

# Sentiment Analysis su cinque prodotti di Angelini Pharma

## Introduzione

Per la realizzazione di questo progetto ho scelto come casa farmaceutica da analizzare [Angelini Pharma](#).

Lo scopo del progetto è quello di fare **Sentiment Analysis** sulle **descrizioni commerciali** di cinque prodotti della casa farmaceutica in questione, estraendole autonomamente tramite scraping dal sito della casa farmaceutica stessa. Questi prodotti sono rispettivamente **Acutil Adulti 55+**, **Energia**, **ThermaCare**, **Amuchina** e **Tachipirina**. Affiancare ai primi quattro prodotti con descrizioni commerciali, quindi mirate alla vendita, il **bugiardino** della Tachipirina, serve ad evidenziare e sottolineare il sentiment tendenzialmente positivo delle descrizioni commerciali rispetto al foglietto illustrativo di un prodotto farmaceutico tradizionale non a scopo commerciale, quindi tendente ad essere neutro.

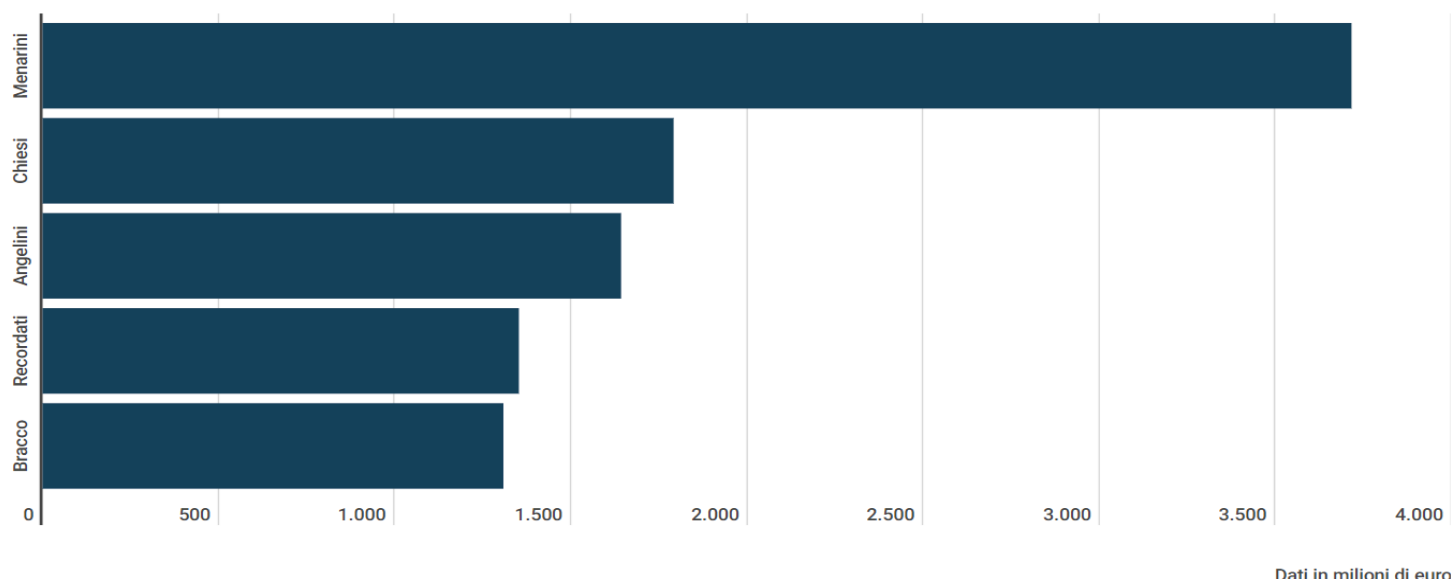
**Angelini Pharma** si distingue come uno dei principali protagonisti nel panorama farmaceutico italiano ed europeo, grazie alla sua posizione di vertice nel mercato, sostenuta da marchi iconici come **Tachipirina** e **Amuchina**.

Parte di [Angelini Industries](#), un'azienda con una solida eredità familiare, Angelini Pharma vanta oltre un secolo di esperienza nel settore e conta oggi più di 1500 dipendenti in Italia, con un significativo contributo al fatturato globale. La sua ricerca innovativa si concentra su malattie del sistema nervoso, dolore e infiammazione, con un'impronta distintiva nel settore pediatrico.

Con impianti produttivi in Italia e all'estero, Angelini Pharma si distingue per la produzione di integratori alimentari, prodotti per l'igiene e la cura della persona.

Ho scelto Angelini perché si posiziona alla terza posizione nella top 5 delle aziende farmaceutiche italiane per fatturato, come mostra questo grafico:

Prime 5 aziende farmaceutiche in Italia per fatturato



## Organizzazione del lavoro:

Per affrontare il problema in questione, ho deciso di suddividere il lavoro in cinque fasi, esclusa l'introduzione:

1. **DataCard e postilla;**
2. **Analisi descrittiva del Dataset;**
3. **Sentiment Analysis;**
4. **Analisi e commento dei risultati;**
5. **Conclusione.**

## 1 DataCard

Per descrivere dettagliatamente il Dataset, il suo contenuto ed il **Preprocessing** ho creato quella che è la DataCard.

Tuttavia nei paragrafi successivi alla DataCard si troverà una breve **postilla** riguardo la scelta dei prodotti effettuata, con degli studi a supporto .

### Farmaceutica Sentiment Analysis

Il Dataset contiene le **descrizioni commerciali di 5 prodotti farmaceutici** scaricati dal sito di Angelini Pharma.

Ho ottenuto le descrizioni dei prodotti farmaceutici tramite Scraping usando Nanonets, questi sono: **Acutil Adulti 55+, Energya, ThermaCare, Amuchina e Tachipirina.**

	Per Tachipirina, ho utilizzato il bugiardino PDF dall'Agenzia Italiana del Farmaco, convertendolo in testo. Successivamente, ho pulito il Dataset rimuovendo le <b>stopwords</b> e le <b>regex</b> , per poi procedere con l'analisi descrittiva, e la Sentiment Analysis.
<b>DATASET LINK</b>  <a href="#">Link Dropbox per scaricare il Dataset in formato .csv o .txt</a>	<b>DATA CARD AUTHOR(S)</b>  <b>Daniele Mariani</b>  <a href="mailto:danyelemariani@gmail.com">danyelemariani@gmail.com</a>  <a href="https://github.com/Mariashish">https://github.com/Mariashish</a>

Dataset Overview	
DATA SUBJECT(S)	CONTENT DESCRIPTION
Descrizioni di prodotti farmaceutici: Acutil 55+ Energya ThermaCare Amuchina Tachipirina	<p>Il Dataset contiene le <b>descrizioni commerciali di 5 prodotti farmaceutici dell'azienda Angelini Pharma</b>.</p> <p>Ogni prodotto può essere trovato nella sezione "Catalogo Prodotti" del sito ufficiale di Angelini Pharma.</p> <p><a href="https://www.angelinipharma.it/dalla-ricerca-ai-farmaci/prodotti/">https://www.angelinipharma.it/dalla-ricerca-ai-farmaci/prodotti/</a></p> <p>Per quanto riguarda Tachipirina, il bugiardino è stato scaricato dal sito dell'AIFA.</p> <p><a href="https://farmaci.agenziafarmaco.gov.it/bancadatifarmaci/farmaco?farmaco=012745">https://farmaci.agenziafarmaco.gov.it/bancadatifarmaci/farmaco?farmaco=012745</a></p>
Sensitivity of Data	
SENSITIVITY TYPE(S)	SECURITY AND PRIVACY HANDLING

Il tipo di dati scaricati sono dati commerciali, disponibili pubblicamente in rete.	<p>I dati scaricati sono resi pubblici sia sul sito di Angelini Pharma che sul sito dell'AIFA.</p> <p><a href="https://www.angelinipharma.it/">https://www.angelinipharma.it/</a></p> <p><a href="https://www.aifa.gov.it/">https://www.aifa.gov.it/</a></p>
RISK TYPE(S)	RISK(S) AND MITIGATION(S)
I rischi non sono direttamente legati all'autore del Dataset dal momento che i dati scaricati sono resi pubblici in rete, tanto più a un utilizzatore terzo che potrebbe usare questi dati in modo malevolo.	Il Dataset potrebbe essere utilizzato da utenti di terze parti per fini non riconducibili o condivisi né dall'autore del Dataset né tantomeno dalla casa farmaceutica Angelini o dall'AIFA.
Dataset Version and Maintenance	
MAINTENANCE STATUS	MAINTENANCE PLAN
Aggiornato all'ultima versione	Il Dataset è attualmente <b>aggiornato</b> e verrà <b>monitorato</b> e aggiornato periodicamente.
	EXPECTED CHANGE(S)
Modificato e aggiornato	Nel caso in cui le descrizioni dei prodotti dovessero essere modificate, il Dataset verrà aggiornato quanto prima possibile.

Example of Data Points	
PRIMARY DATA MODALITY	DATA FIELDS

Dati di tipo testuale, scaricati in formato .txt per poi essere inseriti nel .csv.	<table> <tr> <th>Field Name</th><th>Field Value</th></tr> <tr> <td>Nome Prodotto</td><td>Descrizione Prodotto</td></tr> </table>	Field Name	Field Value	Nome Prodotto	Descrizione Prodotto
Field Name	Field Value				
Nome Prodotto	Descrizione Prodotto				

## Motivations & Intentions

### Motivations

PURPOSE(S)	DOMAIN(S) OF APPLICATION	MOTIVATING FACTOR(S)
Il motivo e gli scopi del Dataset sono per <b>fini didattici</b> , quindi di ricerca e messa alla prova delle proprie competenze	<p>I <b>domini principali</b> di applicazione sono:</p> <p>Machine Learning, Text Mining, Artificial Intelligence, Web Scraping e statistica descrittiva.</p>	I fattori motivanti riguardano la messa in atto delle proprie skills in merito a ciò che si è studiato durante il corso di <b>Text Mining</b> in ambito accademico.

### Intended Use

DATASET USE(S)	SUITABLE USE CASE(S)	UNSUITABLE USE CASE(S)
L'utilizzo del Dataset da parte dell'autore è e sarà sempre a scopo didattico.	<p>Lo scopo principale è quello di portare a termine il percorso didattico dell'esame di Text Mining.</p> <p>Altri casi d'uso potrebbero essere quello di migliorare l'efficienza del marketing e del copywriting aziendale o anche immaginare e prevedere cosa i clienti provino quando vedono pubblicità su questi prodotti.</p>	Un utilizzatore terzo potrebbe utilizzare il Dataset in maniera impropria, non conforme agli utilizzi che ne farebbe l'autore del Dataset.

## Access, Retention, & Wipeout

Access		
ACCESS TYPE	DOCUMENTATION LINK(S)	PREREQUISITE(S)
OpenSource	<p>Tutti i link utili sono i seguenti:</p> <p><a href="#">Link Dataset</a></p> <p><a href="https://github.com/Mariashish">https://github.com/Mariashish</a></p> <p><a href="#">Link HuggingFace</a></p> <p><a href="#">Link MilaNLProc</a></p> <p><a href="#">Nanonets</a></p> <p><a href="#">sentix</a></p>	<p>I prerequisiti per utilizzare e comprendere il Dataset riguardano conoscenze della <b>Data Science</b>, <b>Intelligenza Artificiale</b>, della Statistica e del Machine Learning.</p>

Provenance		
Collection		
METHOD(S) USED	METHODOLOGY DETAIL(S)	SOURCE DESCRIPTION(S)
<p><a href="#">Nanonets</a></p> <p><a href="#">PyMUPDF</a></p>	<p>Per ottenere il Dataset sono state usate tecniche di <b>scraping</b>, in particolare è stato usato un tool chiamato <a href="#">Nanonets</a>, che permette di scaricare da un sito web dato in input tutto il contenuto testuale contenuto, riconoscendo anche il testo nelle immagini tramite <b>OCR</b> (Optical Character Recognition).</p> <p>L'ho fatto per ogni pagina dei prodotti e successivamente ho salvato le descrizioni dei 4 prodotti in file .txt separati, nominati <i>descrizione_n_nomeprodotto.txt</i>, per poi unirli in un unico file .csv denominato <i>descrizioni_prodotti.csv</i>.</p> <p>Ho usato questo metodo piuttosto che il classico Web Scraping con Selenium e BeautifulSoup proprio perchè la maggior parte dei prodotti avevano molte immagini, facendo in questo modo ho evitato di perdere molte informazioni.</p> <p>Per quanto riguarda Tachipirina invece ho scaricato il bugiardino in formato PDF dal sito dell'AIFA e poi tramite una libreria</p>	<p>Tutte le altre informazioni si possono trovare nei link forniti che comprendono la documentazione utile per l'utilizzo degli strumenti. Le principali <b>librerie utilizzate</b> nel progetto sono: pandas, numpy, sklearn, huggingface, tensorflow, gensim, matplotlib, transformers, collections, nltk e textblob.</p>

	Python chiamata <b>PyMUPDF</b> ho estratto il testo dal PDF in formato .txt per poi inserirlo nel csv.	
<b>Collection Criteria</b>		
<b>DATA SELECTION</b>	<b>DATA INCLUSION</b>	<b>DATA EXCLUSION</b>
Come <b>criteri di collezione</b> dati è stato scaricato tutto ciò che era di formato testuale presente nei siti dei prodotti, in modo tale da non perdere informazioni. Come descritto anche nel paragrafo precedente è stata utilizzata anche l'OCR per non perdere le informazioni contenute nelle immagini.	Nel Dataset è stato quindi incluso tutto ciò che era di <b>formato testuale</b> .	Nello scaricamento del Dataset non è stato escluso nulla.

<b>Use in ML or AI Systems</b>		
<b>DATASET USE(S)</b>	<b>NOTABLE FEATURE(S)</b>	<b>USAGE GUIDELINE(S)</b>
Il Dataset nei campi del <b>Machine Learning e dell'Intelligenza Artificiale</b> può essere utilizzato per trainare, testare e valutare un modello. Inoltre può essere utilizzato per fare <b>Sentiment Analysis ed Emotion Anaysis</b> . Inoltre altri utilizzi fanno parte del campo della statistica descrittiva e non.	Per sfruttare al meglio il Dataset bisognerebbe proprio utilizzare tecniche di ML e AI, inoltre il Dataset può essere utilizzato con modelli pre-addestrati. Per farlo ho utilizzato <a href="#">FEEL-IT</a> di <a href="#">MilaNLProc</a> (Fine-Tuning di <a href="#">UmBERTo</a> , <a href="#">BERT</a> ), presenti su <a href="#">Huggin Face</a> .	Il Dataset può essere utilizzato per qualsiasi scopo, tuttavia deve rispettare le normative e gli <b>standard etici</b> . Inoltre deve mantenere le volontà dell'autore del Dataset, espresse anche nei paragrafi precedenti.

<b>Transformations</b>
<b>Synopsis</b>

<p>Le <b>trasformazioni applicate</b> al Dataset sono molteplici e fanno parte del Preprocessing dei dati. Si parte da un controllo iniziale manuale, umano in modo tale da capire se il Dataset fosse omogeneo, conforme, senza troppi errori in fase di scaricamento dei Dati. Successivamente il Dataset è stato inserito in un Dataframe pandas per una migliore manipolazione. Una volta fatta la Tokenizzazione con <i>nltk</i>, sono state rimosse le stopwords e le espressioni regolari, in modo da avere le descrizioni pulite. Inoltre poi sono state create nuove colonne nel Dataframe per facilitare l'analisi descrittiva, come ad esempio per effettuare il conteggio delle frequenze.</p>	<p>Le nuove colonne aggiunte al Dataframe Pandas sono le seguenti:</p> <table><tr><th>Descrizioni_Pulite</th><th>Descrizioni_Tokenizzate</th></tr><tr><td>...</td><td>...</td></tr><tr><td>...</td><td>...</td></tr><tr><td>...</td><td>...</td></tr></table>	Descrizioni_Pulite	Descrizioni_Tokenizzate	...	...	...	...	...	...	<p>Controllo ortografico <b>Tokenizzazione</b> Rimozione Stopwords Rimozione espressioni regolari. Per effettuare queste procedure, ed in generale nel progetto sono state <b>utilizzate le seguenti librerie</b>: pandas, matplotlib, seaborn, transformers, huggingface, nltk, textblob, WordCloud, numpy, transformers, pipeline, gensim e PYMuPDF, Collections. Come lexicon ho utilizzato <a href="#">sentix</a>, un lessico basato sulla lingua italiana.</p>
Descrizioni_Pulite	Descrizioni_Tokenizzate									
...	...									
...	...									
...	...									

Validation Types	
METHOD(S)	DESCRIPTION(S)
<p>I <b>metodi di validazione</b> sono stati i seguenti:</p> <p>Validazione <b>manuale</b></p> <p>Validazione <b>umana</b></p> <p>Validazione <b>automatica</b> tramite gestione e controllo degli errori in Python.</p>	<p>Il Dataset è stato validato nei suoi contenuti da me medesimo per verificare la correttezza dei dati scaricati e la loro coerenza, dal momento che è stato usato uno strumento di scraping per farlo per me una volta impostato.</p> <p>Inoltre tramite vari controlli degli errori lato codice sono state effettuate ulteriori validazioni.</p>


Known Applications & Benchmarks
ML APPLICATION(S)



Il Dataset può ed è stato utilizzato per scopi di Statistica, Machine Learning e Intelligenza Artificiale. Sono stati usati anche modelli pre-addestrati come **MilaNLProc (Fine-Tuning di UmBERTo, BERT)**.

## 1.1 Postilla

I cinque prodotti che ho scelto sono rispettivamente: **Acutil Adulti 55+**, **Energia**, **ThermaCare**, **Amuchina** e **Tachipirina**. Ho scelto di inserire Tachipirina pur non avendo una descrizione “commerciale” fatta dalla casa farmaceutica stessa, perchè è stata proprio Angelini Pharma ad inventarla nel 1958 ed è stato anche il farmaco più venduto negli ultimi anni, come riportato da questo grafico<sup>1</sup> da cui ho estratto la Top 5:

			
TOP 50 FARMACI	VALORE		VARIAZIONE
	2022	2023	2022-2023
TOTALE TOP50 FARMACI	3.503.318.569,67 €	3.474.500.838,47 €	-0,82% ▼
TACHIPIRINA	314.087.576,33 €	308.970.311,36 €	-1,63 % ▼
AUGMENTIN	132.101.673,36 €	148.927.808,59 €	12,74 % ▲
DIBASE	170.147.105,53 €	145.629.300,04 €	-14,41 % ▼
FOSTER	139.681.132,69 €	118.920.010,91 €	-14,86 % ▼
ENTEROGERMINA	98.469.638,71 €	104.569.241,09 €	6,19 % ▲
XANAX	88.183.671,87 €	93.407.827,21 €	5,92 % ▲

Inoltre ho scelto Tachipirina proprio per questo motivo, ovvero che la sua descrizione, essendo presa da un foglietto illustrativo, sarà neutra rispetto alle descrizioni di prodotti commerciali, che devono cercare di vendere il loro prodotto. Questo mi permette di evidenziare e sottolineare questo aspetto.

<sup>1</sup> [Link allo studio](#)

Infine l'ho scelta perchè erano solo 4 i prodotti scaricabili dal sito della casa farmaceutica stessa. La Tachipirina essendo una medicina e non avendo una descrizione commerciale fatta dalla casa farmaceutica stessa può essere scaricata solo tramite la banca dati Agenzia Italiana del Farmaco<sup>2</sup> Per questo, ho scaricato il bugiardino ufficiale della Tachipirina in formato PDF, convertendolo in txt tramite la libreria **PyMuPDF (fitz)** per poi unirlo al csv *descrizioni\_prodotti.csv*.

Per questo prodotto le analisi come WordCloud e bigrammi l'ho fatta separatamente, dal momento che il bugiardino della Tachipirina, rispetto alle descrizioni commerciali degli altri prodotti, è molto più lungo, contenendo quindi una quantità di parole molto più vasta, avrebbe creato troppo **bias** nell'analisi insieme agli altri corpus di testo.

## 2. Analisi descrittiva

In primis fornisco una breve descrizione di ogni prodotto farmaceutico, per avere una visione d'insieme riguardo ogni prodotto che andremo ad analizzare:

**Acutil Adulti 55+<sup>3</sup>:** Integratore formulato per le esigenze degli adulti oltre i 55 anni, supporta la salute mentale e cognitiva. Ho scelto di analizzare la formula Adulti 55+ piuttosto che quella Fosforo (per studenti) o Donne perchè la ritengo più generale, dato che include uomini, donne ed anziani, e maggiormente confrontabile con gli altri prodotti.

**Energya<sup>4</sup>:** Bevanda energetica (integratore) formulata per migliorare la resistenza fisica e mentale, fornendo un rapido boost di energia.

**ThermaCare<sup>5</sup>:** Prodotto che permette la terapia termica avanzata per il sollievo dal dolore muscolare e articolare, utilizzando calore terapeutico costante tramite fasce applicabili sulle zone doloranti.

**Amuchina<sup>6</sup>:** Disinfettante ampiamente utilizzato per la disinfezione di superfici e oggetti, ma anche e soprattutto per le mani. offrendo protezione contro germi e batteri. L'ho scelto perchè questo prodotto è stato fondamentale ma soprattutto vendutissimo durante la pandemia del Covid-19.

**Tachipirina<sup>7</sup>:** Comune analgesico e antipiretico, efficace nel trattamento di febbre e dolori lievi o moderati. Ho scelto la Tachipirina perchè come menzionato in precedenza è il prodotto farmaceutico più venduto di tutti gli ultimi anni.

---

<sup>2</sup> <https://farmaci.agenziafarmaco.gov.it/bancadatifarmaci/cerca-farmaco>

<sup>3</sup> [https://www.acutil.it/prod\\_adulti.asp](https://www.acutil.it/prod_adulti.asp)

<sup>4</sup> <https://www.integratorienergya.it/prodotto-energya/>

<sup>5</sup> <https://www.thermacare.it/>

<sup>6</sup> <https://www.amuchina.it/>

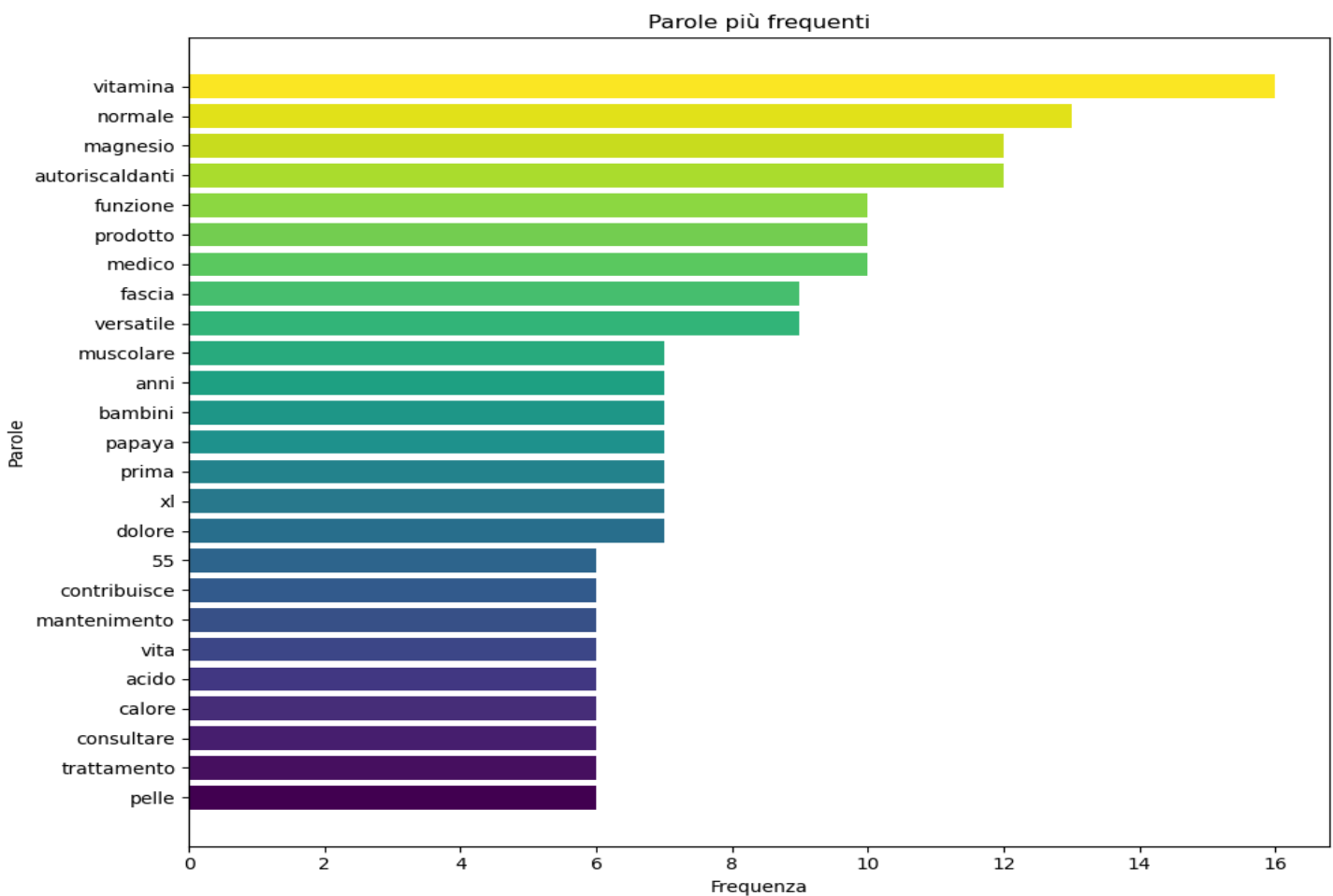
<sup>7</sup> <https://farmaci.agenziafarmaco.gov.it/bancadatifarmaci/farmaco?farmaco=012745>

Per effettuare l'analisi descrittiva mi sono fornito delle seguenti librerie: *pandas*, *numpy*, *nltk*, *wordcloud*, *gensim* e *matplotlib*. Il **preprocessing** è stato effettuato prima dell'analisi descrittiva.

L'analisi è stata riportata principalmente **tramite grafici**, per mostrare con chiarezza le caratteristiche del dataset:

## 1. Conteggio della frequenza delle parole tramite Bag of Words

Per effettuare il conteggio della frequenza delle parole ho utilizzato l'approccio Bag of Words, che a differenza dello String of Words o del Syntactic Parsing non tiene conto dell'ordine e del ruolo grammaticale delle parole.



Come si evince dal grafico, tra le **parole più frequenti** ci sono alcune che spiccano maggiormente, come *vitamina*, *normale*, *magnesio*, *autoriscaldanti*, *funzione* e *medico*. Questo perché i primi due prodotti, ovvero Acutil 55+ e Energya sono due prodotti che mirano al *normale funzionamento* del sistema immunitario e cognitivo, grazie all'utilizzo di *vitamine* e minerali come il *magnesio*. Inoltre la parola *medico* risulta frequente perché sia nei prodotti commerciali che nel bugiardo della Tachipirina ci sono sezioni in cui si esula la responsabilità del prodotto, consigliando sempre di confrontarsi con il proprio medico. *Autoriscaldanti* invece perché il

prodotto ThermaCare nella sua descrizione commerciale sottolinea spesso questa caratteristica del prodotto.

## 2. WordCloud delle parole più frequenti

Le WordCloud permettono di **visualizzare graficamente la frequenza delle parole**.

Più la parola sarà grande maggiore sarà la sua frequenza:



In questa WordCloud dei primi 4 prodotti si rivedono gli stessi risultati analizzati precedentemente, quindi spiccano parole come *normale*, *vitamina*, *autoriscaldanti*, *magnesio* e *funzione*.



## 2.1 WordCloud delle parole più frequenti (Tachipirina)

La WordCloud della Tachipirina ho scelto di analizzarla separatamente per evidenziare la differenza dalle parole contenute nei prodotti commerciali. Infatti come si evince dalla nuvola di parole, i termini maggiormente frequenti sono di “prevenzione”, come ad esempio *effetti indesiderati*, *medico*, *farmacista*, *ogni eventuale* ed *esula*. Questo perché il testo essendo estratto da un bugiardino non mira alla vendita del prodotto quanto più a sottolineare le controindicazioni, gli effetti collaterali e le precauzioni d’uso del prodotto.

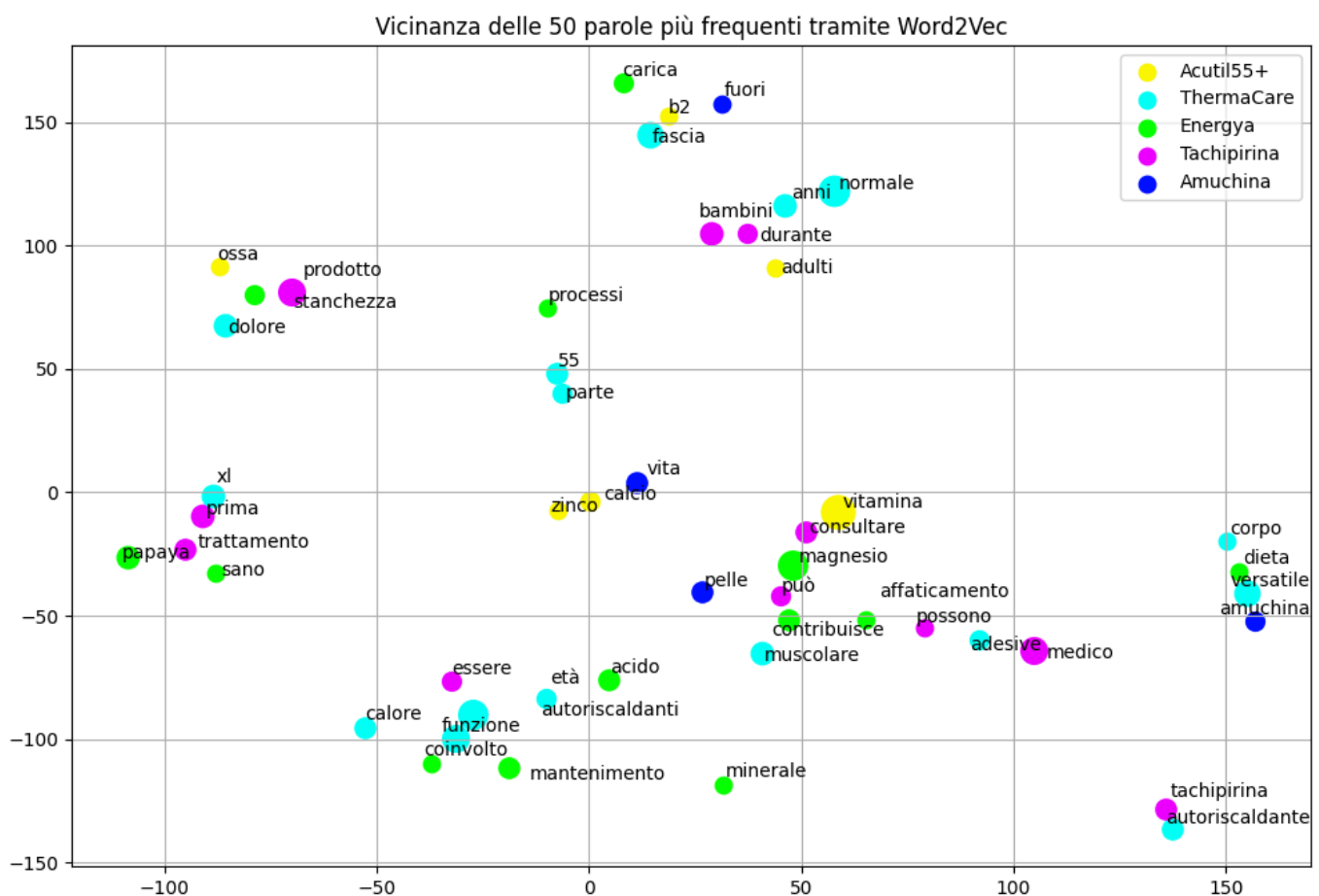
### 3. Word Embedding delle 50 parole più frequenti tramite Word2Vec

Per effettuare il Word Embedding ho deciso di usare Word2Vec, un modello NLP sviluppato da Google. Word2Vec consente di rappresentare le parole del testo come vettori in uno spazio vettoriale. Grazie a questo meccanismo riesce a catturare le relazioni semantiche tra le parole in base al contesto in cui appaiono. L'obiettivo è quello di **mappare parole semanticamente simili** vicine nello spazio vettoriale. Due sono gli approcci principali:

- **Continuous Bag of Words (CBOW)** → Predice una parola dato il contesto;
- **Skip-gram** → Predice le parole del contesto data una parola centrale, usando l'approccio String of Words, che a differenza del Bag of Words tiene conto della posizione delle parole.

Ho deciso di utilizzare Skip-gram dato che risulta essere più performante.

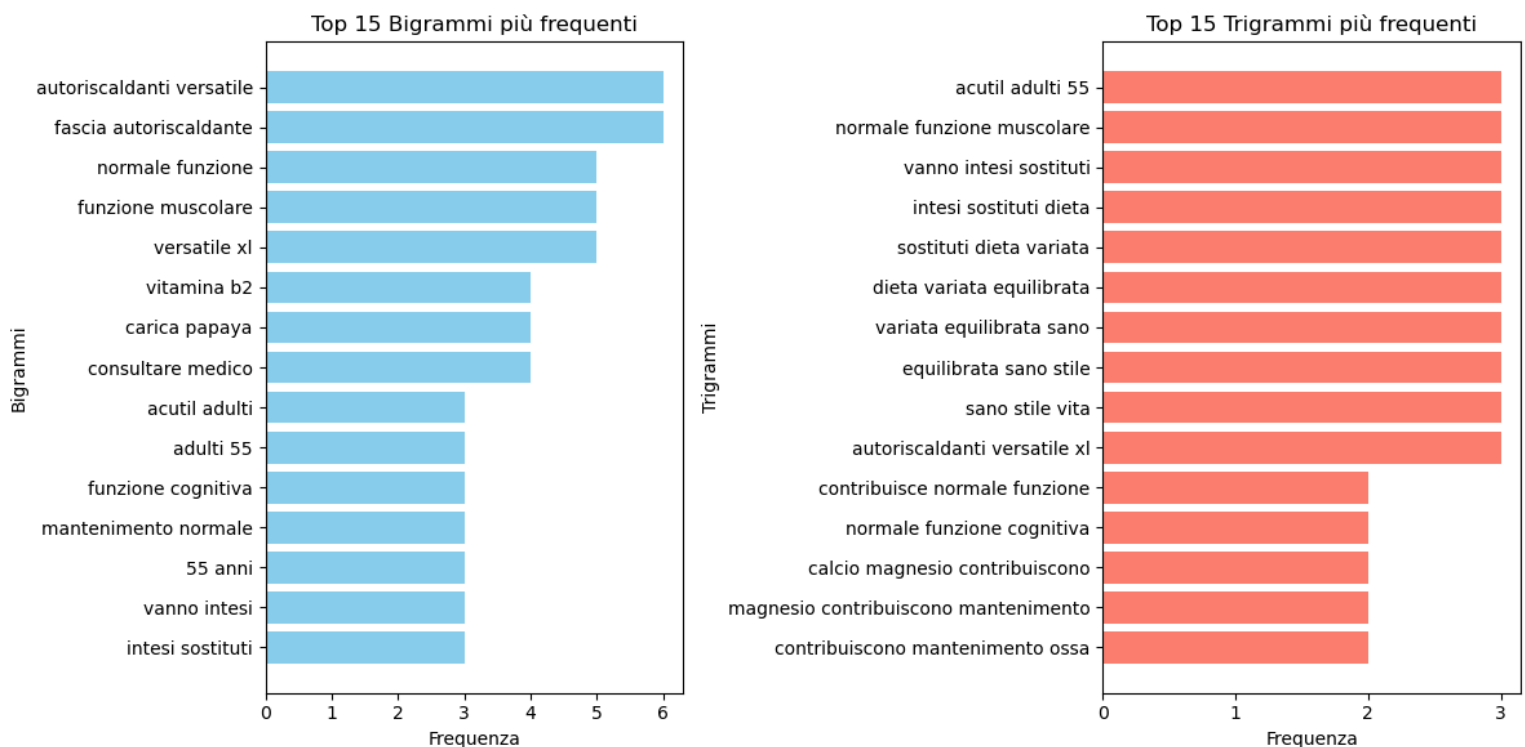
La grandezza dei pallini indica la frequenza di quella parola nel corpus di testo.



Innanzitutto risaltano subito le parole più frequenti, come già analizzato in precedenza. Infatti parole come *vitamina*, *normale* e *prodotto* si distinguono per la loro **grandezza**. Inoltre si evidenzia la **relazione semantica** tra alcuni **cluster** di parole mappati nello stesso spazio vettoriale, come ad esempio *ossa*, *dolore e stanchezza* (in alto a sinistra), *vitamina*, *magnesio*, *pelle*, *muscolare*, *affaticamento* (nella zona centrale) insieme a *zinco* e *calcio*. Inoltre nella parte alta del grafico sono semanticamente simili in base al **contesto** le parole come *bambini*, *adulti ed anni*. Nel complesso la vicinanza di tutte queste parole è sensata dal momento che il contesto in cui potrebbero apparire è molto simile.

#### 4. Top 15 Bigrammi e Trigrammi più frequenti

Il seguente grafico mostra la Top 15 dei bigrammi e trigrammi più frequenti. I bigrammi rappresentano una sequenza di due parole consecutive all'interno di un testo, mentre i trigrammi rappresentano una sequenza di tre parole.

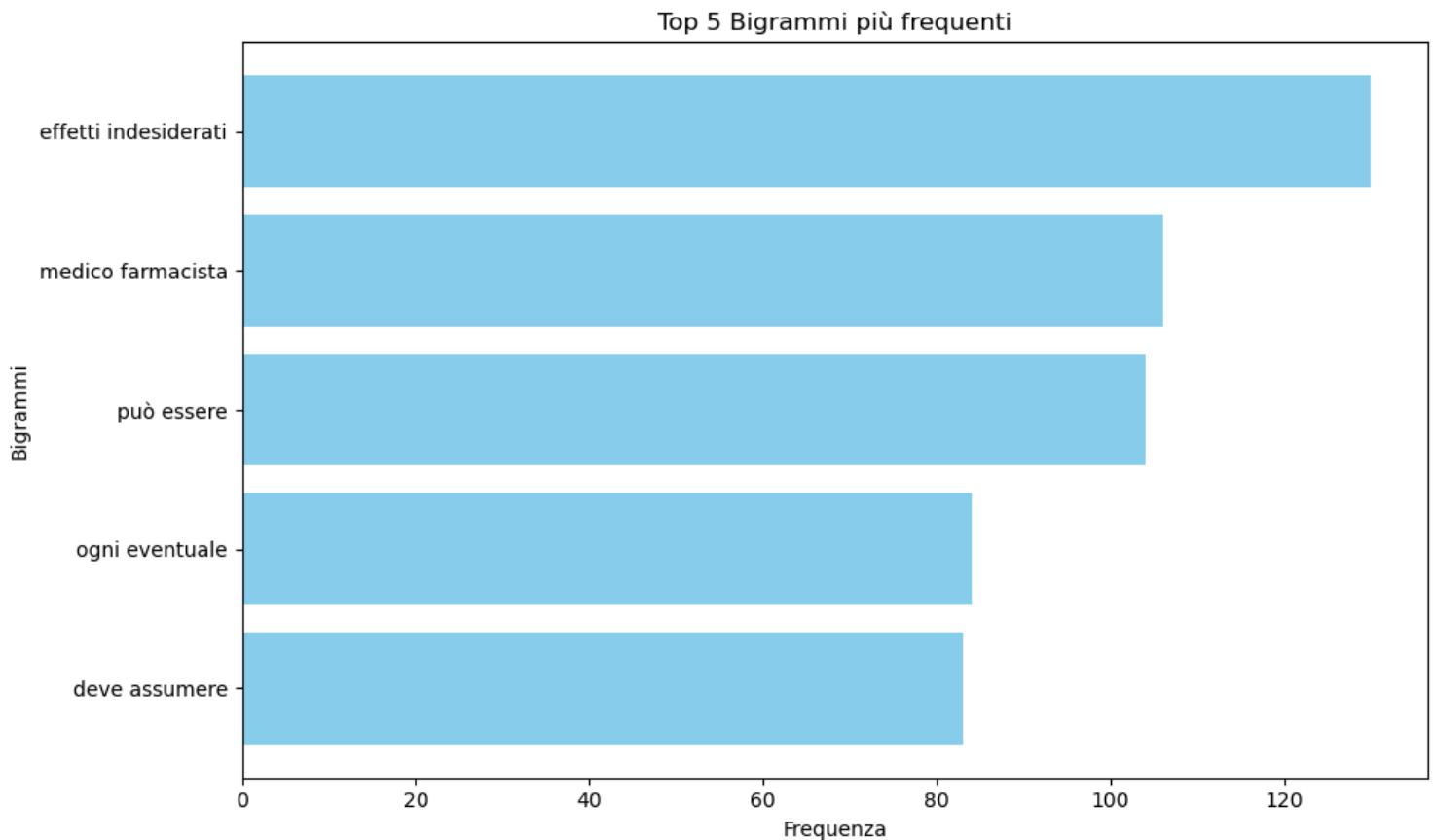


Nei bigrammi e nei trigrammi si presenta lo stesso fenomeno. Le parole più frequenti in precedenza si uniscono formando questi n-grammi. Tra i **bigrammi** troviamo *autoriscaldanti versatili*, *fascia autoriscaldante*, in riferimento al prodotto ThermaCare, e *normale funzione*, *funzione muscolare* in riferimento ai prodotti come Acutil 55+ e Energya. Inoltre si evidenzia anche la frequenza di bigrammi come *adulti 55* e *55 anni*, questo perché i primi due prodotti sono destinati ad un target di clienti su quella fascia d'età.

Per quanto riguarda i **trigrammi**, anche in questo caso spiccano termini associati ai primi due prodotti, che sono per l'appunto integratori destinati a persone sopra i 55

anni d'età. In questo caso vien fuori come nonostante siano descrizioni di prodotti commerciali c'è comunque una parte di testo che funge da precauzione, in modo tale da esulare il prodotto stesso da possibili problemi; vengono infatti utilizzate parole come *non vanno intesi sostituti di una dieta variata e uno stile di vita sano*.

### 3.1 Top 5 Bigrammi più frequenti (Tachipirina)



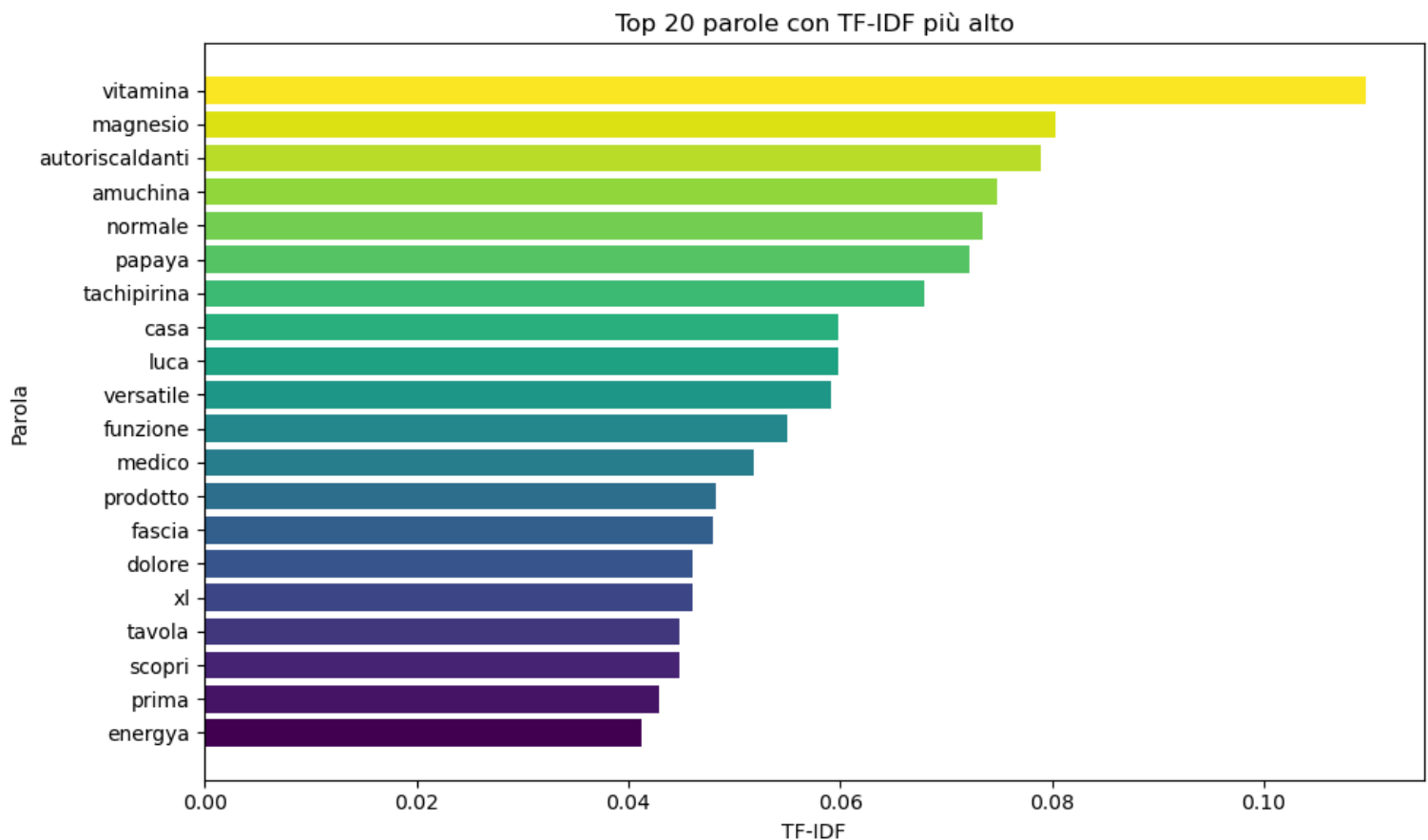
Anche in questo caso ho preferito evidenziare i bigrammi di Tachipirina separatamente, proprio perché da come si evince dal grafico, sono completamente differenti da quelli degli altri prodotti. Infatti troviamo con una frequenza di circa 130 bigrammi il bigramma *effetti indesiderati*. Questo fattore era quasi completamente assente nelle descrizioni commerciali degli altri prodotti.

## 5. Parole con TF-IDF più alto

Il TF-IDF è una tecnica che valuta l'importanza di una parola in un documento in base a due fattori: la frequenza della parola nel documento (TF) e la rarità della parola nel corpus (IDF). È utilizzato per identificare parole chiave o rilevanti in un insieme di documenti. Più una parola è frequente più il TF-IDF si alza, ma se quella parola



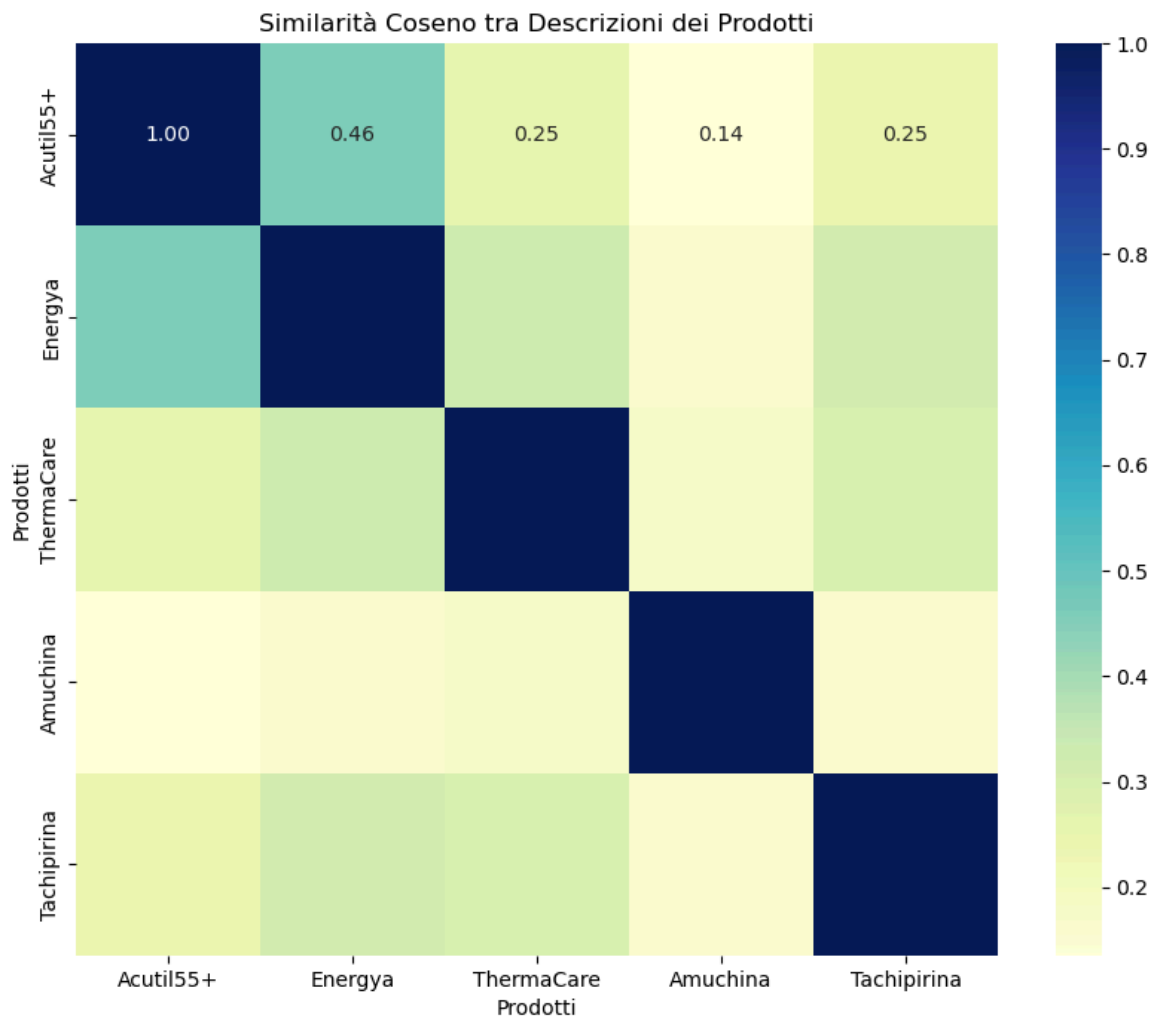
appare in più documenti allora il TF-IDF si abbassa. Più alto è il valore più quella parola è significativa rispetto alle altre.



Quello che si evince dal grafico è che la parola più significativa è *vitamina*, visto che ha il TF-IDF più alto con un valore superiore a 0.10, seguono *magnesio*, *autoriscaldanti*, *amuchina* e *normale* indicando che questi termini sono particolarmente significativi e rilevanti rispetto al resto. Seguono alcune parole contenute nella descrizione di Amuchina e alcune di ThermaCare.

## 6. Heatmap della Similarità Coseno delle descrizioni

La similarità coseno confronta la similarità tra i vettori che rappresentano il contenuto dei testi, per analizzare la similarità tra loro. Ho voluto graficare il risultato tramite la seguente mappa di calore:



Con questa mappa di calore si vuole mettere in mostra la similarità e la non-similarità tra le descrizioni dei prodotti. Gli unici prodotti maggiormente simili tra di loro sono Acutil 55+ ed Energya, infatti sono entrambi integratori che mirano a migliorare la condizione psicofisica degli adulti dai 55 anni in su. Al contrario le altre descrizioni sono poco simili tra di loro, perchè effettivamente sono prodotti diversi tra di loro, che hanno uno scopo ben specifico.

### 3. Sentiment Analysis

La Sentiment Analysis rappresenta il fulcro e l'obiettivo principale del progetto. Questo processo si basa sull'analisi dei testi, in questo caso delle descrizioni commerciali di alcuni prodotti farmaceutici, andando a **far emergere l'inclinazione positiva, neutra o negativa del testo**. Inoltre ho voluto analizzare quella che è l'Emotion Analysis, ovvero l'analisi delle emozioni che suscitano questi testi.

Per effettuare queste analisi ho utilizzato, come descritto nella Datacard, librerie come *nlk* e *TextBlob* e modelli pre-addestrati per quella che è l’NLP (Natural Language Processing, elaborazione del linguaggio naturale). Inoltre ho utilizzato **Feel-IT**<sup>8</sup> di MilanNLProc<sup>9</sup>(Fine-Tuning di UmBERTo<sup>10</sup>) come architetture transformers, ovvero un'architettura di rete neurale artificiale. Basati sul meccanismo di attenzione, i transformers gestiscono sequenze di dati, come il linguaggio naturale, in modo efficiente, catturando relazioni a lungo raggio.

Come lexicon per le analisi tramite *nlk* e *TextBlob*, essendo che di per sé entrambi non sono fatti per l’analisi di testi in italiano, ho utilizzato *sentix*<sup>11</sup>, un lexicon per la Sentiment Analysis in italiano.

Infine Feel-IT mi ha permesso di fare, oltre che la **Sentiment Analysis**, anche l’**Emotion Analysis**, l’analisi delle emozioni che suscita un testo, come ad esempio rabbia, paura, gioia e tristezza.

## 1. NLTK (Natural Language Toolkit)

Dopo aver effettuato la tokenizzazione, cioè il processo di suddivisione di un testo in unità più piccole, chiamate token, come parole, frasi o simboli che aiuta a analizzare e manipolare il testo in modo più efficiente nelle applicazioni di elaborazione del linguaggio naturale e nell’analisi testuale, ho proceduto con l’analisi del sentiment, utilizzando il lexicon *sentix*, come anticipato in precedenza.

I risultati ottenuti tramite NLTK sono riportati nella seguente tabella:

Prodotto	Sentiment
Acutil 55+	Positivo
Energya	Positivo
ThermaCare	Positivo
Amuchina	Positivo
Tachipirina	Negativo

Come si evince dalla tabella dei risultati, per i primi 4 prodotti con le descrizioni commerciali il sentiment risulta essere positivo, perchè come visto anche nelle analisi precedenti, sono presenti parole che rimandano al benessere. Per quanto riguarda invece la descrizione di Tachipirina, essendo estratta dal suo foglietto illustrativo, il sentiment è negativo, dal momento che ci sono per la maggior parte parole che rimandano agli effetti indesiderati, le avvertenze e le precauzioni.

<sup>8</sup> <https://huggingface.co/MilaNLProc/feel-it-italian-sentiment>

<sup>9</sup> <https://huggingface.co/MilaNLProc>

<sup>10</sup> <https://huggingface.co/Musixmatch/umberto-commoncrawl-cased-v1>

[https://huggingface.co/docs/transformers/model\\_doc/bert](https://huggingface.co/docs/transformers/model_doc/bert)

<sup>11</sup> <https://valeriobasile.github.io/twita/sentix.html>

## 2. TextBlob

A differenza di NLTK, TextBlob fornisce oltre al sentiment anche un valore di polarità ed intensità, che ci permetteranno di capire meglio sia il valore effettivo del sentiment sia quanto il testo sia forte o debole, dove 0 indica minima intensità e 1 indica massima intensità.

Anche in questo caso è stato utilizzato il lexicon sentix, e il risultato è stato riportato tramite la seguente tabella:

Prodotto	Sentiment	Polarità	Intensità
Acutil 55+	Positivo	-0.8	0.5
Energya	Positivo	-0.14	0.43
ThermaCare	Positivo	1	0
Amuchina	Positivo	1	0.5
Tachipirina	Neutro	-0.2	0

Grazie a questi nuovi parametri l'analisi si fa più interessante. Per quanto riguarda Acutil 55+ il sentiment risulta positivo, ma la polarità negativa (-0.8). Questo ci fa intendere come ci siano parole che effettivamente vengono associate ad un sentiment negativo, come ad esempio *stanchezza*, *affaticamento* e *irritabilità*. Questo fenomeno si verifica anche nel secondo prodotto ma in una forma molto più leggera. Per quanto riguarda ThermaCare e Tachipirina, hanno un sentiment completamente positivo, ma l'intensità di Amuchina è dello 0.5, questo perché si avverte un'intensità maggiore nei termini che vengono utilizzati.

Infine per Tachipirina il sentiment e l'intensità sono neutri, proprio perché il bugiardino deve essere il più neutro possibile, senza una particolare inclinazione, dal momento che non è mirato alla vendita del prodotto bensì alla descrizione delle caratteristiche, le avvertenze ed i possibili effetti indesiderati. Tuttavia la polarità è negativa (-0.2) proprio perché spesso tra gli effetti indesiderati si trovano parole come legate ad un sentiment negativo.

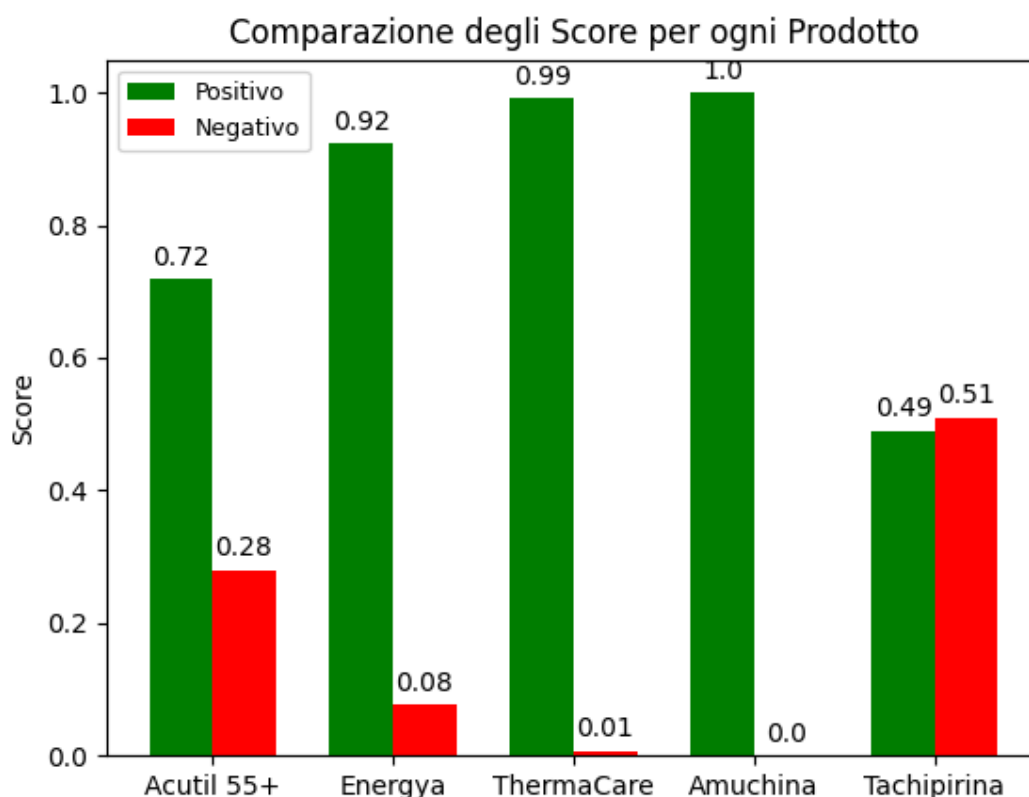
## 3. Sentiment Analysis con Feel-IT

Ho voluto effettuare l'analisi del sentiment anche tramite rete neurale, usando un modello pre-addestrato, Feel-IT<sup>12</sup>. Quest'ultimo è un **modello fine-tuning**, un processo di adattamento di un modello pre addestrato utilizzando dati specifici del dominio per migliorare le prestazioni su una determinata attività, in questo caso la Sentiment Analysis su corpus di testo italiani. Il fine-tuning in questo caso è stato

<sup>12</sup> <https://huggingface.co/MilaNLPProc/feel-it-italian-sentiment>

effettuato dal team di **MilanNLProc**<sup>13</sup> su **UmBERTo**<sup>14</sup> (**BERT**<sup>15</sup>), modello per l'analisi testuale italiano.

I risultati sono stati riportati in forma grafica nella seguente maniera:



Come si può vedere anche qui vengono confermati gli stessi sentiment avuti in precedenza. Uno score preciso sia per il sentiment positivo che negativo ci permette allo stesso tempo di avere una visione più chiara del risultato. Infatti per quanto riguarda Acutil 55+, ha un sentiment prettamente positivo, ma come anticipato nelle altre analisi contiene delle parole che rimandano alla stanchezza e all'affaticamento, di conseguenza il modello le vede come parole associate ad un sentiment negativo. Invece per Energya questo fenomeno si verifica in una percentuale molto più bassa. Per quanto riguarda Tachipirina invece, anche in questo caso risulta avere un sentiment neutro.

#### 4. Emotion Analysis con Feel-IT

Lo stesso modello l'ho utilizzato per fare l'Emotion Analysis, una tecnica simile alla Sentiment Analysis ma che **analizza l'emozione** che suscita un corpus di testo. Tra queste emozioni, FEEL-IT, riconosce la rabbia, la paura, la gioia e la tristezza.

Il grafico che mostra i risultati è il seguente:

<sup>13</sup> <https://huggingface.co/MilanNLProc>

<sup>14</sup> <https://huggingface.co/Musixmatch/umberto-commoncrawl-cased-v1>

<sup>15</sup> [https://huggingface.co/docs/transformers/model\\_doc/bert](https://huggingface.co/docs/transformers/model_doc/bert)

Emotion e Sentiment per ogni Prodotto	
Acutil 55+	tristezza positivo
Energya	paura/timore positivo
ThermaCare	paura/timore positivo
Amuchina	gioia positivo
Tachipirina	paura/timore neutro

Per quanto riguarda i sentiment nell'Emotion Analysis sono gli stessi di prima dato che il modello utilizzato è lo stesso. Mentre le emozioni analizzate quasi tutte negative, tranne per Amuchina. Infatti per i primi tre prodotti le emozioni che suscitano le descrizioni sono tristezza, paura e timore; lo stesso vale per Tachipirina. Per Acutil 55+ la motivazione è la stessa del perchè ha una percentuale di sentiment negativo, cioè che contiene parole effettivamente "tristi" come affaticamento e stanchezza. Per Energya il concetto è simile, essendo che è un prodotto che mira al benessere psicofisico. Per quanto riguarda Amuchina, suscita gioia perchè effettivamente la descrizione è uno spot tra la mamma e il figlio Luca, dove si parla molto di giocare in sicurezza e divertirsi. Infine anche per Tachipirina l'emozione suscitata è paura/timore dato che nel bugiardino si trovano spesso sezioni dedicate agli effetti indesiderati e alle avvertenze.

## 4. Analisi e commento dei risultati

Nel complesso queste analisi sono servite molto per sottolineare la differenza tra descrizioni commerciali dei prodotti, quindi con funzioni **marketing e vendita**, e quelle di prodotti come Tachipirina, quindi disponibili esclusivamente sul sito dell'AIFA e che non mirano alla vendita bensì alla **responsabilizzazione** dei consumatori, **esulando** l'azienda da possibili cause. Di conseguenza in generale mi aspettavo di avere risultati di questo tipo, quindi che per le descrizioni commerciali il sentiment fosse positivo e per quella estratta dal bugiardino della Tachipirina fosse **neutro**.

Tuttavia non mi aspettavo di trovare del sentiment negativo in prodotti come Acutil 55+; ipotizzando una persona con più di 55 anni eviterebbe di comprare il prodotto perchè si sentirebbe “triste” o “vecchia” leggendo la descrizione del prodotto in questione?

Per quanto riguarda invece i metodi utilizzati, quello più affidabile e coerente è stata sicuramente la rete neurale, quindi il modello FEEL-IT. Essendo un modello transformers ha permesso di adattarsi meglio al contesto e ai corpus di testo rispetto a NLTK e TextBlob. In particolare le caratteristiche che hanno fatto la differenza sono state: gli score precisi per ogni prodotto, sia per il sentiment positivo che negativo. Questo ha permesso infatti di spiegare il fenomeno che si era verificato in TextBlob, ovvero che i primi due prodotti avevano un sentiment positivo ma una polarità negativa, in particolare Acutil 55+, dovuta all'utilizzo di parole con un'accezione negativa che però nelle descrizioni servono per descrivere la funzione del prodotto.

Passando ad un'analisi più dettagliata per ogni prodotto, i risultati sono i seguenti:

**Acutil 55+:** Per quello che riguarda questo primo prodotto, i risultati dell'analisi sono alquanto interessanti. Infatti con i vari metodi utilizzati è emerso che il sentiment della sua descrizione, pur essendo positivo, ha una percentuale dello score negativa.

Questo perchè all'interno della descrizione ci sono parole con accezione negativa, come *stanchezza e affaticamento*, che quindi il modello etichetta con un sentiment negativo. Infatti l'emotion in questo caso è tristezza.

Ovviamente essendo che Acutil 55+ è un prodotto che serve a migliorare le capacità psicofisiche nei soggetti che hanno un'età maggiore ai 55 anni, queste parole sono utilizzate per sottolineare le caratteristiche e l'efficacia del prodotto.

A mio parere, dal punto di vista del **marketing**, questo potrebbe essere uno svantaggio; infatti un adulto potrebbe evitare di comprare il prodotto in questione per fattori emotivi, dato che se lo farebbe si sentirebbe più “vecchio” di quanto effettivamente è, e quindi evita di farlo.

**Energia:** Il fenomeno descritto precedente si presenta anche in questo prodotto, tuttavia in forma molto più lieve. Questo è dovuto al fatto che la descrizione di Energia in questo caso contiene una percentuale di parole negative molto più bassa

rispetto ad Acutil 55+. La descrizione di Energya infatti punta maggiormente a sottolineare i benefici che porta il prodotto, usando **espressioni positive** come *ringiovanire* e *ripristinare*. Tuttavia anche in questo caso l'emotion è negativa, ovvero paura/timore.

**ThermaCare:** Per il terzo prodotto il sentiment è positivo, dato che passiamo da integratori ad un prodotto che serve per rilassare i fasci muscolari, di conseguenza la quantità di parole negative è sicuramente minore o meno importante. Tuttavia anche qui l'emotion calcolata è paura/timore, dato che la descrizione contiene parole legate alla sensazione del *dolore*.

**Amuchina:** Il quarto prodotto invece ha un sentiment sempre positivo, come negli altri casi, ma anche un'emotion positiva, ovvero gioia. A mio avviso questo è dovuto al fatto che nello **storytelling** del prodotto viene raccontato lo spot tra una mamma e il figlio Luca. Infatti la mamma sottolinea come Luca possa giocare in sicurezza, divertendosi senza pensare a germi e batteri.

**Tachipirina:** Per quanto riguarda l'ultimo prodotto, Tachipirina, l'esito è differente rispetto ai prodotti precedenti. Partendo dal fatto che ho voluto inserire questo prodotto proprio per sottolineare la differenza tra descrizioni commerciali, mirate alla vendita e alla sponsorizzazione di un prodotto rispetto a quelle inserite nei bugiardini farmaceutici, che servono a descrivere le caratteristiche del prodotto, le avvertenze, le controindicazioni e gli effetti collaterali, il sentiment del prodotto è **neutro**. Questo si evince anche dalla polarità del sentiment, che è 0, o dallo score che è 51% positivo e 49% negativo. Quindi dal mio punto di vista è stato fatto un buon lavoro da chi ha scritto questo foglietto illustrativo, dato che dovrebbe appunto essere il più neutro possibile per esulare l'azienda farmaceutica da possibili conseguenze. Anche in questo caso però l'emotion è paura/timore, dato che ci sono, ovviamente, sezioni che riguardano come descritto in precedenza gli effetti collaterali del prodotto.

## 5. Conclusione

In questo progetto di Sentiment Analysis, condotto sulle descrizioni commerciali di cinque prodotti di Angelini Pharma, ho osservato come il **linguaggio** utilizzato nelle descrizioni commerciali dei primi quattro prodotti e nel foglietto illustrativo del quinto, influisca sulla percezione del **sentiment e delle emozioni** suscitate nei lettori. Attraverso tecniche di scraping, pre processamento e analisi dei dati, ho esaminato prodotti con scopi e target differenti i seguenti prodotti: Acutil Adulti 55+, Energya, ThermaCare, Amuchina e Tachipirina.



L'analisi ha rivelato che, nonostante le descrizioni commerciali siano generalmente positive, contengono parole che possono evocare sentimenti negativi, come **stanchezza** e **affaticamento** ad esempio in Acutil Adulti 55+. Questo evidenzia un potenziale problema di **marketing**, dove un adulto potrebbe evitare l'acquisto per evitare sensazioni di invecchiamento. In Energya, sebbene il fenomeno sia meno pronunciato, si nota comunque la presenza di termini con accezione negativa, nonostante mirino di più a sottolineare gli aspetti positivi del prodotto. ThermaCare, invece, ha un sentiment positivo ma con emozioni di paura/timore legate alla descrizione del **dolore**. Amuchina si distingue per un'emozione positiva di gioia, grazie allo **storytelling familiare** che trasmette sicurezza e divertimento. Tachipirina, con il suo foglietto illustrativo, mantiene un sentiment **neutro**, come previsto, data la natura **informativa e non commerciale** del testo.

L'utilizzo del modello **Feel-IT** si è dimostrato particolarmente efficace nel cogliere queste sfumature, confermando e approfondendo i risultati ottenuti con **NLTK** e **TextBlob**. Infatti la rete neurale ha fornito una visione più chiara e precisa del sentiment e delle emozioni suscitate dai testi.

In conclusione, questo studio ha messo in luce l'importanza della scelta delle parole nelle descrizioni dei prodotti farmaceutici e il loro impatto sulla **percezione del consumatore**. Le aziende devono prestare attenzione non solo al messaggio positivo che intendono trasmettere, ma anche alle implicazioni emotive delle parole utilizzate. Questo approccio può aiutare a migliorare le **strategie di marketing** e la comunicazione con i consumatori, assicurando che i prodotti vengano percepiti in modo ottimale. Future ricerche potrebbero espandere questo studio includendo un numero maggiore di prodotti e confrontando descrizioni di diverse case farmaceutiche per ottenere una visione più completa del panorama della Sentiment Analysis nel settore farmaceutico.