UNIVERSITÀ POLITECNICA DELLE MARCHE

FACOLTÀ DI INGEGNERIA

Corso di Laurea triennale in Ingegneria Biomedica

Analisi di cartelle cliniche elettroniche con tecniche di natural language processing nel campo degli studi sul diabete

Analysis of electronic health records with natural language processing techniques in the field of diabetes studies

Relatore: Dott. Micaela Morettini

Rapporto finale di:
Alessandro Bartelli

Correlatore: Prof. Laura Burattini

Dott. Andrea Tura

Anno Accademico 2021/2022

ABSTRACT

Il natural language processing (NLP) è un insieme di tecniche computazionali informatiche che permettono al computer di analizzare, comprendere e sintetizzare il linguaggio umano scritto o parlato dagli umani.

La ricerca sul natural language processing in relazione agli studi sul diabete sta crescendo rapidamente. Nonostante risultino ancora problemi legati alle ambiguità del linguaggio umano si stanno facendo rapidi progressi in ogni direzione nel campo, sfruttando come base di dati da cui trarre informazioni le cartelle cliniche elettroniche.

Lo scopo di questa tesi è descrivere le basi del natural language processing, la sua storia, il suo funzionamento, per poi comprendere come può essere sfruttato in ambito medico e le sue potenzialità.

Il metodo di ricerca è stato quello di ispezionare articoli in diversi database della letteratura scientifica al fine di avere un'idea globale degli studi effettuati. I dati sono stati ricavati principalmente da PubMed, Scopus e ACM Digital Library, inserendo determinate parole chiave come "natural language processing", "diabetes", "electronic healths record", e simili in modo da avere una visione più ampia ma sempre specifica riguardo l'argomento esaminato.

I risultati hanno portato ad una quindicina di articoli inerenti al target della ricerca e divisi in tre macro ambiti: prediabete, prima fase della malattia; diagnosi di diabete, malattia vera e propria; ipoglicemia, possibile conseguenza per i malati. Gli studi si sono dimostrati coerenti tra loro nei risultati, offrendo un resoconto di tutte le potenzialità del natural language processing e di come può essere usato per poter diventare uno strumento di fondamentale importanza in ambito medico.

Tutti i risultati dimostrano la sua incredibile efficacia se affiancato a strumenti preesistenti e ad un adeguato utilizzo da parte degli operatori sanitari. Il suo potenziale sembra essere grandissimo e largamente sviluppabile, aprendo ad un grande numero di applicazioni, che in futuro potrebbero non limitarsi alla diagnosi ma potrebbero concentrarsi più specificatamente sul trattamento, sulle cure e sulla supervisione dei pazienti, andando a migliorarne lo stile di vita.

INDICE

INTRODUZIONE **III**

Capitolo 1: I**L DIABETE MELLITO**

1.1: Cenni generali 1

1.2: Criteri diagnostici, classificazione e sintomatologia 1

1.3: Eziopatogenesi 3

1.4: Complicanze 4

Capitolo 2: **IL NATURAL LANGUAGE PROCESSING**

2.1: Definizione 7

2.2: Storia 8

2.3: Bert

2.3.1: *Architettura* 10

2.3.2: *Funzionamento* 11

2.4: Natural Language Processing in sanità 15

Capitolo 3: **APPLICAZONI DEL NATURAL LANGUAGE PROCESSING IN RELAZIONE AL DIABETE**

3.1: Natural Language Processing e diabete 19

3.2: Studio sul prediabete 19

3.3: Identificazione automatizzata dei casi di diabete 23

3.4: Rilevamento dell'ipoglicemia dalle EHR nei pazienti con diabete di tipo 2 33

CONCLUSIONI **V**

BIBLIOGRAFIA **VI**

RINGRAZIAMENTI **IX**

INTRODUZIONE

L'attuale ricerca in medicina si basa, in gran parte, su studi controllati che arruolano set di pazienti con patologia specifica i cui dati vengono registrati e sistematizzati secondo protocolli definiti "a priori". Negli ultimi anni è emersa sempre di più la consapevolezza che i dati "reali" forniscono un numero di informazioni maggiore rispetto ai dati degli studi a campione. Per dati "reali" si intendono, per esempio, quei dati ricavabili da cartelle cliniche elettroniche degli ospedali che hanno, però, la caratteristica di essere dei dati non strutturati.

L'estrazione di dati non strutturati per supportare la ricerca medica esiste da molti anni. Ciò richiede l'analisi di estesi set di dati, i cosiddetti "Big Data", e spesso include la revisione manuale di grafici (ad esempio tabelle con valori di laboratorio) per identificare i pazienti ed estrarre attributi specifici. La revisione dei grafici da parte di professionisti qualificati è estremamente costosa e richiede tempo ed è difficile ottenere una buona copertura delle variazioni di lingua e formato. Inoltre, un altro problema riguarda la raccolta di informazioni: storicamente il 50% degli studi clinici fallisce o è ritardato a causa di problemi di reclutamento dei pazienti, mentre alcuni studi faticano a trovare i pazienti necessari per iniziare il lavoro. Allo stesso modo, la ricerca scientifica e medica richiede un'ampia estrazione di letteratura per tenere traccia delle ultime pubblicazioni e risultati. Ogni mese viene pubblicata una grande quantità di informazioni in ogni area della medicina, rendendo sempre più difficile tenersi aggiornati e comprendere il panorama scientifico utilizzando i metodi convenzionali. Per rendere più facile la raccolta di un numero elevatissimo di informazioni in tutti i campi, anche in quello medico, è stata sviluppata la tecnologia detta NLP (Natural Language Processing) che può essere definita come l'applicazione di tecniche computazionali volte ad analizzare e sintetizzare il linguaggio naturale e parlato degli umani.

Il diabete mellito è una malattia metabolica cronica caratterizzata da alti valori di glicemia. In Italia, in base ai dati ISTAT, nel 2020 si stima una prevalenza del diabete pari al 5,9%, che corrisponde a oltre 3,5 milioni di persone. L'International Diabetes Federation (IDF), nel 2021, ha calcolato che, nel mondo, 536,6 milioni di persone tra 20 e 79 anni (il 9,2% degli adulti) siano diabetici e che un ulteriore 1,2 milioni di bambini e adolescenti (0-19 anni) abbia il diabete di tipo 1.

Appare evidente l'importanza di acquisire una grande quantità di dati su questa patologia cronica così diffusa per attuare strategie efficaci per combatterla. Scopo di questo lavoro è descrivere le basi concettuali e l'architettura su cui si fonda l'NLP, partendo da una sua breve storia e quali vantaggi può apportare questa metodologia di acquisizione dei dati per il miglioramento della diagnosi, della informazione e della cura nei pazienti con diabete mellito.

Saranno trattate applicazioni reali volte a risolvere diversi problemi, tra cui diagnosi e controllo, dimostrando i vantaggi di questa tecnologia confrontando i dati; saranno analizzati pro e contro e verranno presentati esempi di algoritmi con lo scopo di comprendere più a fondo il funzionamento di questi sistemi.

CAPITOLO 1
IL DIABETE MELLITO

1.1 Cenni generali

Il diabete mellito è una patologia metabolica cronica nella quale si ha un aumento di glucosio nel sangue, causata da un deficit totale o parziale di insulina e/o dalla sua inefficacia biologica. Questa condizione è nota come iperglicemia. L'insulina è un ormone prodotto dalle cellule Beta del pancreas e ha il compito di controllare la glicemia a livello ematico. Un aspetto comune di tutte le forme di diabete mellito è la presenza di iperglicemia sia a digiuno sia post-prandiale.

Il diabete non è una patologia ereditaria, anche se può esistere una certa predisposizione familiare per lo sviluppo della malattia. Il termine diabete deriva dal greco antico "passare attraverso" e indica un abbondante produzione di urina ed una aumentata ingestione di acqua. Mentre il termine mellito deriva dal latino "miele" e si riferisce al sapore dolce delle urine causato dalla presenza di elevati livelli di glucosio nelle urine.

Questa patologia è ampiamente diffusa soprattutto nei paesi industrializzati e il suo andamento è normalmente cronico e porta molte complicanze sia a breve che a lungo termine. Infatti, è la principale causa di insufficienza renale cronica, cecità, amputazione degli arti inferiori e di patologie cardiovascolari come infarti del miocardio ed ictus cerebrali.

1.2 Criteri diagnostici, classificazione e sintomatologia

Il diabete viene diagnosticato quando si presentano queste condizioni clinico-laboratoristiche (tab. 1.1):

- La glicemia a digiuno è uguale o superiore a 126 mg/dl.
- La glicemia è uguale o superiore a 200 mg/dl alla seconda ora dopo un carico orale di glucosio o in qualsiasi momento della giornata.
- L'emoglobina glicata (HbA1c) è uguale o superiore al 6.5 %.

Le condizioni in cui si ha il rischio di sviluppare il diabete sono le seguenti:

• Glicemia a digiuno fra 100 e 125 mg/di.

• Emoglobina glicata fra 6 e 6.49 %.

• Glicemia dopo due ore dal carico orale di glucosio fra 140 e 199 mg/dl.

Per quanto riguarda le diverse classificazioni le varietà di diabete sono:

• Diabete mellito di tipo 1.

• Diabete mellito di tipo 2.

• Diabete gestazionale.

• Diabete monogenico.

• Diabete secondario ad altra patologia o da farmaci.

Tabella 1.1 (tabella dei valori glicemici)

NOME ESAME	***VALORE NORMALE***	***VALORE ALTERATO***	***RISCHIO DIABETE***	***COME SI FA IL TEST***
GLICEMIA	*65-110 mg/dl*		*Nella norma*	*A digiuno da almeno 8-10 ore*
		100-125 mg/dl	*A rischio diabete (prediabete)*	*A digiuno da almeno 8-10 ore*
		>125 mg/dl in più di un esame	*diabete*	*A digiuno da almeno 8-10 ore*
TEST DI CARICO DEL GLUCOSIO	*140 mg/dl*		*Nella norma*	*A digiuno da almeno 8-10 ore*
		140-200 mg/dl	*A rischio diabete (prediabete)*	*A digiuno da almeno 8-10 ore*
		>200 mg/dl in più di un esame	*diabete*	*A digiuno da almeno 8-10 ore*

Il diabete nella maggior parte dei casi non da particolari sintomi ma i più comuni sono: la sete intensa, la necessità di urinare spesso, la fame costante, la vista offuscata, la perdita improvvisa di peso, la mancanza di energie e la stanchezza estrema.

1.3 Eziopatogenesi

L'eziopatogenesi varia in base alle varie tipologie di diabete.

Il **diabete di tipo 1** ha origine autoimmune ed è causata dalla repentina distruzione delle cellule Beta pancreatiche. La distruzione delle cellule Beta è causata da anticorpi e citochine prodotte delle cellule del sistema immunitario probabilmente in risposta a virus o ad agenti tossici.

Per questo nel diabete di tipo 1 è necessaria la terapia con iniezioni di insulina, perché in poco tempo il pancreas non produce più insulina. Esiste anche una variante del diabete di tipo 1 chiamata LADA, in cui l'attacco autoimmune delle cellule beta pancreatiche è più lento e meno aggressivo e la malattia si sviluppa nel corso di anni.

Generalmente si ha una predisposizione genetica, uno dei vari geni responsabili è il gene localizzato nella regione HLA del cromosoma 6. Viene anche chiamato "diabete giovanile" a causa della prevalenza di incidenza negli adolescenti con picchi tra 7-8 e 13-15 anni. In Italia l'incidenza del diabete di tipo 1 è minore nelle regioni del Centro-Sud e maggiore nel Nord.

La patologia diventa clinicamente rilevante quando circa l'80% delle cellule Beta pancreatiche vengono distrutte.

Il **diabete di tipo 2** ha un'eziologia multifattoriale, infatti ci sono sia fattori genetici, che ambientali; generalmente fa la sua comparsa dopo i 40 anni. Si sviluppa generalmente nell'arco di molti anni ed è causato da una elevata resistenza all'insulina, ossia la ridotta capacità dell'insulina di andare ad agire sui propri tessuti bersaglio come muscoli e fegato, o alla carenza di produzione della stessa. Talora entrano in gioco anticorpi IgG anti-insulina.

La resistenza cronica all'insulina si ha quando il fabbisogno giornaliero di insulina supera le 200 Ui per svariati giorni in assenza di altre patologie. Oltre ad una certa predisposizione familiare i principali fattori di rischio sono: il fumo, l'obesità, la sedentarietà, l'iperalimentazione, bassi livelli di colesterolo HDL, l'età avanzata.

Infatti, l'obesità viscerale è una delle cause principali che portano alla resistenza insulinica, poiché il tessuto adiposo produce una serie di sostanze come la leptina, acidi grassi liberi, TNF-alfa, resistina, e adiponectina che concorrono allo sviluppo dell'insulino-resistenza. Il tessuto adiposo è anche sede di uno stato di infiammazione cronico che porta alla produzione di mediatori chimici come interleuchina 6 e proteina C reattiva che aggravano la resistenza insulinica.

Il **diabete gestazionale** invece è una forma di diabete che si può sviluppare durante il periodo della gravidanza, la sua patogenesi è legata all'insulino-resistenza che talvolta si associa alla gravidanza.
Una leggera insulino-resistenza in gravidanza è un fenomeno fisiologico che si comincia a manifestare nel secondo trimestre e aumenta nel terzo trimestre.
Il **diabete monogenico** è causato da mutazioni del DNA mitocondriale che viene ereditato soltanto dalla madre ma colpisce entrambi i sessi.
Il **diabete** può anche essere **secondario** a numerose patologie genetiche come: la sindrome di Turner, la sindrome di Down, la sindrome di Klinefelter, la sindrome di Huntington.
Inoltre, l'insulino-resistenza può essere aggravata dall'utilizzo di farmaci come gli antinfiammatori steroidei, diuretici tiazidici, beta-bloccanti, acido nicotinico, fenitoina, clozapina e altri, che possono portare sia ad un malfunzionamento delle cellule beta sia al peggioramento dell'insulino-resistenza.
Altre malattie causa di diabete possono essere: l'acromegalia, l'ipogonadismo, il glucagonoma, il feocromocitoma, l'ipertiroidismo.

La popolazione mondiale affetta da diabete viene stimata intorno al 5%, di cui circa il 90% della popolazione diabetica è affetta da diabete di tipo 2. In Italia sono circa 3,5 milioni le persone affette da questa patologia che risulta essere in costante aumento negli ultimi anni. Basti pensare che nel 1980 i malati nel mondo erano circa 108 milioni, nel 2014 erano 422 milioni mentre nel 2021 sono stati stimati 536,6 milioni di malati di diabete.
Questi dati fanno capire quanto sia diffusa questa patologia e quanto sia in costante aumento.

1.4 Complicanze

Le complicanze legate al diabete possono essere sia acute che croniche.
Le più frequenti complicanze acute sono la chetoacidosi nel diabete di tipo 1 e la sindrome iperosmolare non chetosica nel diabete di tipo 2.

La chetoacidosi è causata dalla carenza/assenza di insulina che non permette alle cellule di utilizzare glucosio come fonte energetica, in questa situazione il corpo è costretto ad utilizzare lipidi come fonte energetica. Infatti, la chetoacidosi è caratterizzata da un'eccessiva concentrazione di corpi chetonici nel sangue dovuta sia carenza di insulina sia dalla eccessiva produzione di glucagone.

Il sintomo più comune è la comparsa di uno stato confusionale che può portare al coma e se non prontamente trattato porta alla morte. Anche l'intensità del trattamento può portare a pericolose complicanze prima fra tutte il rischio di ipoglicemia, la quale spesso richiede una immediata ospedalizzazione dal momento che una severa ipoglicemia in un paziente fragile può risultare fatale. La sindrome iperosmolare, invece, può portare al coma soprattutto i pazienti più anziani nei quali la capacità di assumere liquidi è notevolmente ridotta rispetto ai giovani.
Questa situazione causa la quasi impossibilità di compensare le perdite idriche dovute alla diuresi che conduce ad un grave stato di disidratazione cellulare.
Le complicanze croniche si sviluppano invece quando il diabete è curato male o trascurato e ciò determina danni a vari tessuti e organi.
I principali organi e tessuti coinvolti sono: l'occhio, il rene, i nervi, le arterie e il cuore. Per questo motivo il diabete viene considerato come una patologia sistemica.

Le complicanze più frequenti sono:

• La macroangiopatia diabetica porta a sviluppare più precocemente patologie come l'aterosclerosi, la quale è la principale causa di ictus cerebrale e infarto del miocardio.

• La Nefropatia diabetica che affligge il rene e favorisce nel lungo periodo l'insorgere dell'insufficienza renale. Questa complicanza è causata dal progressivo danneggiamento dei vasi arteriosi renali che possono causare il malfunzionamento dei reni o la loro insufficienza.

• La retinopatia diabetica colpisce la retina ed è determinata da un progressivo danneggiamento dei vasi sanguigni che irrorano la retina; inizialmente provoca un peggioramento della vista che se non viene adeguatamente trattata può portare a cecità.

• La neuropatia diabetica che va a colpire il sistema nervoso periferico in svariate forme. I danni al sistema nervoso sono causati dalla persistenza dell'iperglicemia che causa danni alle fibre nervose attraverso alterazioni metaboliche e la compromissione vascolare; possono verificarsi in qualsiasi parte del corpo. I sintomi possono essere: dolore, formicolio e perdita di sensibilità soprattutto negli arti inferiori. Altri danni causati dalla neuropatia diabetica possono essere la disfunzione erettile,

problemi all'apparato digerente, problemi a livello dell'apparato urinario e numerosi altri disturbi.

• Il piede diabetico è causato dal danno ai vasi (danni microvascolari) e dalle lesioni al sistema nervoso. Queste problematiche aumentano il rischio di ulcere, infezioni e amputazioni delle estremità, soprattutto degli arti inferiori. Il rischio di amputazione è di addirittura 20 volte superiore rispetto a chi non soffre di tale patologia.

• Glaucoma. È una patologia oculare causata da un danno cronico dovuto da ipertensione intraoculare che determina progressive alterazioni al nervo ottico, con alterazioni che possono portare anche a cecità se non prontamente trattate.

• Complicanze cardiache come angina pectoris e l'infarto del miocardio sempre causate dai danni ai vasi causati da livelli costanti di iperglicemia.

In molti casi le complicanze sono già presenti durante la prima diagnosi, dato che generalmente essa è in ritardo anche di 5-10 anni dall'inizio dello sviluppo di tale patologia.

Per ovviare a questo problema è molto importante oltre ad uno stile di vita sano, sottoporsi a frequenti controlli per cercare di reagire prontamente ad un eventuale comparsa di diabete (1).

CAPITOLO 2
IL NATURAL LANGUAGE PROCESSING

2.1 Definizione

Il natural language processing si riferisce ad una branca dell'intelligenza artificiale che riguarda la capacità che viene data ai computer di analizzare, comprendere e rappresentare il linguaggio naturale delle persone.

Il NLP, sviluppato su modelli statistici di machine learning e deep learning, rende possibili attività come quelle di traduzione di un testo, risposta a comandi vocali, riassumere velocemente testi di grande volume e ricercare parole specifiche in un articolo piuttosto che in un grande insieme di dati.

La lingua utilizzata dagli esseri umani non è comprensibile in maniera immediata da parte di un calcolatore, per questo negli anni sono stati sviluppati strumenti sempre più precisi e adatti allo svolgimento di questo difficile compito.

I moderni sistemi di NLP hanno sviluppato una grandissima quantità di funzioni, oltre a quelle già elencate è necessario aggiungere le seguenti, per avere un punto di vista più completo delle sue potenzialità:

• Text classification: interpretare un testo ed essere in grado di assegnarlo ad una certa categoria basata sul suo contenuto;

• Sentiment analysis: rilevare l'umore all'interno di un testo, per esempio capire se una recensione sia positiva o negativa;

• Intent monitoring: analisi del testo per prevedere possibili eventi futuri;

• Speech recognition: convertire dati vocali in dati testuali;

• Word sense disambiguation: capire il giusto significato di una parola polisemica all'interno di un testo, per esempio valutare se la parola 'pesca' si riferisca al frutto o all'attività di pescare.

• Co-reference resolution: identificare se e quando due parole si riferiscono alla stessa entità, utile soprattutto per valutare a chi o cosa si riferisce un certo pronome.

Il dialogo tra uomo e macchina è fondamentale per il progresso della scienza, avere una buona collaborazione e la possibilità di far comprendere ai computer informazioni senza necessariamente utilizzare dati strutturati o numeri velocizza incredibilmente i processi di ricerca, sviluppo e le applicazioni in questo senso diventano infinite.
Il numero di informazioni presenti sul web è ormai una quantità così grande che sì offre un ventaglio di opportunità e di scelta enorme, ma al tempo stesso rende difficile la ricerca di particolari informazioni nascoste tra migliaia di dati. [2] [3] [4]

2.2 Storia

Secondo una overview della Stanford Computer Science il campo dell'elaborazione del linguaggio naturale iniziò negli anni '40, dopo la seconda guerra mondiale. A quel tempo, le persone riconoscevano l'importanza della traduzione da una lingua all'altra e speravano di creare una macchina in grado di eseguire questo tipo di traduzione automaticamente. Tuttavia, il compito non era così facile come le persone inizialmente immaginavano. Nel 1958, un ricercatore di nome Noam Chomsky, trovava preoccupante che i modelli di linguaggio riconoscessero frasi senza senso ma grammaticalmente corrette altrettanto rilevanti come frasi senza senso e non grammaticalmente corrette. Chomsky ha riscontrato poi diversi problemi riguardo frasi e parole che potevano riferirsi a più contesti ed essere utilizzate in maniera diversa a seconda della situazione, questa ambiguità è facilmente analizzabile dalla mente umana ma non lo era altrettanto dalle macchine.

In quegli anni i ricercatori si sono divisi in due gruppi con differenti punti di vista riguardo il NLP: i simbolici e gli stocastici. I ricercatori simbolici, o basati su regole, si sono concentrati sui linguaggi formali e sulla generazione della sintassi; questo gruppo era composto da molti linguisti e informatici che consideravano questo ramo l'inizio della ricerca sull'intelligenza artificiale. I ricercatori stocastici erano più interessati ai metodi statistici e probabilistici del NLP, lavorando su problemi di riconoscimento ottico dei caratteri e riconoscimento di schemi tra testi.

Dopo il 1970, i ricercatori si sono ulteriormente divisi, abbracciando nuove aree del NLP man mano che più tecnologia e conoscenza sono diventate disponibili. Una nuova area erano i paradigmi basati sulla logica, i linguaggi incentrati sulla codifica delle regole e il linguaggio nelle logiche matematiche.

La comprensione del linguaggio naturale è stata un'altra area del NLP che è stata particolarmente influenzata del professor Terry Winograd. Il suo programma ha collocato un computer in un mondo di blocchi, consentendogli di manipolare e rispondere alle domande sui blocchi secondo le istruzioni del linguaggio naturale dell'utente. Questo sistema aveva capacità di apprendere e comprendere con incredibile precisione se confrontata con gli strumenti del tempo.
Un'altra fondamentale area del NLP è nata dopo il 1970: la modellazione del discorso. Questa è la parte che esamina gli interscambi tra persone e computer, elaborando idee come la necessità di cambiare "tu" nella domanda dell'operatore con "me" nella risposta del computer.

Dal 1983 al 1993, i ricercatori sono diventati più uniti nel concentrarsi sull'empirismo e sui modelli probabilistici. I ricercatori sono stati in grado di testare alcuni argomenti di Chomsky e altri degli anni '50 e '60, scoprendo che molti argomenti convincenti nel testo non erano empiricamente accurati. Pertanto, nel 1993, i metodi probabilistici e statistici di gestione dell'elaborazione del linguaggio naturale erano i tipi più comuni di modelli. Nell'ultimo decennio, il NLP si è anche concentrato maggiormente sull'estrazione e la rielaborazione di informazioni a causa della grande quantità di informazioni sparse su Internet.

I modelli di NLP basati sull'approccio statistico-probabilistico hanno dominato la scena per oltre 20 anni, con sistemi che si sono costantemente evoluti incrementando le loro performance – due esempi su tutti Google Translate e Microsoft Translator.
Il 2013 è l'anno che più ha rivoluzionato il NLP, grazie allo studio di di Tomáš Mikolov che presenta il modello Word2Vec. Quest'applicazione ha come principio quello di rappresentare le parole come grandezza vettoriale, gettando le fondamenta per i moderni sistemi di NLP – come, per esempio, Bert – basati su campi vettoriali. Questi vettori, o word embedding, vengono ricavati in modo iterativo a partire dall'analisi del contesto di utilizzo delle parole nell'ambito di un corpus di addestramento, e sono in grado di cogliere nelle varie dimensioni la semantica della parola.
Rappresentare le informazioni sotto forma di vettori ha segnato un punto di svolta per l'utilizzo di modelli sempre più simili alla mente umana, sfruttando reti neurali artificiali per lo svolgimento di diverse attività – dalla classificazione di testo, al riconoscimento di entità, all'analisi grammaticale, e al compito della machine translation: nel 2014 Google sviluppa il modello Sequence-to-Sequence (Seq2Seq) basato su word embedding e reti neurali ricorrenti, primo vero modello di neural machine

translation. [5] [6]

Negli ultimi anni sono stati fatti enormi passi avanti e – di pari passo con tutti gli sviluppi tecnologici – il NLP è in continua evoluzione in modo esponenziale, sviluppando sempre nuovi algoritmi in grado di migliorare la versione precedente e fare ulteriori step in avanti, aprendo un grandissimo numero di porte in ogni ambito della scienza.

2.3 Bert

2.3.1 Architettura

Bert (Bidirectional Encoder Representations from Transformers) è un modello di elaborazione del linguaggio naturale, descritto dai ricercatori di Google AI Language, che tramite tecniche di machine learning ha presentato risultati incredibili in moltissimi utilizzi del NLP. Lo sviluppo ha impiegato tempo e numerose risorse ma può essere considerato lo stato dell'arte per quanto riguarda le applicazioni in questo settore.

In particolare, sono stati raggiunti ottimi risultati per ciò che riguarda: la comprensione di parole polisemiche, questo grazie ad una consapevolezza più profonda del contesto di riferimento; la capacità di effettuare ricerche specifiche per determinate parole, in modo da rendere il processo di ricerca più veloce ed efficiente; la capacità di analizzare grandi quantità di dati in poco tempo ed andare a selezionare determinati set di dati non strutturati e dividerli in base all'ambito specifico.

Questo modello si basa sull'architettura del trasformer, ovvero una serie di endocer e decoder avente lo scopo di apprendere una rappresentazione del testo di input, anche se nel suo caso vengono utilizzati solo degli encoder essendo un sistema di pre-training.

Un encoder, noto in italiano con il nome di codificatore, cambia il formato delle informazioni da una forma all'altra, per migliorare la velocità e la precisione durante la trasmissione, per mantenere le informazioni in modo sicuro e per la standardizzazione. Il codificatore potrebbe ridurre le dimensioni di archiviazione effettive convertendo i dati in un altro formato.

Anche i convertitori da digitale ad analogico (DAC) e da analogico a digitale (ADC) sono encoder elettronici.

Un decoder, noto in italiano come decodificatore, svolge le funzioni opposte del codificatore, invertendo il processo di codifica e convertendo le informazioni nel formato precedente o in un altro formato accessibile.

Ad esempio, in elettronica, se un segnale viene codificato utilizzando un convertitore da analogico a digitale per scopi di trasmissione, il ricevitore deve decodificare il segnale utilizzando il convertitore da digitale ad analogico per recuperare il segnale analogico originale.
In questo caso, ADC funge da codificatore e DAC funge da decodificatore.
Bert è spesso descritto come un modello di deep learning pre-addestrato, tuttavia può essere più correttamente definito come un framework, poiché fornisce ai professionisti del machine learning una base per costruire le proprie versioni simil-Bert tramite le quali è possibile soddisfare una vasta gamma di task.
In informatica un framework è un'architettura logica di supporto sulla quale un software può essere progettato e realizzato, facilitandone lo sviluppo da parte del programmatore.

2.3.2 Funzionamento

Il modello di Bert acquisisce ogni parola avente la sua rappresentazione, denominata più propriamente embedding, e la ordina in dei vettori.
Le parole target vengono rappresentate come dei token, il processo di selezione delle parole target viene detto tokenizzazione: è utilizzato per segmentare il testo di input nelle sue parole costituenti (token). In questo modo diventa più semplice convertire i nostri dati in un formato numerico.
Nella fase di tokenizzazione Bert prende la frase di input e deciderà di mantenere ogni parola come una parola intera, dividerla in sottoparole o come ultima risorsa, scomporre la parola in singoli caratteri. Per questo motivo, possiamo sempre rappresentare una parola come, almeno, la raccolta dei suoi singoli caratteri.
Questa suddivisione è data dal fatto che Bert Vocabulary è fisso con una dimensione di ~ 30.000 token, quindi le parole che non fanno parte del vocabolario sono rappresentate come sottoparole e caratteri.

Il word embedding (tradotto letteralmente incorporamento di parole) anche conosciuto come 'rappresentazione distribuita delle parole' permette di memorizzare le informazioni sia semantiche che sintattiche delle parole partendo da un corpus non annotato e costruendo uno spazio vettoriale in cui i vettori delle parole sono più vicini se le parole occorrono negli stessi contesti linguistici, cioè se sono riconosciute come semanticamente più simili.
Fondamentale importanza del processo è l'attention, ovvero una tecnica usata in ambito di

intelligenza artificiale che imita l'attenzione cognitiva.

L'effetto migliora alcune parti dei dati di input mentre diminuisce altre parti: la motivazione è che la rete dovrebbe dedicare maggiore attenzione alle parti piccole, ma importanti dei dati.

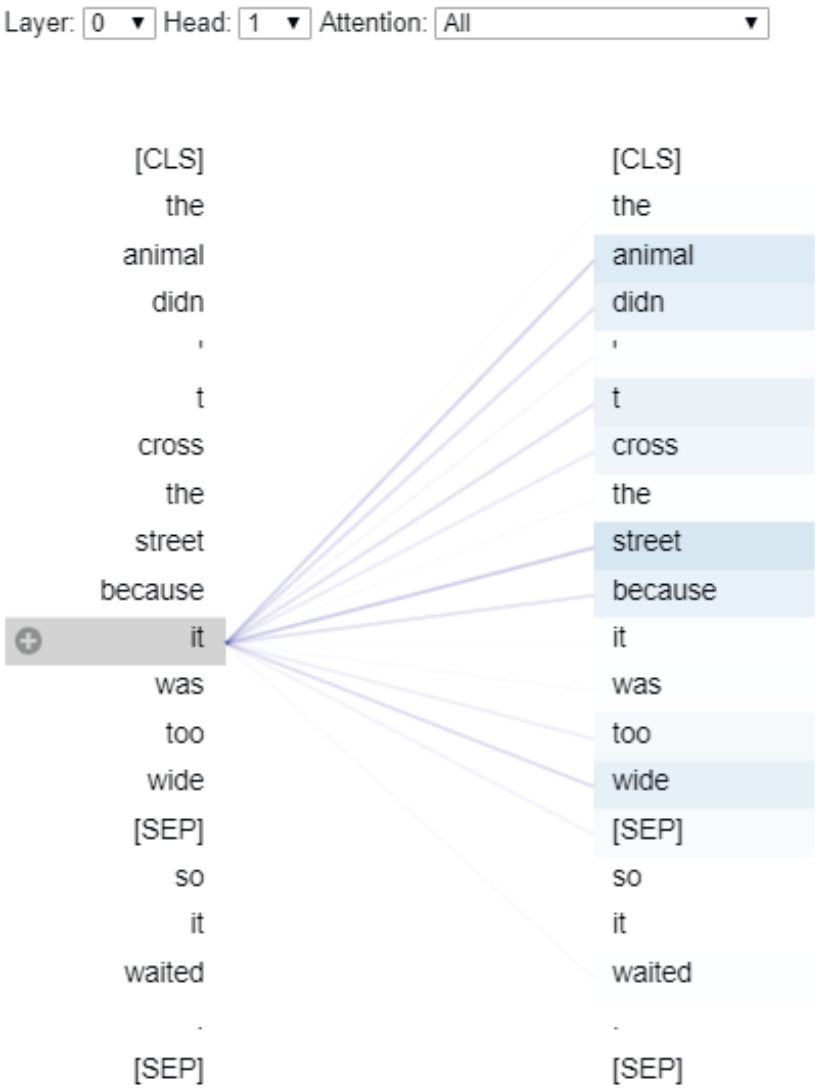

Figura 2.1 Meccanismo di attenzione: collegamenti creati tra una parola e tutte le altre

Nell'illustrazione precedente (figura 2.1), viene mostrato un esempio del meccanismo di attenzione, la parola "*it*" attira tutti gli altri token (parole chiave) e crea una connessione maggiore con "street" e "*animal*", che sono quelli che più si collegano alla parola "it".

Imparare quale parte dei dati è più importante di un'altra dipende dal contesto, e questo viene addestrato da un sistema di ottimizzazione.

Le prestazioni all'avanguardia di Bert si basano su due punti chiave: innanzitutto nuove strategie di pre-training chiamati Masked Language Model (MLM) e Next Sentence Prediction (NSP), in secondo luogo, l'utilizzo di una grande quantità di dati e di potenza di calcolo per la fase di training.

• Masked Language Model: l'architettura Bert analizza le frasi con alcune parole mascherate casualmente (da qui deriva il nome masked) e tenta di prevedere correttamente la parola "nascosta".

Durante il training bidirezionale, che passa attraverso l'architettura del trasformatore Bert, lo scopo del MLM è quello di impedire alle parole target di creare inavvertitamente connessioni errate con altre parole, in quanto tutte vengono esaminate nello stesso istante e nel proprio contesto.

Questo processo, quindi, ha lo scopo di evitare la creazione di un tipo di loop infinito che potrebbe implicare possibili errori nell'apprendimento automatico del linguaggio naturale, alterando il significato della parola.

In termini tecnici, la previsione delle parole di output richiede:

1. Aggiunta di un livello di classificazione sopra l'output dell'encoder.
2. Moltiplicando i vettori di output per la matrice di incorporamento, trasformandoli nella dimensione del vocabolario.
3. Calcolo della probabilità di ogni parola del vocabolario con softmax.

In matematica, una funzione softmax (1), o funzione esponenziale normalizzata, è una generalizzazione di una funzione logistica che comprime un vettore k-dimensionale z di valori reali arbitrari in un vettore k-dimensionale σ(z) di valori compresi in un intervallo (0,1) la cui somma è 1.

La funzione è data da:

$$\sigma : \mathbb{R}^k \rightarrow \left\{ z \in \mathbb{R}^k \mid z_i > 0,\ \sum_{i=1}^{k} z_i = 1 \right\}$$

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \quad per\ j = 1, \dots, k \qquad (1)$$

Dato in input un vettore (1;2;3;4;1;2;3), la funzione softmax restituirà (0,024; 0,064; 0,175; 0,475; 0,024; 0,064; 0,175). Il risultato assegna gran parte del peso al numero 4, il cui valore in uscita risulta essere circa 20 volte maggiore del valore associato a 1. Questo è esattamente ciò per cui la funzione solitamente è usata: mettere in evidenza i valori più grandi e nascondere quelli che sono significativamente più piccoli del valore massimo.

• Next Sentence Prediction: una delle principali innovazioni di Bert è quella di essere in grado di prevedere ciò che un utente dirà dopo una determinata frase.
La frase viene analizzata e viene fatta una previsione della frase che potrebbe essere adatta per rappresentare il suo proseguo.
Una volta calcolate le probabilità ed analizzata la frase successiva il risultato avrà un valore positivo o negativo, in base al grado di affinità tra la previsione e la frase effettivamente inserita.
Durante l'allenamento, il 50% degli input sono una coppia in cui la seconda frase è la frase successiva nel documento originale, mentre nell'altro 50% viene scelta una frase casuale dal corpus come seconda frase. Il presupposto è che la frase casuale sarà disconnessa dalla prima frase.
Per aiutare il modello a distinguere tra le due frasi in training, l'input viene elaborato nel modo seguente prima di entrare nel modello:

1. Un token [CLS] viene inserito all'inizio della prima frase e un token [SEP] viene inserito alla fine di ogni frase;
2. A ciascun token viene aggiunta una frase incorporata che indica la frase A o la frase B;
3. Un'inclusione posizionale viene aggiunta a ciascun token per indicarne la posizione nella sequenza.

Per prevedere se la seconda frase è effettivamente collegata alla prima, vengono eseguiti i seguenti passaggi:
1. L'intera sequenza di input passa attraverso il modello Transformer;

2. L'output del token [CLS] viene trasformato in un vettore, utilizzando un semplice livello di classificazione;

3. Calcolo della probabilità di "IsNextSequence" con softmax.

Durante l'addestramento del modello BERT, Masked LM e Next Sentence Prediction vengono addestrati insieme, con l'obiettivo di ridurre al minimo la funzione di perdita combinata delle due strategie.

La comprensione del contesto da parte di un algoritmo deriva dalla sua capacità di osservare tutte le parole di una frase nello stesso istante e quindi di capire come ogni singola parola influenzi tutte le altre. Con questa metodologia, il ruolo di una determinata parola all'interno di una frase può letteralmente cambiare man mano che la frase si sviluppa.

Bert seleziona una parola target e al contempo prende in considerazione tutte le altre parole presenti nella frase in maniera bidirezionale, di modo che ogni parola sia confrontata con tutte le altre e le venga affidato un certo peso probabilistico di predizione e affinità.

Questo procedimento si può paragonare alla modalità di ragionamento della mente umana, secondo la quale si osserva l'intero contesto di una frase piuttosto che una sola part [7][8].

2.4 Natural Language Processing in sanità

È normale che una tecnologia avanzata come il NLP venga sfruttata in ogni ambito, a maggior ragione in settori di fondamentale importanza come quello sanitario.

Dal momento che una grande parte dei dati "essenziali" agli studi si trova in cartelle cliniche non strutturate, documenti scientifici e articoli tratti da conferenze e convegni di medicina, ottenere un accesso immediato a tali informazioni cliniche può risultare un compito arduo. Riguardo ciò il NLP può essere una chiave unica e necessaria per reperire queste informazioni. I ricercatori possono infatti utilizzare l'intelligenza artificiale per "ordinare" i dati non strutturati – sotto forma di testo libero – in un linguaggio comprensibile al computer.

Questi dati possono quindi fornire indicazioni preziose per la cura dei pazienti, per la ricerca e la diagnosi delle malattie.

Per quanto riguarda l'addestramento dei programmi di NLP in ambito biomedico possono essere usate diverse risorse testuali.
In uno studio condotto da Yanshan Wang e colleghi [9] si sono voluti valutare gli embedding di parole addestrati da quattro diversi corpora: note cliniche, pubblicazioni biomediche, Wikipedia e notizie. Per le prime due sono state utilizzati i dati di EHR non strutturati disponibili presso la Mayo Clinic e gli articoli di PubMed Central. Per le altre due risorse sono stati usati embedding di parole pre-addestrati pubblicamente disponibili, GloVe e Google News.

I risultati mostravano che l'embedding di parole collegato ad un addestramento effettuato su EHR – e quindi utilizzando termini specificatamente biomedici – offriva poi una possibilità di ricerca più adatta a ricercare parole appartenenti a questo settore rispetto che quelli in relazione ad un addestramento su portali più generici. Anche la somiglianza semantica catturata dai prime due è risultata essere più vicina ai giudizi degli esperti umani.
Non esiste, tuttavia, una classificazione globale coerente degli embedding di parole per tutte le applicazioni di NLP ma ognuna avrà le sue specifiche tecniche per rendere il sistema più preciso in relazione al particolare campo di studio.

Per gestire meglio i dati non strutturati nell'ambito della ricerca clinica, lo statunitense Georgetown University Medical Center di Washington D.C. ha adottato strumenti di text-mining (tecnica che utilizza il NLP per trasformare il "testo libero" e non strutturato in dati strutturati) basati sull'intelligenza artificiale applicata alle cartelle cliniche elettroniche.
Lo strumento varato dall' ospedale della capitale USA permette ai medici di cercare rapidamente ciò che desiderano partendo da grandi quantità di dati di letteratura medica, affinché vi sia un supporto efficiente del processo decisionale clinico in tempo reale. I ricercatori del gruppo ospedaliero statunitense Permanente Medical Group hanno addestrato il NLP a ordinare e, attraverso oltre un milione di cartelle cliniche elettroniche ed ecocardiogrammi, identificare alcune abbreviazioni, parole e frasi associate alla stenosi aortica.
In pochi minuti, il software ha identificato più di cinquantamila pazienti affetti da stenosi aortica; un processo che, manualmente, avrebbe richiesto anni per i medici. Questo fa emergere uno dei punti di maggior impiego del NLP, ovvero analizzare, confrontare e classificare una grandissima quantità di dati, in modo da velocizzare enormemente i tempi di ricerca e con il vantaggio di recuperare informazioni che altrimenti sarebbero probabilmente rimaste irraggiungibili.

In questo caso il sistema ha aiutato ad eseguire una diagnosi e quindi ad indirizzare i pazienti verso un programma di screening, monitoraggio di eventuali problemi e di conseguenza anche di cura e riabilitazione.

Esiste una grande quantità di importanti dati clinici nelle cartelle cliniche dei pazienti. Allo stesso tempo però, molti di questi dati rimangono inutilizzati e fuori dalla portata dei ricercatori essendo dati difficilmente reperibili e utilizzabili a causa della loro natura non strutturata.
Le applicazioni basate sul NLP possono trovare la soluzione a questo problema, migliorando l'efficienza e la fattibilità del ricercare e organizzare dati di grandi gruppi di pazienti da studiare senza andare a leggere ogni cartella clinica "manualmente", aumentando così il campo di studio e al tempo stesso velocizzando il processo. Dataset più grandi portano inevitabilmente ad una maggiore precisione di dati e una più alta affidabilità dei risultati.
Il NLP può anche aiutare i ricercatori a superare le limitazioni dei codici di procedura e diagnosi. Codici che, attualmente, non sono progettati per includere dati dettagliati su una specifica condizione medica. Per esempio, lo status medico di un paziente con stenosi aortica moderata o grave è completamente diverso da un paziente con una lieve malattia che colpisce la valvola aortica.
Queste variabili non sono incluse nei codici di diagnosi o di procedura. Inoltre, alcuni codici possono semplicemente etichettare la patologia come "malattia della valvola aortica", che potrebbe essere applicato a un problema clinico completamente diverso dalla stenosi aortica.

In Europa invece, lo strumento di text-mining dell'azienda di Cambridge (UK) Linguamatics utilizza l'elaborazione del linguaggio naturale per ordinare il testo in specifiche "frasi chiave".
Le informazioni estratte possono identificare il miglior corso di trattamento per i pazienti. Utilizzando questa tecnologia, i gruppi di ricerca possono cercare nella letteratura e nelle cartelle cliniche per scoprire i geni associati a determinate malattie e migliorare la loro comprensione dei processi molecolari, al fine di avanzare con il "targeting" dei farmaci.
I motori di ricerca tradizionali (come PubMed) potrebbero cercare e mostrare un grande numero di articoli, la maggior parte dei quali potrebbero non contenere le informazioni che si stanno cercando, dal momento che il risultato della ricerca comprenderebbe tutti gli articoli che contengono quelle parole chiave, ma che magari non rientrano nel contesto di ricerca desiderato.
medici dovrebbero leggere l'abstract o anche il testo completo per identificare uno o due articoli

che contengono le informazioni che stanno cercando, portando inevitabilmente ad un dispendio enorme di tempo e di risorse. Con Linguamatics invece, cinquanta potrebbero essere identificati immediatamente quei due articoli che interessano senza doverne passarne al setaccio altri quarantotto.

Inoltre, Linguamatics utilizza capacità di elaborazione del linguaggio naturale per cercare l'intero testo di un articolo e identificare concetti e relazioni nella letteratura, al fine di fornire cure di alta qualità, analizzando immediatamente il problema e fornendo la soluzione più adeguata confrontando i dati presenti in letteratura. È facile rendersi conto che i tradizionali metodi di ricerca senza capacità di elaborazione del NLP non siano in grado di eseguire gli stessi compiti, né in termini di tempo né di affidabilità. [10]

CAPITOLO 3
APPLICAZIONI DEL NATURAL LANGUAGE PROCESSING IN RELAZIONE AL DIABETE

3.1 Natural Language Processing e diabete

Il diabete mellito è un buon esempio di una malattia che potrebbe trarre vantaggio dalla generazione di real-world evidences (o prove del mondo reale) utilizzando il NLP. Nonostante questo, i primi studi di NLP correlato a studi sul diabete risalgono solo al 2005, anche se si tratta di un settore in grande espansione che negli anni si è arricchito di numerosi articoli e studi.

La ricerca di articoli e studi riguardo NLP e diabete è stata eseguita utilizzando tre principali siti, quali pubmed, scopus e ACM digital library. In questi servizi di ricerca di articoli ho inserito nella barra di ricerca le parole “natural language processing” e “diabetes”. Successivamente per avere una visione più completa sono anche state inserite parole come “EHR”, “electronic health record”, “electronic medical record”, “hyperglycemia”, “diabetes mellitus”.

Dalla ricerca sono risultati pochi articoli, per esempio quelli che contenevano “natural language processing” e “diabetes” risultavano essere nell’ordine della decina. Gli articoli spesso si citavano a vicenda o erano stati scritti più articoli per uno stesso studio. Di seguito ho riportato un esempio di articoli riguardanti tre diverse aree di pertinenza del diabete: prediabete, diabete e ipoglicemia. Lo scopo di questi studi è principalmente quello di descrivere il funzionamento di algoritmi legati alla diagnosi di diabete, spiegando quali sono le caratteristiche principali, come sono stati utilizzati e i risultati che ne sono derivati.

3.2 Studio sul prediabete

Tra i vari studi che mettono in relazione NLP, cartelle cliniche elettroniche e diabete - qui in particolare si tratta di prediabete - in letteratura, ce n’è uno sostenuto dal Johns Hopkins Institute. [11]

Questo report fa riferimento al fatto che gli studi che esaminano le pratiche dei PCP (“Primary Care Providers”, i nostri medici di base) nella gestione dei pazienti con prediabete utilizzando più che altro i dati strutturati delle cartelle cliniche elettroniche (EHR), suggeriscono che i pazienti con prediabete non stanno ricevendo cure basate sull’evidenza. Per questo studio è stato sviluppato e convalidato uno strumento di elaborazione del linguaggio naturale (NLP) per analizzare i dati non strutturati

nelle note EHR per identificare le discussioni sul prediabete e descriverle.

Nelle prime fasi sono stati inclusi gli adulti senza diabete ma con una visita ambulatoriale di persona presso una clinica di cure primarie presso un centro medico accademico (sono stati selezionati 19 centri medici) e almeno un HbA1c 5,7–6,4% tra il 7/1/2016 e il 12 /31/2018.
È stata basata la strategia di ricerca per parola chiave iniziale sull'esperienza clinica degli autori (Tabella 3.1).
Nella fase 1, l'obiettivo era quello di identificare ed estrarre dalle note degli incontri una o più parole chiave che avessero riguardato il prediabete. Per i pazienti che soddisfano i criteri di inclusione/ esclusione ma che non contengono alcuna parola chiave, sono state identificate ulteriori parole chiave.

Tabella 3.1 Parole chiave incluse nella strategia di ricerca e frequenza delle parole chiave corrispondenti alla distcussione clinica sul prediabete.

Parole chiave	**Frequenza di occorrenza in 322 note cliniche. Numero di pazienti (%)**
Prediabete	137 (43)
Iperglicemia	55 (17)
Alterata glicemia a digiuno	56 (17)
Pre-diabete	41 (13)
Alterata tolleranza al glucosio	15 (5)
Elevata emoglobina glicata	14 (4)
Pre-DM	10 (3)
Intolleranza al glucosio	4 (1)
Rischio diabete	2 (1)
Elevata A1c	4 (1)
Glucosio elevato	2 (1)
Elevata glicemia a digiuno	2 (1)
PreDM	1 (0.3)
Rischio aumentato di diabete	0
Aumento del rischio di diabete	0
Elevato rischio di diabete	0
Pre diabete	0
Pre DM	0
Disglicemia	0

Nella fase 2, utilizzando i dati di altre 17 cliniche, sono state estratte le prime annotazioni della visita dal medico dopo che i risultati di laboratorio indicavano il prediabete (n = 1095 incontri) ed è stata applicata la strategia di ricerca delle parole chiave aggiornata (n = 391 incontri).

Due revisori esperti hanno valutato manualmente le note per determinare se contenessero discussioni cliniche sul prediabete.

Sono stati valutati i risultati della classificazione utilizzando sensibilità, specificità, valore predittivo positivo (PPV) e valore predittivo negativo (NPV).

I due revisori hanno esaminato ciascuna nota della fase 2 per descrivere le discussioni sul prediabete: esami di laboratorio ordinati (HbA1c o glicemia a digiuno); consulenza sullo stile di vita;

discussione sul programma di prevenzione del diabete (DPP); discussione/riferimento nutrizionale; discussione o ordinazione/continuazione della metformina.

Questo studio è stato approvato dal Johns Hopkins IRB.

Nella fase 2, 322 delle 391 note degli incontri avevano più di una parola chiave che documentava una discussione sul prediabete.

I risultati della classificazione del NLP e dell'apprendimento automatico erano vicini alle prestazioni umane. I modelli di regressione logistica hanno rivelato una sensibilità di 0,961, una specificità di 0,923, un PPV di 0,967 e un NPV di 0,907.

Le reti neurali convoluzionali (CNN) hanno rivelato una sensibilità di 0,979, specificità di 0,956, PPV di 0,979 e NPV di 0,956.

I PCP comunemente fornivano consulenza sullo stile di vita (78%), esaminavano i dati di laboratorio attuali (67%) e ordinavano esami di laboratorio di follow-up (60%) (Figura 3.2).

Hanno invece discusso o indirizzato i pazienti a un nutrizionista di rado (3%).

Non ci sono state discussioni o rinvii a un DPP. La metformina è stata discussa, prescritta o continuata in < 2% dei pazienti.

Tabella 3.2 Gestione del prediabete documentato negli incontri clinici da parte dei PCP

Risultati	**Frequenza di differenti casi di gestione del diabete. Percentuale (%)**
Discussioni di laboratorio	
Revisione di dati di esami	215 (66.8)
Esami ordinati	196 (60.1)
Discussioni su cambiamenti di vita	
Raccomandazioni sullo stile di vita	250 (77.6)
Segnalazione nutrizionale inserita	10 (3.1)
Discussione di visita nutrizionale	9 (2.8)
Discussione di un programma di prevenzione	0
Discussioni sulla metformina	
Metformina discussa	6 (1.9)
Metformina ordinata	5 (1.6)
Metformina continuata	4 (1.2)

È stato sviluppato e convalidato lo strumento di NLP che identifica le discussioni cliniche sul prediabete da dati EHR non strutturati. I PCP sottoutilizzano il codice di diagnosi del prediabete (dati strutturati); solo il 13% dei pazienti ha ricevuto un codice diagnostico, denotando una mancanza di coscienza riguardo i possibili rischi e non considerando come sia necessario intervenire sin da subito per evitare il peggioramento della situazione. Pertanto, l'utilizzo di dati EHR strutturati non è sufficiente per identificare le visite in cui viene affrontato il prediabete.
I PCP hanno affrontato più comunemente il prediabete attraverso la consulenza sul cambiamento dello stile di vita, anche se si riferivano raramente alla nutrizione o ai DPP. Hanno invece più spesso esaminato e ordinato laboratori di follow-up, anche se i dati suggeriscono bassi tassi di completamento. Questi risultati descrittivi si basano sulla documentazione delle visite dei medici che potrebbero non aver documentato tutti i dettagli delle loro discussioni verbali.
Non sono state considerate discussioni che potrebbero essersi verificate tramite la messaggistica EHR paziente-medico o la documentazione telefonica.

Nonostante siano presenti problematiche e limitazioni – spesso causate dal comportamento dei medici e non da errori della macchina - man mano che la prevenzione del diabete cresce, questo nuovo strumento può aiutare a tenere traccia delle note relative al prediabete al di fuori delle attività identificabili nei dati strutturati, portando numerosi vantaggi nella cura e nel monitoraggio di situazioni a rischio che potrebbero portare conseguenze pericolose al paziente.

3.3 Identificazione automatizzata dei casi di diabete

Per capire più nello specifico come funziona un algoritmo di identificazione a partire da note di EHR è presente in letteratura uno studio condotto da Mayo Clinic Robert D., dal Patricia E. Kern Center for the Science of Health Care Delivery con da una sovvenzione del National Institutes of Health (NIH) in cui viene spiegato com'è progettato uno dei sistemi più funzionali all'identificazione del DM [12].
L'obiettivo di questo studio era quello di creare un algoritmo in grado di identificare pazienti con diabete mellito di tipo 1 o 2 prima di un intervento chirurgico utilizzando i dati clinici disponibili di routine nelle EHR.
Il DM è la principale causa di complicanze microvascolari, infarto del miocardio, ictus, insufficienza cardiaca congestizia e malattia renale allo stadio terminale che spesso si traduce in morte prematura o invalidità. A seguito di ciò i pazienti con DM vengono sottoposti a procedure chirurgiche più

complesse rispetto ai pazienti senza DM, questo deriva dal fatto che è fondamentale tenere sotto controllo più valori che sono vitali per i malati. Numerosi studi hanno infatti dimostrato che le complicanze postoperatorie come ictus, infezioni del tratto urinario, emorragie postoperatorie, trasfusioni, infezioni della ferita e persino la morte sono più comuni tra i pazienti con DM non controllato.

Alla luce di ciò è necessario identificare i pazienti che hanno DM prima del loro episodio di cura e avviare protocolli di cura appropriati per ottimizzare i loro livelli glicemici. L'EHR integrato della Mayo Clinic è conforme a tutte le definizioni di un EHR completo e al suo interno offre ai medici la possibilità di acquisire le caratteristiche demografiche e la storia medica dei pazienti, che include la documentazione dei problemi medici dei pazienti in un formato di testo libero nelle note dei pazienti.

Gli elementi di dati necessari per determinare se un paziente possa essere affetto da DM sono spesso esistiti nell'EHR, ma la mancanza di una chiara identificazione di una diagnosi di DM nell'EHR era un ostacolo all'implementazione. Oltre ai tre valori primari sfruttati per l'identificazione – codici diagnostici (ICD-9-CM); valori delle prove di laboratorio; dati sui farmaci dei pazienti - molti studi hanno indicato che l'elaborazione del linguaggio naturale (NLP) può aumentare considerevolmente l'accuratezza e la precisione durante l'identificazione di problemi di salute documentati nelle note cliniche.

Per la progettazione dell'algoritmo in questione sono state utilizzate regole logiche del primo ordine (se-allora-altro) per implicare la presenza o l'assenza di DM del primo o del secondo tipo.

Questo modello di classificazione è basato su regole aveva una serie di affermazioni logiche che utilizzavano operatori logici "e" ed operatori logici "o". Oltre a modellare l'effettivo processo decisionale umano con affermazioni logiche, l'approccio basato sulla regola se-allora-altro trova un equilibrio tra accuratezza e interpretabilità per problemi di classificazione generale.

Una panoramica del modello di classificazione è fornita dalla figura 3.3 e dalla figura 3.4. Con questo metodo, la nostra regola classificava un paziente come affetto da DM quando l'EHR del paziente conteneva quanto segue:

- 1. Uno o più codici diagnostici ICD-9-CM correlati al diabete ambulatoriale
    - o "o"

• 2. Almeno 1 farmaco ipoglicemizzante segnalato durante la riconciliazione dei farmaci ambulatoriali

- “o”

• 3. Una combinazione di uso di metformina e un valore di laboratorio che supera il valore di soglia massimo

- “o”

• 4. Eventuali annotazioni positive di DM nelle note cliniche del paziente

Contiene:

Forma precisa di DM

Forma precisa di DM2

Diabete

Non contiene:

Forma precisa di DM

Forma precisa di DM2

Diabete

Nella stessa frase di:

Non è stato diagnosticato con, gestazionale, nonni, steroidi, zia, non, no, senza, sorella, fratello, figlia, figlio, figli, mai, famiglia, non riscontrato, non riportato, negativo, possibile, remoto, padre, madre, genitore, storia familiare, no storia

Figura 3.3 Algoritmo per la ricerca delle parole chiave.

Questo algoritmo così strutturato serve a identificare tutte le volte in cui compare un termine relativo al DM e classificarlo, facendo attenzione a non considerare tutte quelle frasi in cui è presente la parola 'diabete' in successione a frasi come 'non è affetto da' oppure 'non ha', e così via; oppure se queste comparse di termini si trovano in relazione ad altri soggetti della famiglia che non siano però il paziente stesso.

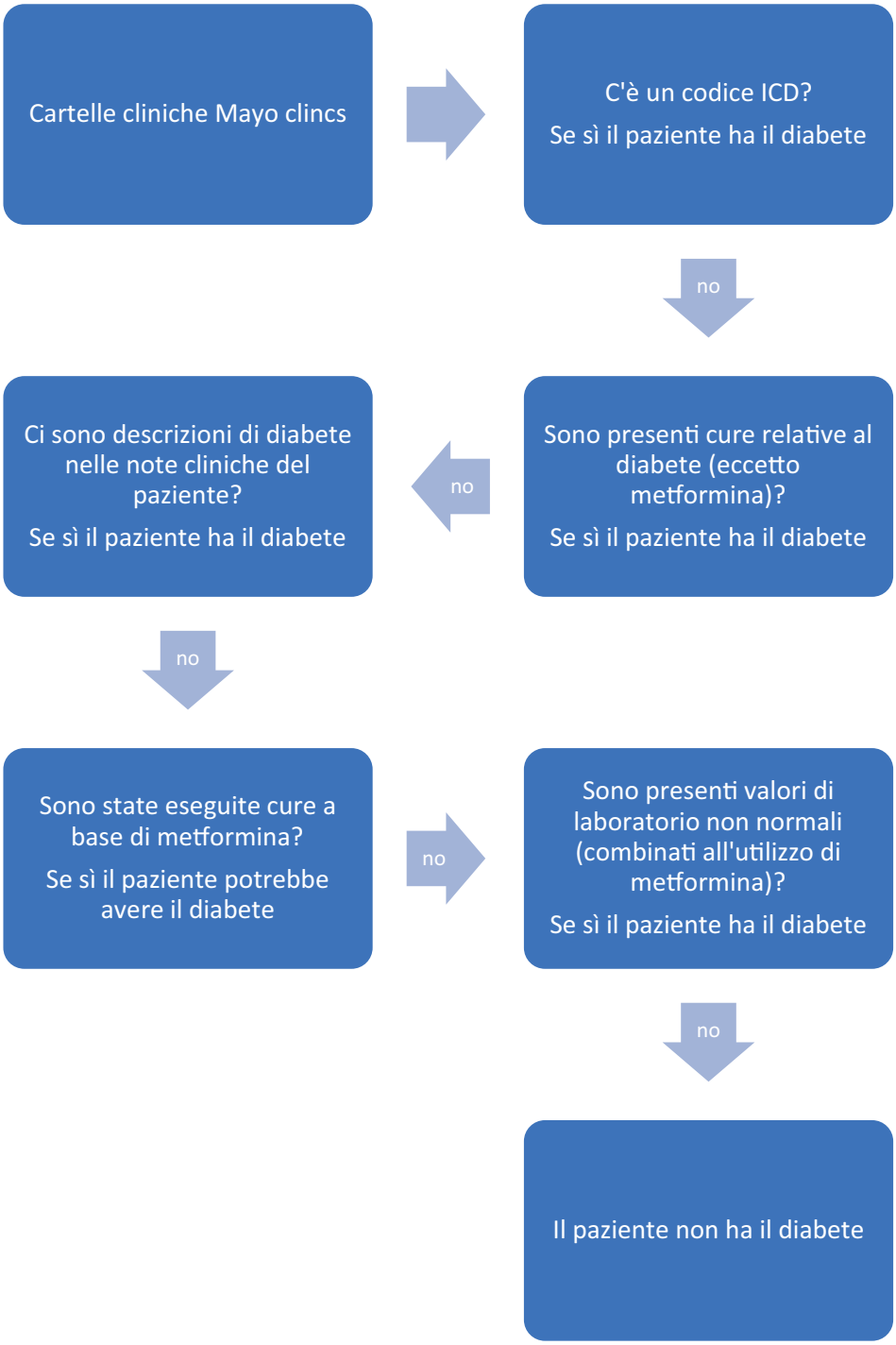

Figura 3.4 Passaggi effettuati dall'algoritmo procedendo per grado di specificità.

L'algoritmo di figura 3.4 come primo passo analizza le cartelle cliniche sulla base dei codici ICD, in questo caso il compito di classificarli come affetti da DM è molto semplice.
In secondo luogo, va ad analizzare i farmaci e le cure presenti segnalate nella cartella clinica del paziente alla ricerca di eventuali sostanze associate soltanto alla cura del diabete. Il terzo passo consiste nella ricerca di parole relative al diabete e la loro analisi, in modo da non creare dei falsi positivi nel caso queste parole siano associate alla negazione o a problemi relativi a familiari.
Gli ultimi step si riferiscono all'uso di metformina, farmaco utilizzato nello specifico per la cura del diabete, e all'analisi dei valori di laboratorio.
Oltre a questo algoritmo ne esistono molti altri sviluppati da altre aziende con diversi scopi. Sembrano essere simili nella costruzione ma c'è una grande differenza soprattutto nell'uso per cui vengono sfruttati rispetto al campione di popolazione che si ha a disposizione: qualcuno è usato solo per un tipo di diabete, qualcuno solo per scopi di ricerca, altri ancora sono specifici per certi gruppi di individui come minoranze etniche o selezionati rispetto ad un certo reddito. Avendo una grande varietà di obiettivi c'è anche incertezza nel poter generalizzare questi algoritmi su larga scala.
L'algoritmo sviluppato dalla Mayo Clinics per poter essere sviluppato ha seguito un approccio iterativo: le informazioni ottenute in ogni fase vengono utilizzate per mettere a punto e migliorare il metodo dell'algoritmo finale, procedendo step-by-step.

Per mettere alla prova questo programma i risultati sono stati confrontati con quelli del processo preesistente, che consiste nella revisione manuale - da parte di operatori specializzati - dell'EHR per identificare una diagnosi di DM effettuata degli infermieri al posto letto che provvedono a documentare le risposte del paziente, informazioni sull'elenco chirurgico, sul regime terapeutico del paziente e su una revisione delle note cliniche.
I risultati sono incredibili, con tutti i parametri quali specificità, sensibilità, precisione, che superano lo 0.99 (su 1). Questo sta a dimostrare come l'algoritmo sia di fondamentale importanza per una revisione accurata, ottenendo statistiche di molto superiori a quelle ottenute con il metodo di lettura manuale da parte di un operatore specializzato. Gli unici casi di errore riguardano un soggetto risultato falso positivo a causa di un errore di scrittura nella cartella clinica e due falsi negativi erroneamente classificati per mancanza di accesso ai dati di origine dei pazienti.
Il diagramma di flusso relativo all'individuazione dei pazienti affetti da DM è riportato di seguito:

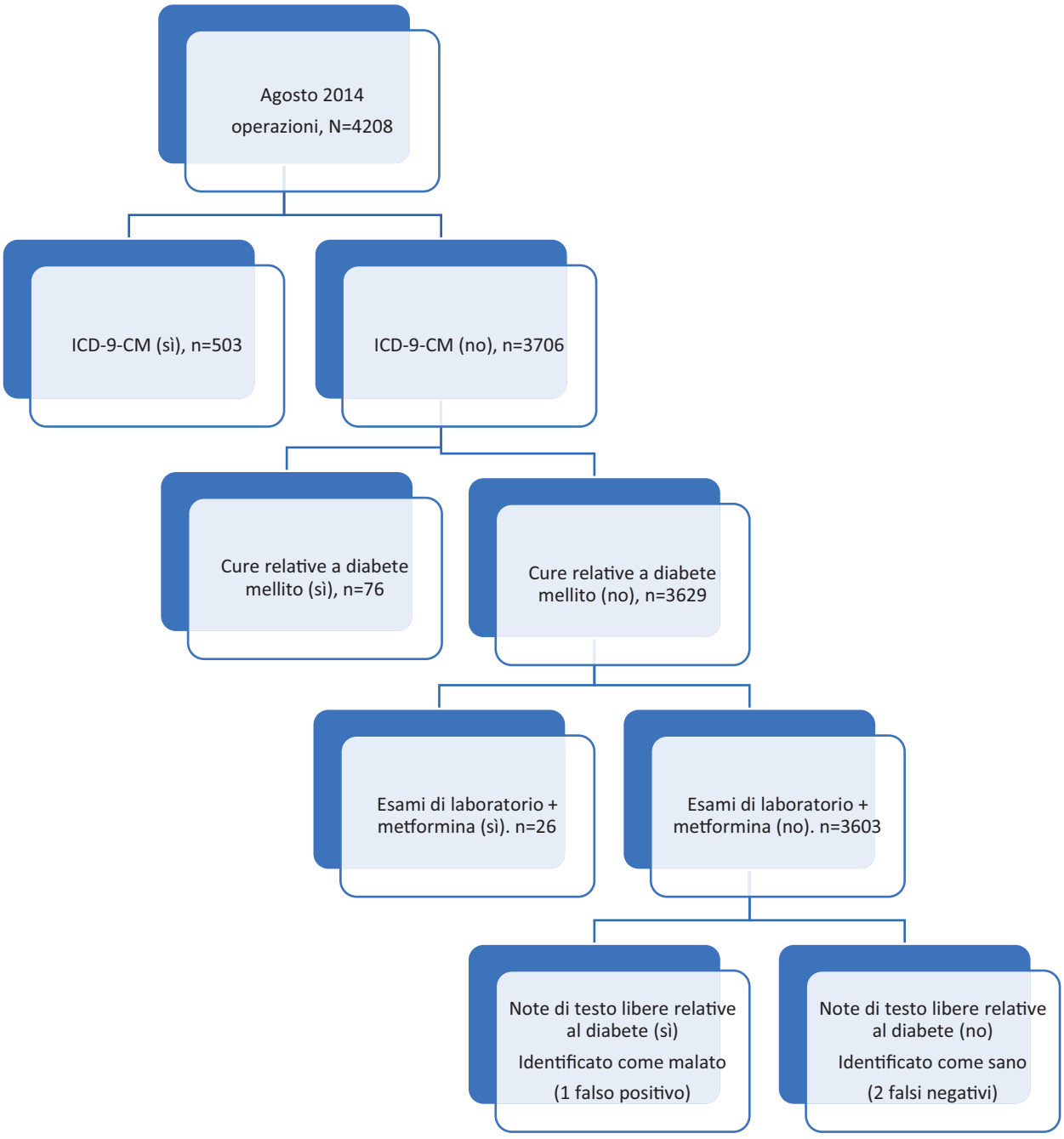

Figura 3.5 Dati relativi alla classificazione effettuata dall'algoritmo su un campione di 4208 pazienti.

Da notare che le regole implementate nell'algoritmo di fenotipizzazione DM proposto non si escludono a vicenda. Diciannove pazienti sono stati identificati come affetti da DM a causa della coesistenza di metformina e un valore di laboratorio anormale del DM e soddisfacevano anche i criteri per il DM sulla base della presenza di una parola chiave correlata al DM nelle loro note cliniche. Uno dei principali punti di forza di questa indagine è la sua inclusione di informazioni provenienti da più domini rilevanti per il DM. Ciò includeva codici diagnostici ICD-9-CM, informazioni sui farmaci, risultati di laboratorio e ricerche per parole chiave che estraggono testo non strutturato contenuto nelle note cliniche.

Oltre a identificare vari stati di affermazione, la negazione ha avuto un ruolo notevole nell'identificare i pazienti che non avevano DM e nel migliorare l'accuratezza dell'algoritmo.
A differenza delle dichiarazioni di conferma, le negazioni sono comunemente indicate con termini come no , NA e unknown, quindi inizialmente si è iniziato con questi 3 termini di base e poi perfezionato in modo iterativo il sistema aggiungendo più clausole di negazione, come fratelli, uno dei quali ha il tipo di diabete, insignificante per il diabete mellito, non ha una storia di diabete, screening del diabete: glicemia a digiuno = NA e screening del diabete: sconosciuto se mai verificato e molti altri durante il processo di validazione iterativa.
Questa combinazione ha attribuito 79 casi aggiuntivi identificati dall'algoritmo, risultando in prestazioni superiori a quelle di altri approcci che non includevano dati non strutturati. Pertanto, ne consegue che un tale sistema di NLP di analisi del testo che utilizza semplici parole chiave per negare o convalidare la presenza di DM sarà un sistema altamente portatile che può essere implementato in altri ambienti EHR. Tutto questo ha anche qualche punto negativo da discutere: ovviamente questo algoritmo è stato sviluppato come caso specifico di ricerca, molte istituzioni sanitarie potrebbero non avere tutti gli elementi e i dati proposti necessari al successo di questa strategia, sono necessarie ulteriori validazioni in altri contesti sanitari per comprendere la generalizzabilità del metodo.
Tuttavia, l'efficacia nei risultati dimostra come queste tecniche combinate che sfruttano dati non strutturati tramite il NLP siano la giusta strada da seguire per avere un sistema sempre più organizzato e prestazionale nell'individuare casi di malattie come il DM, potendo sviluppare questa tecnologia in moltissimi ambiti della sanità.

A sostegno di ciò è presente in letteratura un altro studio relativo all'identificazione del diabete dall'EHR utilizzando tecniche di NLP, sfruttando un algoritmo di ricerca molto simile a quello appena descritto ma che comunque presenta qualche differenza da menzionare.

Lo studio in questione - condotto da Le Zheng e colleghi [13] – prende in considerazione una grandissima mole di dati, evidenziando come algoritmi testati per un piccolo campione della popolazione possano anche essere adattati su scale più ampie mantenendo la loro efficacia. I dati presi in esame derivano da cartelle cliniche elettroniche che comprendono circa il 95% della popolazione dello stato del Maine (Stati Uniti), contando su informazioni provenienti da 35 ospedali, 34 centri sanitari qualificati e più di 400 ambulatori, per un totale di oltre due milioni di pazienti nel periodo che va dal 1° luglio 2012 al 30 giugno 2014.

Le note cliniche derivano da più di cento diversi tipi di referti, inclusi anamnesi, referti fisici, riepiloghi delle dimissioni e referti di emergenza.

Come per l'algoritmo sviluppato da Mayo Clinic anche in questo caso il primo passo è quello di eliminare le informazioni fuorvianti come la negazione del diabete: non sono da considerare le frasi in cui il paziente nega di avere il diabete o si parla della sua storia familiare, per non correre il rischio di creare dei falsi positivi. Un secondo passaggio è fondamentale per determinare i fattori di rischio di DM riconosciuti e i termini di NLP nei dati strutturati.

I dati clinici nella base di conoscenza del sistema relativi al DM sono stati derivati dai codici ICD-9-CM, dalla terminologia SNOMED CT, e titoli restituiti dalla query di "diabete" applicata ad un vocabolario controllato come Medical Subject Headings. Questi termini sono stati ulteriormente tokenizzati, combinati e filtrati per ottenere altri token più semplici. Su 742 termini individuati, 72 sono risultati significativamente associati alla diagnosi di DM. Sono stati identificati anche 22 farmaci – su 36 analizzati – come associati alla diagnosi di DM.

Sono stati creati set di dati gold standard per lo sviluppo del modello e per definire il punto di cutoff, ovvero stabilire un confine che analizzando i dati riesce ad effettuare una valutazione per distinguere i positivi dai negativi. I dati gold standard per allenare l'algoritmo sono stati estratti da un campione di 200 pazienti affetti da DM e 1000 pazienti senza diagnosi di DM.

Nel modello finale sono state mantenute in totale 100 caratteristiche discriminanti del DM, inclusi dati demografici (n=2), fattori di rischio (n=5), storia clinica (n=1), farmaci (n=20) e dati clinici estratti dalla PNL termini (n=72). Le prime 30 sono rappresentate in figura 3.6.

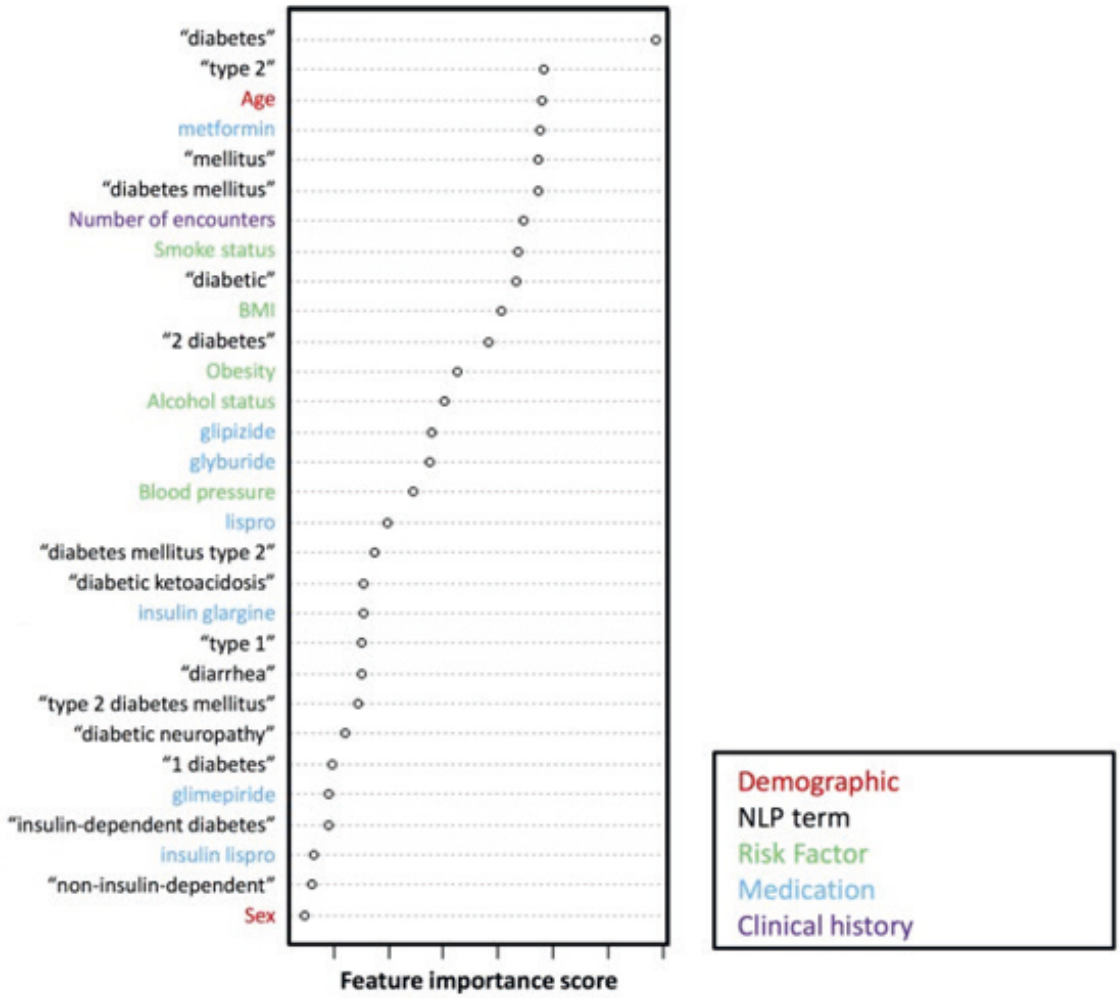

Figura 3.6 Prime 30 caratteristiche discriminanti del DM utilizzate per l'algoritmo.

Questo algoritmo di codifica mantiene più o meno le stesse caratteristiche di quello sviluppato da Mayo Clinics ed anche i risultati sono in linea, dimostrando l'incredibile efficacia di questi sistemi nella valutazione e nella classificazione di dati selezionati da linguaggio naturale.

Come per altri sistemi e altri utilizzi, l'algoritmo ha aumentato le prestazioni dei sistemi tradizionali – basati solo sull'analisi di dati strutturati – incrementando il numero di diagnosi di DM di un 8.97%. Il momento in cui la prima diagnosi di DM di un paziente è stata identificata dai codici ICD è stato valutato e confrontato con il momento in cui il DM è stato identificato dall'algoritmo. Il 30.46% dei pazienti sono stati identificati dal software basato sulla NLP prima che un codice ICD DM fosse annotato nella cartella clinica (differenza di tempo media = 48 giorni).

In particolare, il 19,86% dei pazienti è stato identificato dal caso di NLP 3 mesi o più prima di essere identificato da un codice DM ICD. Anche in questo caso numerosi pazienti con DM erroneamente classificati come sani dalla ricerca effettuata sui codici sono stati invece identificati come positivi sfruttando il NLP, risultando in un'analisi più complessa e veritiera dell'effettivo stato di salute dei pazienti.
Una ragione ulteriore del perché ci siano certe discordanze tra codici e malattia effettiva potrebbe stare nel fatto che un paziente sia stato ricoverato di emergenza in ospedale per problematiche cliniche più pericolose per la vita, trascurando le informazioni relative a un caso di diabete.

Esistono numerosi algoritmi differenti, come quello di Asma Ahmed Abokhzam e colleghi [14], basato sulle foreste casuali. Le foreste casuali sono un insieme di alberi decisionali singoli che si basano su caratteristiche casuali, è come se ognuno avesse un suo punto di vista e formuli una sua ipotesi per prevedere un risultato. Unendo tutte le previsioni si ottiene un risultato più calibrato e preciso.

Nonostante queste differenze, però, tutti si basano sullo stesso principio: quello di effettuare prima una preselezione sui dati, eliminando duplicati ed escludendo informazioni non rilevanti che potrebbero portare errori; successivamente - dopo aver addestrato l'algoritmo su vocabolari medici - i dati non strutturati vengono immessi e l'accuratezza e la precisione di questi sistemi nel rilevare i pazienti positivi supera il 90/95%. Il punto di forza che però accomuna tutti questi algoritmi è sicuramente quello di avere un quadro più completo dei pazienti a rischio, effettuando diagnosi più precise che convergono in risultati più affidabili.
Per secondo è evidente una riduzione notevole del tempo impiegato per analizzare e ricercare dati di pazienti, nonché di effettuare una diagnosi anticipata fondamentale per intraprendere tempestivamente programmi di cura e interventi sullo stile di vita che hanno un impatto a lungo termine per ritardare la progressione e prevenire le complicanze del diabete.
Dall'altro lato - come per altri algoritmi di questo tipo – sono presenti inevitabili errori di classificazione. Qualche caso 'borderline' deriva da una mancanza di informazioni relative a farmaci utilizzati, stile di vita, fattori di rischio.
Le principali limitazioni, infatti, derivano dalla diagnosi effettuata dal medico e dalla qualità ed esaustività della documentazione delle note cliniche.

3.4 Rilevamento dell'ipoglicemia dalle EHR nei pazienti con diabete di tipo 2

Una componente critica e ben riconosciuta della gestione del diabete è il rischio di ipoglicemia; gli episodi di ipoglicemia grave (SH) definiti come richiedenti il ricovero o la visita al pronto soccorso in pazienti con diabete di tipo 2, possono essere identificati attraverso cartelle cliniche elettroniche strutturate o dati amministrativi. D'altro canto, l'ipoglicemia non grave (NSH) che non richiede assistenza per il recupero, viene spesso segnalata durante le visite ambulatoriali ma è effettivamente rintracciabile solo se viene inserito un codice di diagnosi di ipoglicemia.

Il NLP anche in questo caso gioca un ruolo fondamentale, dando la possibilità di estrarre i dati di testo non strutturati in modo da migliorare il rilevamento dell'ipoglicemia.

Uno studio condotto da Inc. ADM-H. si concentra a descrivere un modello per l'estrazione di dati di NSH in pazienti con diabete di tipo 2 utilizzando sia i codici diagnostici che il NLP per creare un modello predittivo per gli eventi di SH. [15]

In questo studio è stata utilizzata una versione leggermente modificata di un algoritmo molto simile a quello di figura 3.4, per identificare i pazienti del Cleveland Clinic Health System con diabete di tipo 2 nel periodo 2005-2017.

Un punto fondamentale per lo sviluppo dell'algoritmo è la conoscenza del programma cTAKS.

Quest'ultimo è stato utilizzato per scomporre le frasi in altre frasi per identificare i concetti del sistema di linguaggio medico unificato correlati all'ipoglicemia. Le espressioni regolari sono state scritte per classificare la polarità (evento o nessun evento) delle frasi in base al pattern matching. Usando l'embedding di parole, le frasi rimanenti sono state classificate con un pattern matching aggiuntivo. Il pattern matching è l'azione di controllo della presenza di un certo motivo (pattern) all'interno di un oggetto composito. In concreto l'espressione si può riferire al riconoscimento di pattern all'interno di una stringa di caratteri, in modo da comprendere e classificare le parole.

Il cTAKS è un sistema modulare di componenti che combinano tecniche basate su regole di apprendimento automatico volte all'estrazione di informazioni dalla narrativa clinica. I set di dati "gold standard" per le etichette linguistiche e i concetti clinici vengono creati su contenuti che sono un sottoinsieme di note cliniche della Mayo Clinic EHR. Le metriche di valutazione standard vengono utilizzate per misurare la qualità dei "gold standard" e le prestazioni di cTAKS.

L’attuale versione open source è composta dai seguenti componenti:

• Rilevatore di confine di frasi

• Tokenizzatore

• Normalizzatore

• Tagger di parte del discorso (POS).

• Analizzatore superficiale

• Annotatore di riconoscimento dell’entità con nome (NER), inclusi annotatori di stato e negazione.

Il rilevatore di limiti di frasi estende lo strumento di rilevamento di frasi di OpenNLP (un sistema di NLP accessibile e utilizzabile da tutti). In sostanza prevede se un punto, un punto interrogativo o un punto esclamativo siano la fine di una frase. Il tokenizer cTAKS è costituito da due sottocomponenti: il primo divide il flusso di testo interno della frase sullo spazio e sulla punteggiatura; il secondo, il tokenizzatore dipendente dal contesto, unisce i token per creare token di data, frazione, misura, titolo della persona, intervallo, numero romano e tempo applicando regole per ciascuno di questi tipi. Il normalizzatore cTAKS è un wrapper (strumento informatico che provvede all’estrazione delle informazioni da documenti HTML e nella rappresentazione delle informazioni estratte in formato XML) attorno a un componente degli strumenti lessicali SPECIALIST chiamato “norma”, che fornisce una rappresentazione per ogni parola nel testo di input che è normalizzata rispetto a una serie di proprietà lessicali, tra cui “caso alfabetico, inflessione, varianti di ortografia, punteggiatura, parole stop, segni, simboli e legature”.

La normalizzazione consente di mappare più menzioni della stessa parola che non hanno le stesse rappresentazioni di stringa nei dati di input. Il tagger cTAKS POS e il parser superficiale sono wrapper attorno ai moduli di OpenNLP per queste attività, che comprendono in particolare un’analisi automatica della struttura morfologica delle parole sulla base di grammatica e lessico di una lingua data. Il componente cTAKS NER implementa un algoritmo di ricerca nel dizionario indipendente dalla terminologia all’interno di una finestra di ricerca per la frase nominale. Attraverso la ricerca nel dizionario, ogni entità denominata viene mappata su un concetto dalla terminologia. È utilizzato un dizionario che include i concetti SNOMED (Systematized Nomenclature Of Medicine) guidati da ampie consultazioni con ricercatori clinici e professionisti.

Ogni termine del dizionario appartiene a uno dei seguenti tipi semantici come definito in: disturbi/malattie con un gruppo separato per segni/sintomi, procedure, anatomia e farmaci, quest'ultimo include termini dall'Orange Book che hanno un codice RxNORM. Il componente NER non risolve le ambiguità risultanti dall'identificazione di più termini nello stesso intervallo di testo.
L'annotatore di negazione implementa l'algoritmo NegEx, che è un approccio basato su pattern per trovare parole e frasi che indicano la negazione vicino a menzioni di entità nominate. L'annotatore di stato utilizza un approccio simile per trovare parole e frasi rilevanti che indicano lo stato di un'entità denominata.
Ogni entità denominata scoperta appartiene a uno dei tipi semantici del dizionario e ha attributi per l'intervallo di testo associato all'entità denominata, il codice terminologico a cui l'entità denominata è mappata ('concept ' attributo), se l'entità nominata è negata (attributo 'negazione') e lo stato associato all'entità nominata con un valore di "corrente, storia di, storia familiare di, possibile" (attributo 'status'). Qualsiasi evento futuro è considerato ipotetico; quindi, il valore dello stato sarà impostato su 'possibile'. [16]

Il modello di previsione dei rischi potenziali per SH è stato creato utilizzando le seguenti variabili: sesso, etnia, reddito medio, storia di comorbidità e variabili dipendenti dal tempo, tra cui età, tipo di assicurazione, emoglobina glicosilata e farmaci per il diabete. Dopo la revisione delle cartelle cliniche, 1.111 dei 1.200 eventi selezionati casualmente classificati come NSH dal programma di NLP sono stati confermati (93% valore predittivo positivo).
Dal 2005 al 2017, 10.205 eventi NSH sono stati acquisiti da codici e 14.763 eventi da NLP, con una sovrapposizione di soli 5 eventi. Tra 204.517 pazienti senza codici per NSH, l'evidenza di NSH è stata trovata in 7.035 (3,4%) utilizzando NLP.
Nonostante questo, i risultati del NLP per identificare SH erano sostanzialmente gli stessi che risultavano dall'utilizzo dei soli codici ICD, da questo si evince come il NLP in questo specifico caso non offre prestazioni aggiuntive rispetto a quelle già precedentemente a disposizione.

Il risultato dello studio fa emergere come l'applicazione del NLP nelle note di EHR ha migliorato il rilevamento di NSH e che quest'ultimo è un predittore significativo per l'SH.
Il rilevamento aumenta con il NLP anche perché il NSH può essere segnalato dai pazienti, mentre gli operatori sanitari potrebbero non inserire un codice di diagnosi di ipoglicemia se non sia

effettivamente classificata come grave, evidenziando il beneficio del NLP.

Tuttavia, la previsione di SH utilizzando codici di diagnosi non è migliorata con l'aggiunta del NLP.
Tutto questo, però, può portare a grandi benefici per il paziente a cui viene diagnosticato il NSH, potendo essere monitorato più facilmente, tenuto sotto osservazione e anche preventivamente indirizzato in un percorso di cura e terapia.

A confermare i vantaggi portati dal NLP per la scoperta di ipoglicemia nei pazienti è una revisione di Yaguang Zheng e colleghi [17] pubblicata ad aprile 2022.
Nel loro lavoro di rassegna hanno confrontato otto articoli e i principali risultati, evidenziando gli stessi aspetti positivi già riportati in precedenza fornendo però dati aggiuntivi.
Per esempio – secondo la review - la combinazione di codici NLP e ICD-9 o ICD-10 ha aumentato significativamente l'identificazione di eventi ipoglicemici rispetto ai singoli metodi; i tassi di prevalenza dell'ipoglicemia erano del 12,4% per i codici di classificazione internazionale delle malattie, del 25,1% per un algoritmo NLP e del 32,2% per algoritmi combinati.
Questo fornisce una prova di supporto al lavoro riportato precedentemente di Misra-Hebert e colleghi, dimostrando come il NLP sia di fondamentale importanza per aiutare la diagnosi di eventuali casi di ipoglicemia, soprattutto se affiancato ai codici ICD.
Tutti gli strumenti di NLP non sono da considerarsi come sostituti dei metodi tradizionali, al contrario è stato dimostrato che è proprio il lavoro combinato delle tecnologie che porta i risultati migliori, per questo qualunque sviluppo non può che portare miglioramenti, dal momento che non modifica ciò che è già presente e collaudato, ma gli si affianca per incrementare le prestazioni e ridurre i tempi e l'impiego di energie.

CONCLUSIONI

Il NLP è uno degli strumenti informatici più utili per una quantità enorme di funzioni, in ogni ambito della scienza. Questa tecnologia è in continua espansione e negli ultimi anni si sta evolvendo in maniera esponenziale, attirando sempre più l'attenzione da parte di studiosi e sviluppatori.

Se nel primo decennio degli anni 2000 il NLP offriva poche potenzialità in ambito biomedico, negli ultimi anni gli studi e le ricerche in questo campo si stanno moltiplicando, evidenziando la sua efficacia e la sua incredibile utilità nell'affiancare il personale sanitario nei processi di scoperta di eventuali malattie, nella loro diagnosi, classificazione, nella cura e nei percorsi di monitoraggio dei pazienti. Scopo di questa tesi era quello di fornire una visione generale del natural language processing e riportare i suoi utilizzi più rilevanti in ambito medico in relazione a cartelle cliniche elettroniche e diabete.

Gli studi analizzati dimostrano questo e di come con lo sviluppo di algoritmi e software informatici si possano risolvere numerosi problemi e soprattutto si possono ridurre incredibilmente i tempi e l'impiego di risorse per compiti che un calcolatore riesce a portare a termine in poco tempo e con maggiore precisione. Per esempio, la ricerca di parole chiave in un testo o di articoli in un sistema di ricerca viene effettuata autonomamente e con grande precisione dal sistema, in grado di comprendere cosa è scritto e se può essere utile e inerente alla ricerca effettuata. I risultati evidenziano la potenzialità e l'efficacia di questi algoritmi, sia per la ricerca, sia per la diagnosi, sia per il supporto decisionale che viene fornito agli operatori sanitari.

Gran parte degli studi riguardanti il diabete si concentrano principalmente sulla sua identificazione. In molti articoli presenti in letteratura, oltretutto, sono presenti modelli di sviluppo per nuovi algoritmi che però non hanno poi un uso pratico testato realmente, ma rimangono come una teorizzazione. Questo potrebbe avere diverse possibili spiegazioni. Una è che gli strumenti di elaborazione del linguaggio naturale che sono stati sviluppati non sono stati resi disponibili ai potenziali utenti. Un altro è che gli stessi strumenti che sono stati sviluppati non erano effettivamente quelli di cui gli utenti avevano bisogno. In ogni caso, è necessaria una maggiore collaborazione e cooperazione tra sviluppatori e utenti per garantire che le risorse dedicate alla progettazione e alla valutazione di questa sofisticata tecnologia siano utilizzate nel modo più efficace a beneficio dei pazienti e del pubblico in generale.

Tuttavia, il settore del NLP in relazione al diabete negli ultimi anni sta sempre più prendendo piede

anche per quanto riguarda i percorsi di cura e le terapie da effettuare per migliorare la qualità della vita dei malati. L'uso del NLP per analizzare le note cliniche ha migliorato l'acquisizione di eventi ipoglicemici non documentati o persi utilizzando la classificazione internazionale delle malattie (ICD-9 e ICD-10) e test di laboratorio.

Il trattamento del diabete spesso implica discussioni estese tra i pazienti e gli operatori sanitari che coinvolgono molteplici aspetti del processo assistenziale, compresi i cambiamenti nello stile di vita, gli obiettivi della cura, le reazioni avverse ai farmaci, le barriere alle cure. Queste discussioni tendono ad essere rappresentate in minima parte in dati strutturati/discreti e di conseguenza sono impossibili da studiare o monitorare su scala di popolazione senza una soluzione di elaborazione del linguaggio naturale. Tutto questo può essere superato con metodologie e tecniche in grado di estrarre conoscenza da dati a disposizione per fare predizioni su dati o eventi nel futuro. L'elemento chiave di queste tecniche consiste nella capacità di apprendere modelli partendo dai dati a disposizione. A loro volta, tali modelli sono in grado di operare sui nuovi dati fornendo diverse capacità.

Il NLP ha il potenziale per avere un impatto significativo sulla misurazione e sugli interventi per migliorare la qualità della cura del diabete, aiutando a definire le popolazioni a rischio che trarrebbero il massimo beneficio da interventi mirati e cure specializzate. È un'area di ricerca molto attiva e negli ultimi anni sta significativamente migliorando la sua precisione consentendo ai modelli di incorporare informazioni non presenti in altre fonti di dati, segnando un vero e proprio punto di svolta nel settore.

BIBLIOGRAFIA

1) Pontieri G, Russo M.A., Frati L. Patologia generale e fisiopatologia generali. Cap. 54: Il diabete mellito. 2015

2) Hirschberg J., Manning C.D. Advances in natural language processing. Agosto 2019

3) Adriens G., Hahn U., Parallel Natural Language Processing

4) Eisenstein J. Introduction to natural language processing. 2019

5) Gallagher S., Rafferty A., Wu A., Story of natural language processing. 2004

6) Nadkarni P.M., Ohno-Machado L., Chapman W.W., Natural language processing: an introduction. Journal of the American Medical Informatics Association, volume 18, numero 5, settembre 2011

7) Devlin J.,Chang M.W., Lee K., Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018

8) Shin J., Lee Y., Jung K. Effective Sentence Scoring Method Using BERT for Speech Recognition 2019

9) Wang Y., Liu S., Afzal N., Rastegar-Mojarad M., Wang L, Shen F., Kingsbury P., Liu H., A comparison of word embeddings for the biomedical natural language processing. Journal of Biomedical Informatics. Novembre 2018

10) Demner-Fushman D., Chapman W.W., McDonald C.J. What can natural language processing do for clinical decision support? Journal of Biomedical Informatics. Volume 42, 2009

11) Tseng E., Schwartz JL., Rouhizadeh M., Maruthur NM. Analysis of Primary Care Provider Electronic Health Record Notes for Discussions of Prediabetes Using Natural Language Processing Methods. J Gen Intern Med. 19 gennaio 2021

12) Upadhyaya SG., Murphree DH Jr., Ngufor CG., Knight AM., Cronk DJ., Cima RR., Curry TB., Pathak J., Carter RE., Kor DJ. Automated Diabetes Case Identification Using Electronic Health Record Data at a Tertiary Care Facility. Mayo Clin Proc Innov Qual Outcomes. 28 aprile 2017

13) Zheng L., Wang Y., Hao S., Shin AY., Jin B., Ngo AD., Jackson-Browne MS., Feller DJ., Fu T., Zhang K., Zhou X., Zhu C., Dai D., Yu Y., Zheng G., Li YM., McElhinney DB., Culver DS., Alfreds ST., Stearns F., Sylvester KG., Widen E., Ling XB. Web-based Real-Time Case Finding for the Population Health Management of Patients With Diabetes Mellitus: A Prospective

Validation of the Natural Language Processing-Based Algorithm With Statewide Electronic Medical Records. JMIR Med Inform. 11 novembre 2016

14) Abokhzam, A.A., Gupta, N.K. & Bose, D.K. Efficient diabetes mellitus prediction with grid based random forest classifier in association with natural language processing. 25 marzo 2021

15) Misra-Hebert AD., Milinovich A., Zajichek A., Ji X., Hobbs TD., Weng W., Petraro P., Kong SX., Mocarski M., Ganguly R., Bauman JM., Pantalone KM., Zimmerman RS., Kattan MW. Natural Language Processing Improves Detection of Nonsevere Hypoglycemia in Medical Records Versus Coding Alone in Patients With Type 2 Diabetes but Does Not Improve Prediction of Severe Hypoglycemia Events: An Analysis Using the Electronic Medical Record in a Large Health System. Diabetes Care. Agosto 2020

16) Savova GK., Masanz JJ., Ogren PV., Zheng J., Sohn S., Kipper-Schuler KC., Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. settembre-ottobre 2010

17) Zheng, Yaguang, Dickson, Vaughan V., Blecker, Saul, Ng, Jason M., Rice, Campbell B., Melkus, Deramo G., Shenkar, Liat, Mortejo, Marie Claire R., Johnson, Stephen B, Identifying Patients with Hypoglycemia Using Natural Language Processing: Systematic Literature Review. 1° aprile 2022

RINGRAZIAMENTI

Un ringraziamento alla mia professoressa, che mi ha seguito e che mi ha fatto appassionare alla sua materia, proponendomi una tesi interessante e stimolante dal punto di vista didattico.

Un ringraziamento speciale ai miei genitori, agli amici, ai colleghi, in particolare a chi mi è sempre stato vicino e a chi mi ha supportato lungo il percorso, spronandomi a dare il massimo e facendo in modo di farmi vivere anni di gioia e soddisfazioni, accompagnate da tranquillità e felicità.