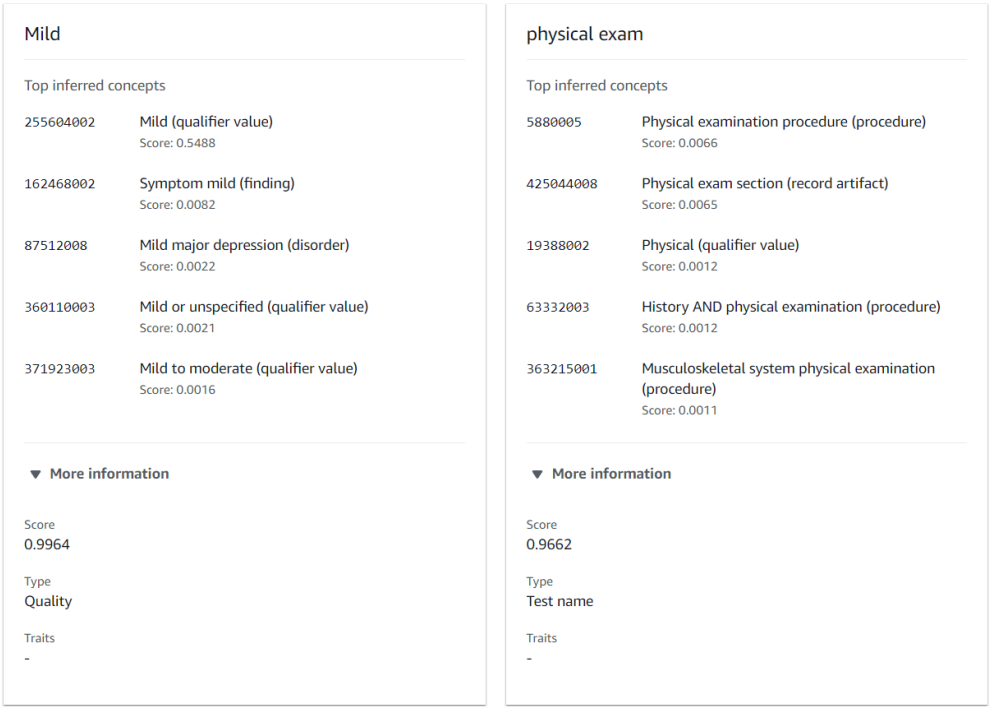


Figura 2.26: La pagina 4 dei risultati dell'analisi rispetto al servizio InferSNOMEDCT dell'esempio

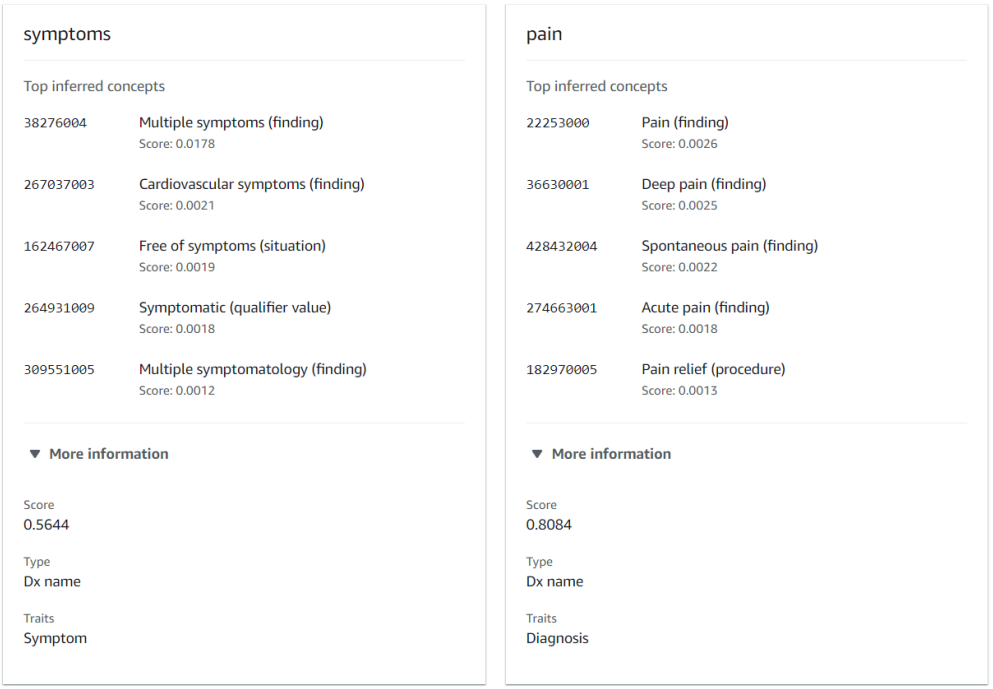


Figura 2.27: La pagina 5 dei risultati dell'analisi rispetto al servizio InferSNOMEDCT dell'esempio

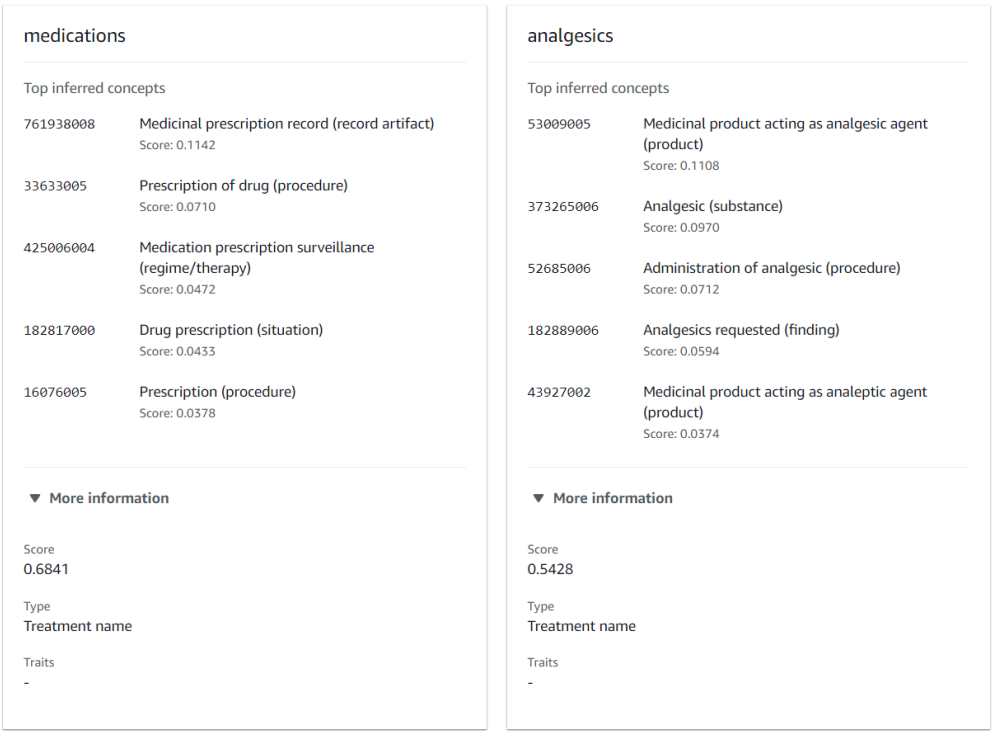


Figura 2.28: La pagina 6 dei risultati dell'analisi rispetto al servizio InferSNOMEDCT dell'esempio

CAPITOLO 3

La Sentiment Analysis con Google

Il capitolo corrente delinea un'analisi dettagliata dei servizi offerti da Google Cloud, utili nell'ambito della Sentiment Anlaysis. In particolare, il focus è rivolto sull'approfondimento del servizio API Natural Language, di cui verrà fornita una chiara esposizione del suo funzionamento, accompagnato da esempi pratici che chiariranno ulteriormente i concetti. Sarà, inoltre, prestata particolare attenzione riguardo all'impiego di tale servizio nel contesto del marketing.

In seguito, si riserverà un'apposita sezione all'analisi del servizio API Healthcare Natural Language, un'applicazione appositamente progettata per il settore medico. In questo contesto, si esplorerà attentamente il suo funzionamento, il tutto supportato da esempi esaustivi che contribuiranno alla comprensione complessiva.

3.1 Cos'è l'API Natural Language

L'API Natural Language è un servizio offerto da Google che sfrutta avanzate tecniche di machine learning per l'estrazione di informazioni da testi non strutturati. Questo strumento si basa sull'analisi delle entità, permettendo l'individuazione e la categorizzazione di elementi all'interno di documenti, come e-mail, chat o post sui social media. Inoltre, grazie all'analisi del sentiment, è in grado di comprendere le opinioni dei clienti, fornendo input utili per scoprire prodotti strategici e valutare l'esperienza utente.

Un'applicazione pratica dell'API Natural Language consiste nella classificazione dei contenuti su diverse piattaforme mediatiche, con la possibilità di fornire suggerimenti migliorati sui contenuti e una più efficace segmentazione degli annunci pubblicitari.

Una delle caratteristiche interessanti di questo servizio è la possibilità di addestrare modelli di machine learning personalizzati e di alta qualità senza richiedere una vasta competenza nell'ambito del machine learning. Ciò è reso possibile grazie all'utilizzo di Vertex AI per il linguaggio naturale, basato su AutoML. In questo modo, è possibile creare modelli adatti alle esigenze specifiche con uno sforzo minimo.

3.1.1 Gli insight dell'API Natural Language

L'API Natural Language di Google fornisce una gamma diversificata di insight relativi a un documento, attraverso molteplici tipologie di analisi, tra cui:

- analisi delle entità;

- analisi del sentimento relativo al documento;
- analisi del sentimento relativo alle entità;
- analisi della sintassi;
- analisi della categoria del testo;

Nel seguito esamineremo in dettaglio ciascuna di queste funzionalità.

Analisi delle entità

L'API Natural Language utilizza l'analisi delle *entità* per analizzare il testo fornito e riconoscere entità rilevanti, come nomi propri di personaggi pubblici, punti di riferimento e altre informazioni note. Le entità riconosciute sono generalmente suddivise in due categorie: nomi propri che corrispondono a entità uniche, come persone o luoghi, e nomi comuni, noti anche come "nominali" nell'ambito dell'elaborazione del linguaggio naturale. Una regola pratica da seguire è considerare un termine come entità se si tratta di un nome.

Le entità identificate vengono restituite con le relative posizioni all'interno del testo originale, indicizzate tramite offset. Queste informazioni sono ottenute attraverso l'uso del metodo `analyzeEntities`.

L'analisi delle entità fornisce un insieme di entità rilevate, assieme a parametri associati, come il tipo di entità, la rilevanza dell'entità rispetto all'intero testo e le posizioni all'interno del testo in cui si fa riferimento alla stessa entità. Le entità sono ordinate in base al loro punteggio di "salience", che riflette la loro rilevanza all'interno del testo generale, con quelle più importanti elencate per prime.

La struttura, in formato JSON, della risposta all'analisi delle entità è la seguente:

```
      {
        "entities": [
          {
            "name": "Lawrence of Arabia",
            "type": "WORK_OF_ART",
            "metadata": {
              "mid": "/m/0bx0l",
              "wikipedia_url": "http://en.wikipedia.org/wiki/Lawrence_ofArabia(film)"
            },
            "salience": 0.75222147,
            "mentions": [
              {
                "text": {
                  "content": "Lawrence of Arabia",
                  "beginOffset": 1
                },
                "type": "PROPER"
              },
              {
                "text": {
                  "content": "film biography",
                  "beginOffset": 39
                },
                "type": "COMMON"
              },
              ...
```

```
            ]
          }
        ],
        "language": "en"
      }
```

Quindi, analizzando un entità compaiono vari campi quali:

- *Type*: che indica il tipo di entità (ad esempio, se si tratta di una persona, una località, un bene di consumo e così via). Queste informazioni consentono di distinguere e/o disambiguare le entità e possono essere utilizzate per scrivere pattern o estrarre informazioni. Ad esempio, un valore type può aiutare a distinguere entità con nomi simili come "Lawrence d'Arabia" che è codificata come WORK_OF_ART (film).
- *Metadata*: che contiene informazioni di origine sul repository delle conoscenze dell'entità. Questo campo può contenere i seguenti sottocampi:
  - *Wikipedia_url*: se presente, contiene l'URL di Wikipedia relativo a questa entità.
  - *Mid*: se presente, contiene un identificatore MID (generato automaticamente) dell'entità. I valori mid rimangono univoci per tutte le lingue; pertanto, possono essere utilizzati per collegare entità tra lingue diverse.
- *Salience*: indica l'importanza o la pertinenza dell'entità rispetto al testo del documento. Questo punteggio può aiutare il recupero delle informazioni e il riepilogo dando la priorità alle entità fondamentali. I punteggi più vicini a 0.0 sono meno importanti, mentre i punteggi più vicini a 1.0 sono molto importanti.
- *Mentions*: indicano le posizioni di offset all'interno del testo in cui è menzionata un'entità. Una menzione dell'entità può essere di due tipi: PROPER o COMMON, quindi con nome proprio o nome comune.

Analisi del sentimento relativo al documento

L'API Natural Language offre un'analisi del *sentimento* per determinare l'atteggiamento complessivo (positivo o negativo) espresso nel documento fornito. Il sentimento viene rappresentato attraverso due valori numerici: lo score e la magnitude. Per ottenere questa analisi, si utilizza il metodo `analyzeSentiment`.

La struttura del risultato di questo tipo di analisi in formato JSON è la seguente:

```
{
  "documentSentiment": {
    "score": 0.2,
    "magnitude": 3.6
  },
  "language": "en",
   "sentences": [
    {
      "text": {
        "content": "Four score and seven years ago our fathers brought forth
        on this continent a new nation, conceived in liberty and dedicated to
        the proposition that all men are created equal.",
        "beginOffset": 0
      },
      "sentiment": {
        "magnitude": 0.8,
```

```
            "score": 0.8
          }
        },
        ...
    }
```

I valori dei campi sono i seguenti:

- *DocumentSentiment*: che comprende il sentimento generale del documento, con i seguenti campi:
  - *Score*: indica il punteggio del sentimento, variante da -1.0 (negativo) a 1.0 (positivo), rappresentante la tendenza emotiva globale del testo.
  - *Magnitude*: il quale indica la forza complessiva dell'emozione (sia positiva che negativa) all'interno del testo, tra 0.0 e +inf. A differenza di score, magnitude non è normalizzato; ogni espressione di emozione all'interno del testo (sia positiva che negativa) contribuisce alla magnitude di testo. Pertanto, blocchi di testo più lunghi potrebbero avere dimensioni maggiori.
- *Language*: contiene la lingua del documento, trasmessa nella richiesta iniziale o rilevata automaticamente se assente.
- *Sentences*: raccoglie un elenco delle frasi estratte dal documento originale, come il seguente campo:
  - *Sentiment*: include i valori di sentimento a livello di frase associati ad ogni singola frase, che contengono gli attributi di score e magnitude come descritto precedentemente.

Analisi del sentimento relativo alle entità

L'analisi del *sentimento* dell'*entità* offerta dall'API Natural Language combina sia l'analisi dell'entità che l'analisi del sentimento e tenta di determinare il sentimento (positivo o negativo) espresso in merito alle entità all'interno del testo. Il sentimento dell'entità è rappresentato dal punteggio numerico e dai valori di grandezza ed è determinato per ogni menzione di un'entità. I punteggi vengono, quindi, aggregati in un ambito di andamento e di magnitudine per un'entità. Le richieste di analisi del sentimento delle entità vengono inviate all'API Natural Language utilizzando il metodo `analyzeEntitySentiment`.

La struttura del risultato di questo tipo di analisi in formato JSON è la seguente:

```
{
  "entities": [
    {
      "name": "R&B music",
      "type": "WORK_OF_ART",
      "metadata": {},
      "salience": 0.5306305,
      "mentions": [
        {
          "text": {
            "content": "R&B music",
            "beginOffset": 7
          },
          "type": "COMMON",
```

```
                "sentiment": {
                  "magnitude": 0.9,
                  "score": 0.9
                }
              }
            ],
            "sentiment": {
              "magnitude": 0.9,
              "score": 0.9
            }
          },
          ...
        ],
        "language": "en"
      }
```

Analisi della sintassi

L'API Natural Language mette a disposizione un insieme di strumenti altamente efficaci per condurre l'analisi sintattica di testi. Per avviare questo processo, ci si avvale del metodo `analyzeSyntax`. Grazie a tale metodo, l'API elabora il testo fornito con l'obiettivo di estrarre frasi e token significativi. L'esecuzione di una richiesta di analisi sintattica genera una risposta che aderisce al seguente formato, includendo sia le frasi che i token individuati:

```
      {
        "sentences": [
          ... Array of sentences with sentence information
        ],
        "tokens": [
          ... Array of tokens with token information
        ]
      }
```

Le sentence appaiono nel formato:

```
        "sentences": [
          {
            "text": {
              "content": "Four score and seven years ago our fathers brought forth on
                         this continent a new nation, conceived in liberty and
                         dedicated to the proposition that all men are created
                         equal.",
              "beginOffset": 0
            }
          },
          ...
        ],
        "language": "en"
```

I token appaiono, invece, nel formato:

```
       "tokens": [
          {
            "text": {
              "content": "The",
              "beginOffset": 4
            },
```

```
            "partOfSpeech": {
              "tag": "DET",
            },
            "dependencyEdge": {
              "headTokenIndex": 2,
              "label": "DET"
            },
            "lemma": "The"
          },
          {
            "text": {
              "content": "only",
              "beginOffset": 8
            },
            "partOfSpeech": {
              "tag": "ADJ",
            },
            "dependencyEdge": {
              "headTokenIndex": 2,
              "label": "AMOD"
            },
            "lemma": "only"
          },
          ...
        ]
```

Qui per ogni token, quindi, vengono riportati vari dati, quali:

- *Text*: contiene i dati di testo associati a questo token, con i seguenti campi secondari:
  - *BeginOffset*: rappresenta l'offset dei caratteri (a base zero) all'interno del testo fornito.
  - *Content*: contiene il contenuto testuale effettivo del testo originale.
- *PartOfSpeech*: fornisce informazioni grammaticali, inclusi dettagli morfologici relativi al token, come tempo, persona, numero, genere, e così via.
- *Lemma*: contiene la "radice" su cui si basa questa parola, permettendo la standardizzazione dell'uso delle parole all'interno del testo. Ad esempio, le parole "scrivere" e "scrittura" si basano sullo stessa lemma. Forme plurali e singolari sono anch'esse basate sullo stesso lemma: sia "casa" che "case" si riferiscono alla stessa forma.
- *DependencyEdge*: identifica la relazione tra le parole di una frase contenente un token. Queste informazioni possono essere preziose per la traduzione, l'estrazione di informazioni e la sintesi. Questo campo contiene i seguenti sottocampi:
  - *HeadTokenIndex*: fornisce l'indice (a base zero) del token principale al quale il token attuale è legato. Un token non ha un proprio indice principale.
  - *Label*: indica il tipo di dipendenza tra questo token e il token principale.

Analisi della categoria del testo

L'API Natural Language permette anche di eseguire l'analisi di un documento e ottenere un elenco di *categorie* di contenuti rilevanti applicabili al testo presente in esso. Per la classificazione dei contenuti di un documento, è possibile utilizzare il metodo

`classifyText`. Questa funzionalità consente di ottenere una valutazione accurata e dettagliata delle categorie di contenuti presenti nel testo, fornendo informazioni preziose per comprendere il contesto e la natura del documento in esame. Bisogna tenere presente che, però, questa funzionalità supporta soltanto la lingua inglese.

La struttura del risultato dell'analisi in formato JSON è la seguente:

```
{
  "categories": [
    {
      "name": "/Internet & Telecom/Mobile & Wireless/Mobile Apps & Add-Ons",
      "confidence": 0.6499999761581421
    },
    ...
  ]
}
```

3.1.2 Come funziona l'API Natural Language

Google mette a disposizione una serie di strumenti tramite l'API Natural Language per l'analisi dei testi. Tra questi, risulta particolarmente immediata e semplice da utilizzare la console disponibile direttamente sul sito web, la quale non richiede alcuna competenza nello sviluppo specifico.

Nella Figura 3.1 si può osservare l'area in cui inserire il testo all'interno della console. Qui è già presente un testo predefinito fornito da Google. Per condurre l'analisi del testo, basterà premere il pulsante "Analyze", il quale restituirà immediatamente i risultati dell'analisi in un'area sottostante della console, come mostrato nella sezione precedente.

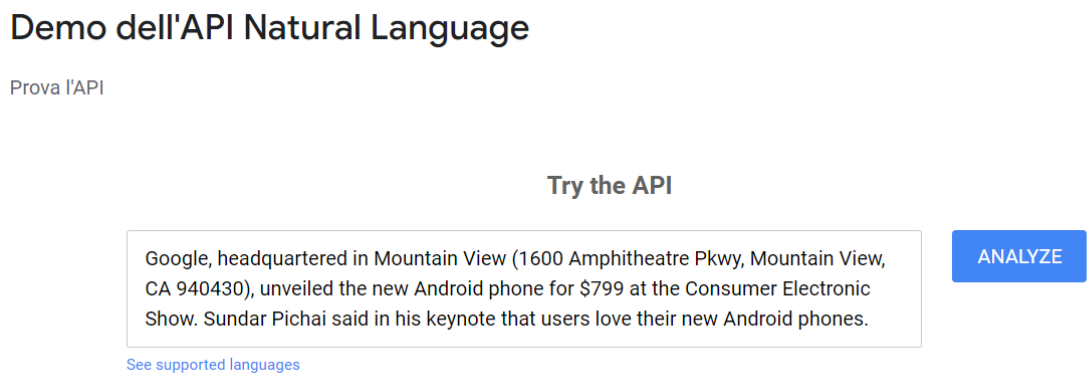


Figura 3.1: La console dell'API Natural Language

I risultati dell'analisi vengono visualizzati nell'area mostrata nella Figura 3.2. Qui, si trovano i pulsanti che consentono di navigare tra i diversi tipi di insight, mostrando i dati specifici correlati ad ognuno di essi.

Un altro modo per utilizzare l'API Natural Language è attraverso il Google Cloud SDK, utilizzando le funzioni da riga di comando. Per procedere con questa modalità, è necessario configurare l'interfaccia a riga di comando "gcloud", creare un progetto su Google Cloud, attivare l'API Natural Language e, infine, eseguire l'analisi delle entità mediante il seguente comando:

```
gcloud ml language analyze-entities --content="Michelangelo Caravaggio, Italian
    painter, is known for 'The Calling of Saint Matthew'."
```

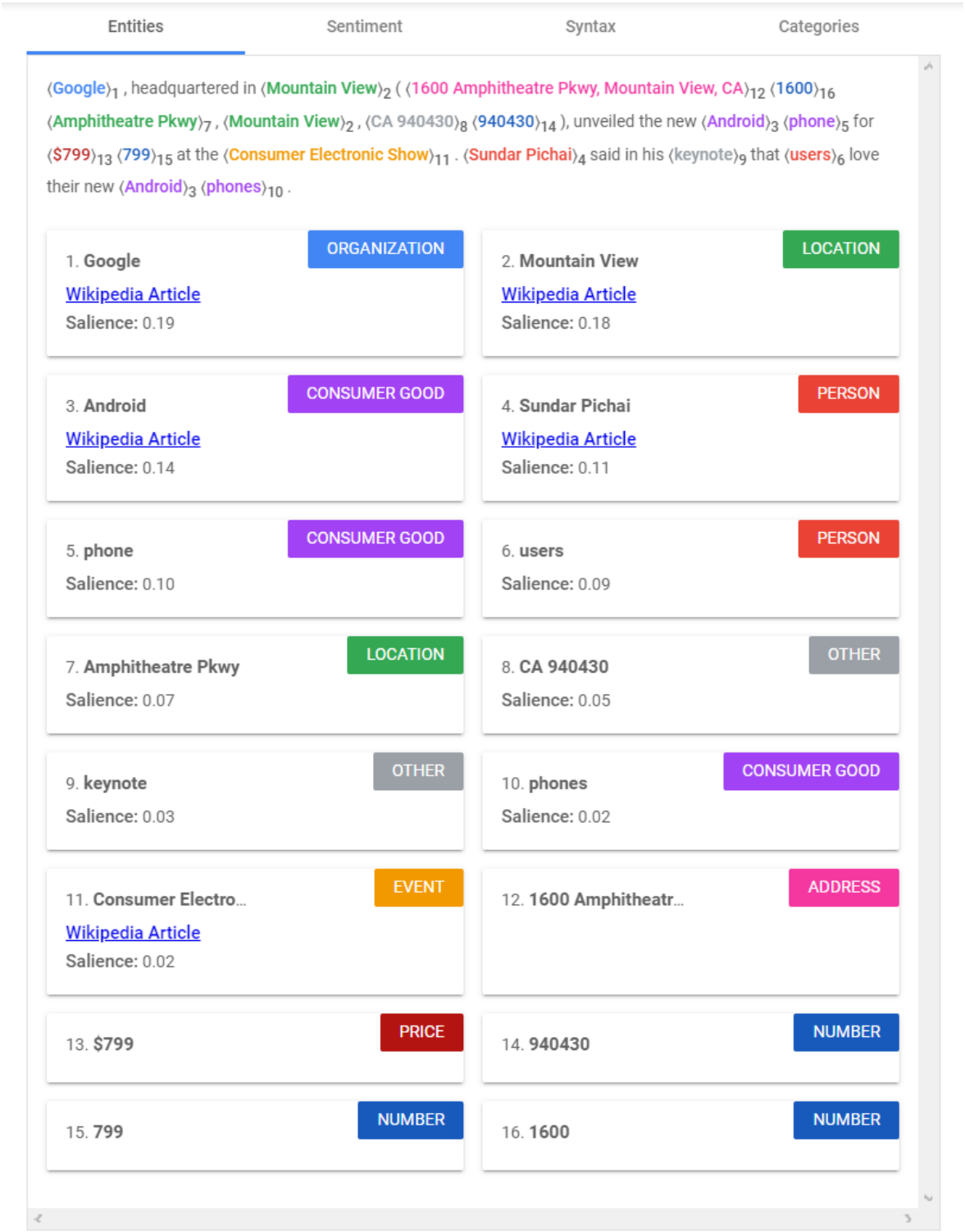


Figura 3.2: I risultati dell'API Natural Language

Come si può intuire, quindi, il testo che verrà analizzato sarà quello dopo il `content`.

Google Cloud mette a disposizione una vasta gamma di librerie per utilizzare l'API Natural Language con diversi linguaggi di programmazione, tra cui Go, Node.js, Java e Python. Tuttavia, per poter procedere con l'utilizzo, è necessario seguire un processo di configurazione. Questo processo comprende l'installazione di Google Cloud CLI, la creazione o selezione di un progetto su Google Cloud, l'attivazione dell'API Cloud Natural Language e la generazione di credenziali di autenticazione locali per il proprio Account Google, tutte operazioni da eseguire tramite la riga di comando. Un esempio di come utilizzare la libreria sviluppata specificamente per Python è disponibile nel repository al seguente indirizzo:

`https://github.com/Walter-Di-Sabatino/API-Natural-Language-Example.git`

Nell'esempio, importando il metodo `language_v1` dalla libreria `google.cloud`, sono state sviluppate delle funzioni che consentono di analizzare gli insight forniti dall'API Natural Language e di stampare i risultati ottenuti. Queste funzioni sono:

- `analyze_entities_sentiment`
- `analyze_syntax`
- `get_total_sentiment`
- `get_text_category`

3.1.3 Esempi con l'API Natural Language

Al fine di illustrare il funzionamento dell'API Natural Language è stato scelto il seguente testo in inglese:

> John Smith is a software engineer who lives happly in New York City. He enjoys playing videogames and loves listening to rock music. His favorite book series is 'The Lord of the Rings', and he has a pet Labrador named Max. If you you would like to contact him his email is: fictionalEmail@gmail.com.

Anche in questo caso è stata selezionata la lingua inglese in modo tale da poter sfruttare appieno le capacità di analisi dell'API Natural Language. Di seguito, è riportata la traduzione letterale del testo:

> John Smith è un ingegnere informatico che vive felicemente a New York. Si diverte a giocare ai videogiochi e ama ascoltare la musica rock. La sua serie di libri preferita è "Il Signore degli Anelli" e ha un Labrador di nome Max. Se volete contattarlo, il suo indirizzo e-mail è: fictionalEmail@gmail.com.

Per facilitare la presentazione dei risultati, verranno forniti i file JSON di risposta e le immagini della sezione della console contenente l'analisi per ciascun insight.

Risultati dell'analisi delle entità

Il risultato dell'analisi delle *entità* all'interno della console è quello riportato all'interno della Figura 3.3; invece l'analisi in formato JSON, che non verrà mostrata tutta per brevità, appare così:

```
{
  "entities": [
    {
      "mentions": [
        {
          "text": {
            "beginOffset": 0,
            "content": "John Smith"
          },
          "type": "PROPER"
        },
        {
          "text": {
            "beginOffset": 16,
            "content": "software engineer"
          },
          "type": "COMMON"
        }
      ],
      "metadata": {},
      "name": "John Smith",
      "salience": 0.89666426,
      "type": "PERSON"
    },
    {
      "mentions": [
        {
          "text": {
            "beginOffset": 54,
            "content": "New York City"
          },
          "type": "PROPER"
        }
      ],
      "metadata": {
        "mid": "/m/02_286",
        "wikipedia_url": "https://en.wikipedia.org/wiki/New_York_City"
      },
      "name": "New York City",
      "salience": 0.028159358,
      "type": "LOCATION"
    },
    {
      "mentions": [
        {
          "text": {
            "beginOffset": 146,
            "content": "book series"
          },
          "type": "COMMON"
        }
      ],
      "metadata": {},
      "name": "book series",
      "salience": 0.018803356,
      "type": "WORK_OF_ART"
    },
    ...
```

```
        ],
        "language": "en"
      }
```

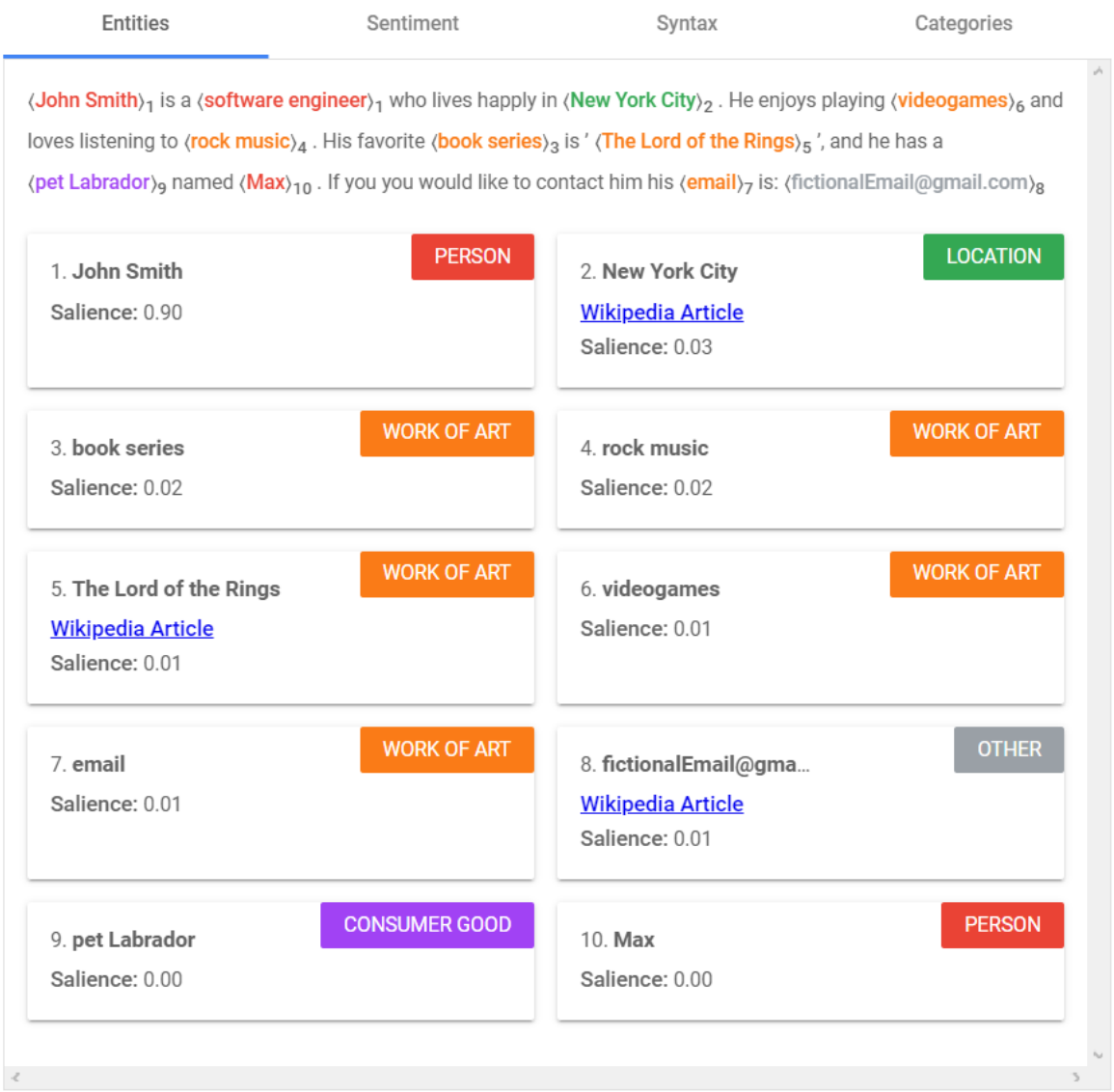


Figura 3.3: I risultati dell'analisi delle entità dell'esempio

I risultati dell'analisi sono in parte corretti: il servizio riesce a distinguere correttamente le entità, ma commette degli errori nei tipi di alcune parole. Ad esempio, riguardo ai termini "email", "pet labrador" e "Max" possiamo osservare che: "email" viene considerata di tipo WORK_OF_ART, anche se sarebbe più sensato inserirla nei tipi OTHER o, al massimo, CONSUMER_GOOD. Allo stesso modo, "pet labrador" viene categorizzato come CONSUMER_GOOD, quando sarebbe più opportuno considerarlo come OTHER. Infine, "Max" viene identificato come PERSON, mentre dovrebbe essere trattato nello stesso modo di "pet labrador", essendo il nome del cane del protagonista. Quest'ultima analisi, tuttavia, può essere considerata opinabile.

Risultati dell'analisi del sentimento

Il risultato dell'analisi del *sentimento* all'interno della console è quello riportato all'interno della Figura 3.4. Invece l'analisi rispetto alle entità in formato JSON, che non verrà mostrata tutta per brevità, appare così:


```
{
  "entities": [
    {
      "mentions": [
        {
          "sentiment": {
            "magnitude": 0.1,
            "score": 0.1
          },
          "text": {
            "beginOffset": 0,
            "content": "John Smith"
          },
          "type": "PROPER"
        },
        {
          "sentiment": {
            "magnitude": 0.1,
            "score": 0.1
          },
          "text": {
            "beginOffset": 16,
            "content": "software engineer"
          },
          "type": "COMMON"
        }
      ],
      "metadata": {},
      "name": "John Smith",
      "salience": 0.89666426,
      "sentiment": {
        "magnitude": 1.4,
        "score": 0.1
      },
      "type": "PERSON"
    },
    {
      "mentions": [
        {
          "sentiment": {
            "magnitude": 0.1,
            "score": 0.1
          },
          "text": {
            "beginOffset": 54,
            "content": "New York City"
          },
          "type": "PROPER"
        }
      ],
      "metadata": {
        "mid": "/m/02_286",
        "wikipedia_url": "https://en.wikipedia.org/wiki/New_York_City"
```

```
            },
            "name": "New York City",
            "salience": 0.028159358,
            "sentiment": {
              "magnitude": 0.1,
              "score": 0.1
            },
            "type": "LOCATION"
          },
          {
            "mentions": [
              {
                "sentiment": {
                  "magnitude": 0.3,
                  "score": 0.3
                },
                "text": {
                  "beginOffset": 146,
                  "content": "book series"
                },
                "type": "COMMON"
              }
            ],
            "metadata": {},
            "name": "book series",
            "salience": 0.018803356,
            "sentiment": {
              "magnitude": 0.3,
              "score": 0.3
            },
            "type": "WORK_OF_ART"
          },
          ...
        ],
        "language": "en"
      }
```


L'analisi del sentimento rispetto al documento in formato JSON invece è la seguente:


```
      {
        "documentSentiment": {
          "magnitude": 2.0,
          "score": 0.3
        },
        "language": "en",
        "sentences": [
          {
            "sentiment": {
              "magnitude": 0.0,
              "score": 0.0
            },
            "text": {
              "beginOffset": 0,
              "content": "John Smith is a software engineer who lives happly in New York
                   City."
            }
          },
          {
            "sentiment": {
              "magnitude": 0.9,
```

```
                "score": 0.9
              },
              "text": {
                "beginOffset": 69,
                "content": "He enjoys playing videogames and loves listening to rock music."
              }
            },
            {
              "sentiment": {
                "magnitude": 0.8,
                "score": 0.8
              },
              "text": {
                "beginOffset": 133,
                "content": "His favorite book series is The Lord of the Rings, and he has a pet Labrador named Max."
              }
            },
            {
              "sentiment": {
                "magnitude": 0.2,
                "score": -0.2
              },
              "text": {
                "beginOffset": 227,
                "content": "If you you would like to contact him his email is: fictionalEmail@gmail.com."
              }
            }
          ]
        }
```


In generale, l'analisi del sentimento fornita dal servizio sembra essere corretta. Lo SCORE totale di 0.3 indica una certa positività del documento, vicina alla neutralità, principalmente a causa della descrizione delle passioni del protagonista. Infatti, le frasi e le entità legate agli hobby del protagonista sono quelle che hanno una valutazione maggiormente positiva.

Risultati dell'analisi della sintassi

Il risultato dell'analisi della *sintassi* all'interno della console è quello riportato in maniera parziale all'interno della Figura 3.5. Invece l'analisi in formato JSON, che non verrà mostrata tutta per brevità, appare così:


```
        {
          "language": "en",
          "sentences": [
            {
              "text": {
                "beginOffset": 0,
                "content": "John Smith is a software engineer who lives happly in New York City"
              }
            }
          ],
          "tokens": [
            {
```

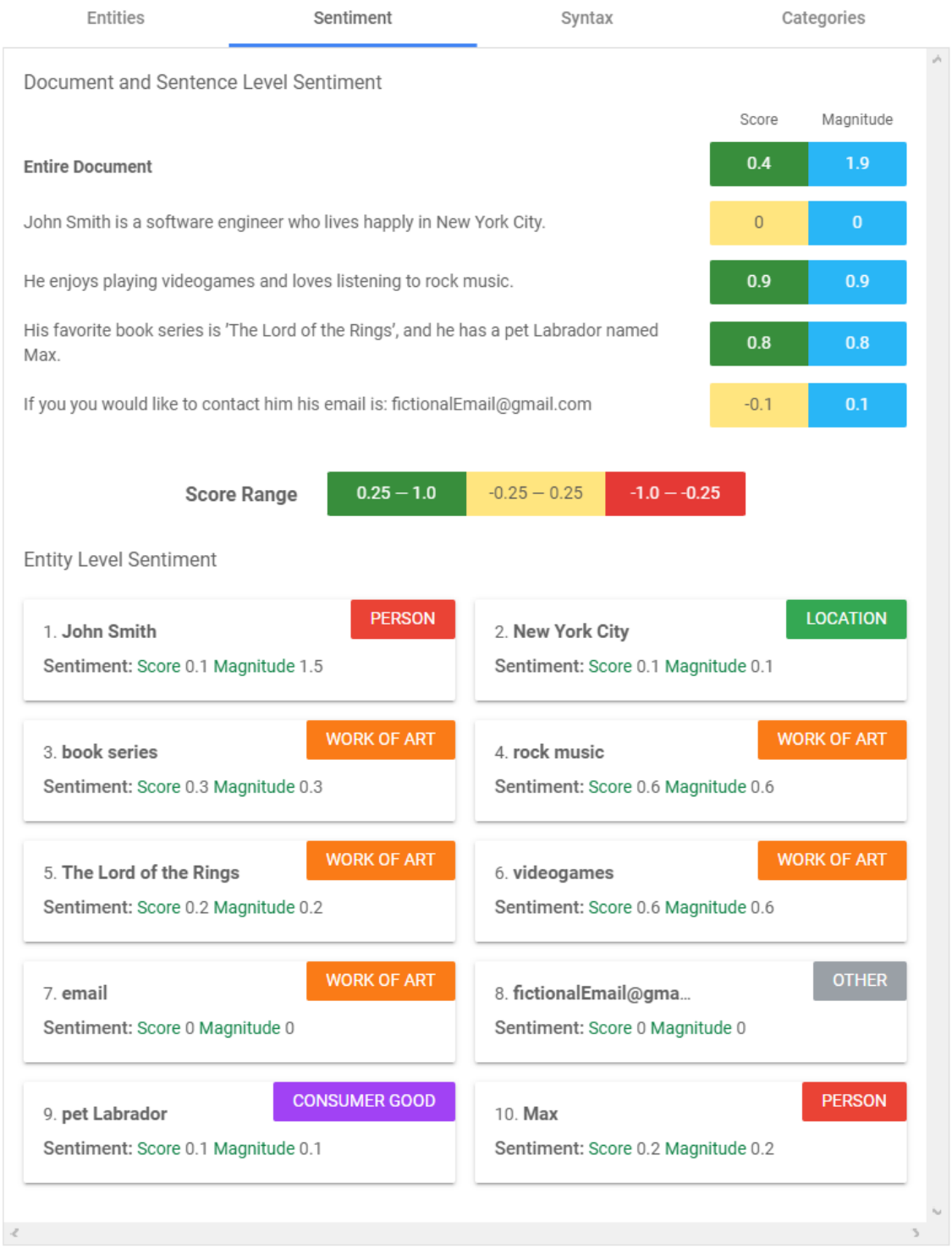


Figura 3.4: I risultati dell'analisi del sentimento dell'esempio

```
            "dependencyEdge": {
              "headTokenIndex": 1,
              "label": "NN"
            },
            "lemma": "John",
            "partOfSpeech": {
              "aspect": "ASPECT_UNKNOWN",
              "case": "CASE_UNKNOWN",
              "form": "FORM_UNKNOWN",
              "gender": "GENDER_UNKNOWN",
              "mood": "MOOD_UNKNOWN",
              "number": "SINGULAR",
              "person": "PERSON_UNKNOWN",
              "proper": "PROPER",
              "reciprocity": "RECIPROCITY_UNKNOWN",
              "tag": "NOUN",
              "tense": "TENSE_UNKNOWN",
              "voice": "VOICE_UNKNOWN"
            },
            "text": {
              "beginOffset": 0,
              "content": "John"
            }
          },
          {
            "dependencyEdge": {
              "headTokenIndex": 2,
              "label": "NSUBJ"
            },
            "lemma": "Smith",
            "partOfSpeech": {
              "aspect": "ASPECT_UNKNOWN",
              "case": "CASE_UNKNOWN",
              "form": "FORM_UNKNOWN",
              "gender": "GENDER_UNKNOWN",
              "mood": "MOOD_UNKNOWN",
              "number": "SINGULAR",
              "person": "PERSON_UNKNOWN",
              "proper": "PROPER",
              "reciprocity": "RECIPROCITY_UNKNOWN",
              "tag": "NOUN",
              "tense": "TENSE_UNKNOWN",
              "voice": "VOICE_UNKNOWN"
            },
            "text": {
              "beginOffset": 5,
              "content": "Smith"
            }
          },
          ...
        ]
      }
```


Dai risultati dell'analisi, possiamo dedurre che il servizio offerto da Google Cloud è in grado di effettuare l'analisi della sintassi in maniera precisa e semplice.

Figura 3.5: I risultati dell'analisi della sintassi dell'esempio

Risultati dell'analisi delle categorie del testo

Il risultato dell'analisi delle *categorie* del testo all'interno della console è quello riportato all'interno della Figura 3.6. Invece l'analisi in formato JSON appare così:

```
{
    "categories": [
        {
          "name": "/Arts & Entertainment"
          "confidence": 0.65
        }
    ]
}
```

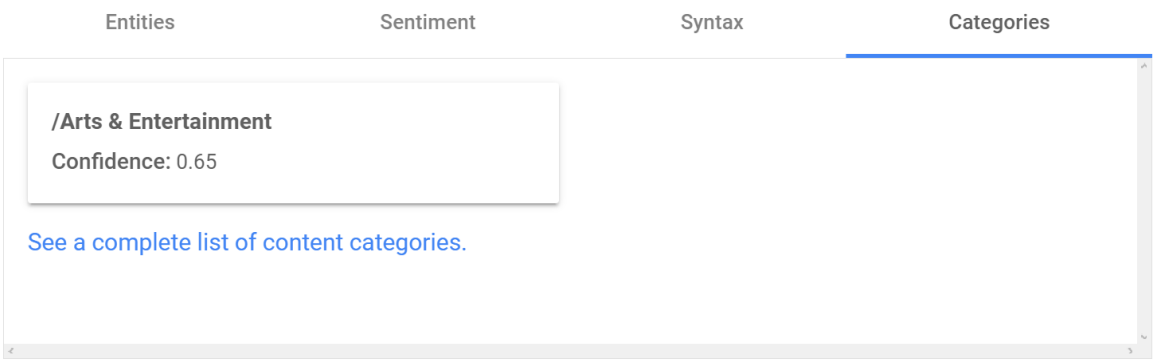


Figura 3.6: I risultati dell'analisi delle categorie dell'esempio

Considerando che, nel testo analizzato, vengono trattati gli hobby del protagonista, che riguardano principalmente musica e libri, il risultato delle categorie come "/Arts & Entertainment" può essere considerato corretto.

3.2 Cos'è l'API Healthcare Natural Language

L'API Healthcare Natural Language, offerta da Google Cloud, rappresenta un servizio avanzato basato su machine learning, progettato per estrarre insight dal testo medico. Con questo servizio, è possibile ottenere analisi in tempo reale di informazioni nascoste all'interno di testi medici non strutturati, come cartelle cliniche o richieste di risarcimento assicurativo.

L'utilizzo dell'API Healthcare Natural Language consente di automatizzare l'estrazione di insight medici comprensibili da documenti medici, semplificando notevolmente il processo di analisi. Inoltre, AutoML Entity Extraction for Healthcare facilita la creazione di modelli personalizzati per l'estrazione delle conoscenze nelle applicazioni del settore sanitario e delle scienze biologiche, eliminando la necessità di possedere competenze avanzate di programmazione.

Le funzionalità offerte dall'API includono:

- Estrarre informazioni relative a concetti medici, come malattie, farmaci, dispositivi medici, procedure e attributi clinicamente pertinenti.
- Mappare concetti medici a vocabolari medici standard, come RxNorm, ICD-10, MeSH e CT SNOMED (questa funzione è solo per utenti statunitensi).

- Estrarre insight medici dal testo e integrali con i prodotti per l'analisi dei dati in Google Cloud

Quindi, grazie all'API Healthcare Natural Language, è possibile ottenere una comprensione approfondita dei dati medici, facilitando la creazione di soluzioni avanzate per il settore sanitario e le scienze biologiche. Bisogna considerare, però, che questo servizio supporta soltanto la lingua inglese.

3.2.1 Gli insight dell'API Healthcare Natural Language

L'API Healthcare Natural Language sfrutta modelli adattabili al contesto al fine di identificare ed estrarre entità, relazioni e attributi di rilievo nel contesto medico. Ciascuna entità testuale viene associata a un termine nel vocabolario medico. Per ottenere tale profondità di analisi all'interno di testi medici, si fa ricorso al metodo `analyzeEntities`.

La struttura della risposta della richiesta di analisi in formato JSON è la seguente:

```
{
 "entityMentions": [
   {
     "mentionId": "1",
     "type": "MEDICINE",
     "text": {
       "content": "Insulin regimen human"
     },
     "linkedEntities": [
       {
         "entityId": "UMLS/C3537244"
       },
       ...
     ],
     "temporalAssessment": {
       "value": "CURRENT",
       "confidence": 0.87631082534790039
     },
     "certaintyAssessment": {
       "value": "LIKELY",
       "confidence": 0.9999774694442749
     },
     "subject": {
       "value": "PATIENT",
       "confidence": 0.99999970197677612
     },
     "confidence": 0.41636556386947632
   },
   {
     "mentionId": "2",
     "type": "MED_DOSE",
     "text": {
       "content": "5 units",
       "beginOffset": 22
     },
     "confidence": 0.56910794973373413
   },
   ...
 ],
 "entities": [
   {
```

```
            "entityId": "UMLS/C3537244",
            "preferredTerm": "Insulins",
            "vocabularyCodes": [
              "MSH/D061385",
              "MTH/NOCODE"
            ]
          },
          ...
        ],
        "relationships": [
          {
            "subjectId": "1",
            "objectId": "2",
            "confidence": 0.53775161504745483
          },
          ...
        ]
      }
```


La risposta quindi contiene i campi:

- *EntityMention*: che sono occorrenze di entità mediche nel testo medico di origine. Ogni menzione di entità contiene i seguenti sottocampi:
  - *MentionId*: che indica un identificatore univoco per un'entità indicata nella risposta.
  - *Type*: che indica la categoria medica della menzione dell'entità.
  - *Text*: che comprende i campi "textContent", che descrive l'estratto del testo medico contenente la menzione dell'entità, e "offset", che identifica la posizione della menzione dell'entità nel testo medico di origine.
  - *TemporalAssessment*: specifica in che modo l'entità collegata si riferisce alla menzione dell'entità; i valori possibili sono: CURRENT, CLINICAL_HISTORY, FAMILY_HISTORY, UPCOMING o OTHER.
  - *CertaintyAssessment*: negazione o qualifica del concetto medico; i valori possibili sono: LIKELY, SOMEWHAT_LIKELY, UNCERTAIN, SOMEWHAT_UNLIKELY, UNLIKELY o CONDITIONAL.
  - *Subject*: specifica l'argomento a cui si riferisce il concetto medico; i valori possibili sono: PATIENT, FAMILY_MEMBER o OTHER.
  - *MentionId*: un elenco di concetti medici che potrebbero essere correlati a questa menzione. Le entità collegate specificano "entityId", che collega un concetto medico a un'entità in "entities".
- *Entities*: descrive i concetti medici derivati dai campi delle entità collegate. Ogni entità viene descritta mediante i seguenti campi:
  - *EntityId*: un identificatore univoco del campo "linkedEntities".
  - *PreferredTerm*: termine preferito per il concetto medico.
  - *VocabularyCodes*: la rappresentazione del concetto medico nei vocabolari medici supportati.
- *Relationships*: definiscono le relazioni dirette tra menzioni di entità.

- *Confidence*: un'indicazione della fiducia del modello nella relazione, espressa come numero compreso tra 0 e 1.

Oltre ai campi elencati, la risposta potrebbe contenere anche il campo "additionalInfo", in cui viene indicata qualsiasi descrizione aggiuntiva relativa al tipo di menzione dell'entità.

3.2.2 Come funziona l'API Healthcare Natural Language

Google Cloud mette a disposizione diversi strumenti per utilizzare l'API Natural Language for Healthcare. Tra questi, uno utile per testare direttamente l'API sul sito di Google Cloud è la console inclusa nella documentazione.

Nella Figura 3.7, viene mostrata la sezione dedicata alla visualizzazione del testo da analizzare. All'interno di questa area, il servizio offre diversi testi di esempio che possono essere facilmente selezionati utilizzando i pulsanti posizionati sopra il campo di testo. Questi pulsanti includono:

- *Sample medical record*, che rappresenta un esempio di cartella clinica;
- *Sample research paper*, che costituisce un esempio di documento di ricerca;
- *Sample lab form*, che illustra un esempio di modulo di laboratorio;
- *Custom*: questa opzione consente agli utenti di inserire il proprio testo personalizzato per l'analisi.

Una volta selezionato il pulsante "Run", evidenziato nella Figura 3.7, verranno generati i risultati dell'analisi e gli stessi verranno visualizzati nella schermata illustrata dalla Figura 3.8. Per accedere alle informazioni specifiche di ciascuna parola, è sufficiente cliccarvi sopra. All'interno di questa schermata, sarà possibile individuare diversi pulsanti che permettono di accedere a varie sezioni dettagliate dell'analisi, quali:

- *Knowledge panel*: in questa sezione sarà possibile visualizzare il testo con le parole rilevanti evidenziate e ottenere analisi dettagliate per ciascuna di esse.
- *Relationships*: qui saranno visualizzate tutte le relazioni esistenti tra le parole, compresi gli oggetti collegati ai soggetti.
- *JSON*: in questa sezione verrà mostrata la risposta dell'analisi nel formato JSON.

Purtroppo, l'utilizzo di questa API è limitato all'accesso tramite terminale, mediante strumenti come Curl o Powershell. Non sono disponibili, infatti, librerie fornite da Google Cloud per facilitare l'integrazione dell'API nei linguaggi di programmazione più comuni.

3.2.3 Esempi con l'API Healthcare Natural Language

Al fine di illustrare il funzionamento dell'API Healthcare Natural Language è stato scelto il testo in inglese a seguire:

> Patient John Smith, 45-year-old male, presented with severe headaches and dizziness at the ER. Noted family history of migraines. Mild hypertension observed during the physical exam. Further lab tests requested to determine the

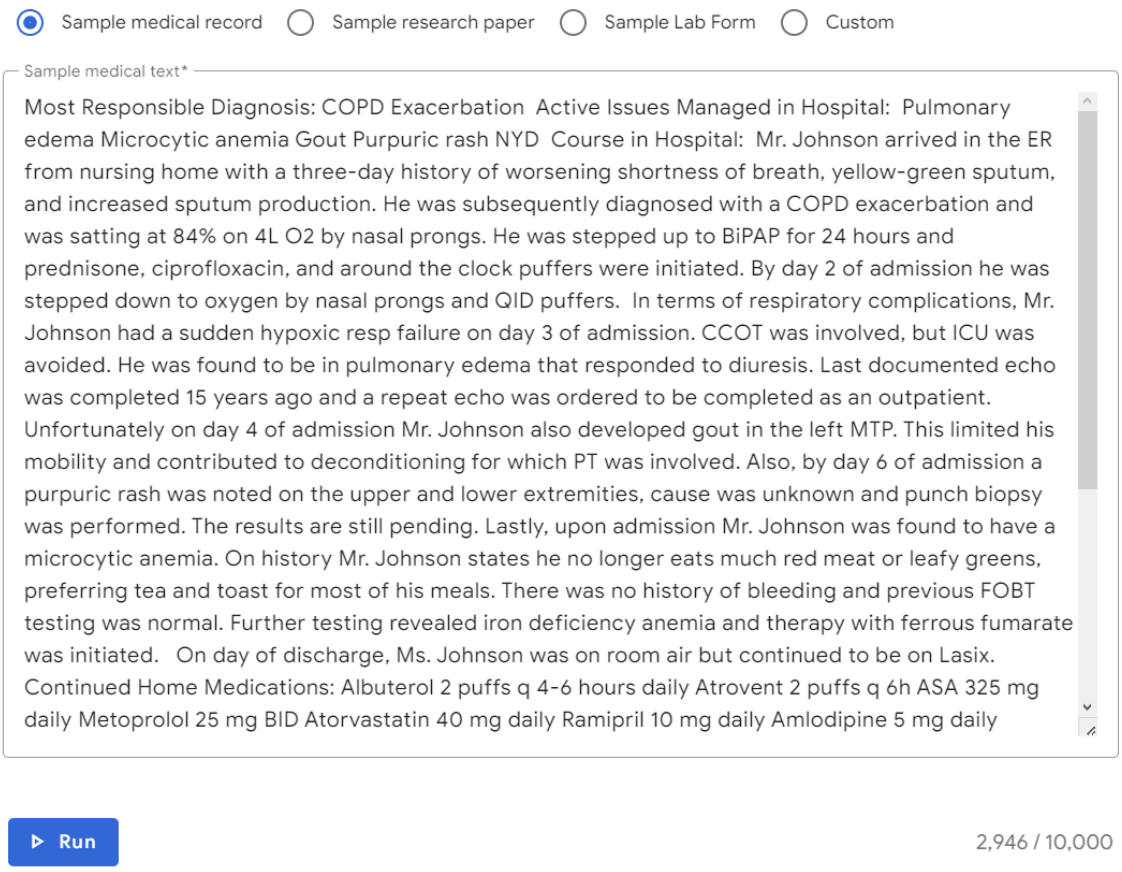


Figura 3.7: La console dell'API Healthcare Natural Language

> underlying cause of symptoms. Administered acetaminophen for immediate pain relief. Goal: identify the underlying cause and establish an appropriate treatment plan, which may involve prescription medications such as triptans or analgesics.

Anche in questo caso è stata selezionata la lingua inglese in modo tale da poter sfruttare appieno le capacità di analisi dell'API Healthcare Natural Language. Di seguito, è riportata la traduzione letterale del testo:

> Il paziente John Smith, uomo di 45 anni, si è presentato al Pronto Soccorso con forti mal di testa e vertigini. Anamnesi familiare di emicrania. Durante l'esame fisico è stata osservata una lieve ipertensione. Sono stati richiesti ulteriori esami di laboratorio per determinare la causa dei sintomi. Somministrazione di acetaminofene per alleviare immediatamente il dolore. Obiettivo: identificare la causa sottostante e stabilire un piano di trattamento appropriato, che può comportare la prescrizione di farmaci come triptani o analgesici.

Per facilitare la presentazione dei risultati, verrà fornito il file di risposta JSON e le immagini dell'analisi all'interno della console.

I risultati dell'analisi all'interno della console appaiono nelle Figure 3.9, 3.10 , 3.11. In formato JSON, invece, il risultato appare in maniera seguente:

```
1 {
```

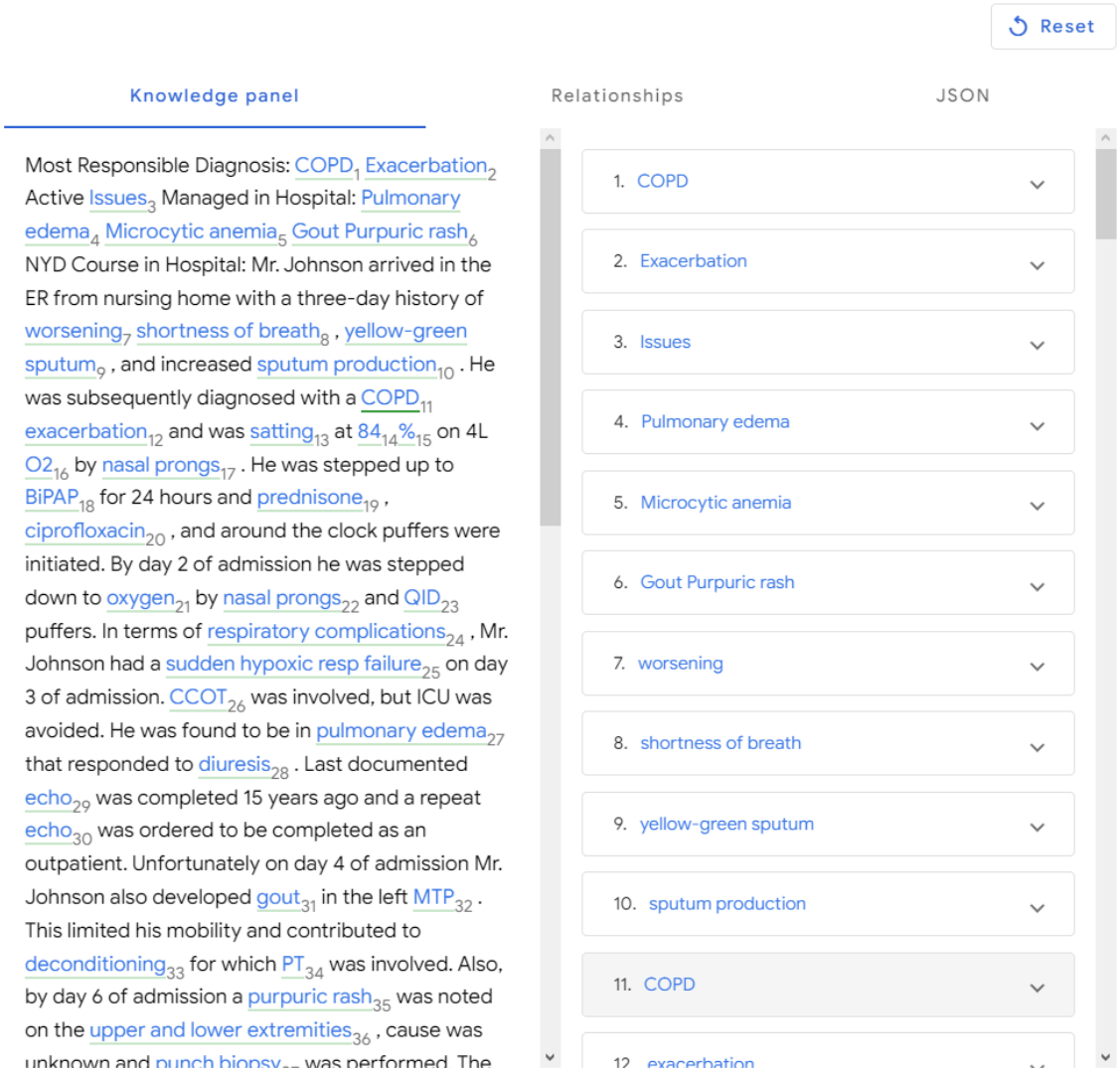


Figura 3.8: I risultati dell'API Healthcare Natural Language

```
        "entityMentions": [
            {
                "mentionId": "1",
                "type": "SEVERITY",
                "text": {
                    "content": "severe",
                    "beginOffset": 53
                },
                "linkedEntities": [
                    {
                        "entityId": "UMLS/C0205082"
                    },
                    {
                        "entityId": "UMLS/C5203119"
                    }
                ],
                "confidence": 0.9433214068412781
            },
            {
```

```
                    "mentionId": "2",
                    "type": "PROBLEM",
                    "text": {
                        "content": "headaches",
                        "beginOffset": 60
                    },
                    "linkedEntities": [
                        {
                            "entityId": "UMLS/C0018681"
                        }
                    ],
                    "temporalAssessment": {
                        "value": "CURRENT",
                        "confidence": 0.998397946357727
                    },
                    "certaintyAssessment": {
                        "value": "LIKELY",
                        "confidence": 0.9996469616889954
                    },
                    "subject": {
                        "value": "PATIENT",
                        "confidence": 0.9996985197067261
                    },
                    "confidence": 0.9953935742378235
                },
                {
                    "mentionId": "3",
                    "type": "PROBLEM",
                    "text": {
                        "content": "dizziness",
                        "beginOffset": 74
                    },
                    "linkedEntities": [
                        {
                            "entityId": "UMLS/C0012833"
                        }
                    ],
                    "temporalAssessment": {
                        "value": "CURRENT",
                        "confidence": 0.9865706562995911
                    },
                    "certaintyAssessment": {
                        "value": "LIKELY",
                        "confidence": 0.9993785619735718
                    },
                    "subject": {
                        "value": "PATIENT",
                        "confidence": 0.9991985559463501
                    },
                    "confidence": 0.997984766960144
                },
                {
                    "mentionId": "4",
                    "type": "PROBLEM",
                    "text": {
                        "content": "migraines",
                        "beginOffset": 119
                    },
                    "linkedEntities": [
```

```
                        {
                            "entityId": "UMLS/C0149931"
                        }
                    ],
                    "temporalAssessment": {
                        "value": "FAMILY_HISTORY",
                        "confidence": 0.9565732479095459
                    },
                    "certaintyAssessment": {
                        "value": "LIKELY",
                        "confidence": 0.9993579983711243
                    },
                    "subject": {
                        "value": "FAMILY_MEMBER",
                        "confidence": 0.8756473660469055
                    },
                    "confidence": 0.9907057285308838
                },
                {
                    "mentionId": "5",
                    "type": "SEVERITY",
                    "text": {
                        "content": "Mild",
                        "beginOffset": 130
                    },
                    "linkedEntities": [
                        {
                            "entityId": "UMLS/C0239574"
                        },
                        {
                            "entityId": "UMLS/C0278138"
                        },
                        {
                            "entityId": "UMLS/C1837545"
                        },
                        {
                            "entityId": "UMLS/C1839253"
                        },
                        {
                            "entityId": "UMLS/C1840417"
                        }
                    ],
                    "confidence": 0.919769287109375
                },
                {
                    "mentionId": "6",
                    "type": "PROBLEM",
                    "text": {
                        "content": "hypertension",
                        "beginOffset": 135
                    },
                    "linkedEntities": [
                        {
                            "entityId": "UMLS/C0020538"
                        }
                    ],
                    "temporalAssessment": {
                        "value": "CURRENT",
                        "confidence": 0.9851529598236084
```

```
                    },
                    "certaintyAssessment": {
                        "value": "LIKELY",
                        "confidence": 0.9993747472763062
                    },
                    "subject": {
                        "value": "PATIENT",
                        "confidence": 0.9996469616889954
                    },
                    "confidence": 0.9869211316108704
                },
                {
                    "mentionId": "7",
                    "type": "LABORATORY_DATA",
                    "text": {
                        "content": "lab tests",
                        "beginOffset": 192
                    },
                    "linkedEntities": [
                        {
                            "entityId": "UMLS/C0022885"
                        }
                    ],
                    "temporalAssessment": {
                        "value": "CURRENT",
                        "confidence": 0.7327945232391357
                    },
                    "certaintyAssessment": {
                        "value": "LIKELY",
                        "confidence": 0.9969125986099243
                    },
                    "subject": {
                        "value": "PATIENT",
                        "confidence": 0.9996373057365417
                    },
                    "confidence": 0.6547127962112427
                },
                {
                    "mentionId": "8",
                    "type": "PROBLEM",
                    "text": {
                        "content": "symptoms",
                        "beginOffset": 249
                    },
                    "linkedEntities": [
                        {
                            "entityId": "UMLS/C1457887"
                        }
                    ],
                    "temporalAssessment": {
                        "value": "CURRENT",
                        "confidence": 0.8834100365638733
                    },
                    "certaintyAssessment": {
                        "value": "LIKELY",
                        "confidence": 0.9272956848144531
                    },
                    "subject": {
                        "value": "PATIENT",
```

```
                        "confidence": 0.9982303380966187
                    },
                    "confidence": 0.9858320355415344
                },
                {
                    "mentionId": "9",
                    "type": "MEDICINE",
                    "text": {
                        "content": "acetaminophen",
                        "beginOffset": 272
                    },
                    "linkedEntities": [
                        {
                            "entityId": "UMLS/C0000970"
                        }
                    ],
                    "temporalAssessment": {
                        "value": "CURRENT",
                        "confidence": 0.9108323454856873
                    },
                    "certaintyAssessment": {
                        "value": "LIKELY",
                        "confidence": 0.9983591437339783
                    },
                    "subject": {
                        "value": "PATIENT",
                        "confidence": 0.9996469616889954
                    },
                    "confidence": 0.98540860414505
                },
                {
                    "mentionId": "10",
                    "type": "PROBLEM",
                    "text": {
                        "content": "pain",
                        "beginOffset": 300
                    },
                    "linkedEntities": [
                        {
                            "entityId": "UMLS/C0030193"
                        }
                    ],
                    "temporalAssessment": {
                        "value": "CURRENT",
                        "confidence": 0.9670794010162354
                    },
                    "certaintyAssessment": {
                        "value": "LIKELY",
                        "confidence": 0.9656879305839539
                    },
                    "subject": {
                        "value": "PATIENT",
                        "confidence": 0.9995003342628479
                    },
                    "confidence": 0.8425870537757874
                },
                {
                    "mentionId": "11",
                    "type": "PROCEDURE",
```

```
                    "text": {
                        "content": "treatment",
                        "beginOffset": 378
                    },
                    "linkedEntities": [
                        {
                            "entityId": "UMLS/C0087111"
                        },
                        {
                            "entityId": "UMLS/C1533734"
                        }
                    ],
                    "temporalAssessment": {
                        "value": "CURRENT",
                        "confidence": 0.771186888217926
                    },
                    "certaintyAssessment": {
                        "value": "LIKELY",
                        "confidence": 0.9956434369087219
                    },
                    "subject": {
                        "value": "PATIENT",
                        "confidence": 0.9991032481193542
                    },
                    "confidence": 0.5349583029747009
                },
                {
                    "mentionId": "12",
                    "type": "MEDICINE",
                    "text": {
                        "content": "medications",
                        "beginOffset": 425
                    },
                    "linkedEntities": [
                        {
                            "entityId": "UMLS/C0013227"
                        }
                    ],
                    "temporalAssessment": {
                        "value": "CURRENT",
                        "confidence": 0.6968519687652588
                    },
                    "certaintyAssessment": {
                        "value": "LIKELY",
                        "confidence": 0.7919047474861145
                    },
                    "subject": {
                        "value": "PATIENT",
                        "confidence": 0.9994842410087585
                    },
                    "confidence": 0.8463497161865234
                },
                {
                    "mentionId": "13",
                    "type": "MEDICINE",
                    "text": {
                        "content": "triptans",
                        "beginOffset": 445
                    },
```

```
                    "linkedEntities": [
                        {
                            "entityId": "UMLS/C1567966"
                        }
                    ],
                    "temporalAssessment": {
                        "value": "UPCOMING",
                        "confidence": 0.6642533540725708
                    },
                    "certaintyAssessment": {
                        "value": "LIKELY",
                        "confidence": 0.8490233421325684
                    },
                    "subject": {
                        "value": "PATIENT",
                        "confidence": 0.9996469616889954
                    },
                    "confidence": 0.9893652200698853
                },
                {
                    "mentionId": "14",
                    "type": "MEDICINE",
                    "text": {
                        "content": "analgesics",
                        "beginOffset": 457
                    },
                    "linkedEntities": [
                        {
                            "entityId": "UMLS/C0002771"
                        }
                    ],
                    "temporalAssessment": {
                        "value": "CURRENT",
                        "confidence": 0.6786859035491943
                    },
                    "certaintyAssessment": {
                        "value": "LIKELY",
                        "confidence": 0.8334541916847229
                    },
                    "subject": {
                        "value": "PATIENT",
                        "confidence": 0.9996469616889954
                    },
                    "confidence": 0.9866154193878174
                }
            ],
            "entities": [
                {
                    "entityId": "UMLS/C0000970",
                    "preferredTerm": "acetaminophen",
                    "vocabularyCodes": [
                        "LNC/LP14712-1",
                        "LNC/MTHU003399",
                        "MSH/D000082",
                        "MTH/NOCODE",
                        "NCI/C198",
                        "RXNORM/161",
                        "VANDF/4017513"
                    ]
```

```
                },
                {
                    "entityId": "UMLS/C0002771",
                    "preferredTerm": "Analgesics",
                    "vocabularyCodes": [
                        "LNC/LP134119-9",
                        "LNC/LP31483-8",
                        "LNC/MTHU015955",
                        "MEDLINEPLUS/4059",
                        "MSH/D000700",
                        "MTH/NOCODE",
                        "NCI/C241",
                        "NCI/C29631",
                        "VANDF/4021580"
                    ]
                },
                {
                    "entityId": "UMLS/C0012833",
                    "preferredTerm": "Dizziness",
                    "vocabularyCodes": [
                        "HPO/HP:0002321",
                        "LNC/LA7428-1",
                        "MSH/D004244",
                        "MTH/NOCODE",
                        "NCI/C37943",
                        "OMIM/MTHU018247"
                    ]
                },
                {
                    "entityId": "UMLS/C0013227",
                    "preferredTerm": "Pharmaceutical Preparations",
                    "vocabularyCodes": [
                        "LNC/LA16120-0",
                        "LNC/LA20271-5",
                        "LNC/LP116700-8",
                        "LNC/LP18046-0",
                        "LNC/LP30459-9",
                        "LNC/LP72931-6",
                        "LNC/MTHU008870",
                        "LNC/MTHU038481",
                        "LNC/MTHU044418",
                        "MEDLINEPLUS/31",
                        "MSH/D004364",
                        "MTH/NOCODE",
                        "NCI/C459"
                    ]
                },
                {
                    "entityId": "UMLS/C0018681",
                    "preferredTerm": "Headache",
                    "vocabularyCodes": [
                        "HPO/HP:0002315",
                        "ICD9CM/784.0",
                        "LNC/LA15854-5",
                        "LNC/LA15903-0",
                        "LNC/LP217220-5",
                        "LNC/LP74908-2",
                        "LNC/MTHU020860",
                        "LNC/MTHU053651",
```

```
                        "MEDLINEPLUS/273",
                        "MSH/D006261",
                        "MTH/NOCODE",
                        "NCI/C34661",
                        "OMIM/MTHU036348",
                        "OMIM/MTHU036635"
                    ]
                },
                {
                    "entityId": "UMLS/C0020538",
                    "preferredTerm": "Hypertensive disease",
                    "vocabularyCodes": [
                        "HPO/HP:0000822",
                        "ICD9CM/401-405.99",
                        "ICD9CM/997.91",
                        "LNC/LA14293-7",
                        "LNC/LA7444-8",
                        "LNC/LP74941-3",
                        "LNC/MTHU020789",
                        "MEDLINEPLUS/34",
                        "MSH/D006973",
                        "MTH/005",
                        "MTH/NOCODE",
                        "NCI/C3117",
                        "OMIM/MTHU002068",
                        "OMIM/MTHU014957",
                        "OMIM/MTHU058377",
                        "OMIM/MTHU070458"
                    ]
                },
                {
                    "entityId": "UMLS/C0022885",
                    "preferredTerm": "Laboratory Procedures",
                    "vocabularyCodes": [
                        "MEDLINEPLUS/210",
                        "MTH/NOCODE",
                        "NCI/C25294"
                    ]
                },
                {
                    "entityId": "UMLS/C0030193",
                    "preferredTerm": "Pain",
                    "vocabularyCodes": [
                        "HPO/HP:0012531",
                        "ICD9CM/338-338.99",
                        "LNC/LA17107-6",
                        "LNC/LA27491-2",
                        "LNC/LA7460-4",
                        "LNC/LP75331-6",
                        "LNC/LP75342-3",
                        "LNC/MTHU021175",
                        "LNC/MTHU029813",
                        "MEDLINEPLUS/351",
                        "MSH/D010146",
                        "MTH/306",
                        "MTH/NOCODE",
                        "NCI/C25271",
                        "NCI/C3303",
                        "OMIM/MTHU033713"
```

```
                    ]
                },
                {
                    "entityId": "UMLS/C0087111",
                    "preferredTerm": "Therapeutic procedure",
                    "vocabularyCodes": [
                        "LNC/LP267221-2",
                        "LNC/LP267354-1",
                        "LNC/LP75725-9",
                        "LNC/LP75791-1",
                        "LNC/LP94823-9",
                        "LNC/MTHU021207",
                        "LNC/MTHU021209",
                        "MSH/D013812",
                        "MTH/U003371",
                        "NCI/C49236"
                    ]
                },
                {
                    "entityId": "UMLS/C0149931",
                    "preferredTerm": "Migraine Disorders",
                    "vocabularyCodes": [
                        "HPO/HP:0002076",
                        "ICD9CM/346",
                        "ICD9CM/346.9",
                        "LNC/LA15141-7",
                        "MEDLINEPLUS/3157",
                        "MSH/D008881",
                        "MTH/485",
                        "MTH/NOCODE",
                        "NCI/C89715",
                        "OMIM/157300",
                        "OMIM/MTHU000096",
                        "OMIM/MTHU000541",
                        "OMIM/MTHU068828"
                    ]
                },
                {
                    "entityId": "UMLS/C0205082",
                    "preferredTerm": "Severe (severity modifier)",
                    "vocabularyCodes": [
                        "HPO/HP:0012828",
                        "LNC/LA28568-6",
                        "LNC/LA32395-8",
                        "LNC/LA6750-9",
                        "MTH/NOCODE",
                        "NCI/C14158",
                        "NCI/C70667"
                    ]
                },
                {
                    "entityId": "UMLS/C0239574",
                    "preferredTerm": "Low grade fever",
                    "vocabularyCodes": [
                        "HPO/HP:0011134",
                        "MTH/NOCODE",
                        "NCI/C35292"
                    ]
                },
```

```
                {
                    "entityId": "UMLS/C0278138",
                    "preferredTerm": "Mild pain",
                    "vocabularyCodes": [
                        "LNC/LA15111-0",
                        "MTH/NOCODE"
                    ]
                },
                {
                    "entityId": "UMLS/C1457887",
                    "preferredTerm": "Symptoms",
                    "vocabularyCodes": [
                        "ICD9CM/780-789.99",
                        "LNC/LP185406-8",
                        "LNC/LP75708-5",
                        "LNC/MTHU021540",
                        "LNC/MTHU048142",
                        "MTH/NOCODE",
                        "NCI/C4876"
                    ]
                },
                {
                    "entityId": "UMLS/C1533734",
                    "preferredTerm": "Administration procedure",
                    "vocabularyCodes": [
                        "ICD10PCS/3",
                        "LNC/LA20296-2",
                        "MTH/NOCODE",
                        "NCI/C25409"
                    ]
                },
                {
                    "entityId": "UMLS/C1567966",
                    "preferredTerm": "Triptans",
                    "vocabularyCodes": [
                        "MSH/D014363",
                        "MTH/NOCODE"
                    ]
                },
                {
                    "entityId": "UMLS/C1837545",
                    "preferredTerm": "Nystagmus, mild",
                    "vocabularyCodes": [
                        "OMIM/MTHU001863"
                    ]
                },
                {
                    "entityId": "UMLS/C1839253",
                    "preferredTerm": "Mild scoliosis",
                    "vocabularyCodes": [
                        "OMIM/MTHU006842",
                        "OMIM/MTHU034731"
                    ]
                },
                {
                    "entityId": "UMLS/C1840417",
                    "preferredTerm": "Mild sclerosis",
                    "vocabularyCodes": [
                        "OMIM/MTHU017793"
```

```
                    ]
                },
                {
                    "entityId": "UMLS/C5203119",
                    "preferredTerm": "Intensity and Distress 5",
                    "vocabularyCodes": [
                        "MTH/NOCODE",
                        "NCI/C159939"
                    ]
                }
            ],
            "relationships": [
                {
                    "subjectId": "10",
                    "objectId": "9",
                    "confidence": 0.9987768530845642
                },
                {
                    "subjectId": "2",
                    "objectId": "1",
                    "confidence": 0.9997064471244812
                },
                {
                    "subjectId": "6",
                    "objectId": "5",
                    "confidence": 0.9996469616889954
                }
            ],
            "text": "Patient John Smith, 45-year-old male, presented with severe headaches
                 and dizziness at the ER. Noted family history of migraines. Mild
                hypertension ob-served during the physical exam. Further lab tests
                requested to determine the underlying cause of symptoms. Administered
                acetaminophen for immediate pain relief. Goal: identify the underlying
                cause and establish an appropriate treatment plan, which may involve
                prescription medications such as triptans or analgesics.",
            "responseUri": "https://healthcare.googleapis.com/v1/projects/4808913407/
                locations/us-central1/services/nlp:analyzeEntities"
        }
```


Come possiamo osservare, la risposta del servizio è incredibilmente dettagliata e completa. L'API è in grado di individuare tutte le parole rilevanti nel testo medico, evidenziandole e analizzandole, fornendo anche dati utili provenienti da importanti dizionari medici.

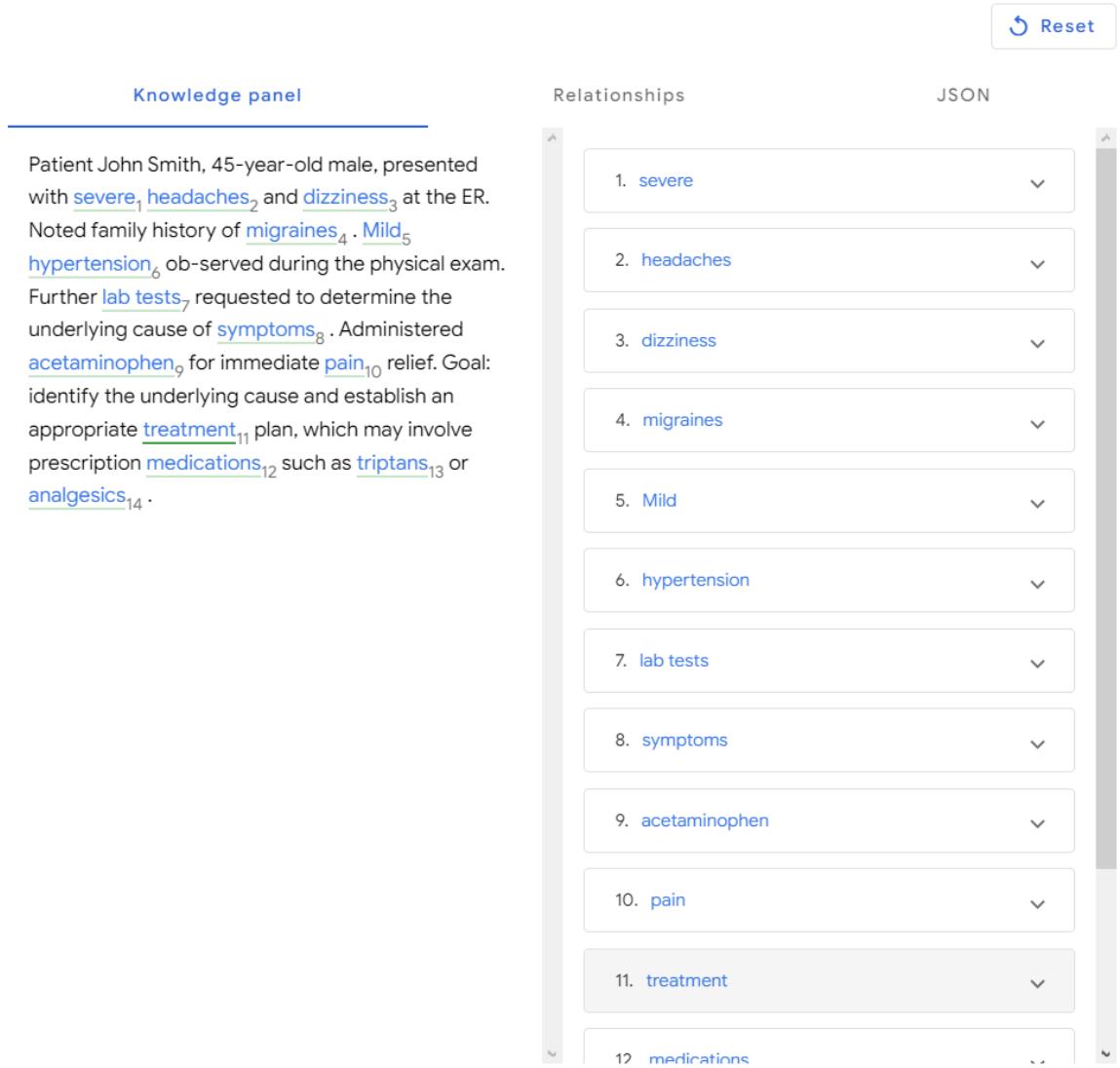


Figura 3.9: I risultati dell'analisi

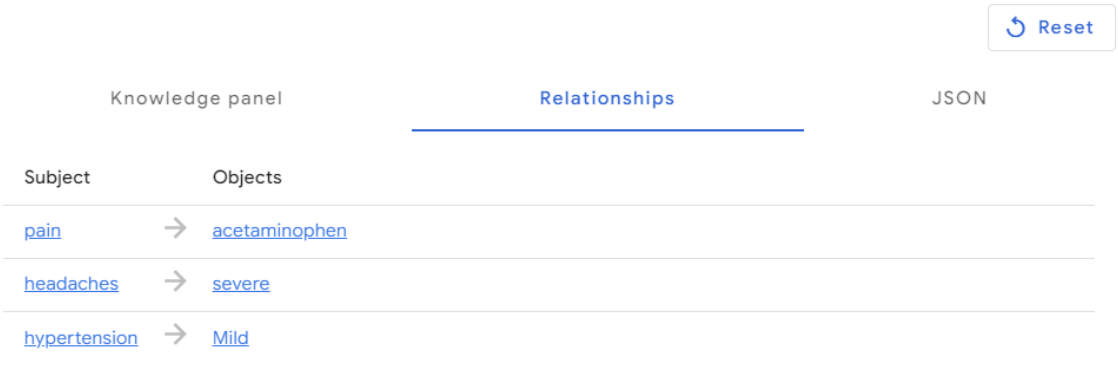


Figura 3.10: Le relationships dell'analisi

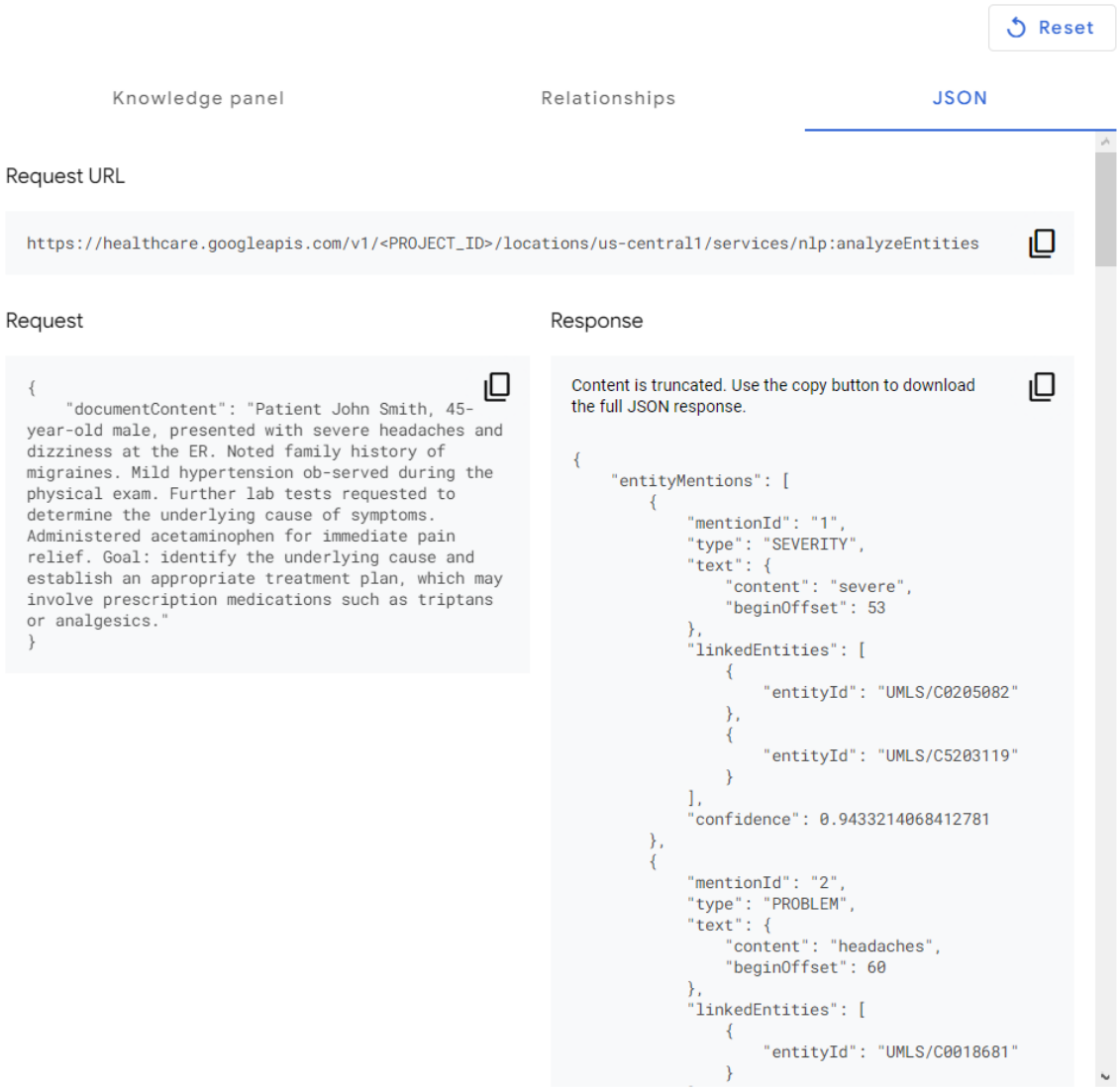


Figura 3.11: I file JSON dell'analisi

CAPITOLO 4

La Sentiment Analysis con Azure

Questo capitolo approfondisce in maniera dettagliata i servizi offerti da Azure relativi alla Sentiment Analysis e all'analisi testuale. Un'attenzione particolare è stata rivolta al servizio di Intelligenza Artificiale denominato AI Language, il quale costituisce un macro-servizio comprendente una vasta gamma di strumenti per condurre analisi approfondite del testo. Nel seguito di questo capitolo, sarà presentato in dettaglio il funzionamento di tale servizio, con un'apposita sezione dedicata all'illustrazione di esempi concreti dedicati sia all'ambito del marketing che all'ambito medico.

4.1 Cos'è Azure AI Language

Il servizio AI Language, offerto da Azure, la piattaforma di cloud computing di Microsoft, costituisce un insieme completo e avanzato di risorse che sfruttano appieno le capacità della tecnologia Natural Language Understanding (NLU), utilizzata per estrarre in maniera efficiente e precisa informazioni di valore da testi non strutturati. Questa robusta suite di strumenti è progettata per svolgere una vasta gamma di compiti sofisticati. Questi includono l'identificazione di frasi chiave, l'estrazione accurata di informazioni sensibili e personali (come PII), oltre al riconoscimento e alla classificazione di entità specifiche. Inoltre, il servizio offre la possibilità di personalizzare i modelli di estrazione delle entità in base ai domini di documenti specifici, consentendo un'adeguata adattabilità.

Un ulteriore punto di forza del servizio consiste nella sua abilità di supportare l'analisi di testi di ambito medico. Questi strumenti specializzati forniscono un valore significativo nell'ambito della medicina, in quanto sono in grado di analizzare tipologie di testi quali note mediche, sintesi di dimissioni, documenti clinici e cartelle cliniche elettroniche. Questa analisi mirata è in grado di estrarre informazioni dettagliate e rilevanti dall'enorme mole di dati presenti in questi documenti medici, fornendo, così, un supporto prezioso agli operatori sanitari e ai professionisti del settore.

4.1.1 Gli insight di Azure AI Language

Il servizio AI Language fornito da Azure offre un ampio spettro di insight estratte da documenti attraverso molteplici forme di analisi, tra cui:

- analisi delle "linked entity";

- analisi delle "named entity";
- analisi dei dati PII;
- analisi delle frasi chiave;
- analisi del sentimento;
- analisi della lingua;
- analisi delle informazioni mediche;

Nel seguito, esamineremo in modo approfondito ciascuna di queste funzionalità, mettendo in luce il modo in cui ciascuna contribuisce a offrire un'analisi esaustiva e significativa dei testi trattati.

Analisi delle "linked entity"

L'analisi delle *"linked entity"* offerta da Azure consente di identificare ed estrarre entità specifiche, come persone, luoghi, organizzazioni, date e quantità, all'interno di testi o documenti. Tuttavia, ciò che distingue l'analisi delle "linked entity" da un'analisi delle entità tradizionale è la capacità di collegare queste entità a fonti esterne per arricchire ulteriormente le informazioni estratte. In pratica, le "linked entity" creano un collegamento tra le entità menzionate nel testo e le risorse esterne, come database o ontologie. Principalmente, queste risorse esterne saranno URL di Wikipedia.

La struttura, in formato JSON, della risposta di questo tipo di analisi è la seguente:

```
{
    "documents": [
        {
            "id": "id__544",
            "entity": [
                {
                    "bingId": "9958ca5c-ea31-4e71-8a17-bd1e7839c723",
                    "name": "Los Angeles",
                    "matches": [
                        {
                            "text": "Los Angeles",
                            "offset": 31,
                            "length": 11,
                            "confidenceScore": 0.1
                        }
                    ],
                    "language": "en",
                    "id": "Los Angeles",
                    "url": "https://en.wikipedia.org/wiki/Los_Angeles",
                    "dataSource": "Wikipedia"
                },
                ...
            ],
            "warnings": []
        }
    ],
    "errors": [],
    "modelVersion": "2021-06-01"
}
```

Nel contesto attuale, emergono i seguenti campi di interesse:

1. *"bingId"*: indica un codice univoco assegnato all'entità in base ai dati provenienti da Bing.
2. *"offset"*: presente anche in altre analisi, indica il punto iniziale dell'entità nel testo.
3. *"confidenceScore"*: presente anche in altre analisi, denota il grado di sicurezza con cui il servizio è in grado di riconoscere una specifica entità.
4. *"url"*: segnala l'URL associato all'entità individuato durante la ricerca.
5. *"dataSource"*: specifica la fonte da cui sono stati recuperati i dati.

Analisi delle "named entity"

L'analisi delle *"named entity"*, resa disponibile tramite Azure, si avvicina maggiormente all'analisi delle entità tradizionale. In questo processo, mediante l'uso avanzato dell'elaborazione del linguaggio naturale (NLP), vengono estratti dal testo oggetti o concetti specifici, quali nomi di persone, organizzazioni, luoghi, date, numeri di telefono e altro ancora. Queste entità vengono, inoltre, categorizzate per tipologia, fornendo ulteriore contesto all'analisi. Questo processo consente, quindi, di acquisire una comprensione più profonda del contenuto testuale.

La struttura, in formato JSON, della risposta di questo tipo di analisi è la seguente:

```
{
    "documents": [
        {
            "id": "id__467",
            "entity": [
                {
                    "text": "wife",
                    "category": "PersonType",
                    "offset": 3,
                    "length": 4,
                    "confidenceScore": 0.99
                },
                {
                    "text": "Los Angeles",
                    "category": "Location",
                    "subcategory": "GPE",
                    "offset": 31,
                    "length": 11,
                    "confidenceScore": 1
                },
                {
                    "text": "first",
                    "category": "Quantity",
                    "subcategory": "Ordinal",
                    "offset": 51,
                    "length": 5,
                    "confidenceScore": 0.8
                },
                ...
            ],
            "warnings": []
        }
```

```
            ],
            "errors": [],
            "modelVersion": "2021-06-01"
        }
```

Analisi dei dati PII

L'analisi dei dati *PII* (Personally Identifiable Information), fornita da Azure, segna un passo avanti nella gestione dei dati sensibili. Utilizzando avanzate tecnologie NLP (elaborazione del linguaggio naturale) e IA (Intelligenza Artificiale), consente l'individuazione e la classificazione automatica di informazioni personali, come nomi, indirizzi e numeri di carta di credito all'interno dei testi. Questo non solo rafforza la sicurezza dei dati, ma offre anche una visione più chiara delle informazioni raccolte, promuovendo la conformità alle normative sulla privacy e la fiducia dei clienti.

La struttura, in formato JSON, della risposta di questo tipo di analisi è la seguente:

```
        {
            "documents": [
                {
                    "redactedText": "Hello, my name is ***********. I lost my Credit card
                        on ***********, and I would like to request its cancellation. The
                        last purchase I made was of a Chicken parmigiana dish at Contoso
                        Restaurant, located near the Hollywood Museum, for $40. Below is
                        my personal information for validation:\n        Profession:
                        **********\n        Social Security number is 123-45-6789\n
                               Date of birth: ********\n        Phone number:
                        ************\n        Personal address:
                        ************************************\n        Linked email
                        account: ***************************\n        Swift code:
                        ***********",
                    "id": "id__1619",
                    "entity": [
                        {
                            "text": "Mateo Gomez",
                            "category": "Person",
                            "offset": 18,
                            "length": 11,
                            "confidenceScore": 0.99
                        },
                        {
                            "text": "August 17th",
                            "category": "DateTime",
                            "subcategory": "Date",
                            "offset": 56,
                            "length": 11,
                            "confidenceScore": 0.8
                        },
                        ...
                    ],
                    "warnings": []
                }
            ],
            "errors": [],
            "modelVersion": "2021-01-15"
        }
```

Come è evidente, il risultato dell'analisi non solo fornisce un'analisi delle entità PII, categorizzandole, ma presenta anche una versione del testo in cui tali informazioni vengono opportunamente oscurate per garantire la protezione della privacy.

Analisi delle frasi chiave

L'analisi delle *frasi chiave*, offerta da Azure, permette di identificare le porzioni di testo più significative e rappresentative di un documento. In pratica, il sistema è in grado di estrarre le frasi che catturano l'essenza del contenuto, fornendo un'anteprima sintetica, ma completa, dell'informazione contenuta. Questa capacità di rilevare le frasi chiave non solo semplifica l'analisi di testi lunghi e complessi, ma consente anche una maggiore efficienza nella ricerca e nell'individuazione di contenuti rilevanti.

La struttura, in formato JSON, della risposta di questo tipo di analisi è la seguente:

```
{
    "documents": [
        {
            "id": "id__2175",
            "keyPhrases": [
                "1234 Hollywood Boulevard Los Angeles CA",
                "social security number",
                "following email address",
                ...
            ],
            "warnings": []
        }
    ],
    "errors": [],
    "modelVersion": "2022-10-01"
}
```

Analisi del sentimento

L'analisi del *sentimento*, fornita da Azure, è un servizio che, attraverso l'impiego di algoritmi di apprendimento automatico, consente di comprendere e valutare le emozioni e le opinioni espresse nei testi. Il sistema è in grado di distinguere tra sentimenti positivi, negativi e neutri, offrendo un'analisi dettagliata delle reazioni degli utenti o dei clienti. Questo strumento trova applicazione in diverse aree, come il monitoraggio dell'opinione pubblica sui prodotti o servizi di un'azienda, la valutazione dell'efficacia di una campagna di marketing o, addirittura, l'identificazione di tendenze emergenti.

La struttura, in formato JSON, della risposta di questo tipo di analisi è la seguente:

```
{
    "documents": [
        {
            "id": "id__3169",
            "sentiment": "mixed",
            "confidenceScores": {
                "positive": 0.73,
                "neutral": 0.03,
                "negative": 0.25
            },
            "sentences": [
                ...
                {
```

```
                        "sentiment": "negative",
                        "confidenceScores": {
                            "positive": 0,
                            "neutral": 0.02,
                            "negative": 0.98
                        },
                        "offset": 40,
                        "length": 84,
                        "text": "I found the zipper a little bit difficult to get up &
                             down due to the side rushing. ",
                        "targets": [
                            {
                                "sentiment": "negative",
                                "confidenceScores": {
                                    "positive": 0,
                                    "negative": 1
                                },
                                "offset": 52,
                                "length": 6,
                                "text": "zipper",
                                "relations": [
                                    {
                                        "relationType": "assessment",
                                        "ref": "#/documents/0/sentences/1/assessments/
                                            0"
                                    }
                                ]
                            }
                        ],
                        "assessments": [
                            {
                                "sentiment": "negative",
                                "confidenceScores": {
                                    "positive": 0,
                                    "negative": 1
                                },
                                "offset": 72,
                                "length": 9,
                                "text": "difficult",
                                "isNegated": false
                            }
                        ]
                    },
                    ...
                ],
                "warnings": []
            }
        ],
        "errors": [],
        "modelVersion": "2022-11-01"
    }
```


Qui, come è possibile notare, l'analisi del sentimento non si limita soltanto al livello del documento, bensì si estende anche a quello delle singole frasi. Infatti, l'output dell'analisi include i dati relativi al sentimento, accuratamente annotati nei campi "confidenceScore", per ciascuna frase. All'interno di questa analisi, vengono, inoltre, evidenziati gli "assessment", ovvero le valutazioni effettuate riguardo a specifiche entità presenti nel testo, contribuendo, così, a delineare un sentimento complessivamente positivo o negativo.

Analisi della lingua

L'analisi della *lingua* tramite Azure è mirata all'identificazione della lingua principale all'interno di un testo. Questa funzionalità si basa su algoritmi avanzati per determinare con precisione la lingua predominante, risultando particolarmente utile in scenari multilingue. Tale capacità di identificare la lingua di origine può essere impiegata per ottimizzare processi come la traduzione automatica, contribuendo a migliorare la qualità delle traduzioni e l'efficienza complessiva delle comunicazioni multilingue.

La struttura, in formato JSON, della risposta di questo tipo di analisi è la seguente:

```
{
    "documents": [
        {
            "id": "id__4022",
            "detectedLanguage": {
                "name": "Spanish",
                "iso6391Name": "es",
                "confidenceScore": 0.7
            },
            "warnings": []
        }
    ],
    "errors": [],
    "modelVersion": "2022-10-01"
}
```

Analisi delle informazioni mediche

Il servizio di analisi delle *informazioni mediche* offerto da Azure è in grado di effettuare un'accurata estrazione di dati cruciali da documenti clinici. Tale processo di estrazione mirata consente ai professionisti della salute di identificare diagnosi, trattamenti e tendenze in modo più efficiente, fornendo, così, un supporto per decisioni informate e personalizzate.

Oltre a semplificare la gestione dei dati medici, questo servizio si rivela un'opportunità preziosa nell'analisi delle ricerche mediche su vasta scala. La sua capacità di identificare con precisione parole chiave e concetti all'interno dei testi scientifici facilita una ricerca più mirata e approfondita, accelerando, così, la scoperta di nuovi trattamenti, protocolli e informazioni diagnostiche.

La struttura, in formato JSON, della risposta di questo tipo di analisi è la seguente:

```
{
    "documents": [
        {
            "id": "id__5722",
            "entity": [
                {
                    "offset": 53,
                    "length": 9,
                    "text": "2/14/2001",
                    "category": "Date",
                    "confidenceScore": 1
                },
                {
                    "offset": 63,
                    "length": 11,
```

```
                            "text": "12:00:00 AM",
                            "category": "Time",
                            "confidenceScore": 0.98
                        },
                        {
                            "offset": 77,
                            "length": 23,
                            "text": "CORONARY ARTERY DISEASE",
                            "category": "Diagnosis",
                            "confidenceScore": 1,
                            "name": "Coronary Artery Disease",
                            "links": [
                                {
                                    "dataSource": "UMLS",
                                    "id": "C1956346"
                                },
                                {
                                    "dataSource": "AOD",
                                    "id": "0000005327"
                                },
                                ...
                            ]
                        },
                        ...
                    ],
                    "warnings": []
                }
            ],
            "errors": [],
            "modelVersion": "2022-08-15-preview"
        }
```


Come è possibile constatare, dalla struttura stessa dell'esito di questa analisi, ciascuna entità medica individuata non solo viene categorizzata, ma viene altresì collegata al campo "links". Quest'ultimo campo presenta ulteriori sotto-campi in cui è specificata una "dataSource", ovvero una fonte dati, a cui fare riferimento per ottenere una comprensione più dettagliata dell'entità identificata. In aggiunta, è presente anche un "id" che può essere utilizzato per localizzare e richiamare l'entità all'interno di questa risorsa di dati.

4.1.2 Come funziona Azure AI Language

Azure offre una gamma di strumenti essenziali per l'analisi dei testi, con il più intuitivo denominato Language Studio, visibile nella Figura 4.1. Questo servizio ci consente di accedere alle diverse console degli insight individuali, i quali sono ospitati nelle sezioni identificate come "Classify text" ed "Extract information".

La configurazione delle console per ciascun insight individuale è uniforme; è una composta da una sezione dedicata all'inserimento del testo, visibile nella Figura 4.2, e da un'altra sezione in cui vengono visualizzati i risultati, evidenziata nella Figura 4.3. Nella prima sezione è possibile trovare dei pratici pulsanti posizionati sotto l'area di inserimento che forniscono dei testi predefiniti. Nella seconda sezione, in aggiunta, è presente un pulsante che consente di visualizzare la risposta in formato JSON.

Azure fornisce, inoltre, una gamma completa di strumenti dedicati all'analisi dei testi, comprensivi di librerie ottimizzate per lo sviluppo di applicazioni in vari linguaggi di programmazione, quali C#, Python, Java e Javascript. Inoltre, questa piattaforma

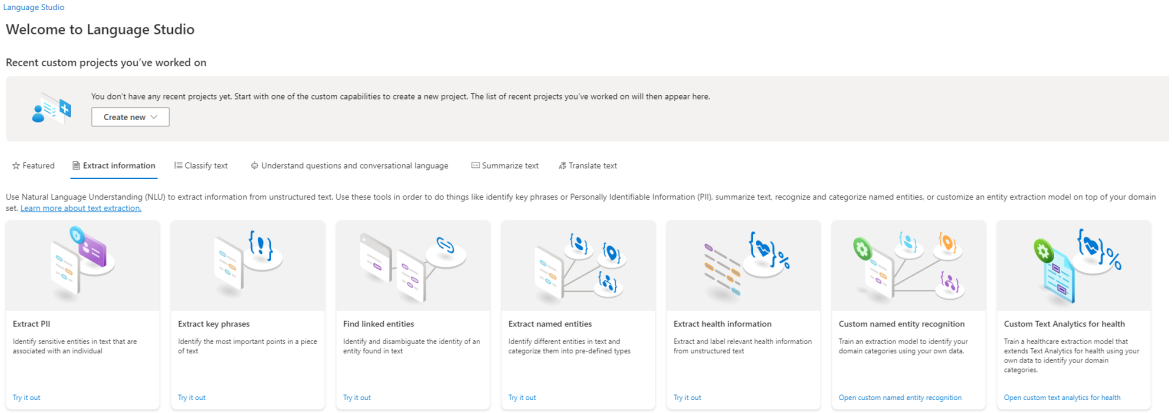


Figura 4.1: Language studio

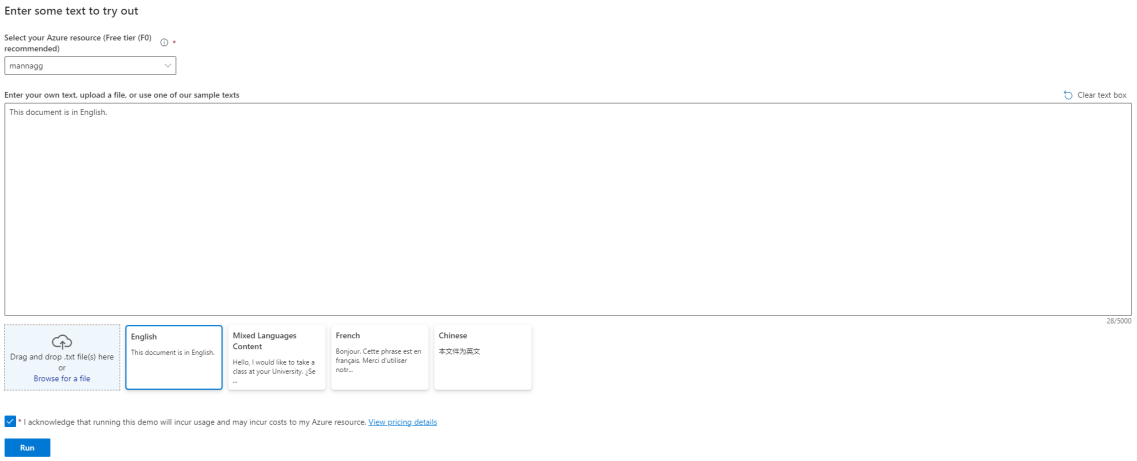


Figura 4.2: La console dell'insight

offre anche la flessibilità di utilizzare chiamate REST API per l'accesso alle funzionalità di analisi. Un esempio pratico di come utilizzare le librerie in Python è illustrato nel repository reperibile al seguente URL:

`https://github.com/Walter-Di-Sabatino/AI-Language-Example.git`

In questo esempio, l'importazione di metodi dalla libreria Azure e la creazione di un oggetto `client` hanno permesso lo sviluppo di funzioni apposite. Queste consentono di analizzare le informazioni significative estratte dal testo e di visualizzare in modo chiaro i risultati ottenuti. Le funzioni implementate includono:

- `analyze_sentiment`
- `analyze_entity`
- `analyze_linked_entity`
- `analyze_pii`
- `detect_primary_language`
- `analyze_sentiment`
- `analyze_key_phrases`

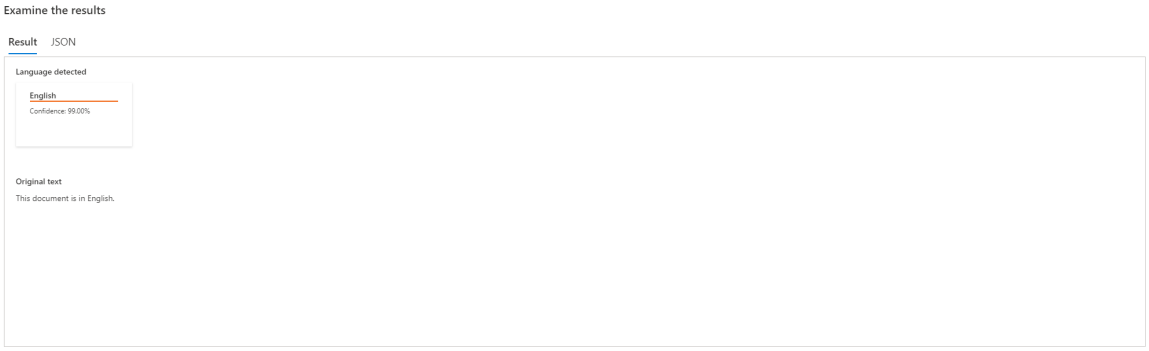


Figura 4.3: I risultati dell'analisi dell'insight

Ciascuna di esse, quindi, come si può intuire, è dedicata all'analisi di un singolo insight.

4.2 Esempi con Azure AI Language

In questa sezione, saranno presentati esempi d'utilizzo dell'AI Language di Azure. In tali esempi verrà condotta un'analisi sia su un testo "generico", includendo l'analisi di tutti gli insight, che su un testo medico. In quest'ultimo caso, si farà uso dell'insight fornito dal servizio dedicato all'estrazione di dati medici.

4.2.1 Analisi di un testo

Al fine di illustrare il funzionamento dell'AI Laguage di Azure è stato scelto il testo seguente in inglese:

> John Smith is a software engineer who lives happly in New York City. He enjoys playing videogames and loves listening to rock music. His favorite book series is 'The Lord of the Rings', and he has a pet Labrador named Max. If you you would like to contact him his email is: fictionalEmail@gmail.com.

È stata selezionata la lingua inglese in modo tale da sfruttare appieno le capacità di analisi del servizio. Di seguito, è riportata la traduzione letterale del testo:

> John Smith è un ingegnere informatico che vive felicemente a New York. Si diverte a giocare ai videogiochi e ama ascoltare la musica rock. La sua serie di libri preferita è "Il Signore degli Anelli" e ha un Labrador di nome Max. Se volete contattarlo, il suo indirizzo e-mail è: fictionalEmail@gmail.com.

Per agevolare la presentazione dei risultati, saranno mostrate esclusivamente le immagini delle console ad essi relativi e i corrispondenti file JSON.

Risultati dell'analisi delle "linked entity"

Il risultato dell'analisi delle *"linked entity"* all'interno della console è quello riportato all'interno della Figura 4.4. Invece l'analisi rispetto alle entità, in formato JSON, appare come mostrato nel listato successivo:

Figura 4.4: I risultati dell'analisi delle "linked entity"


```
{
    "documents": [
        {
            "id": "id__445",
            "entity": [
                {
                    "bingId": "0753ace5-7704-daf0-650f-85e6de72168c",
                    "name": "John Smith (explorer)",
                    "matches": [
                        {
                            "text": "John Smith",
                            "offset": 0,
                            "length": 10,
                            "confidenceScore": 0.03
                        }
                    ],
                    "language": "en",
                    "id": "John Smith (explorer)",
                    "url": "https://en.wikipedia.org/wiki/John_Smith_(explorer)",
                    "dataSource": "Wikipedia"
                },
                {
                    "bingId": "60d5dc2b-c915-460b-b722-c9e3485499ca",
                    "name": "New York City",
                    "matches": [
                        {
                            "text": "New York City",
                            "offset": 54,
                            "length": 13,
                            "confidenceScore": 0.22
                        }
                    ],
                    "language": "en",
                    "id": "New York City",
                    "url": "https://en.wikipedia.org/wiki/New_York_City",
                    "dataSource": "Wikipedia"
                },
                {
                    "bingId": "14e05ec2-edce-7b74-9725-4d99b75243b7",
                    "name": "The Lord of the Rings (film series)",
                    "matches": [
```

```
                                {
                                    "text": "The Lord of the Rings",
                                    "offset": 162,
                                    "length": 21,
                                    "confidenceScore": 0.04
                                }
                            ],
                            "language": "en",
                            "id": "The Lord of the Rings (film series)",
                            "url": "https://en.wikipedia.org/wiki/The_Lord_of_the_Rings_(
                                film_series)",
                            "dataSource": "Wikipedia"
                        },
                        {
                            "bingId": "b13b9c99-0331-e767-33de-2d895157883b",
                            "name": "Labrador Retriever",
                            "matches": [
                                {
                                    "text": "Labrador",
                                    "offset": 203,
                                    "length": 8,
                                    "confidenceScore": 0.17
                                }
                            ],
                            "language": "en",
                            "id": "Labrador Retriever",
                            "url": "https://en.wikipedia.org/wiki/Labrador_Retriever",
                            "dataSource": "Wikipedia"
                        },
                        {
                            "bingId": "257b909c-a141-f1b5-f328-da73c7e69510",
                            "name": "Max (software)",
                            "matches": [
                                {
                                    "text": "Max",
                                    "offset": 218,
                                    "length": 3,
                                    "confidenceScore": 0.07
                                }
                            ],
                            "language": "en",
                            "id": "Max (software)",
                            "url": "https://en.wikipedia.org/wiki/Max_(software)",
                            "dataSource": "Wikipedia"
                        }
                    ],
                    "warnings": []
                }
            ],
            "errors": [],
            "modelVersion": "2021-06-01"
        }
```


I risultati dell'analisi delle "linked entity" possono essere considerati accurati, poiché il servizio analizza in modo appropriato tutte le entità e fornisce collegamenti a pagine di Wikipedia come riferimenti.

Risultati dell'analisi delle "named entity"

Il risultato dell'analisi delle *"named entity"* all'interno della console è quello riportato all'interno della Figura 4.5. Invece l'analisi rispetto alle entità, in formato JSON, appare come mostrato nel listato successivo:

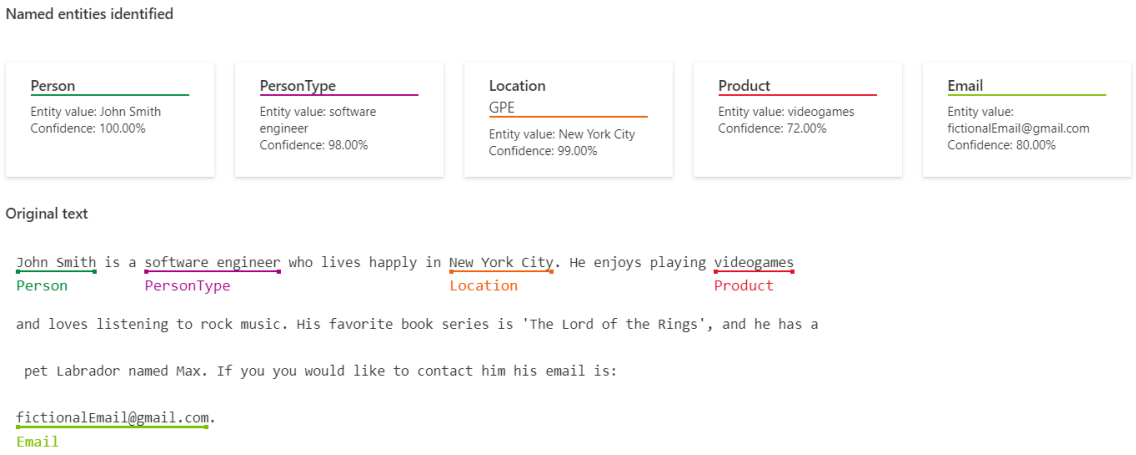


Figura 4.5: I risultati dell'analisi delle "named entity"


```
{
    "documents": [
        {
            "id": "id__943",
            "entity": [
                {
                    "text": "John Smith",
                    "category": "Person",
                    "offset": 0,
                    "length": 10,
                    "confidenceScore": 1
                },
                {
                    "text": "software engineer",
                    "category": "PersonType",
                    "offset": 16,
                    "length": 17,
                    "confidenceScore": 0.98
                },
                {
                    "text": "New York City",
                    "category": "Location",
                    "subcategory": "GPE",
                    "offset": 54,
                    "length": 13,
                    "confidenceScore": 0.99
                },
                {
                    "text": "videogames",
                    "category": "Product",
                    "offset": 87,
                    "length": 10,
                    "confidenceScore": 0.72
                },
```

```
                        {
                            "text": "fictionalEmail@gmail.com",
                            "category": "Email",
                            "offset": 274,
                            "length": 24,
                            "confidenceScore": 0.8
                        }
                    ],
                    "warnings": []
                }
            ],
            "errors": [],
            "modelVersion": "2021-06-01"
        }
```


I risultati ottenuti da questo tipo di analisi risultano essere parziali. Infatti, sebbene il servizio riesca ad analizzare con precisione alcune entità e a identificarne correttamente i tipi, ne tralascia altre, come, ad esempio, "rock music", "labrador" o "The Lord Of The Rings", che non vengono opportunamente considerate.

Risultati dell'analisi dei dati PII

Il risultato dell'analisi dei dati *PII* all'interno della console è quello riportato all'interno della Figura 4.6. Invece l'analisi rispetto alle entità, in formato JSON, appare come mostrato nel listato successivo:

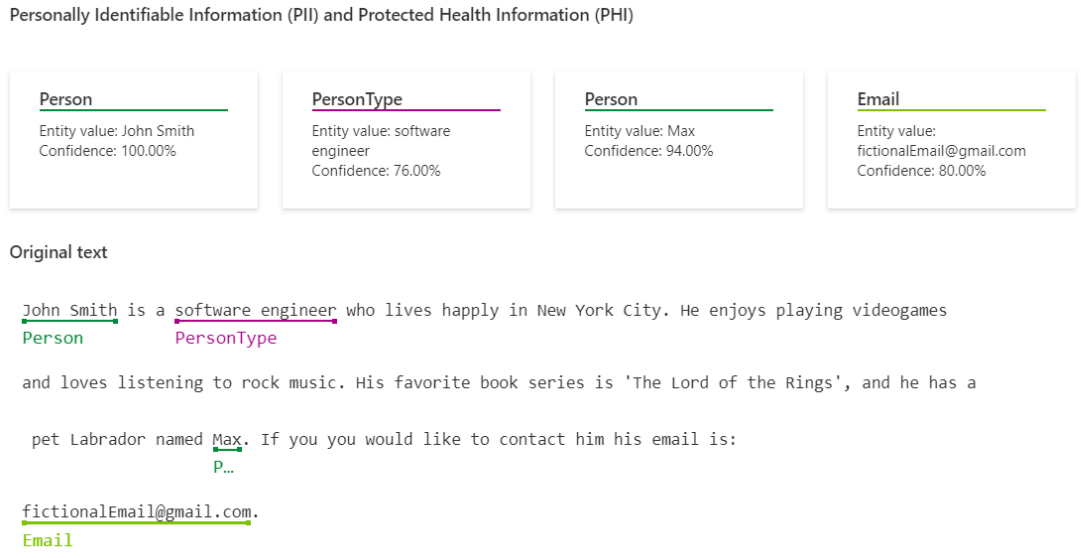


Figura 4.6: I risultati dell'analisi dei dati PII


```
        {
            "documents": [
                {
                    "redactedText": "********** is a ********** who lives happly in
                         New York City. He enjoys playing videogames and loves listening
                        to rock music. His favorite book series is 'The Lord of the Rings'
                        , and he has a pet Labrador named ***. If you you would like to
                        contact him his email is: **********.",
                    "id": "id__1437",
```

```
                    "entity": [
                        {
                            "text": "John Smith",
                            "category": "Person",
                            "offset": 0,
                            "length": 10,
                            "confidenceScore": 1
                        },
                        {
                            "text": "software engineer",
                            "category": "PersonType",
                            "offset": 16,
                            "length": 17,
                            "confidenceScore": 0.76
                        },
                        {
                            "text": "Max",
                            "category": "Person",
                            "offset": 218,
                            "length": 3,
                            "confidenceScore": 0.94
                        },
                        {
                            "text": "fictionalEmail@gmail.com",
                            "category": "Email",
                            "offset": 274,
                            "length": 24,
                            "confidenceScore": 0.8
                        }
                    ],
                    "warnings": []
                }
            ],
            "errors": [],
            "modelVersion": "2021-01-15"
        }
```

I risultati di questa analisi sono precisi; il servizio dimostra di essere capace di esaminare in maniera appropriata tutte le informazioni personali contenute nel testo. L'unico punto che potrebbe essere considerato un errore è l'identificazione della parola "Max" come una Persona, nonostante essa si riferisca al cane del protagonista.

Risultati dell'analisi delle frasi chiave

Il risultato dell'analisi delle *frasi chiave* all'interno della console è quello riportato all'interno della Figura 4.7. Invece l'analisi rispetto alle entità, in formato JSON, appare come mostrato nel seguente listato:

```
        {
            "documents": [
                {
                    "id": "id__1916",
                    "keyPhrases": [
                        "New York City",
                        "favorite book series",
                        "John Smith",
                        "software engineer",
                        "rock music",
```

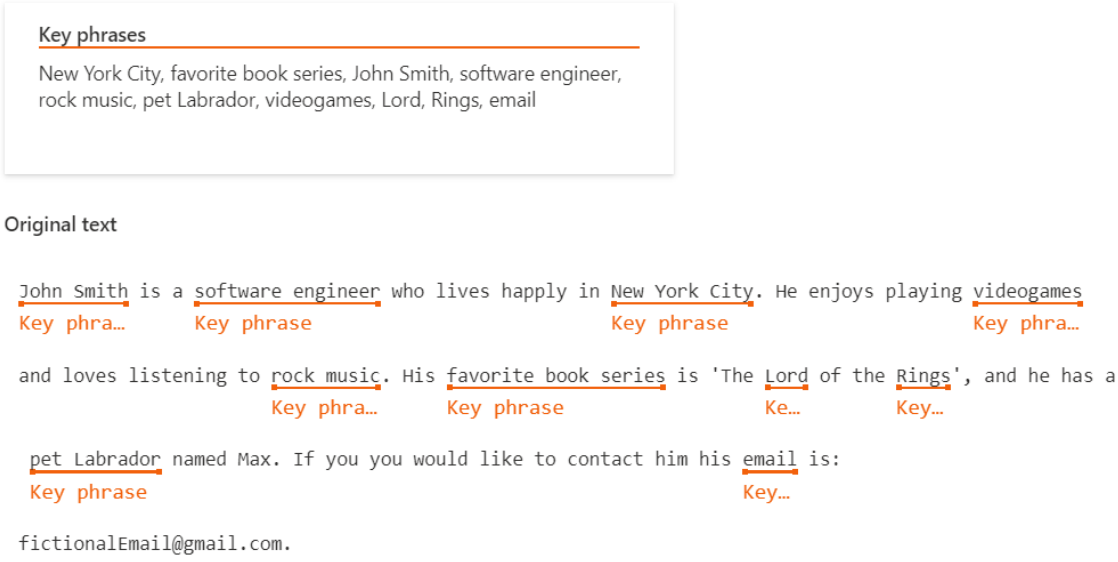


Figura 4.7: I risultati dell'analisi delle frasi chiave

```
                        "pet Labrador",
                        "videogames",
                        "Lord",
                        "Rings",
                        "email"
                    ],
                    "warnings": []
                }
            ],
            "errors": [],
            "modelVersion": "2022-10-01"
        }
```

Le frasi chiave individuate sono generalmente corrette; l'unico errore è la separazione delle parole "Lord" e "Rings" che, in realtà, fanno parte di un unico titolo.

Risultati dell'analisi del sentimento

Il risultato dell'analisi del *sentimento* all'interno della console è quello riportato all'interno delle Figure 4.8 e 4.9. Invece l'analisi rispetto alle entità, in formato JSON, appare come mostrato nel seguente listato:

```
        {
            "documents": [
                {
                    "id": "id__2402",
                    "sentiment": "positive",
                    "confidenceScores": {
                        "positive": 0.76,
                        "neutral": 0.23,
                        "negative": 0.01
                    },
                    "sentences": [
                        {
```

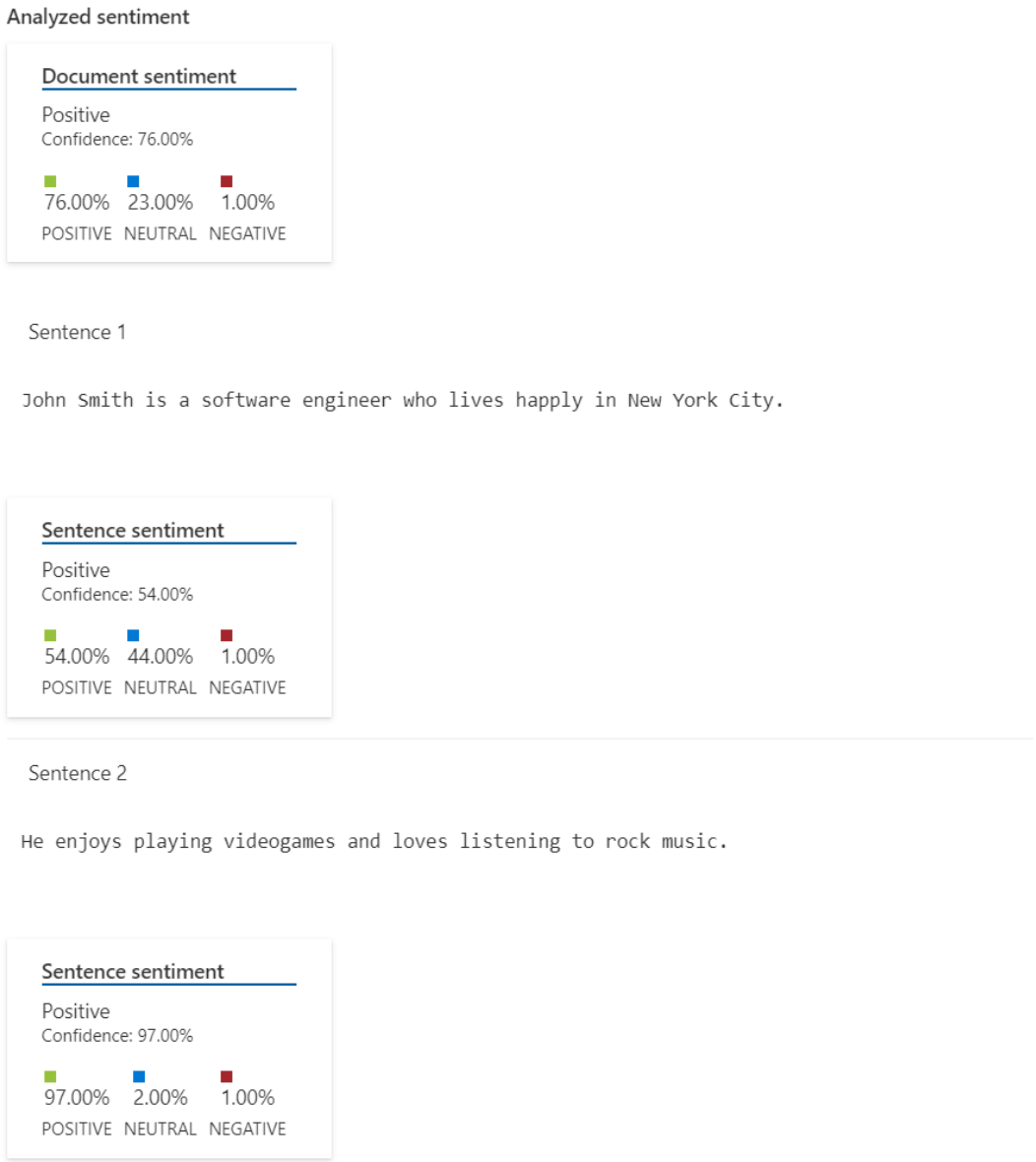


Figura 4.8: Risultati dell'analisi del sentimento (prima parte)


```
                        "sentiment": "positive",
                        "confidenceScores": {
                            "positive": 0.54,
                            "neutral": 0.44,
                            "negative": 0.01
                        },
                        "offset": 0,
                        "length": 69,
                        "text": "John Smith is a software engineer who lives happly in
                             New York City. ",
                        "targets": [],
                        "assessments": []
                    },
                    {
                        "sentiment": "positive",
                        "confidenceScores": {
                            "positive": 0.97,
```

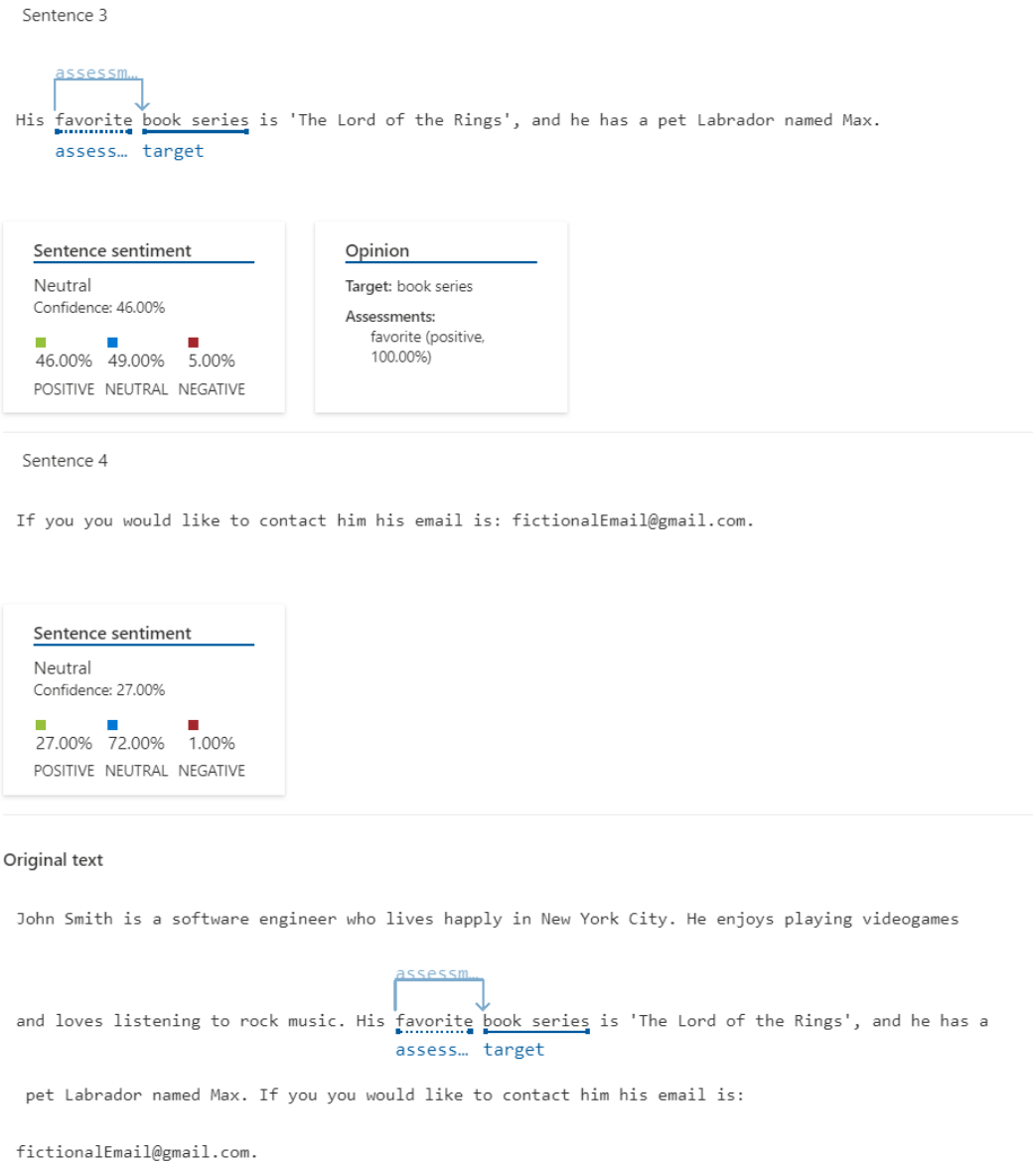


Figura 4.9: Risultati dell'analisi del sentimento (seconda parte)


```
                            "neutral": 0.02,
                            "negative": 0.01
                        },
                        "offset": 69,
                        "length": 64,
                        "text": "He enjoys playing videogames and loves listening to
                            rock music. ",
                        "targets": [],
                        "assessments": []
                    },
                    {
                        "sentiment": "neutral",
                        "confidenceScores": {
                            "positive": 0.46,
                            "neutral": 0.49,
                            "negative": 0.05
                        },
```

```
                            "offset": 133,
                            "length": 90,
                            "text": "His favorite book series is 'The Lord of the Rings',
                                and he has a pet Labrador named Max. ",
                            "targets": [
                                {
                                    "sentiment": "positive",
                                    "confidenceScores": {
                                        "positive": 1,
                                        "negative": 0
                                    },
                                    "offset": 146,
                                    "length": 11,
                                    "text": "book series",
                                    "relations": [
                                        {
                                            "relationType": "assessment",
                                            "ref": "#/documents/0/sentences/2/assessments/
                                                0"
                                        }
                                    ]
                                }
                            ],
                            "assessments": [
                                {
                                    "sentiment": "positive",
                                    "confidenceScores": {
                                        "positive": 1,
                                        "negative": 0
                                    },
                                    "offset": 137,
                                    "length": 8,
                                    "text": "favorite",
                                    "isNegated": false
                                }
                            ]
                        },
                        {
                            "sentiment": "neutral",
                            "confidenceScores": {
                                "positive": 0.27,
                                "neutral": 0.72,
                                "negative": 0.01
                            },
                            "offset": 223,
                            "length": 76,
                            "text": "If you you would like to contact him his email is:
                                fictionalEmail@gmail.com.",
                            "targets": [],
                            "assessments": []
                        }
                    ],
                    "warnings": []
                }
            ],
            "errors": [],
            "modelVersion": "2022-11-01"
        }
```

Anche l'analisi del sentimento risulta essere accurata. Infatti il testo risulta mostrare più un sentimento positivo che neutro o negativo, soprattutto grazie alla descrizione delle passioni del protagonista. In particolare, anche l'analisi del sentimento delle singole frasi risulta essere adeguato. Ad esempio, nella frase "Il suo ciclo di libri preferito è 'Il Signore degli Anelli', e ha un cane Labrador di nome Max.", viene attribuito, grazie alla prima parte della frase, un punteggio positivo del 46%; invece, grazie alla seconda parte, viene attribuito un punteggio neutro del 49%. Il punteggio negativo, invece, risulta essere solo del 5%.

Risultati dell'analisi della lingua

Il risultato dell'analisi della *lingua* all'interno della console è quello riportato all'interno della Figura 4.10. Invece l'analisi rispetto alle entità, in formato JSON, appare come mostrato nel seguente listato:

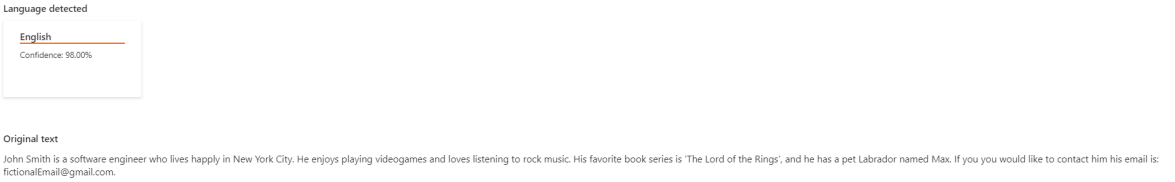


Figura 4.10: I risultati dell'analisi della lingua

```
{
    "documents": [
        {
            "id": "id__2885",
            "detectedLanguage": {
                "name": "English",
                "iso6391Name": "en",
                "confidenceScore": 0.98
            },
            "warnings": []
        }
    ],
    "errors": [],
    "modelVersion": "2022-10-01"
}
```

Come si può notare il servizio identifica con facilità e precisione la lingua utilizzata all'interno del testo.

4.2.2 Analisi di un testo medico

Al fine di illustrare il funzionamento dell'analisi di un testo medico nel servizio AI Language di Azure è stato scelto il seguente testo in inglese:

> Patient John Smith, 45-year-old male, presented with severe headaches and dizziness at the ER. Noted family history of migraines. Mild hypertension observed during the physical exam. Further lab tests requested to determine the underlying cause of symptoms. Administered acetaminophen for immediate pain relief. Goal: identify the underlying cause and establish an appropriate

> treatment plan, which may involve prescription medications such as triptans or analgesics.

Anche in questo caso è stata selezionata la lingua inglese in modo tale da poter sfruttare appieno le capacità di analisi del servizio. Di seguito, è riportata la traduzione letterale del testo:

> Il paziente John Smith, uomo di 45 anni, si è presentato al Pronto Soccorso con forti mal di testa e vertigini. Anamnesi familiare di emicrania. Durante l'esame fisico è stata osservata una lieve ipertensione. Sono stati richiesti ulteriori esami di laboratorio per determinare la causa dei sintomi. Somministrazione di acetaminofene per alleviare immediatamente il dolore. Obiettivo: identificare la causa sottostante e stabilire un piano di trattamento appropriato, che può comportare la prescrizione di farmaci come triptani o analgesici.

Il risultato dell'analisi dei dati *medici* all'interno della console è quello riportato all'interno della Figura 4.11. Invece l'analisi rispetto alle entità, in formato JSON, la quale non verrà mostrata tutta per brevità, appare come mostrato nel seguente listato:

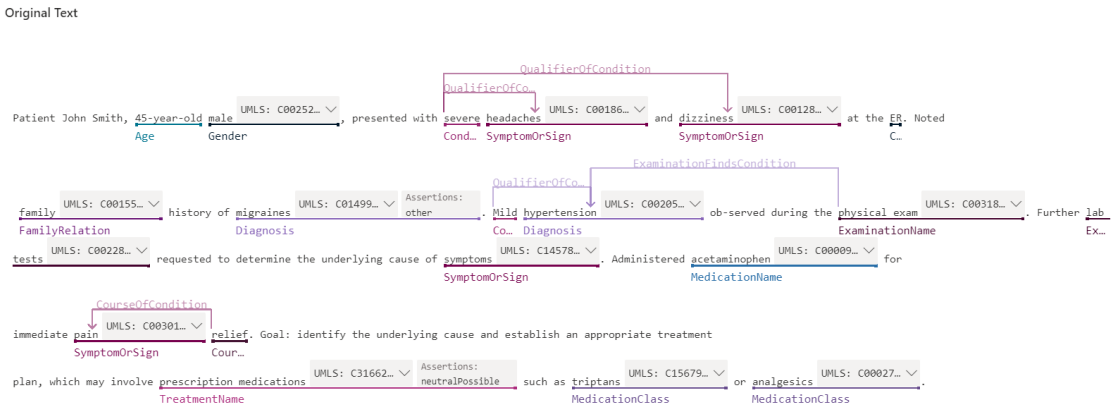


Figura 4.11: I risultati dell'analisi dei dati medici

```
{
    "documents": [
        {
            "id": "id__893",
            "entity": [
                {
                    "offset": 20,
                    "length": 11,
                    "text": "45-year-old",
                    "category": "Age",
                    "confidenceScore": 1
                },
                {
                    "offset": 32,
                    "length": 4,
                    "text": "male",
                    "category": "Gender",
                    "confidenceScore": 1,
                    "name": "Male population group",
                    "links": [
```

```
                                {
                                    "dataSource": "UMLS",
                                    "id": "C0025266"
                                },
                                {
                                    "dataSource": "AOD",
                                    "id": "0000026918"
                                },
                                ...
                            ]
                        },
                        {
                            "offset": 53,
                            "length": 6,
                            "text": "severe",
                            "category": "ConditionQualifier",
                            "confidenceScore": 1
                        },
                        {
                            "offset": 60,
                            "length": 9,
                            "text": "headaches",
                            "category": "SymptomOrSign",
                            "confidenceScore": 0.99,
                            "name": "Headache",
                            "links": [
                                {
                                    "dataSource": "UMLS",
                                    "id": "C0018681"
                                },
                                {
                                    "dataSource": "AOD",
                                    "id": "0000006115"
                                },
                                ...
                            ]
                        },
                        {
                            "offset": 74,
                            "length": 9,
                            "text": "dizziness",
                            "category": "SymptomOrSign",
                            "confidenceScore": 1,
                            "name": "Dizziness",
                            "links": [
                                {
                                    "dataSource": "UMLS",
                                    "id": "C0012833"
                                },
                                {
                                    "dataSource": "AOD",
                                    "id": "0000004419"
                                },
                                ...
                            ]
                        },
                        {
                            "offset": 91,
                            "length": 2,
```

```
                        "text": "ER",
                        "category": "CareEnvironment",
                        "confidenceScore": 1
                    },
                    {
                        "offset": 101,
                        "length": 6,
                        "text": "family",
                        "category": "FamilyRelation",
                        "confidenceScore": 0.98,
                        "name": "Family",
                        "links": [
                            {
                                "dataSource": "UMLS",
                                "id": "C0015576"
                            },
                            {
                                "dataSource": "AOD",
                                "id": "0000006914"
                            },
                            ...
                        ]
                    },
                    {
                        "offset": 119,
                        "length": 9,
                        "text": "migraines",
                        "category": "Diagnosis",
                        "confidenceScore": 0.99,
                        "assertion": {
                            "association": "other"
                        },
                        "name": "Migraine Disorders",
                        "links": [
                            {
                                "dataSource": "UMLS",
                                "id": "C0149931"
                            },
                            {
                                "dataSource": "AOD",
                                "id": "0000006120"
                            },
                            ...
                        ]
                    },
                    ...
                ],
                "warnings": []
            }
        ],
        "errors": [],
        "modelVersion": "2022-08-15-preview"
    }
```


Anche in questa circostanza, il servizio offre un'analisi accurata evidenziando tutti i termini pertinenti al campo medico, oltre a fornire un ricco assortimento di risorse dati.

CAPITOLO 5

Confronto tra i tre provider

L'obiettivo principale di questo capitolo è condurre un confronto approfondito tra i tre provider che sono stati precedentemente analizzati in questo documento. In particolare, concentreremo la nostra attenzione sulle performance degli strumenti offerti dai servizi dedicati all'analisi del sentiment dei documenti e all'analisi dei dati medici all'interno dei testi medici. Tale confronto ci consentirà di identificare i punti di forza e di evidenziare i candidati più promettenti. Oltre alla valutazione delle performance, esamineremo la facilità d'uso offerta da ciascun provider. Questo aspetto sarà analizzato da varie prospettive, tra cui la presenza di documentazione, il design dell'interfaccia utente e il supporto a servizi esterni. In definitiva, quindi, l'obiettivo è quello di fornire una panoramica completa che aiuti a comprenderne le differenze e a individuare la soluzione più adatta alle specifiche esigenze.

5.1 Analisi della prestazione dei servizi

In questa sezione, condurremo un'approfondita analisi sulle performance dei servizi offerti da Amazon, Google e Azure precedentemente esaminati. In particolare alcuni di questi sono specializzati nell'analisi di documenti, mentre altri si concentrano sull'analisi di dati medici. A tale scopo, esamineremo con precisione e metteremo a confronto gli strumenti dedicati a ciascun tipo di analisi che compongono i servizi offerti dai fornitori. Questi confronti saranno basati su metriche chiave e scenari d'uso diversificati, allo scopo di offrire un quadro completo delle capacità di ciascun fornitore coinvolto.

5.1.1 Analisi delle entità

Tutti e tre i servizi dedicati all'analisi del sentimento sono provvisti di strumenti utili per l'analisi delle entità. Nei servizi forniti da AWS e Google è presente un singolo strumento per ciascuno. Nel caso del servizio offerto da Azure, invece, l'analisi delle entità è suddivisa in due strumenti distinti: l'analisi delle "linked entity", che stabilisce collegamenti tra le entità e i documenti esterni, e l'analisi delle "named entity", un tipo di analisi più comune e generalizzata.

Eseguendo un riferimento agli esempi di analisi precedentemente trattati all'interno di questo documento (Figure 2.5, 3.3, 4.4, 4.5), è possibile affermare quanto segue:

- *Amazon Comprehend:* È in grado di condurre un'analisi delle entità che, di norma, restituisce un minor numero di risultati ma che si contraddistinguono per una migliore precisione nel delineare i tipi delle singole entità.
- *API Natural Language di Google:* Conduce un'analisi delle entità che, in genere, produce la maggior quantità di risultati tra i servizi. Tuttavia presenta una precisione limitata nell'identificazione dei tipi di entità.
- *AI Language di Azure:* Questo servizio, riguardo all'identificazione delle entità, si trova nel mezzo. Esso fornisce un numero medio di risultati con buona precisione nell'analisi dei tipi di entità. Inoltre, grazie all'analisi delle "linked entity", è in grado di recuperare entità collegate a documenti e fonti esterne, prevalentemente pagine Wikipedia.

5.1.2 Analisi del sentimento

Anche in questa circostanza, tutti e tre i servizi dedicati all'analisi di documenti esaminati precedentemente sono provvisti di strumenti per l'analisi del sentimento. Nei servizi di Azure e Google vi è, per ciascuno, solo un singolo strumento per effettuare l'analisi del sentimento, la quale viene effettuata sull'intero documento, sulle singole frasi e, nel caso di Google, anche sulle singole entità. Invece, nel servizio offerto da AWS, se ne possono trovare due distinti: uno per l'analisi generale del sentimento e un altro per il Targeted Sentiment, ovvero l'analisi mirata, in cui si analizza il sentimento di ogni singola entità.

A questo punto, quindi, facendo riferimento agli esempi di analisi precedentemente trattati all'interno di questo documento (Figure 2.9, 2.10, 2.11, 3.4, 4.8, 4.9), è possibile affermare quanto segue:

- *Amazon Comprehend:* conduce un'analisi del sentimento che non sempre risulta essere del tutto corretta, essendo incline all'attribuzione di punteggi troppo elevati di sentimento neutro. Nonostante ciò, comunque, riesce a giustificare i suoi risultati grazie all'analisi delle singole entità tramite l'analisi del Targeted Sentiment.
- *API Natural Language di Google:* il servizio riesce, generalmente, ad effettuare in maniera opportuna, e migliore, l'analisi del sentimento, attribuendo generalmente dei punteggi che possono essere ritenuti corretti per le frasi, le entità, e l'intero documento. Inoltre, questo servizio adotta un sistema di punteggio differente rispetto agli altri, che si basano sulla "Confidence" per ogni tipo di sentimento. Questo punteggio varia in un intervallo da -1 a 1. Nel caso in cui il punteggio sia tra -1 e -0.25 indica un sentimento negativo; se è compreso tra -0.25 e 0.25 denota un sentimento neutro; se si colloca tra 0.25 e 1 rappresenta un sentimento positivo.
- *AI Language di Azure:* anche in questo caso il servizio risulta essere nella correttezza dei risultati nel mezzo. Infatti, sia per l'analisi delle frasi che per l'intero documento il servizio mostra una comprensione solida dei sentimenti principali, fornendo, anche, un approfondimento grazie all'individuazione degli "assessment" delle frasi.

5.1.3 Analisi dei dati PII

In questo caso non tutti i servizi dedicati all'analisi del sentimento che sono stati analizzati hanno a disposizione degli strumenti atti all'analisi dei dati PII (informazioni di

identificazione personale). L'unico che non ne è dotato, tra quelli visti precedentemente, è il servizio offerto da Google.

Perciò, facendo riferimento agli esempi di analisi precedentemente trattati all'interno di questo documento (Figure 2.8, 4.6), è possibile affermare quanto segue:

- *Amazon Comprehend:* nonostante l'analisi generalmente fornisca un numero minore di risultati questo servizio risulta essere il più preciso nella analisi dei dati PII riuscendo a distinguerli in maniera opportuna tra tutte le entità presenti all'interno del testo.

- *AI Language di Azure:* anche questa soluzione si dimostra generalmente accurata nell'analisi dei dati PII, presentando risultati che, tuttavia, talvolta non sono pienamente precisi. Inoltre, a differenza dell'altro servizio, è in grado di individuare il "PersonType" degli individui, come, ad esempio, la loro occupazione, e, mediante l'output in formato JSON, offre anche una versione del testo modificata in cui tutti i dati PII sono oscurati.

5.1.4 Analisi delle frasi chiave

Anche in questo caso il servizio dedicato all'analisi del sentimento offerto da Google non mette a disposizione strumenti per effettuare l'analisi delle frasi chiave, i quali, invece, sono presenti nei servizi offerti da AWS e Azure.

Pertanto, facendo riferimento agli esempi di analisi precedentemente discussi all'interno di questo documento (Figure 2.6, 4.7), si può affermare quanto segue:

- *Amazon Comprehend:* tramite il suo strumento di analisi, il servizio riesce a generare una lista di frasi chiave in modo soddisfacente, identificandole correttamente, anche se con alcune leggere imprecisioni. Nonostante queste limitazioni, le frasi chiave riescono, comunque, a fornire una visione adeguata del contenuto testuale.

- *AI Language di Azure:* analogamente al servizio precedente, anche questo fornisce una lista adeguata di frasi chiave, individuate in modo corretto, sebbene con alcune imprecisioni. Tuttavia, queste frasi riescono a garantire una visione appropriata del testo.

5.1.5 Analisi della lingua

Questa analisi può essere effettuata con tutti i servizi dedicati all'analisi del sentimento che sono stati analizzati precedentemente. Tuttavia, è opportuno sottolineare che per il servizio di Google non esiste uno strumento specifico. In effetti, per identificare la lingua principale di un documento mediante Google, è necessario consultare gli output in formato JSON derivanti dall'analisi delle entità o dall'analisi del sentimento per il documento. In questa circostanza, tuttavia, i punteggi che indicano la precisione dell'individuazione non saranno inclusi.

Quindi, nel caso specifico, considerando gli esempi di analisi precedentemente illustrati (Figure 2.7, 4.10), si può affermare che tutti e tre i servizi sono in grado di fornire strumenti che, in termini di risultati, presentano una sostanziale parità. Infatti, i punteggi di "Confidence" ottenuti dall'analisi con AWS e con Azure presentano lievi scostamenti che sono fondamentalmente trascurabili, considerando che il risultato finale è fondamentalmente sempre lo stesso.

5.1.6 Analisi della sintassi

Non tutti i servizi dedicati all'analisi del sentimento sono dotati di strumenti per effettuare l'analisi della sintassi. In particolare, tra quelli che sono stati analizzati precedentemente, Azure è l'unico che non ne è fornito. Al contrario, AWS e Google offrono questa funzionalità.

Osservando gli esempi di analisi precedentemente discussi all'interno di questo documento (Figure 2.12, 3.5), possiamo giungere, quindi, alle seguenti conclusioni:

- *Amazon Comprehend:* il servizio è in grado di fornire un'analisi sintattica del testo, identificando in modo semplice a quale parte del discorso appartiene ogni singola parola.
- *API Natural Language di Google:* l'analisi sintattica offerta da questo servizio può essere considerata la migliore. Oltre a garantire un'analisi sempre corretta e puntuale, vengono fornite numerose informazioni, come la morfologia delle parole, le dipendenze tra le parole, il lemma e la parte del discorso.

5.1.7 Analisi delle categorie

L'unico servizio dedicato all'analisi del sentimento dotato di uno strumento per analizzare le categorie di un testo è quello offerto da Google. Quindi, riflettendo sugli esempi di analisi condotti in precedenza (Figura 3.6), possiamo affermare che il servizio, selezionando possibili categorie da una vasta lista, cerca costantemente di identificare quelle che meglio rappresentano la realtà, sebbene, a volte, possa incorrere in imprecisioni. Inoltre, esso assegna anche un indicatore di "Confidence" appropriato a evidenziare il grado di sicurezza associato a ciascuna assegnazione. Questo approccio conferisce ulteriore trasparenza e consapevolezza al processo di categorizzazione, consentendo agli utenti di valutare attentamente i risultati ottenuti.

5.1.8 Analisi dei dati medici

Tutti i provider che sono stati analizzati precedentemente hanno a disposizione servizi per effettuare l'analisi di dati medici; tuttavia, ognuno di essi ha delle caratteristiche uniche che lo contraddistinguono.

Quindi, prendendo come riferimento gli esempi effettuati in precedenza all'interno del documento (Figure 2.17-2.28, 3.9-3.11, 4.11) possiamo stabilire che:

- *Amazon Comprehend Medical:* questo servizio si distingue come uno dei più completi nell'analisi di dati medici. Esso offre una gamma diversificata di analisi, tra cui:
  - *Analisi delle entità mediche generali*, che mira a individuare in modo generico tutti i dati pertinenti per una corretta valutazione medica.
  - *Analisi delle entità in relazione ai concetti RxNorm*, basati sul database della US National Library of Medicine, che riguardano principalmente tipologie di farmaci.
  - *Analisi delle entità in accordo con i concetti ICD-10-CM*, collegando le entità ai codici della versione 2019 della International Classification of Diseases, 10th Revision, Clinical Modification, che riguardano principalmente condizioni mediche.

  - *Analisi delle entità rispetto ai concetti SNOMED CT*, associandoli ai concetti del Systematized Nomenclature of Medicine, Clinical Terms, che riguardano concetti medici generali.

  In ciascuno di questi contesti, il servizio dimostra di essere in grado di identificare e classificare con successo le entità rilevanti e di stabilire collegamenti significativi tra di loro, al fine di migliorare la comprensione del testo medico.

- *API Healthcare Natural Language di Google:* l'analisi dei dati medici che viene effettuata da questo servizio risulta sempre essere efficace. Ogni singola entità medica individuata viene accuratamente identificata e dotata di una classificazione appropriata con valutazioni pertinenti, comprensive anche dei relativi codici medici di riferimento. Inoltre, il servizio offre una schermata dedicata alle relazioni tra le entità individuate, facilitando, così, una visione approfondita delle connessioni tra gli elementi.

- *AI Language di Azure:* quest'ultimo servizio, come gli altri, riesce ad individuare in maniera opportuna tutte le entità mediche rilevanti presenti all'interno di un testo medico. Per ognuna di queste entità il servizio riesce a classificarle, stabilendo la loro tipologia, e a collegarle tra loro, affiancando ad esse, anche in questo caso, tutti i codici medici utili al loro approfondimento.

5.2 Facilità di utilizzo dei servizi

Un elemento critico da considerare nell'impiego di un servizio è la sua usabilità, per questo motivo verrà esaminato con attenzione tale aspetto all'interno del presente paragrafo. In questa sezione, condurremo un'analisi dettagliata di questo aspetto per tutti i servizi che sono stati precedentemente esaminati. Tale analisi si focalizzerà sulla valutazione della disponibilità di documentazione e risorse che vengono fornite per agevolare l'uso dei servizi. Inoltre, verranno esaminate l'interfaccia utente e l'integrazione dei servizi con strumenti esterni, poiché queste componenti giocano un ruolo determinante nell'orientare la valutazione complessiva.

5.2.1 Documentazione e risorse

Tutti i provider dei servizi offrono una documentazione per semplificare il loro utilizzo; tuttavia, la qualità, chiarezza e disponibilità per ognuno di essi varia abbastanza; per questo motivo, possiamo trarre le seguenti conclusioni:

- *Amazon:* questo fornitore, sia per i servizi dedicati all'analisi di documenti che per quelli relativi ai dati medici offre una quantità appropriata di documentazione. Questi documenti sono presentati in modo estremamente efficace, contribuendo a rendere l'esperienza dell'utente più intuitiva e semplificando notevolmente l'utilizzo degli strumenti di analisi. L'intera documentazione è facilmente raggiungibile tramite una pagina web ben strutturata e priva di confusione, la quale è direttamente accessibile dalla home page di ciascun servizio.

- *Google:* anche questo fornitore offre un'opportuna documentazione per i suoi servizi. Nel caso dell'API Natural Language, la documentazione è sempre accessibile tramite una pagina web collegata direttamente dalla home page del servizio. Tuttavia, la

strutturazione di questa pagina non risulta sempre chiara a causa di una nomenclatura poco definita delle risorse. La situazione è analoga per l'API Healthcare Natural Language, con la differenza che questo servizio non possiede una propria home page, essendo integrato all'interno della documentazione del pacchetto di servizi di Google denominato API Cloud Healthcare.

- *Azure:* la documentazione per i servizi offerti da questo provider è altrettanto adeguata, con accesso diretto dalla home page del servizio Language Studio e dalle singole console di analisi. Tuttavia, pur essendo esaustiva, la documentazione è ospitata in una pagina web che occasionalmente può apparire sovraffollata di informazioni e con alcuni documenti che potrebbero mancare di una struttura perfettamente organizzata.

5.2.2 Interfaccia utente

L'interfaccia utente di un servizio rappresenta uno degli elementi cruciali per rendere l'utilizzo dello stesso più agevole e intuitivo. Per tale ragione, procederemo ora a condurre un'analisi approfondita di questo aspetto in relazione ai provider precedentemente esaminati.

- *Amazon:* il provider offre un'interfaccia di buona qualità per la sue console, la quale risulta intuitiva e con un accesso agevole sia ai singoli risultati per ciascun tipo di analisi, che alle relative risposte fornite in formato JSON. Inoltre, il fornitore mette costantemente a disposizione una sezione dedicata della console, in cui i risultati sono evidenziati direttamente all'interno del testo. Ciò che è stato precedentemente menzionato vale sia per Amazon Comprehend (Figure 2.2, 2.3, 2.4) che per Amazon Comprehend Medical (Figure 2.13, 2.14, 2.15, 2.16). Tuttavia, quest'ultimo presenta un'interfaccia leggermente più articolata, in quanto, nel testo, con i risultati evidenziati, vengono anche messi alla luce i collegamenti presenti tra le singole entità.

- *Google:* per questo fornitore, si evidenziano notevoli differenze nella struttura dell'interfaccia utente tra i suoi servizi API Natural Language e API Healthcare Natural Language. Nel dettaglio, la console del primo servizio (Figure 3.1, 3.2), che è facilmente raggiungibile dalla home page, è presentata in formato demo e offre un'esperienza intuitiva. Questa consente l'accesso ai risultati dell'analisi in maniera agevole. Tuttavia, purtroppo, risulta un po' meno immediato accedere alla risposta in formato JSON, in quanto questa opzione è disponibile soltanto attraverso l'utilizzo diretto dell'API di Google.

  La console relativa all'API Healthcare Natural Language (Figure 3.7, 3.8), invece, si distingue per una struttura differente ma altrettanto intuitiva, e che permette l'accesso dei risultati in formato JSON. Tuttavia, essa è collocata all'interno della documentazione del servizio Google API Cloud Healthcare. Questo approccio può rendere più complesso comprendere dove esattamente sfruttarla, richiedendo un'analisi più attenta della documentazione per individuarla.

- *Azure:* Per questo provider, ciascuno degli strumenti di analisi testuale presenti in AI Language dispone di una console privata (Figure 4.2, 4.3) accessibile direttamente dalla piattaforma Language Studio (Figura 4.1). Questa console è uniformemente

progettata per ciascun tipo di analisi, con variazioni soltanto nella rappresentazione dei risultati, in base alla specifica tipologia di analisi eseguita. Inoltre, per ogni console, vi è facile accesso anche ai risultati in formato JSON, migliorando ulteriormente la flessibilità dell'esperienza utente.

Quindi, l'uniformità nella struttura del servizio garantisce un'interfaccia utente di alta qualità e di facile comprensione. In definitiva, grazie a questa coerenza, Language Studio si configura come una piattaforma ben organizzata e intuitiva da utilizzare.

5.2.3 Integrazione con strumenti esistenti

L'integrazione senza intoppi con strumenti esistenti e flussi di lavoro è assolutamente fondamentale per servizi software analoghi a quelli che abbiamo analizzato in precedenza. Ecco perché, in questa sezione, procederemo a una comparazione approfondita di questo aspetto rispetto ai provider precedentemente esaminati.

- *Amazon:* questo provider offre un'ampia gamma di integrazioni con servizi esterni, sia per Amazon Comprehend che per Amazon Comprehend Medical. Per quanto riguarda quest'ultimo, l'integrazione si estende alla console a riga di comando e alle librerie Python e Java. Per quanto riguarda Amazon Comprehend, oltre alle integrazioni menzionate in precedenza, vi è anche un'interazione con il servizio .NET. È evidente, quindi, come Amazon metta a disposizione strumenti robusti per semplificare in modo significativo il flusso di lavoro dei suoi utenti.

- *Google:* il servizio API Natural Language di Google si distingue per l'elevato numero di integrazioni con servizi esterni. Questo servizio è stato, infatti, integrato con un'ampia varietà di librerie, tra cui Go, Node.js, Java, Python e C++. Un aspetto notevole è che il servizio può essere utilizzato direttamente anche all'interno della Google Cloud Command Line, rendendo l'esperienza ancora più fluida e intuitiva. Purtroppo, per quanto riguarda il servizio API Healthcare Natural Language, l'integrazione è più limitata. In questo caso, è possibile utilizzarlo solo attraverso strumenti a linea di comando, come Curl o Powershell, che, pur avendo meno opzioni, comunque consentono di sfruttarne le funzionalità.

- *Azure:* anche questo provider si distingue per la sua vasta gamma di integrazioni con servizi esterni per le sue soluzioni. Ad esempio, gli strumenti di analisi di AI Language possono essere impiegati in combinazione con un ampio spettro di librerie, tra cui C#, Python, Java e Javascript. Inoltre, Azure offre la flessibilità aggiuntiva di utilizzare chiamate alle API REST per accedere alle funzionalità di analisi. Questo approccio versatile consente agli utenti di adattare l'utilizzo alle proprie esigenze, arricchendo l'esperienza complessiva.

CAPITOLO 6

Etica e Sentiment Analysis

All'interno di questo capitolo verranno presentate delle considerazioni prettamente etiche riguardo allo sviluppo di sistemi dedicati alla Sentiment Analysis e di Intelligenza Artificiale. Prima verranno analizzate le implicazioni riguardanti la privacy e la sicurezza che questi sistemi generano, mettendo alla luce, quindi, le potenziali problematiche legate alla protezione dei dati e alla riservatezza delle persone coinvolte. Successivamente verrà effettuata un'analisi di tipo socio-culturale, in cui si metterà in evidenza l'importanza di una raccolta di dati per questi sistemi che sia il più possibile equa e corretta verso le persone. Infine, si analizzerà l'importanza che risiede nella trasparenza e nelle regolamentazioni per i sistemi di Sentiment Analysis e di IA, analizzando anche delle normative esistenti e in fase di sviluppo. Tutto ciò al fine di garantire una visione completa delle sfide e delle opportunità legate a questi avanzati sistemi tecnologici.

6.1 Privacy dei dati e sicurezza

Un aspetto di grande preoccupazione legato all'impiego dell'analisi del sentiment è rappresentato dalla delicata questione della privacy e della sicurezza dei dati. Il rischio principale sorge dall'utilizzo di sistemi di Sentiment Detection, i quali raccolgono costantemente informazioni senza necessariamente ottenere un consenso esplicito da parte delle persone analizzate. Questo scenario potrebbe facilmente trasformarsi in una violazione della sfera privata, poiché molte persone potrebbero preferire che le loro emozioni non vengano dedotte e analizzate.

In aggiunta, i sistemi automatizzati di analisi del sentiment potrebbero essere adottati come strumenti di sorveglianza su vasta scala, sia da aziende che da enti governativi, senza il dovuto consenso da parte dei soggetti sottoposti all'analisi. Infatti, spesso, le persone non sono pienamente consapevoli del fatto che i dati che condividono pubblicamente online possano essere utilizzati in modi che contrastano con i loro interessi personali; tale incompletezza le può portare a compiere scelte incaute riguardo alla propria privacy.

Questo problema potrebbe avere conseguenze significative, specialmente in contesti in cui un'autorità governativa sia autoritaria. In tali situazioni, l'uso di sistemi di analisi del sentiment potrebbe accentuare drasticamente le restrizioni sulle libertà di espressione e di protesta. È, quindi, evidente che l'analisi del sentiment, se non regolamentata adeguatamente, possa compromettere seriamente la privacy individuale e le libertà civili.

In questi scenari, gli strumenti di analisi del sentiment potrebbero fungere da mezzi non solo di sorveglianza, ma anche di tutela della sicurezza pubblica, consentendo l'individuazione di individui che potrebbero essere etichettati come "sospetti" o che giustifichino una maggiore attenzione investigativa. A questo punto, le informazioni raccolte da tali sistemi potrebbero persino essere impiegate in situazioni di interrogatorio, in cui l'assegnazione di emozioni possibili per trarre deduzioni potrebbe incrinare i diritti umani.

Questa pratica entra in diretta contraddizione con il diritto a non autoincriminarsi sancito nel contesto del diritto internazionale; esso sancisce il diritto di ogni individuo a "non essere costretto a testimoniare contro se stesso o a confessare la propria colpevolezza". L'utilizzo delle informazioni ottenute da tali analisi per trarre deduzioni potrebbe, quindi, arrecare un impatto negativo su questi diritti fondamentali.

In aggiunta, questi strumenti, possono anche essere sfruttati per manipolare il comportamento delle persone in base alle loro emozioni e sentimenti. Ad esempio, è noto che quando una persona è triste allora tende maggiormente a fare acquisti. Quindi, individuare i momenti in cui si è più suscettibili alla suggestione al fine di inculcare idee su cosa comprare, chi votare o chi disapprovare, attraverso mezzi come fake news o pubblicità mirate, può avere implicazioni pericolose. D'altra parte, però, l'identificazione di come soddisfare le esigenze individuali per migliorare, ad esempio, la loro conformità alle misure di salute pubblica in una pandemia globale, o per aiutare le persone a smettere di fumare, può essere vista in una luce più positiva.

Nell'utilizzo dei sistemi di Analisi del Sentimento, è importante considerare, inoltre, non solo la privacy dei singoli individui, ma anche quella di gruppi interi. Il concetto di privacy collettiva diventa particolarmente rilevante nell'ambito delle "soft-biometrics", ovvero dei tratti e delle preferenze dedotti dall'Analisi del Sentimento, che non mirano all'identificazione individuale, bensì all'individuazione di gruppi di individui con caratteristiche simili. Questo concetto potrebbe essere sfruttato banalmente per scopi commerciali mirati a specifici gruppi o, nel caso di governi autoritari, per discriminare certe fasce sociali.

Di conseguenza, è imperativo che tutti i sistemi di Sentiment Analysis siano opportunamente regolamentati e implementati al fine di poter bilanciare al meglio il loro potenziale positivo, ovvero la comprensione delle opinioni e le emozioni delle persone, e la loro possibile minaccia per la privacy e i diritti umani. Tutto ciò al fine di garantire il massimo rispetto dei valori fondamentali della società.

In aggiunta, è di vitale importanza promuovere la consapevolezza pubblica e l'educazione riguardo alle implicazioni etiche legate all'Intelligenza Artificiale e all'analisi dei sentimenti. Con l'IA sempre più integrata nella nostra vita quotidiana, è, infatti, imperativo che le persone comprendano appieno i rischi e i benefici connessi a questa tecnologia.

6.2 Implicazioni socio-culturali

Quindi la Sentiment Analysis, pur offrendo notevoli vantaggi nell'analisi delle emozioni e delle opinioni attraverso i testi, solleva importanti questioni di natura socio-culturale. Uno dei punti cruciali riguarda la mancanza di dati disaggregati utilizzati per effettuare l'analisi. Spesso la società ha trattato gruppi diversi in maniera ineguale, basandosi su caratteristiche come razza, genere, reddito e lingua, creando strutture sociali e di potere disuguali. Tuttavia, persino quando i pregiudizi non sono consapevoli, si tende spesso

a trascurare le esigenze specifiche dei singoli gruppi. Ad esempio, la mancanza di dati disaggregati per le donne ha portato a esiti negativi in diversi aspetti della loro vita, come la salute, il reddito, la sicurezza e il successo. Questa carenza si riflette in modo ancora più marcato, ad esempio, per le persone transgender.

Pertanto, coloro che sono coinvolti nello sviluppo di strumenti per l'analisi del sentiment dovrebbero considerare attentamente il valore della disaggregazione al fine di creare strumenti il più possibile equi. Per raggiungere questo obiettivo, è essenziale, quindi, seguire diverse linee guida:

- Durante la fase di creazione di dataset per l'analisi, diventa fondamentale raccogliere annotazioni da gruppi diversificati.
- Quando si testano ipotesi o si traggono conclusioni sull'utilizzo del linguaggio, è opportuno valutare tali ipotesi in modo dettagliato per ciascun gruppo demografico rilevante, per verificare che esse siano corrette.
- Nello sviluppo di sistemi automatici di previsione, è cruciale che la valutazione delle prestazioni sia anch'essa disaggregata per ciascun gruppo demografico coinvolto.

Un'altra questione è l'invisibilità intersezionale nella ricerca. Con "intersectionality" si intende il modo complesso in cui gli aspetti di gruppi diversi, come razza, status socioeconomico, neurodiversità e genere si sovrappongono, amplificando la discriminazione. Infatti, gli individui che si identificano con più identità di determinati gruppi spesso non vengono visti come membri prototipici di nessuno e, quindi, sono soggetti alla cosiddetta "invisibilità intersezionale". Questo significa che le sfumature e le complessità delle loro identità non sono adeguatamente prese in considerazione.

L'effetto di questa invisibilità può essere notato nella ricerca e nella società in generale. Le esperienze uniche di coloro che sono colpiti dall'invisibilità intersezionale possono non essere adeguatamente comprese o rappresentate nei dati e nelle analisi. Questo può portare a un'ignoranza delle sfide e delle disuguaglianze che tali individui affrontano quotidianamente. Pertanto, nello sviluppo di strumenti per l'analisi del sentiment, è anche necessario considerare le sfumature che caratterizzano la "intersectionality" al fine di poter creare dei sistemi che siano il più possibile inclusivi.

Inoltre, nello sviluppo di questi sistemi, bisogna anche considerare i fenomeni di reificazione ed essenzializzazione culturale. Infatti, alcune variabili demografiche sono essenzialmente, o in gran parte, costrutti sociali. Pertanto, il lavoro sulla disaggregazione dei dati può talvolta rafforzare le false convinzioni che esistano differenze innate tra i diversi gruppi o che alcune caratteristiche siano fondamentali per l'appartenenza a una categoria sociale. È, quindi, indispensabile contestualizzare il lavoro sulla disaggregazione durante lo sviluppo di dataset per l'analisi. Ad esempio, è importante comprendere che, anche se la razza è un costrutto sociale, l'impatto delle percezioni e dei comportamenti delle persone sulla razza porta a conseguenze molto reali.

Particolare attenzione deve essere dedicata anche verso persone caratterizzate da neurodiversità, alessitimia (analfabetismo emotivo) e autismo, spesso ignorate. Queste persone, infatti, sono spesso caratterizzate da una difficoltà nell'espressione e percezione delle emozioni e dei sentimenti. Pertanto, nello sviluppo di sistemi di Sentiment Analysis, bisogna avere una particolare cura verso di loro. Attualmente, i sistemi che vengono sviluppati sono rivolti prevalentemente verso persone neurotipiche. È cruciale, quindi, prestare maggiore attenzione alla ricerca verso le persone neurodiverse. Un passo

importante potrebbe consistere, ad esempio, nell'annotare accuratamente i dati raccolti, specificando se i partecipanti sono neurotipici o neurodiversi. Questo consentirebbe di utilizzare tali dati per sviluppare sistemi che siano veramente inclusivi e adatti alle esigenze di tutte le persone.

La raccolta di dati disaggregati, quindi, sebbene fondamentale, deve essere affrontata con sensibilità e contestualizzazione. La complessità delle implicazioni socio-culturali di questa pratica sottolinea l'importanza di considerare il contesto più ampio e i reali impatti delle analisi del sentiment nel panorama sociale e umano. Questa comprensione va oltre la mera raccolta di dati demografici, poiché solleva questioni profonde riguardo al consenso informato e al rispetto della privacy, autonomia e dignità delle persone coinvolte.

In modo particolare, nella progettazione di questionari inclusivi per la raccolta di dati o nell'attribuzione delle persone a gruppi sociali si richiede un approccio particolarmente riflessivo. Ad esempio, anche quando vengono fornite caselle di testo per l'auto-rapporto, dove inserire dati autonomamente, l'analisi successiva spesso non tiene conto di essi o li combina in modi che sfuggono al controllo dei partecipanti. Quindi, l'adozione di metodologie per la raccolta di dati che garantiscano la veridicità e l'inclusività degli stessi è cruciale al fine di evitare l'abuso, la cancellazione e la perpetuazione di stereotipi.

Inoltre, bisogna fare attenzione anche all'utilizzo di sistemi di analisi automatica. Questi strumenti, ad esempio, possono essere impiegati per tentare di dedurre automaticamente statistiche di gruppo a livello aggregato, cercando di estrarre informazioni quali razza o genere da indizi come il linguaggio utilizzato, o associazioni storiche tra nomi e generi al fine di condurre analisi disaggregate. Tuttavia, questi approcci sono gravidi di problemi etici, come il misgendering, l'essenzializzazione e la reificazione. È, inoltre, importante riconoscere che in passato le persone sono state marginalizzate proprio a causa delle loro appartenenze sociali, il che solleva giustificate e serie preoccupazioni riguardo all'uso di metodi che potrebbero ulteriormente contribuire ad abusi, cancellazioni e alla perpetuazione di stereotipi dannosi.

Perciò, lo sviluppo di sistemi per la Sentiment Analysis, o anche generalmente di Intelligenza Artificiale, non richiede soltanto una conoscenza in ambito tecnologico, ma anche un profonda attenzione verso aspetti socio-culturali, come diversità, inclusività e equità. Tutto ciò al fine di garantire che tali tecnologie non solo siano all'avanguardia dal punto di vista tecnico, ma anche che rispettino i valori umani e contribuiscano a un mondo migliore.

6.3 Trasparenza e regolamentazioni

Nello sviluppo dei sistemi per la Sentiment Analysis, un aspetto cruciale è, senza dubbio, la trasparenza. Questo concetto implica la necessità di condividere una serie di elementi fondamentali nello sviluppo di questi sistemi, come gli strumenti utilizzati, le pratiche adottate per comprendere il modello, i dati su cui è stato addestrato e il processo di categorizzazione degli errori e delle distorsioni, insieme alla loro frequenza. Tutto ciò al fine di costruire fiducia verso gli utenti, e anche permettere a esperti esterni di valutare eventuali pregiudizi o inesattezze nei risultati del modello.

Pertanto, la trasparenza nello sviluppo di questi sistemi aiuta a garantire che tutte le parti interessate possano capire chiaramente il funzionamento del sistema stesso, compreso come prende decisioni e elabora i dati. Questo concetto è legato direttamente ad altri due: l'esplicabilità e l'interpretabilità.

L'esplicabilità si concentra nel fornire motivazioni comprensibili per le decisioni prese da un sistema di Intelligenza Artificiale per la Sentiment Analysis. Dall'altra parte, l'interpretabilità si riferisce alla prevedibilità degli output di un modello in base ai suoi input. Quindi, mentre l'esplicabilità e l'interpretabilità sono cruciali per raggiungere la trasparenza, da sole non la comprendono completamente.

Di conseguenza, sommariamente, la trasparenza nello sviluppo di questi tipi di sistemi è necessaria per varie necessità, come:

- costruire la fiducia dei clienti e dei dipendenti;
- garantire sistemi equi ed etici;
- rilevare e risolvere potenziali distorsioni dei dati;
- migliorare l'accuratezza e le prestazioni dei sistemi;
- garantire la conformità alle nuove normative sull'IA, come l'EU AI Act.

Quindi, la trasparenza riveste un ruolo fondamentale per riuscire a comprendere appieno quali siano i limiti di un sistema di Intelligenza Artificiale, anche progettato per la Sentiment Analysis. Infatti, dovendo utilizzare questi sistemi per prendere decisioni critiche, è imperativo comprendere i meccanismi di ragionamento del sistema. Ad esempio, un modello di Intelligenza Artificiale progettato per individuare il cancro, anche se ha un margine di errore solo dell'1%, potrebbe mettere a rischio la vita di una persona. In situazioni di questo genere, quindi, l'IA e gli esseri umani devono collaborare, e il compito diventa molto più facile quando il modello può spiegare come ha raggiunto una determinata decisione. La trasparenza dell'Intelligenza Artificiale, quindi, la rende un supporto attivo nel poter prendere decisioni critiche calcolate.

Nonostante la trasparenza abbia dei chiari benefici, non tutti gli algoritmi possono vantare tale caratteristica a causa di alcune loro debolezze fondamentali. In particolare:

- I modelli "trasparenti" sono più suscettibili a possibili attacchi da parte di hacker, visto che questi ultimi, avendo una maggiore conoscenza degli algoritmi, possono individuare vulnerabilità con maggiore facilità.
- Un'ulteriore preoccupazione collegata alla trasparenza di questi sistemi è la protezione degli algoritmi proprietari. Infatti, è stato dimostrato che interi algoritmi possono essere sottratti semplicemente analizzandone le spiegazioni.
- La progettazione degli algoritmi "trasparenti" si rivela più intricata, specialmente nel caso di modelli complessi con milioni di parametri.
- Un altro problema è che non tutti i metodi di trasparenza sono affidabili. Essi possono generare risultati diversi ogni volta che vengono eseguiti. Questa mancanza di affidabilità e ripetibilità può ridurre la fiducia nel sistema e ostacolare gli sforzi di trasparenza.

Dunque, per implementare la trasparenza nell'ambito dell'Intelligenza Artificiale, compreso il trovare un equilibrio tra obiettivi organizzativi contrastanti, occorre anche una collaborazione e un apprendimento continui tra dirigenti e dipendenti. Ciò richiede una chiara comprensione dei requisiti di un sistema da prospettive aziendali, utente

e tecniche. In tal senso è necessario preparare i dipendenti a contribuire attivamente nell'individuare risposte o comportamenti errati nei sistemi di AI.

Perciò, nello sviluppo di tali sistemi, oltre a concentrarsi su caratteristiche come utilità e novità, vi è anche il bisogno di concentrarsi nello sviluppo di sistemi che siano sicuri, affidabili e robusti, ponendo la trasparenza in cima alle priorità fin dalle fasi progettuali iniziali.

Per poter ottenere dei sistemi di Sentiment Analysis e di Intelligenza Artificiale che siano etici, trasparenti e che rispettino la privacy, vi è anche la necessità di sviluppare regolamentazioni adeguate. Queste dovrebbero cercare di bilanciare l'innovazione tecnologica con la protezione dei diritti umani e della privacy.

Alcune regolamentazioni già sono state sviluppate. Un esempio importante è il GDPR (General Data Protection Regulation), un regolamento dell'Unione Europea in materia di trattamento dei dati personali e di privacy. Esso è entrato in vigore il 25 maggio 2018, e, come si può immaginare, ha influenzato profondamente lo sviluppo di algoritmi per la raccolta dati, come quelli utilizzanti nei sistemi per la Sentiment Analysis.

Il GDPR ha introdotto importanti misure di tutela dei dati degli utenti. In particolare, ha reso obbligatorio per aziende ed organizzazioni, sia all'interno che all'esterno dell'UE, che raccolgono, elaborano o conservano dati personali di cittadini europei, ottenere il consenso esplicito degli utenti per l'utilizzo di tali informazioni, le quali, per il regolamento, devono anche essere trattenute con integrità e riservatezza. Inoltre è richiesto che queste entità mettano a disposizione del pubblico la loro informativa sulla privacy. Questo documento deve spiegare in maniera esplicita e semplice come verranno utilizzati i dati degli utenti, quali specificatamente verranno prelevati e fino a quando verranno trattenuti.

L'obiettivo principale di queste misure è garantire una maggiore trasparenza e controllo agli utenti sul modo in cui le loro informazioni personali vengono gestite da aziende e organizzazioni. Il GDPR rappresenta, quindi, un punto di riferimento significativo nell'ambito delle regolamentazioni che riguardano la protezione dei dati personali e la privacy nell'era digitale.

Un'altra normativa molto importante ed innovativa è l'EU AI act, proposta nell'Aprile 2021, che costituisce la prima regolamentazione dedicata all'Intelligenza Artificiale. Questa iniziativa mira a regolamentare l'Intelligenza Artificiale al fine di garantire condizioni migliori per lo sviluppo e l'utilizzo di questa tecnologia innovativa.

La priorità per il Parlamento Europeo nello sviluppo di tale regolamentazione è quella di assicurarsi che i sistemi di Intelligenza Artificiale utilizzati nell'UE siano sicuri, trasparenti, tracciabili, non discriminatori e rispettosi dell'ambiente. Questo aspetto è di particolare rilevanza anche nel contesto della Sentiment Analysis e di tecnologie affini, che potrebbero presentare implicazioni ambientali significative. L'elaborazione di algoritmi complessi richiede, infatti, considerevoli risorse energetiche, con il potenziale di impattare negativamente sull'ambiente. Ad esempio, ricerche recenti hanno dimostrato che l'allenamento di Gpt-3 di OpenAi ha portato al consumo di ben 700.000 litri di acqua e che una singola conversazione con questo chatbot equivale all'incirca al consumo di una bottiglia di acqua. Di conseguenza, vi è un grande impulso verso lo sviluppo di "Green AI", con un minor consumo energetico, e che vengono sviluppate con la consapevolezza della necessità di bilanciare l'innovazione tecnologica con la salvaguardia dell'ambiente.

Comunque, in linea con l'EU AI Act, si sottolinea anche l'importanza di supervisionare i sistemi di Intelligenza Artificiale attraverso il coinvolgimento umano, piuttosto che affidarsi completamente all'automazione, al fine di evitare conseguenze indesiderate o

dannose. Il documento, inoltre, si impegna anche a suddividere dei regolamenti in base ai livelli di rischio dell'IA. In particolare:

- I sistemi con rischio inaccettabile sono vietati e comprendono manipolazione cognitiva, classificazione sociale e identificazione biometrica in tempo reale. Alcune eccezioni potrebbero essere considerate, come l'identificazione biometrica post-riconoscimento per reati gravi, previa autorizzazione giudiziaria.
- I sistemi ad alto rischio sono sistemi di Intelligenza Artificiale che possono avere un impatto negativo sulla sicurezza o sui diritti fondamentali. Essi sono suddivisi in due categorie:
  - Sistemi utilizzati in prodotti come giocattoli, aviazione, automobili, dispositivi medici, etc.
  - Sistemi in otto aree specifiche che devono essere registrati in database dell'UE, come identificazione biometrica, gestione delle infrastrutture critiche, istruzione, forze dell'ordine, migrazione, etc.

  Questi sistemi saranno valutati prima di essere messi sul mercato e durante il loro utilizzo.
- I sistemi a rischio limitato devono soddisfare requisiti minimi di trasparenza per consentire decisioni informate agli utenti, specialmente nei casi di generazione o manipolazione di contenuti multimediali.
- L'IA generativa, invece, come ChatGPT, deve rispettare requisiti di trasparenza, rivelando che i suoi contenuti sono stati generati da un'Intelligenza Artificiale, prevenendo contenuti illegali e pubblicando riepiloghi dei dati con diritti d'autore utilizzati per l'addestramento.

Quindi, nel complesso, l'impatto di questa regolamentazione sarà destinato a influenzare profondamente l'ambito dello sviluppo di sistemi di Intelligenza Artificiale, compresi quelli utilizzati per l'Analisi dei Sentimenti. L'EU AI Act, insieme al GDPR, rappresentano, perciò, dei pilastri portanti di un futuro in cui l'Intelligenza Artificiale non solo sarà caratterizzata dalla trasparenza, ma sarà anche guidata da principi etici e rigorose norme di sicurezza.

Conclusioni

Il lavoro svolto all'interno di questa tesi ha esaminato in maniera dettagliata tre dei servizi più significativi che forniscono strumenti per condurre la Sentiment Analysis, un ambito dell'elaborazione del linguaggio naturale che ha l'obiettivo di individuare i sentimenti espressi dal linguaggio e che, nel corso degli ultimi anni, ha acquisito crescente importanza in molti settori.

La sezione introduttiva è stata cruciale nel delineare e approfondire il campo della Sentiment Analysis, offrendo una panoramica generale della sua natura e fornendone una visione storica. In seguito, sono stati descritti i principali tipi di Sentiment Analysis, che si distinguono per complessità e profondità. Infine, sono state esplorate le diverse aree di applicazione di questa tecnologia, cioè Healthcare, Bussiness Intelligence, Reccomender Systems e Governement Intellingence, evidenziando le molteplici opportunità che la Sentiment Analysis può fornire in questi campi. L'obiettivo, quindi, è stato proporre una visione completa di questa tecnologia, mettendo in luce la sua versatilità e l'ampia portata delle sue applicazioni.

Successivamente, nelle tre sezioni seguenti, sono stati analizzati i singoli servizi che offrono strumenti di Sentiment Analysis in base al provider, sia per il settore aziendale che per l'ambito sanitario. Il primo è stato AWS, poi Google e infine Azure. Per ognuno di questi, tranne Azure, si è prima considerato il servizio predisposto al buisness, analizzando gli insight forniti, il suo funzionamento e presentando esempi pratici con risultati. Successivamente, è stato fatto lo stesso per il servizio di analisi dei testi medici. Per Azure, questa distinzione appare soltanto negli esempi, poiché esso offre un servizio unico che copre sia l'analisi per il bussiness che per l'healthcare.

Dopo aver fatto ciò, sono stati messi a confronto i tre provider in base alla loro performance e alla facilità di utilizzo. Per il primo punto, semplicemente, si è confrontata ogni tipo di analisi offerta dagli strumenti dei provider, definendone i pregi e i difetti. Per quanto riguarda la facilità d'uso, invece, si sono confrontati vari aspetti, come la documentazione e le risorse fornite dai provider, le interfacce utente e l'integrazione con strumenti esistenti.

Infine, si è analizzata la Sentiment Analysis dal punto di vista prettamente etico, esaminando i problemi che può generare nell'ambito della privacy e della sicurezza, le implicazioni socio culturali che produce e l'importanza della trasparenza e regolamentazioni nel suo ambito, mettendo, quindi, in evidenza l'importanza di un utilizzo etico e cosciente di questa tecnologia.

La Sentiment Analysis si prospetta come una tecnologia dal futuro incredibilmente promettente. Infatti essa, con l'avanzamento delle tecniche di deep learning, dell'Intelligenza Artificiale e l'aumento della quantità e qualità di dati disponibili, è destinata a migliorare la sua precisione nella rilevazione dei sentimenti. Ciò può soltanto giovare al suo possibile utilizzo all'interno di numerosi settori aumentandone la versatilità e utilità.

Ovviamente, nel processo di sviluppo, come è stato detto, sarà anche imperativo considerare attentamente le esigenze dei singoli individui al fine di preservare in modo attivo i loro diritti fondamentali.

Quindi, in maniera definitiva, possiamo dire che la Sentiment Analysis si trova soltanto all'alba del suo processo di sviluppo e che, per ancora lungo tempo, potrà offrire delle ampie opportunità di utilizzo.

Bibliografia

AA. VV. (2018), «Elezioni 2018: come voteranno gli italiani secondo i social network», *Wired.it*.

AA. VV. (2020), «Emotional Entanglement: China's emotion recognition market and its implications for human rights», *aritlce19.org*.

Adamec, C. e Viden, I. (1947-1948), «Polls Come to Czechoslovakia», *The Public Opinion Quarterly*.

Bo Pang, L. L. e Vaithyanathan, S. (2002), «Thumbs up? Sentiment Classification using Machine Learning Techniques», Rap. tecn., Department of Computer Science Cornell University and IBM Almaden Research Center.

Crescenzi, C. (2023), «L'uso di ChatGpt equivale al consumo di migliaia di litri d'acqua», *Wired.it*.

Dey, L. e Haque, S. K. M. (2008), «Opinion mining from noisy text data», *Publication History*.

Fegiz, P. L. (1947), «Italian Public Opinion», *The Public Opinion Quarterly*.

Gupta, S. (2018), «Applications of Sentiment Analysis in Business», *Towards Data Science*.

Knutson, A. L. (1945), «Japanese Opinion Surveys: The Special Need and the Special Difficulties», *The Public Opinion Quarterly*.

Lawton, G. (2023), «AI transparency: What is it and why do we need it?», *TechTarget*.

Mika Mäntylä, D. G. e Kuutila, M. (2016), «The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers», Rap. tecn., University of Oulu and Institute of Software Technology, University of Stuttgart.

Mohammad, S. M. (2021a), «Ethics Sheet for Automatic Emotion Recognition and Sentiment Analysis», Rap. tecn., National Research Council Canada.

Mohammad, S. M. (2021b), «Ethics Sheets for AI Tasks», Rap. tecn., National Research Council Canada.

Parlamento Europeo (2023), «Normativa sull'IA: la prima regolamentazione sull'intelligenza artificiale», *europarl.europa.eu*.

Samarati, M. (2018), «GDPR: I 6 principi della protezione dei dati», *it governance*.

Subramaniam, A. (2022), «Why is AI-Driven Patient Sentiment Analysis important for Healthcare?», *KANINI*.

Turney, P. D. (2002), «Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews», Rap. tecn., Institute for Information Technology National Research Council of Canada.

Yilmaz, B. (2022), «5 Use Cases/Applications of Medical Sentiment Analysis in 2023», *AI Multiple*.

Siti Web consultati

- ACM digital library – `dl.acm.org`
- AI Multiple – `research.aimultiple.com`
- Arxiv – `arxiv.org`
- ARTICLE 19 – `article19.org`
- AWS – `aws.amazon.com`
- Azure – `azure.microsoft.com`
- Google Cloud – `cloud.google.com`
- Google Scholar – `scholar.google.com`
- Inside Marketing – `insidemarketing.it`
- it governance – `itgovernance.eu`
- Jstor – `jstor.org`
- KANINI – `kanini.com`
- Medium – `medium.com`
- Parlamento Europeo – `europarl.europa.eu`
- ResearchGate – `researchgate.net`
- TechTarget – `techtarget.com`
- Wikipedia – `wikipedia.org`
- Wired – `wired.it`

Ringraziamenti

Voglio dedicare quest’ultima sezione a tutti coloro che hanno reso possibile il completamento di questa tesi. Per primi voglio ringraziare i miei genitori senza i quali non avrei potuto affrontare gli studi universitari. Li ringrazio per il costante incoraggiamento e la loro fiducia nei miei confronti senza cui, sicuramente, non sarei riuscito a completare questo percorso. Mi scuso, anche, per le poche chiamate che vi ho fatto mentre ero ad Ancona: di solito il cellulare è troppo distante e non ho voglia di alzarmi dalla sedia. Vi voglio un gran bene.

Ringrazio i miei zii e cugini che hanno sempre mostrato un grande sostegno nei miei confronti. Grazie per le belle giornate passate al mare o nel Nord Italia e anche per le cene in cui, qualche volta, diciamo che si sono suscitati vivaci dibattiti di idee. La vostra presenza e il vostro affetto mi hanno dato la forza di andare avanti con lo studio.

Un grande grazie lo faccio al mio gemello Nicola, letteralmente il mio "fratm carnal". Ti ringrazio per esserci sempre, per avermi aiutato costantemente con le mie fisime mentali, per essere sempre stato un supporto continuo e spero che l’anno prossimo riusciremo ad andare a vedere un GP di F1 insieme.

Ringrazio mia nonna Esterina e mio nonno Valder che, anche se non ci sono più, so che mi hanno sempre sostenuto e hanno sempre sperato che riuscissi ad avere un futuro migliore del loro. Mi mancate.

Ringrazio tutti i miei amici di università con cui ho condiviso questi ultimi tre anni. Ringrazio Laura, Alessandra, Luca, Valeria, Edoardo, Sara e Alessio con cui ho passato le mie giornate a studiare e a distrarci durante le lezioni. Un ringraziamento speciale va a Giansimone, il mio primo amico dell’università, conosciuto quasi casualmente su Whatsapp durante la pandemia, col quale, in questi ultimi tre anni, ho lavorato su mille progetti e condiviso una quantità incredibile di pazienza nello studio, ma anche tanto divertimento. Bro, senza il tuo aiuto non ce l’avrei fatta.

Ringrazio i miei amici di sempre: Giulia, Chiara, Giorgia, Maria, Angelica, Caterina, Lina, Elio e i due Gianluca, che mi hanno costantemente supportato e con cui ho sempre passato delle belle serate teramane. Un ringraziamento speciale va a Matteo che, grazie alle nostre serate su Discord, insieme anche a mio fratello, mi ha sempre aiutato a superare lo stress universitario con ironia e ottimismo. Vediamo di continuare a farle anche per la magistrale.

Vorrei, infine, ringraziare il mio relatore, il Professor Domenico Ursino, che, con grande disponibilità, anche nel mese di agosto, mi ha dato l’occasione di svolgere una tesi e un tirocinio su argomenti che mi hanno appassionato profondamente. La sua guida e i

suoi consigli sono stati fondamentali per il completamento del mio percorso accademico; quindi lo ringrazio davvero.