



UNIVERSITÀ DEGLI STUDI DI MILANO - BICOCCA

Scuola di Scienze

Dipartimento di Informatica, Sistemistica e Comunicazione

Corso di laurea in Informatica

Enhancing irony detection with affective information

Relatore: Prof.ssa Elisabetta Fersini

Co-Relatore: Dott.ssa Debora Nozza

Relazione della prova finale di:

Gianluca Giudice

Matricola 830694

Anno Accademico 2019-2020

Contents

1	Introduzione	3
1.1	Descrizione del problema	3
1.2	Approccio al problema	3
1.3	Sintesi dei risultati	4
2	Stato dell'arte	5
2.1	Sarcasmo e sentiment analysis	5
2.2	Difficoltà e problemi	6
2.2.1	Ironia nel testo	6
2.2.2	Importanza del contesto	6
2.3	Linguistica del sarcasmo	6
2.3.1	Definizione formale	6
2.3.2	Tipi di sarcasmo	7
2.4	Dataset utilizzati	7
2.5	Approcci adottati	7
2.5.1	Rule based	8
2.5.2	Semi supervisionati	8
2.5.3	Supervisionati	8
2.5.4	Deep Learning	8
2.6	Conclusioni	8
3	Sistema realizzato	10
3.1	Descrizione del sistema proposto	10
3.2	Dati in input	12
3.3	Rappresentazione del testo	12
3.3.1	Rappresentazione Bag-of-word	12
3.3.1.1	Tokenization	12
3.3.1.2	Filtering	13
3.3.1.3	Stemming	13
3.3.1.4	Vector encoding	14
3.3.2	Rappresentazione mediante Transformer	14
3.3.2.1	BERT	14
3.3.2.2	Sentence-BERT	15
3.4	Caratteristiche linguistiche	15
3.4.1	PP (Pragmatic particles)	15
3.4.1.1	Emoticons	16
3.4.1.2	Acronimi	16

3.4.1.3	Espressioni onomatopoeiche	16
3.4.1.4	Punteggiatura	16
3.4.1.5	Estrazione particelle pragmatiche	16
3.4.2	POS (Part Of Speech)	17
3.4.3	EMOT (Emotional Features)	17
3.4.4	Sintesi features utilizzate	18
4	Modelli supervisionati	20
4.1	Decision Tree	21
4.2	Naive Bayes	22
4.3	Multinomial Naive Bayes	23
4.4	Bayesian Network	23
4.5	Support Vector Machine	25
4.5.1	Iperpiano	25
4.5.2	Iperpiano ottimo	25
4.5.3	Vettori di supporto	26
4.5.4	SVM non lineare	26
4.6	Strumenti utilizzati	28
4.6.1	Scikit-learn	28
4.6.2	Weka	28
5	Campagna sperimentale	29
5.1	Dataset	29
5.2	10-Folds Cross-Validation	29
5.3	Misure di performance	30
5.4	Esperimenti	32
5.4.1	Caratteristiche linguistiche	33
5.4.1.1	Weighted average	33
5.4.1.2	Confusion matrix	35
5.4.2	BOW + Caratteristiche linguistiche	36
5.4.2.1	Weighted average	36
5.4.2.2	Confusion matrix	38
5.4.3	BOW vs BERT vs Sentence-BERT	39
5.4.3.1	Weighted average - Accuracy	39
5.4.3.2	Weighted average - F measure	39
5.4.4	Analisi lessico con PCA	40
5.4.4.1	Top k componenti principali	40
5.4.4.2	Features con contributo maggiore	41
5.4.4.3	Conclusioni	41
6	Conclusioni e sviluppi futuri	42
	Bibliography	42

Chapter 1

Introduzione

1.1 Descrizione del problema

La sentiment analysis è un campo dell'elaborazione del linguaggio naturale (NLP) che si occupa di costruire sistemi per l'analisi del testo, ha il fine di identificare e classificare l'informazione riguardo il sentimento e l'opinione espressa nello stesso. Si basa sui principali metodi di linguistica computazionale e di analisi testuale. L'analisi del sentiment è utilizzata in molteplici settori: dalla politica ai mercati azionari, dal marketing alla comunicazione, dall'analisi dei social media alla valutazione delle preferenze del consumatore.

Il riconoscimento automatico dell'ironia nei contenuti generati da utenti, è uno dei compiti più complessi per quanto riguarda l'elaborazione del linguaggio naturale. Tuttavia è di fondamentale importanza per tutti i sistemi di sentiment analysis, in quanto facendo uso dell'ironia è possibile invertire completamente la polarità di una propria opinione, facendola passare da positiva a negativa e viceversa. Diventa pertanto cruciale sviluppare dei sistemi che siano consapevoli del fenomeno e in grado di riconoscerlo.

L'ironia è un tema studiato in diverse discipline, come la linguistica, filosofia e psicologia, ma è difficile da definire formalmente, soprattutto per questo motivo ne è difficile il riconoscimento. Nonostante ciò ci sono basi teoriche che suggeriscono il ruolo importante della sfera emotiva nell'uso dell'ironia, quindi un fattore chiave per riconoscerlo. Con questo si intende anche un uso indiretto e non esplicito del carico emotivo in ciò che si vuole comunicare.

I social network in generale, e twitter nello specifico, sono ampiamente utilizzati come fonte di informazione per comprendere la web reputation di un brand, perciò si prestano bene per la sperimentazione di modelli computazionali per il riconoscimento dell'ironia, essendo di fatti una grande risorsa per quanto riguarda i dati testuali generati da utenti.

1.2 Approccio al problema

Si può considerare il riconoscimento dell'ironia come un problema di classificazione. Una frase potrà quindi essere classificata come appartenente alla classe ironica o non ironica. Si noti che per come viene approcciato il problema in questa tesi, l'ironia è considerata al pari del sarcasmo.

A questo scopo, la classificazione avviene mediante tecniche di machine learning (una branca dell'intelligenza artificiale). Nello specifico vengono creati dei modelli supervisionati tramite l'utilizzo di algoritmi, i quali fanno uso di una grande quantità di dati per estrarre dei pattern presenti in quest'ultimi.

Come già detto, twitter è una risorsa che fornisce moltissimi contenuti generati da utenti. Viene sfruttato questo aspetto collezionando i dati utilizzati dai modelli matematici, così da far apprendere la macchina con lo scopo di riconoscere l'ironia nel testo.

Il codice scritto in python, utilizzato per l'estrazione delle features, il training dei modelli, la creazione della matrice su cui è stata fatta PCA, oltre ai notebooks creati per l'analisi dei report e il risultato della PCA, è disponibile in questo repository¹ su github.

1.3 Sintesi dei risultati

Il risultato degli esperimenti mostra come per il riconoscimento dell'ironia si ottengono i risultati migliori combinando tutte le features considerate. Inoltre le performance variano decisamente a seconda dell'algoritmo utilizzato per creare i classificatori. Nel caso preso in considerazione, l'algoritmo Support Vector Machine con un kernel lineare risulta essere il migliore tra quelli analizzati.

¹Irony detection repository: www.github.com/gianlucagiudice/irony-detection

Chapter 2

Stato dell'arte

Abstract

L'ironia è figura retorica molto importante, è uno strumento usato nella comunicazione in cui un individuo esprime l'opposto di ciò che realmente intende. Il riconoscimento automatico dell'ironia ha attirato l'attenzione nell'ambito natural language processing (NLP) e machine learning (ML) così da dinvetare un ambito di ricerca.

2.1 Sarcasmo e sentiment analysis

Facendo riferimento a [1], un'analisi dal punto di vista della linguistica computazionale sull'ironia, può fornire approfondimenti data-driven facendo luce riguardo i meccanismi nascosti nell'uso di questo fenomeno, potenzialmente migliorando la comprensione per quanto riguarda i processi mentali umani, la linguistica e la comunicazione.

Tralasciando l'aspetto più filosofico, il riconoscimento dell'ironia ha implicazioni pratiche per quanto riguarda i sistemi di sentiment analysis, ed in effetti i modelli allo stato dell'arte in questo ambito, puntano proprio a portare benefici verso questi sistemi. Sentiment analysis vuol dire estrarre l'opinione di un individuo o della società riguardo un particolare evento o argomento. La presenza di ironia nei messaggi, è responsabile di errori nella classificazione in cui cadono i metodi allo stato dell'arte per la creazione di questi sistemi. Ciò avviene in quanto usando l'ironia viene invertita la polarità dell'opinione espressa nella frase, ingannando in questo modo gli approcci che non tengono in considerazione del fenomeno.

In questi anni i social network, soprattutto Twitter e Facebbok, hanno acquisito molta importanza in questo ambito, dal momento che gli individui condividono i propri pensieri, emozioni e opinioni su queste piattaforme. Di fatto, dato l'enorme contenuto di messaggi generati, questo è un esempio di Big Data. In diverse aree c'è un grande interesse nell'estrarre l'informazione riguardo l'opinione contenuta in questi dati generati da utenti, così da conoscere, ad esempio, la reputazione relativa ad un prodotto piuttosto che l'opinione politica.

Per dare un'idea di quanto possano essere utili sistemi di questo tipo, come fatto notare nel survey [2], nel 2014 D. K. Tayal e altri [3] hanno introdotto un sistema per riconoscere il sarcasmo e identificare la polarità espressa nei tweets politici. Il loro lavoro include la collezione del data set, data pre-processing e la classificazione. Attraverso un approccio supervisionato hanno concluso che da questi dati, comprendendo la componente sarcastica, è stato possibile fare una previsione sul risultato delle elezioni in modo efficiente.

2.2 Difficoltà e problemi

2.2.1 Ironia nel testo

Come indicato nel survey [4], il riconoscimento dell'ironia è un compito complesso non solo per quanto riguarda le macchine, ma in alcuni casi anche per gli uomini. Questo è dovuto all'ambiguità intrinseca del sarcasmo. Inoltre un approccio unicamente bag-of-words tipicamente utilizzato nell'ambito NLP, non ottiene grandi performance in quanto usando il sarcasmo si possono esprimere opinioni negative facendo uso di parole positive. Tuttavia anche considerare altre features oltre a bow può non risolvere completamente il problema, si pensi a messaggi molto corti in cui non è proprio possibile estrarre delle features.

Bisogna anche tenere in considerazione che nel testo è più difficile cogliere l'ironia rispetto che nel parlato, in quanto non si hanno a disposizione features come il tono della voce, il linguaggio del corpo e le espressioni del volto. In questo ambito sono stati condotti degli studi da Tepperman e altri [5], in cui si vuole riconoscere l'ironia nel parlato estraendo dallo spettro del suono features come l'intonazione, le pause etc.

2.2.2 Importanza del contesto

Un altro aspetto importante, analizzato da Joshi e altri [6] è il ruolo fondamentale del contesto quando si tratta di riconoscere il sarcasmo. Pur essendo questo un fattore chiave, Wallace in [1] fa notare come non venga tenuto conto del contesto nei modelli allo stato dell'arte.

Per contesto si intende l'informazione che va oltre il contenuto del testo di per sé. Oltre ad una conoscenza o semantica in generale, esistono altre informazioni legate al contesto come dettagli sull'autore (ad esempio viene tenuto conto del sesso in [5]), a chi viene rivolto il messaggio e l'ambito in generale. In questo campo è stato condotto uno studio da Rajadesingan e altri [7], che hanno sviluppato un modello SCUBA (sarcasm using behavioral modeling), dove per riconoscere il sarcasmo vengono analizzati anche i tweets passati di uno specifico utente.

2.3 Linguistica del sarcasmo

2.3.1 Definizione formale

Con lo scopo di meglio riconoscere il sarcasmo, nel survey [8] ne viene data una definizione formale: Una frase sarcastica può essere rappresentata da una sestupla $\langle S, H, C, u, P, P' \rangle$ dove S è chi parla, H chi ascolta, C il contesto, u la frase, P il senso letterale e P' ciò che si vuole far intendere. Ad esempio, se il cliente di un hotel dice "You room service was fantastic!" e il direttore è consapevole del fatto che il servizio è stato peggiore rispetto al solito, utilizzando la rappresentazione 6-tupla si ha:

S : Cliente

H : Direttore

u : "You room service was fantastic!"

P : Il servizio è stato *ottimo*

P' : Il servizio è stato *peggiore*

Si noti come da questa definizione formale, emerge chiaramente che il contesto è un fattore chiave per cogliere il sarcasmo della frase.

2.3.2 Tipi di sarcasmo

Come fatto notare nel survey [6], esistono diversi tipi di sarcasmo. Camp nel suo lavoro [9] ne considera quattro diversi:

1. **Propositional:** Un'affermazione come "Il tuo piano sembra fantastico" può essere interpretata come sarcastica o meno, dipende dal contesto.
2. **Embedded:** Nella frase è presente una contraddizione intrinseca nella forma stessa delle parole utilizzate. Ad esempio considerando "Dato che si è rivelato così un bravo amico, ho deciso di smettere completamente di vederlo", emerge una chiara contraddizione.
3. **Like-prefixed:** In questo tipo di sarcasmo viene implicitamente negato ciò che si afferma. Ad esempio "Come se la polizia troverà davvero l'assassino".
4. **Illocutionary:** Sono casi in cui situazioni o azioni non relative al testo possono svelare il reale significato inteso dalla frase. Ad esempio "Grazie per avermi tenuto la porta" può considerarsi una frase sarcastica solo conoscendo le circostanze (la porta non è stata realmente tenuta aperta).

2.4 Dataset utilizzati

La qualità del dataset è un fattore cruciale per il riconoscimento dell'ironia/sarcasmo. Il più delle volte vengono considerati dataset in lingua inglese, ma alcuni lavori sono stati fatti anche considerando lingue differenti, come l'italiano nel modello proposto da Barbieri e altri in [10].

Come riportato in [6], anche la dimensioni dei messaggi cambia nei vari studi affrontati, andando a considerati testi brevi (tweets, recensioni prodotti amazon) o più lunghi (sottotitoli di una serie TV o conversazioni di un call center). Volendo valutare le performance dei vari modelli, è necessario che il dataset sia etichettato.

Nel caso dei tweets sono state adottate due strategie diverse. La prima (crowdsourcing) consiste nell'annotare manualmente ogni tweets, questo compito è incaricato ad alcune persone che in alcuni casi sono stati proprio esperti del linguaggio. La seconda tecnica invece (self-tagging), si basa sull'uso degli hashtag. Assumendo che se un utente utilizza l'hashtag *#sarcasm* o *#not*, allora ha l'intenzione di esprimere un messaggio ironico, questo semplifica nettamente la collezione del dataset. Con questa tecnica si ha il vantaggio di creare un dataset di grandi dimensioni con meno sforzo e in un tempo minore rispetto alla prima tecnica, inoltre l'utente stesso viene tratto come il miglior giudice per annotare un suo messaggio come ironico o meno.

Inoltre per quanto riguarda la tecnica crowdsourcing, Joshi e altri [6] hanno fatto notare l'influenza di differenti culture quando si tratta di etichettare i messaggi ironici. Infatti i loro risultati mostrano come dato un dataset americano da far annotare sia a persone americane che indiane, i secondi concordavano maggiormente rispetto agli americani, e gli indiani avevano difficoltà riguardo studi o entità a loro sconosciuti (trattandosi di un dataset americano).

Si noti inoltre che l'ironia è un fenomeno non molto frequente quando si ha a che fare con il testo, questo si riflette in un dataset non bilanciato.

2.5 Approcci adottati

Nello stato dell'arte il problema viene formulato come un task di classificazione, quindi dato un tweet riconoscerlo come ironico o non ironico. Esistono diversi approcci per questo tipo di problema, ma la maggior parte dello stato dell'arte utilizza modelli supervisionati.

2.5.1 Rule based

Un approccio non supervisionato è stato proposto da Bharti e altri [11] in cui vengono usate due regole di classificazione. La prima consiste nel creare un parsing tree basato sul lessico a partire dalla frase, vengono quindi individuate le espressioni che trasportano del sentiment. Se una locuzione negativa è presente in una frase positiva, allora la frase è classificata come sarcastica. La seconda regola di decisione è costituita da un algoritmo che mira a catturare le espressioni iperboliche individuando interiezioni (come "wow") e intensificatori (come "assolutamente").

2.5.2 Semi supervisionati

Tra gli approcci semi supervisionati troviamo l'esempio di algoritmo introdotto da Davidov e altri [12] per riconoscere il sarcasmo nei tweets e recensioni di prodotti Amazon. Prima si identificano dei pattern presenti nelle frasi, dopodiché questi pattern vengono automaticamente estratti dal dataset e la feature principale consiste nel vettore in cui viene riportata la presenza (o assenza) del pattern all'interno del testo in input.

2.5.3 Supervisionati

L'approccio generale consiste nel creare modelli supervisionati, dove partendo dal dataset, dopo una fase di pre-processing, vengono estratte delle features. In letteratura sono state utilizzate diverse caratteristiche per il riconoscimento del sarcasmo, un elenco esaustivo è presente in [8]. Qualche esempio di features ricorrenti tra i modelli allo stato dell'arte sono: bag-of-words; n-gram; part of speech tags; punteggiature; interiezioni. Una volta estrarre le features vengono utilizzati i classici algoritmi di machine learning (SVM, Naive Bayes, Decision tree) per effettuare il training dei modelli, vengono poi testati con la tecnica cross-validation per valutare il più performante.

Come esempio di modelli supervisionati possiamo trovare il lavoro svolto da Farías e altri [13] in cui vengono estratte alcune delle features elencate sopra, considerate altre ulteriori caratteristiche e creati classificatori con gli algoritmi DT e SVM.

Ulteriori approcci supervisionati, come nel caso di Fersini e altri [14], sfruttano tecniche Ensemble, nello specifico Bayesian Model Averaging (BMA), in cui diversi classificatori vengono combinati considerando ognuno di questi durante la classificazione per ottenere risultati più affidabili.

2.5.4 Deep Learning

Come fatto notare in [6], le tecniche basate su architetture deep learning hanno acquisito popolarità nel campo NLP, di conseguenza alcuni approcci si basano su questi modelli per riconoscere automaticamente l'ironia. Un esempio è il modello proposto da Joshi [15] in cui viene utilizzata la similarità tra embeddings delle parole come features per i modelli. Un ulteriore esempio è il modello proposto da Bharti e altri [16] che consiste in una combinazione di Convolutional Neural Network, Recurrent Neural Network e Deep Neural Network mostrando un miglioramento rispetto ad altri modelli allo stato dell'arte.

2.6 Conclusioni

Il riconoscimento dell'ironia è una delle sfide principali nell'ambito della sentiment analysis, e negli ultimi anni ha acquisito man mano più importanza. Si tratta di un campo nel data mining che richiede maggiori approfondimenti.

Ad oggi i modelli allo stato dell'arte hanno performance peggiori nel riconoscere l'ironia rispetto agli umani. Come fatto notare da Wallace in [1], il problema principale degli esistenti modelli di machine learning, può essere una poca considerazione del contesto.

Wallace fa notare come tutti i modelli allo stato dell'arte approcciano il task di riconoscimento dell'ironia/sarcasmo come un problema di classificazione. Non vi sono approcci che hanno lo scopo di determinare il tipo di sarcasmo o ricostruire il significato latente, ovvero ciò che si vuole far intendere, della frase.

Chapter 3

Sistema realizzato

3.1 Descrizione del sistema proposto

In questa tesi, per riconoscere l'ironia nel testo, vengono proposti dei modelli supervisionati di machine learning. Partendo dai dati a disposizione si utilizzano delle tecniche di preprocessing che consistono nel pulire il testo di partenza e prepararlo per essere dato in input ai diversi algoritmi per creare i classificatori. Inoltre si estraggono delle features da ogni messaggio, le quali più sono discriminanti nel riconoscimento dell'ironia, meglio distinguono la classe di appartenenza.

Il dataset viene diviso in due parti: training set e test set. La prima ha dimensione decisamente maggiore e viene utilizzata per far apprendere uno specifico classificatore, l'altra per testarlo. Questo è importante in quanto un determinato modello dovrà riconoscere l'ironia in un generico documento e non solo tra quelli a disposizione. Si noti che entrambe le porzioni di dati sono composte sia dal testo già codificato e processato che le features estratte, hanno solo scopi diversi quando si tratta di creare il classificatore.

Vengono quindi creati vari modelli fornendo in input il dataset di training, con lo scopo di far apprendere la macchina. Avendo costruito più classificatori che utilizzano diversi algoritmi per apprendere dai dati, è possibile confrontarne le performance per valutare il migliore tra tutti, ovvero quello che meglio distingue tra documenti ironici e non ironici.

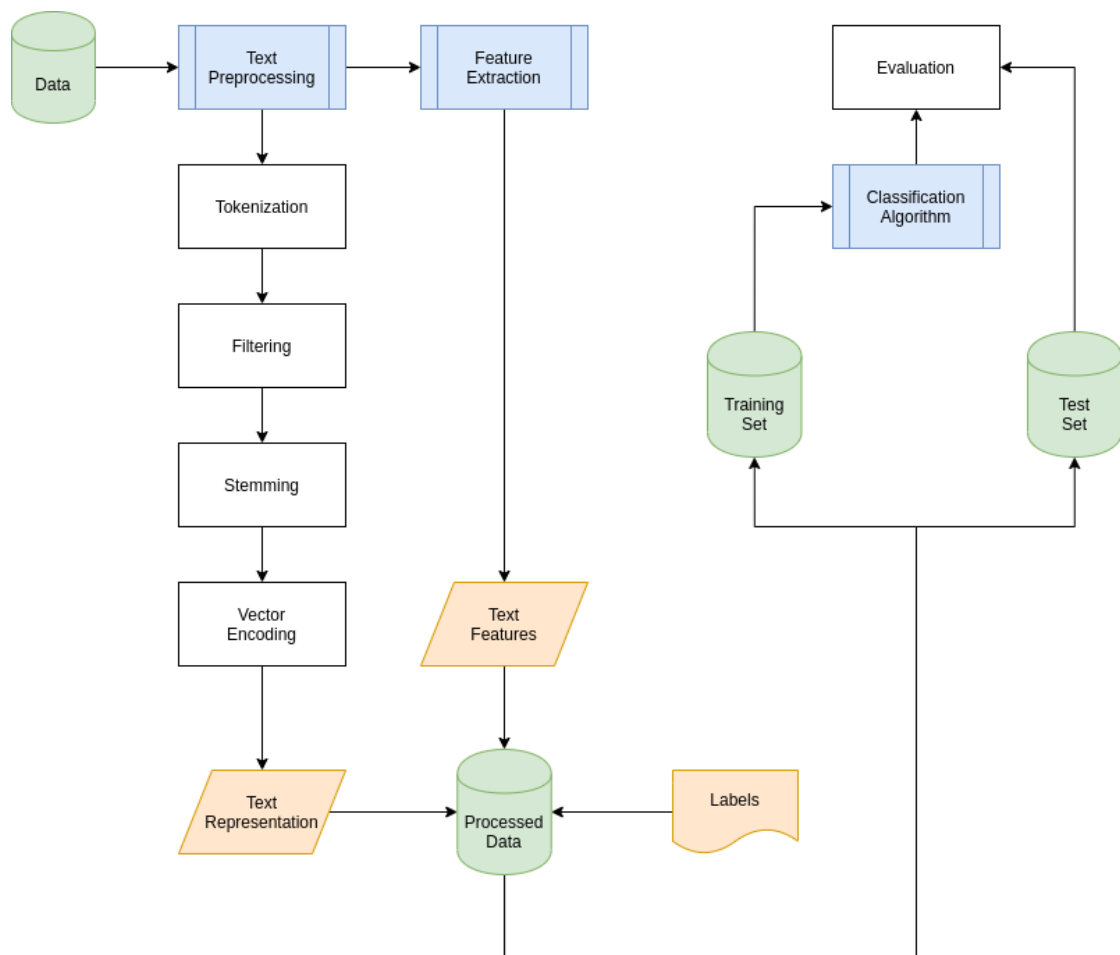


Figure 3.1: Workflow.

3.2 Dati in input

Trattandosi di modelli supervisionati, per il training dei classificatori è necessaria una collezione di testi (in questo caso tweet) con le relative etichette associate ad ogni documento. Il testo e le labels compongono il dataset, questo aspetto è cruciale in quanto una volta aver opportunamente processato i tweets, saranno proprio questi i dati forniti in input al classificatore. La fase di training ha lo scopo di far apprendere al classificatore, così facendo sarà poi in grado di riconoscere l'ironia in testi che non ha mai visto prima.

Tweet	Label associata
Lorem ipsum dolor sit amet, consectetur adipiscing elit	ironico
Cras ornare turpis ut odio finibus, viverra porta felis lobortis	ironico
Suspendisse in ex a felis porta convallis	non ironico
Nam laoreet, lacus at ullamcorper iaculis, id interdum nisl elit nec elit.	non ironico

Table 3.1: Esempio di dataset a disposizione.

3.3 Rappresentazione del testo

Il corpora a disposizione è definito "crudo" e non può essere direttamente utilizzato per la fase di training. Prima è necessario codificare il testo per ottenere una rappresentazione numerica da dare in input all'algoritmo. A questo scopo sono state utilizzate due diverse tecniche

3.3.1 Rappresentazione Bag-of-word

La prima tecnica di rappresentazione utilizzata è boolean bag-of-words. Il testo viene codificato come una matrice booleana composta dai documenti sulle righe e i tokens sulle colonne. I tokens sono tutte quelle parole appartenenti ad un dizionario costruito a partire dal corpora, tutte le operazioni eseguite sono spiegate di seguito.

3.3.1.1 Tokenization

La tokenizzazione è un passo preliminare per l'elaborazione computazionale del testo. Tokenizzare vuol dire dividere le sequenze di caratteri in unità minime di analisi dette "token". Un approccio potrebbe essere considerare un token come una sequenza di caratteri delimitata da spazi, tuttavia una tale definizione lascia spazio a numerose eccezioni, come ad esempio la presenza di punteggiatura. Per questa operazione viene utilizzata la libreria nltk [17] in grado di gestire correttamente tutti questi casi limite, oltre a poter sfruttare una Twitter-aware tokenization adattandosi quindi bene al dominio in questione.



Figure 3.2: Text tokenization.

I vari tokens estratti vengono convertiti in lowercase e sono considerati validi solo se in match con il pattern regex `[a-z]+`. Così facendo non si considerano gli hashtag, la punteggiatura, le emoji e tutto ciò che non è considerato come parola.

Il risultato ottenuto è una lista di tokens validi contenuti in ogni tweet:

Tweet	Tokens lowercase
Lorem ipsum dolor sit amet, consectetur adipiscing elit	$[t_1, t_2, t_3, t_4, t_5]$
Cras ornare turpis ut odio finibus, viverra porta felis lobortis	$[t_6, t_7, t_1, t_8, t_9, t_4]$
Suspendisse in ex a felis porta convallis	$[t_{10}, t_6, t_{11}, t_{12}]$
Nam laoreet, lacus at ullamcorper iaculis.	$[t_{13}, t_{14}, t_8, t_{15}]$

Table 3.2: Esempio di tokens associati ad ogni tweet.

Da qui si crea l'insieme di tokens univoci a partire dai tokens presenti in ogni tweet.

3.3.1.2 Filtering

L'insieme di termini univoci ottenuto è filtrato in base ai seguenti criteri:

- **Occorrenza minima:** Viene fissato un valore di threshold (per gli esperimenti eseguiti è 10) e si scartano tutte quelle parole che compaiono tra tutti i documenti meno del valore scelto.
- **Stopwords + RT:** Vengono scartate tutte le parole più comuni che sono generalmente presenti in una frase. Queste sono un insieme predefinito contenente parole come gli articoli, proposizioni, pronomi e verbi ausiliari.

Es: $\{a, the, of, is, into, it, \dots\}$

Dato il dominio preso in considerazione, è possibile che un utente compia l'operazione di reTweet. Questo è codificato nei messaggi dall'abbreviazione "rt". Non ritenendolo un aspetto rilevante per il riconoscimento dell'ironia, viene anch'essa considerata una stopwords, e di conseguenza non trattato come token valido.

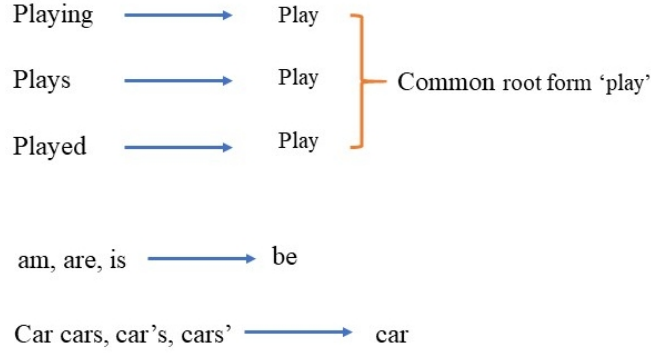
- **Irony e Ironic:** Essendo il dataset a disposizione etichettato mediante la tecnica self-tagging (section 5.1), è presente l'hashtag `#irony` in tutti i tweet ironici. Pertanto nell'insieme dei tokens validi verranno escluse le parole *irony* e *ironic*, così da non avere un bias nei dati sotto questo aspetto. Nel caso contrario i dati usati per creare i modelli avrebbero a disposizione una componente che ben distingue i tweet ironici, ma che nella realtà non sarebbe presente, in quanto non è decisamente corretto assumere che ogni testo ironico contenga una delle due parole.

3.3.1.3 Stemming

Lo stemming è il processo di riduzione della forma flessa di una parola alla sua forma radice.

Con questa tecnica è possibile ridurre il numero di tokens, dal momento che, ad esempio, le stesse parole con suffisso diverso verranno mappate nella medesima radice. Inoltre non è necessario che una parola dopo il processo di stemming venga trasformata in un'altra di senso compiuto, quello che conta è la semantica del token.

¹Fonte: www.datacamp.com/community/tutorials/stemming-lemmatization-python.



Using above mapping a sentence could be normalized as follows:

the boy's cars are different colors → the boy car be differ color

Figure 3.3: Words stemming.¹

3.3.1.4 Vector encoding

L'insieme delle parole ottenute costituisce il dizionario del corpora. Si procede costruendo una matrice booleana composta dai tweets sulle righe e le parole del dizionario sulle colonne.

Sia $D = \{w_1, w_2, \dots, w_m\}$ il dizionario individuato dall'insieme delle n parole univoche.

Sia $T_i = \{t_{i0}, t_{i1}, \dots, t_{ik}\}$, $1 \leq i \leq n$, l' i -esimo tweet composto dai k tokens opportunamente stemmatizzati presenti nello stesso.

Viene definita la matrice $M_{n,m}$

$$M[i, j] = \begin{cases} 1 & \text{se } w_j \in T_i \\ 0 & \text{altrimenti} \end{cases} \quad t.c. \ 1 \leq i \leq n, \ 1 \leq j \leq m$$

3.3.2 Rappresentazione mediante Transformer

Nell'ambito dei modelli di deep learning, per Transformer si intende un particolare tipo di architettura della rete neurale. Questi modelli hanno il vantaggio di essere parallelizzabili, ciò comporta una diminuzione di tempo per la fase di training. Inoltre fanno uso di alcuni particolari meccanismi, per fare un'analogia con quello che è il processo cognitivo si può far riferimento all'attenzione (Da qui il titolo del paper in cui Google introduce il modello BERT "Attention Is All You Need" [18]), per meglio comprendere la semantica della frase [19].

3.3.2.1 BERT

BERT (Bidirectional Encoder Representations from Transformers) è un modello di deep learning creato da Google. A partire da un documento, una volta generati i tokens, è possibile fornire questi come input al modello per generare word embeddings, ovvero un vettore di numeri reali associato ad ogni token. Questo modello viene sfruttato per utilizzare gli embeddings come features, sostituendoli alla rappresentazione del testo BOW.

La rete neurale è composta da 12 hidden layers, ciascuno di questi fornisce in output un vettore di 768 elementi. Sorge quindi il problema della scelta del vettore da utilizzare, inoltre per gli embeddings vengono generati per ciascuno dei tokens. Bisogna pertanto adottare la corretta pooling strategy:

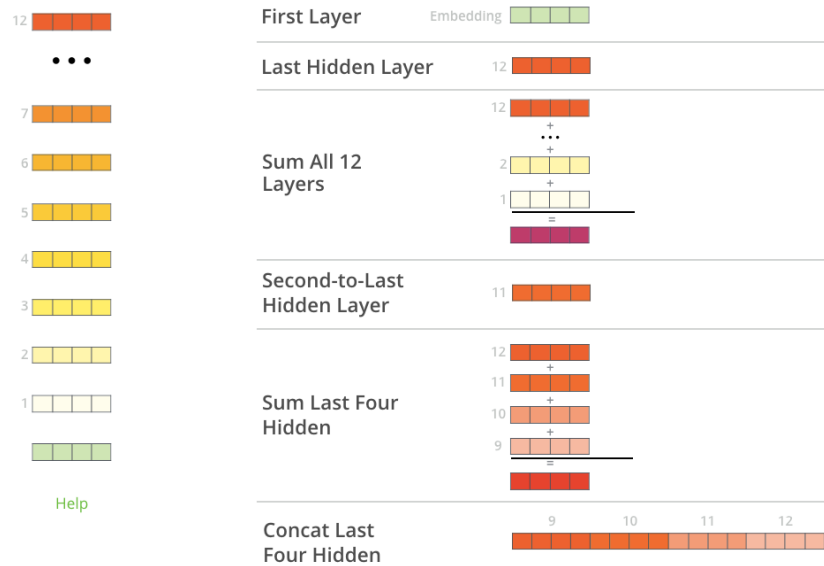


Figure 3.4: Pooling strategy. [19]

Nel caso in questione, si è scelto di concatenare gli ultimi 4 hidden layers e mediare tutti gli embeddings dei tokens presenti nel documento [?]. In tutto si ottiene un vettore di 3072 elementi, il quale sarà utilizzato come features dei modelli in sostituzione della rappresentazione BOW [20].

3.3.2.2 Sentence-BERT

Sentence-BERT [21] è una messa a punto della rete neurale BERT, permette di generare embeddings semanticamente validi relativi ad intere frasi.

3.4 Caratteristiche linguistiche

La scelta di features ad alto contenuto informativo ed indipendenti tra loro, è un passo fondamentale per un corretto riconoscimento di pattern nei dati. A questo scopo vengono estratte delle caratteristiche linguistiche dal testo che sono rilevanti per il riconoscimento dell'ironia. Più le caratteristiche sono discriminanti quando si esprime un messaggio ironico, meglio si potranno distinguere le due classi. Queste features, insieme alla codifica testuale, verranno usate per creare il modello.

3.4.1 PP (Pragmatic particles)

Per catturare il senso non letterale racchiuso in ogni messaggio, vengono considerati alcuni elementi linguistici di seguito elencati.

3.4.1.1 Emoticons

Le emoticons sono riproduzioni stilizzate di quelle principali espressioni facciali umane che esprimono un'emozione. In accordo con l'assunzione che esista una relazione tra il sentimento espresso e l'ironia, queste vengono distinte in due categorie:

- **Positive:** Es. ':-)', ':-D'
- **Negative:** Es. ':-(', ':-D'

Viene quindi considerata l'occorrenza di emoticons positive e negative presenti in un tweet.

3.4.1.2 Acronimi

Gli acronimi sono un ulteriore strumento della comunicazione non verbale. Come le emoticons vengono presi in considerazione dal momento che possono esprimere sentimenti positivi o negativi, infatti possiamo anch'essi dividerli in due categorie:

- **Positivi:** Es. 'ROLF' (Rolling On Floor Laughing), 'LHO' (Laughing Head Off)
- **Negativi:** Es. 'BM' (Bad Manner)

Anche in questo caso si considera l'occorrenza di acronimi positivi e negativi.

3.4.1.3 Espressioni onomatopeiche

L'uso di espressioni onomatopeiche come 'boom' o 'clap' possono essere utilizzate nell'esprimere un messaggio ironico. Tuttavia in questo caso viene presa in considerazione solo la frequenza di queste espressioni, al contrario di come visto in precedenza per le emoticons e gli acronimi, dove si valutava anche la polarità.

3.4.1.4 Punteggiatura

La punteggiatura nei social media non segue le convenzioni ortografiche: Es.: 'Do you hear us now ?????'

Per questa ragione viene contata la frequenza di:

- **Virgole**
- **Punti di domanda**
- **Punti esclamativi**

3.4.1.5 Estrazione particelle pragmatiche

Di seguito viene mostrato un esempio di come le particelle pragmatiche vengono estratte da un messaggio:

```
Tweet: ':-D lmao! RT @KarlDetkenProDJ iPad Humor Steve, I'ma let you finish ,
        but Moses had the greatest tablet of all time ;) clap clap'
Emot (-, +) >>>> [0, 1]
Init (-, +) >>>> [0, 1]
Onom (#)    >>>> [2]
Punct       >>>> {'', ':', '!', '?': 1, '': 0}
```

Listing 3.1: Esempio di tweet processato per estrarre le particelle pragmatiche.

3.4.2 POS (Part Of Speech)

La struttura delle frasi che costituiscono il tweet, può essere un indicatore per il riconoscimento dell'ironia. A questo scopo viene tenuta in considerazione la frequenza dei POS (part-of-speech) tag relativi ai tokens parole presenti in un messaggio.

Per associare i POS tag ad ogni tweet viene utilizzato il POS tagger Noah's ARK [22] specifico per il dominio considerato. Si procede con una tokenizzazione del tweet per poi associare ad ognuno di questi il corretto tag. Di seguito la lista di tags considerati:

Nominale

- **N**: Nome comune
- **O**: Pronome personale
- **^**: Nome proprio
- **S**: Nominale + Possessiva
- **Z**: Nome proprio + Possessiva

Closed-class words

- **D**: Determiner
- **P**: Pre/Post-posizione;
Congiunzione subordinante
- **&**: Congiunzione coordinante
- **T**: Verb particles
- **X**: Existential there

Open-class words

- **V**: Verbo
- **A**: Aggettivo
- **R**: Avverbio
- **!**: Interiezioni

Composti

- **L**: Nominale + Verbo (e.g.: I'm)
- **M**: Nome proprio + Verbo
- **Y**: X + Verbo

Di seguito è mostrato un esempio di POS tags associati al tweet passando per i tokens:

```
Tweet: 'Need a second opinion? Who gave you the first one?'
Tokens >>> [Need], [a], [second], [opinion], [?], [Who],
           [gave], [you], [the], [first], [one], [?]
Tags     >>> ['V', 'D', 'A', 'N', ',', 'O', 'V', 'O', 'D', 'A', '$', ', ', '']
Freq.    >>> {'N': 1, 'O': 2, '^': 0, 'S': 0, 'Z': 0, 'V': 2,
           'A': 2, 'R': 0, '!': 0, 'D': 2, 'P': 0, '&': 0,
           'T': 0, 'X': 0, 'L': 0, 'M': 0, 'Y': 0}
```

Listing 3.2: Esempio di tweet con POS tag associate.

Come si può notare, il POS tagger associa dei tag che non sono stati elencati sopra, i verranno semplicemente ignorati e non utilizzati come features per il modello.

3.4.3 EMOT (Emotional Features)

Alcune teorie psicologiche suggeriscono l'esistenza di emozioni primarie. Per questo motivo sono stati utilizzati i lessici EmoLex e EmoSentNet, in cui vengono riportate le sfere emotive a cui appartengono le parole. I lessici si possono intendere come diversi insiemi che caratterizzano le emozioni primarie, dove ognuno di questi insiemi è composto dalle parole che sono considerate

appartenenti a quella specifica categoria emozionale. Inoltre gli insiemi considerati non devono necessariamente essere disgiunti.

Qualche esempio di seguito:

$$\begin{aligned} birthday &\in JOY \wedge birthday \in SURPRISE \wedge birthday \in ANTICIPATION \\ honor &\in TRUST \\ stair &\notin (JOY \cup SURPRISE \cup ANTICIPATION \cup TRUST) \end{aligned}$$

Come feature aggiuntiva per il riconoscimento dell'ironia, dato un tweet si considerano tutti i tokens validi che rappresentano le parole. Quindi si procede consultando i lessici per ognuna delle parole, così da recuperare le sfere emotive a cui è essa associata. Una volta eseguita l'operazione per tutte queste, si calcola la frequenza totale di ogni emozione all'interno del tweet.

Le due risorse si basano su differenti teorie che identificano ognuna le proprie emozioni primarie, per questo motivo si combina il risultato di entrambe sommando le frequenze ottenute da ognuna delle due risorse.

Emozione	Lessico
Anger	EmoLex, EmoSenticNet
Anticipation	EmoLex
Disgust	EmoLex, EmoSenticNet
Fear	EmoLex, EmoSenticNet
Joy	EmoLex, EmoSenticNet
Sadness	EmoLex, EmoSenticNet
Suprise	EmoLex, EmoSenticNet
Trust	EmoLex

Table 3.3: Emozioni primarie dei lessici.

Un esempio di come vengono estratte le features emozionali da un tweet:

```
Tweet: 'I nominate @MrsStephenFry for a Shorty Award in because... why
not? http://bit.ly/shorty'
Words      >>> ['nominate', 'shorty', 'award']
EmoLex      >>> {'anger': 0, 'anticipation': 1, 'disgust': 0, 'fear': 0,
                'joy': 1, 'sadness': 0, 'surprise': 1, 'trust': 1}
EmoSenticNet >>> {'anger': 0, 'disgust': 0, 'joy': 3, 'sadness': 0,
                'surprise': 0, 'fear': 0}
Combined    >>> {'sadness': 0, 'anticipation': 1, 'fear': 0, 'anger': 0,
                'joy': 4, 'disgust': 0, 'trust': 1, 'surprise': 1}
```

Listing 3.3: Esempio di estrazione delle emozioni da un tweet.

3.4.4 Sintesi features utilizzate

Viene sotto riportata la tabella delle features utilizzate con il relativo numero:

Oltre alle caratteristiche linguistiche, bisogna prendere in considerazione la rappresentazione del testo. Nel caso BOW dipende dalla cardinalità del dizionario. In questo caso, dagli esperimenti è emerso che, dopo l'opportuno filtraggio dei tokens 3.3.1.2, considerando tutto il dataset a disposizione il dizionario contiene 3967 tokens univoci.

Nel caso di BERT invece, le features dipendono dalla pooling strategy adottata. Dal momento che ognuno dei 12 hidden layers della rete neurale fornisce in output un vettore composto da

Caratteristica	#features
Emoticons	2
Acronimi	2
Espressioni onomatopeiche	1
Punteggiatura	3
Part Of Speech	17
Emotional features	8
	33

Table 3.4: Features utilizzate.

768 elementi, e negli esperimenti eseguiti si è scelto di concatenare gli ultimi 4 layers, nella rappresentazione del testo mediante BERT vengono considerate 3072 features.

Chapter 4

Modelli supervisionati

Il machine learning è lo studio di algoritmi che migliorano automaticamente attraverso l'esperienza, ed è sottoinsieme dell'intelligenza artificiale. Il machine learning crea modelli matematici basati sui dati, i quali vengono utilizzati per far apprendere alla macchina, con lo scopo di fare previsioni o decidere riguardo dati che non sono mai stati visti prima.

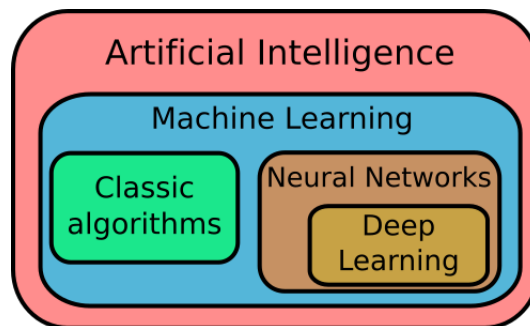


Figure 4.1: Diagramma intelligenza artificiale.¹

Nello specifico i modelli supervisionati fanno parte di quell'insieme di algoritmi "classici" che per apprendere dai dati ed estrarre dei pattern da questi, oltre ai dati di training in input, hanno bisogno delle relative etichette. Questa informazione aggiuntiva viene usata nella fase di apprendimento e indica a priori la relativa classe di appartenenza delle istanze che compongono il dataset. Nel caso dell'ironia si fa riferimento ad algoritmi supervisionati di machine learning per risolvere un problema di classificazione, ovvero dato un tweet classificarlo come "ironico" o "non ironico".

Esistono diversi algoritmi per creare modelli supervisionati, quelli che sono stati utilizzati negli esperimenti sono:

- Decision Tree
- Multinomial Naive Bayes
- Bayesian Network
- Support Vector Machine

¹Fonte: <http://medium.com/@vashkelis/machine-learning-in-5-minutes-9a6cce23ce7e>.

4.1 Decision Tree

Un albero di decisione è un modello predittivo con una struttura ad albero. Ogni nodo interno rappresenta una feature, un arco verso un nodo figlio rappresenta un possibile valore per quella proprietà, ovvero una decisione, e una foglia il valore predetto. Il processo consiste in una sequenza di test, comincia sempre dal nodo radice, e procede verso il basso. A seconda dei valori rilevati in ciascun nodo, si otterrà un path che porterà ad un nodo foglia e quindi una classificazione.

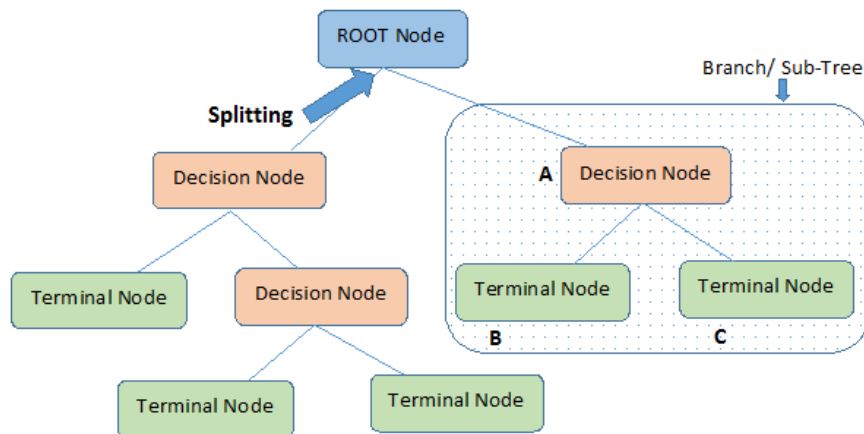


Figure 4.2: Albero di decisione.²

Gli alberi di decisione hanno il vantaggio della semplicità. Infatti è possibile seguire il percorso radice-foglia per analizzare come la macchina è giunta a una determinata decisione basandosi sui dati in input. Questo processo rappresenta di fatto la logica utilizzata dalla macchina per interpretare i dati.

Un problema degli alberi di decisione è che tendono ad andare in overfitting.

²Fonte: <http://medium.com/datadriveninvestor/decision-tree-algorithm-with-hands-on-example-e6c2afb40d38>.

4.2 Naive Bayes

L'algoritmo di classificazione Naive Bayes si basa sull'applicazione del teorema di Bayes:

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}$$

Contestualizzato per la creazione di un modello per un task di classificazione, si ha la tupla composta dalle features di un particolare documento identificato da:

$$\mathbf{x} = (x_1, \dots, x_n)$$

Applicando il teorema di Bayes, rappresentando y come la classe a cui appartiene un individuo, si ottiene:

$$P(y|\mathbf{x}) = \frac{P(\mathbf{x} | y)P(y)}{P(\mathbf{x})}$$

Si noti che la parte destra dell'uguaglianza è la probabilità a priori, ed è stimabile utilizzando il dataset a disposizione. In particolare il numeratore rappresenta la probabilità:

$$P(\mathbf{x} \cap y) = P(x_1, \dots, x_n, y)$$

Viene sviluppata la probabilità congiunta utilizzando la chain rule per l'applicazione ripetuta della probabilità condizionata:

$$\begin{aligned} P(x_1, \dots, x_n, y) &= P(x_1 | x_2, \dots, x_n, y)P(x_2, \dots, x_n, y) \\ &= P(x_1 | x_2, \dots, x_n, y)P(x_2 | x_3, \dots, x_n, y)P(x_3, \dots, x_n, y) \\ &= \dots \\ &= P(x_1 | x_2, \dots, x_n, y)P(x_2 | x_3, \dots, x_n, y) \dots P(x_{n-1} | x_n, y)P(x_n | y)P(y) \end{aligned} \tag{4.1}$$

Il classificatore preso in questione è considerato "naive", in quanto si assume che le features siano indipendenti tra loro, pertanto:

$$P(x_i | x_{i+1}, \dots, x_n, y) = P(x_i | y)$$

Si procede quindi sviluppando l'equazione per il calcolo della probabilità a posteriori:

$$P(y | x_1, x_2, \dots, x_n) = \frac{P(x_1 | y) \dots P(x_n | y)P(y)}{P(x_1) \dots P(x_n)}$$

A questo punto, date le features di un individuo, è possibile calcolare la probabilità che questo appartenga ad una particolare classe y . Inoltre, fissate le features, il denominatore è costante. Si considera quindi:

$$P(y | x_1, x_2, \dots, x_n) \propto P(y) \prod_{i=1}^n P(x_i | y)$$

Un individuo viene classificato come quella classe y che massimizza la probabilità a posteriori:

$$\hat{y} = \arg \max_y P(y) \prod_{i=1}^n P(x_i | y)$$

4.3 Multinomial Naive Bayes

Nell'algoritmo Naive Bayes, avendo dati n -dimensionali e k classi, la probabilità a posteriori si ottiene conoscendo $P(x_i | y_j)$ per ogni $1 \leq n, 1 \leq j \leq k$

Queste probabilità vanno stimate dal dataset, e nella variante Multinomial Naive Bayes si considera una distribuzione binomiale in più variabili (da qui multinomiale). Si hanno quindi i parametri $\theta_y = (\theta_{y1}, \theta_{y2}, \dots, \theta_{yn})$, dove $\theta_{yi} = P(x_i | y)$ vengono stimati in questo modo:

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

Dove, indicando con T il training set:

$$N_{yi} = \sum_{x \in T} x_i \quad N_y = \sum_{i=1}^n N_{yi}$$

Il parametro $\alpha : 0 < \alpha \leq 1$ è chiamato di smoothing, e ha lo scopo di gestire quei casi in cui una features non è presente nel training set, così da non ottenere probabilità uguali a 0.

4.4 Bayesian Network

I classificatori naive Bayes si basano sull'assunzione di indipendenza delle caratteristiche. Ciò permette di semplificare le cose quando si considera l'equazione (4.1) come risultato dell'applicazione della chain rule. Tuttavia questa è un'assunzione piuttosto forte.

A questo scopo i classificatori Bayesian Network introducono una struttura dati per far fronte a questo problema. Viene pertanto considerato un DAG (Directed acyclic graph) in cui le n features compongono i nodi del grafo, e gli archi rappresentano la dipendenza tra le variabili.

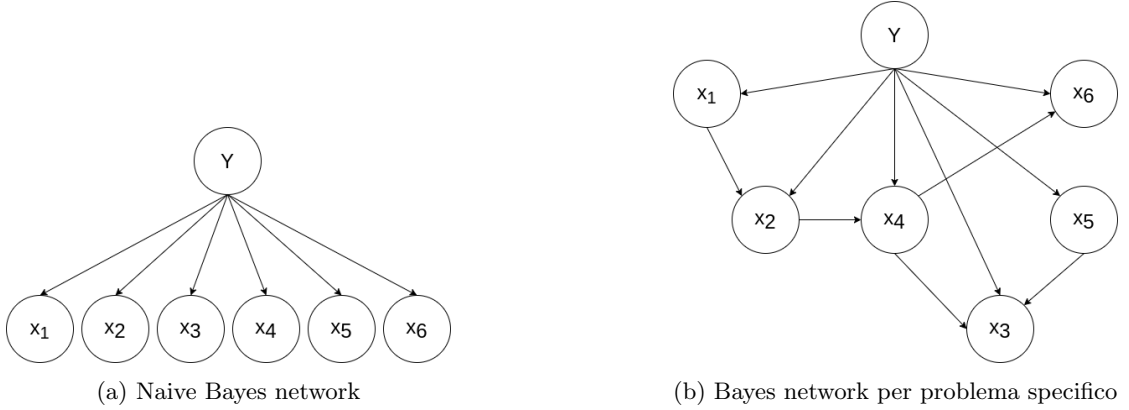


Figure 4.3: Dipendenza tra variabili.

Considerando le n features indipendenti tra loro si ha:

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i)$$

Senza questa assunzione si applica la chain rule per la probabilità congiunta, la quale è computazionalmente onerosa:

$$P\left(\bigcap_{i=1}^n X_i\right) = \prod_{i=1}^n P\left(X_i \mid \bigcap_{j=1}^{i-1} X_j\right)$$

Tuttavia si sfrutta il grafo diretto aciclico G così definito:

- $G = (V, E)$
- $V = \{X_1, \dots, X_n\}$
- $E = \{(U, V) \mid (U, V) \in V^2 \wedge P(U, V) = P(V \mid U)P(U) = P(U \mid V)P(V)\}$
- $E = \{(U, V) \mid (U, V) \in V^2 \wedge P(U, V) \neq P(U)P(V)\}$

Il quale rappresenta le dipendenze tra variabili. Quindi nello sviluppo della probabilità congiunta, si considerano solo quei fattori con un arco presente nel grafo. Pertanto, indicando con $\pi(u) = \{v \mid (v, u) \in E\}$, $v, u \in V$ si ha:

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i \mid X_{i+1}, \dots, X_n) = \prod_{i=1}^n P(X_i \mid \pi(X_i))$$

Così facendo viene semplificata la probabilità congiunta, pur non assumendo la mutua indipendenza tra le variabili.

4.5 Support Vector Machine

Support vector machine è un'altro algoritmo di machine learning supervisionato, in cui gli n individui sono rappresentati nella forma:

$$(\vec{x}_1, y_1), \dots, (\vec{x}_n, y_n)$$

Dove $y_i \in \{1, -1\}$ rappresenta la classe di appartenenza di un individuo. Ogni \vec{x}_i è un vettore reale k -dimensionale, con tante dimensioni quante sono le features.

4.5.1 Iperpiano

L'obiettivo dell'algoritmo è trovare l'iperpiano ottimo che che meglio separa le due classi. Nel caso 2-dimensionale, l'iperpiano è individuato da una retta, tuttavia è facilmente generalizzare il concetto al caso 3-dimensionale e così via:

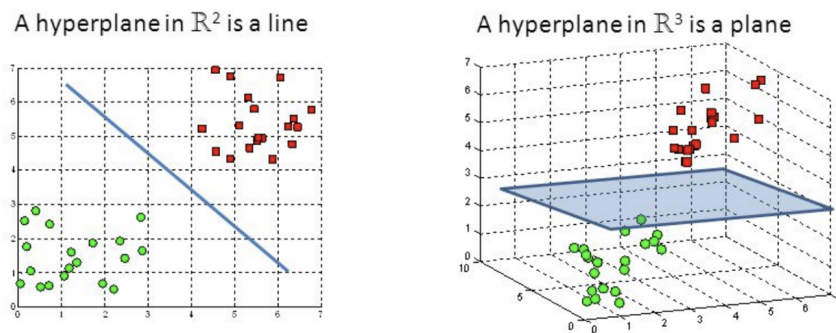


Figure 4.4: Iperpiano ottimo.³

4.5.2 Iperpiano ottimo

Per iperpiano ottimo si intende quello che massimizza la distanza tra l'iperpiano e il più vicino punto \vec{x}_i di ciascuna classe. Indicando con \vec{w} il vettore normale all'iperpiano, ogni iperpiano può essere scritto come l'insieme dei punti \vec{x} che soddisfano l'equazione:

$$\vec{w} \cdot \vec{x} - b = 0$$

Vengono considerati i due iperpiani paralleli che separano le due classi, la quale distanza tra i due è la più grande possibile. La regione compresa tra questi è detta margine, e l'iperpiano ottimo si trova nel mezzo tra i due. I due iperpiani possono essere scritti come:

$$\vec{w} \cdot \vec{x} - b = 1$$

$$\vec{w} \cdot \vec{x} - b = -1$$

Essendo \vec{w} il vettore normale all'iperpiano, la distanza tra i due è $\frac{2}{\|\vec{w}\|}$, pertanto per massimizzare la distanza si vuole minimizzare $\|\vec{w}\|$.

³Fonte: www.towardsdatascience.com/a-friendly-introduction-to-support-vector-machines-svm-925b68c5a079.

Si considerano inoltre i vincoli per esprimere che ogni punto deve cadere nel lato corretto rispetto al margine:

$$\begin{aligned}\vec{w} \cdot \vec{x} - b &\geq 1, \text{ se } y_i = 1 \\ \vec{w} \cdot \vec{x} - b &\leq -1, \text{ se } y_i = -1\end{aligned}$$

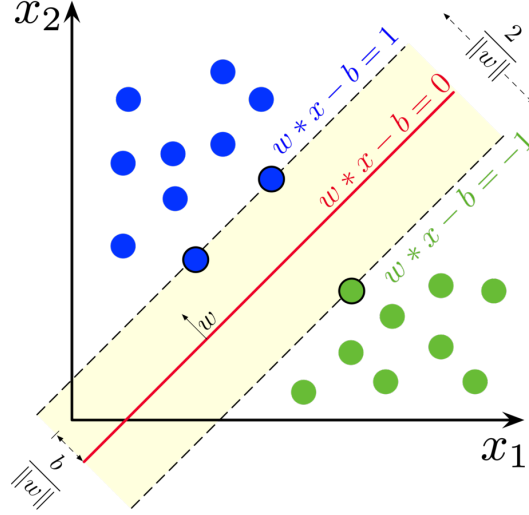


Figure 4.5: Margine SVM.⁴

4.5.3 Vettori di supporto

I due iperpiani che individuano il margine, sono trovati a partire dai vettori che appartengono al dataset. Quei punti che permettono di individuare le frontiere del margine, ovvero i due iperpiani paralleli, sono i vettori di supporto. Questi punti del dataset sono fondamentali, in quanto per costruzione cadono sul bordo del margine, pertanto lo spostamento o la rimozione di uno di essi, avrebbe come risultato due iperpiani diversi che definiscono il margine, quindi un differente iperpiano ottimo.

4.5.4 SVM non lineare

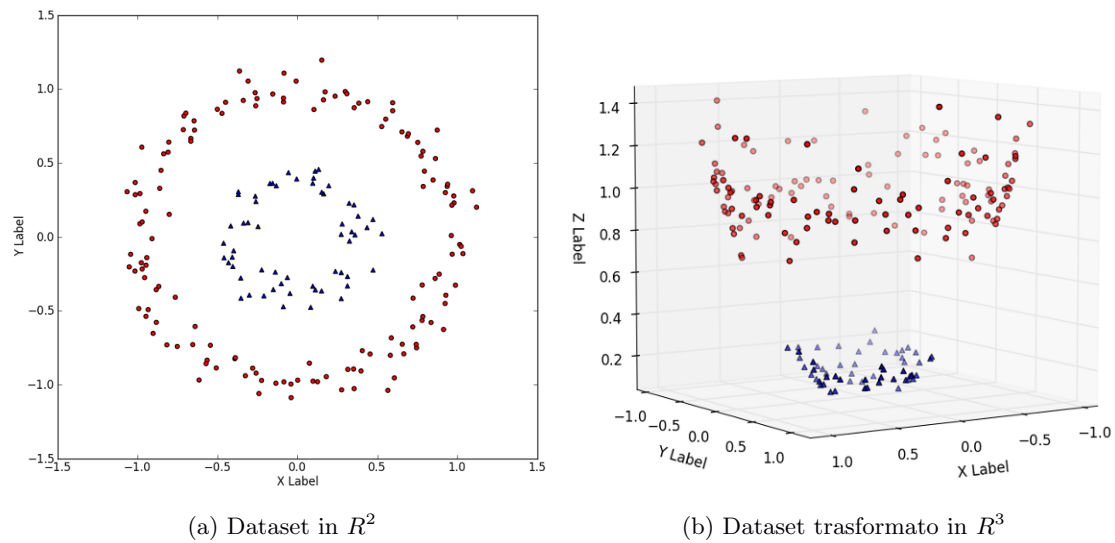
Fino ad ora sono stati presi in considerazione solo dataset linearmente separabili. Tuttavia questo può non essere sempre possibile. L'idea è quindi di mappare i dati in uno spazio con una dimensione in più, con lo scopo di ottenere la separazione lineare utilizzando un iperpiano nello spazio con una dimensione aggiuntiva.

Sotto viene preso come esempio un dataset in R^2 non separabile utilizzando un iperpiano. Si considerano quindi la trasformazione:

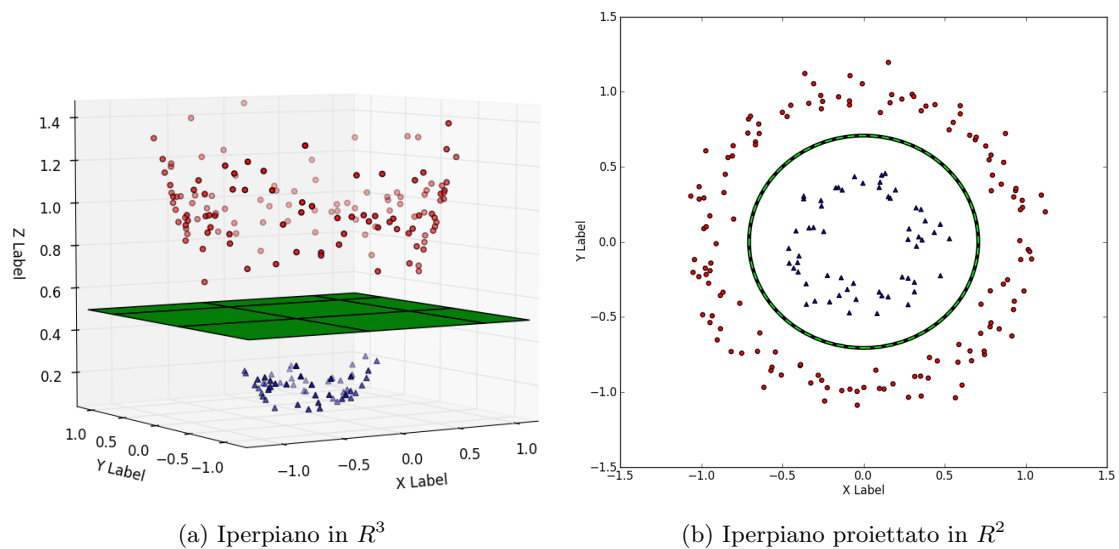
$$\begin{aligned}\phi: R^2 &\rightarrow R^3 \\ \phi(\mathbf{X}) &= [x_1, x_2, x_1^2 + x_2^2]^T\end{aligned}$$

Utilizzando tale trasformazione, il dataset di partenza viene mappato in uno spazio con una dimensione in più:

⁴Fonte: www.wikipedia.org/wiki/File:SVM_margin.png.

Figure 4.6: Aggiunta di una nuova dimensione.⁵

Dopo la trasformazione, come si può notare dalla figura 4.6b, è possibile trovare un iperpiano in R^3 che ben separa le due classi:

Figure 4.7: Aggiunta di una nuova dimensione.⁶

Nella figura 4.7b, è riportata la linea di decisione per le due classi, ovvero l'iperpiano trovato in R^3 e poi proiettato in R^2 .

⁵Fonte: www.medium.com/@vivek.yadav/why-is-gradient-descent-robust-to-non-linearly-separable-data-a50c543e8f4a.

⁶Fonte: <https://www.eric-kim.net/eric-kim-net/posts/1/imgs/data.2d.to.3d.hyperplane.png>.

4.6 Strumenti utilizzati

Per l'implementazione degli algoritmi, sono stati utilizzati degli strumenti che mettono a disposizione la possibilità di effettuare il training dei modelli.

4.6.1 Scikit-learn

Scikit-learn [23] è un modulo di python open source per il machine learning. Utilizzando questo modulo, è stato scritto il codice python per creare i modelli con gli algoritmi:

- Decision Tree
- Multinomial Naive Bayes
- Support Vector Machine - Kernel lineare

Inoltre per ogni algoritmo sono stati creati più modelli considerando diverse l'insieme potenza delle features. Quindi sono stati eseguiti i seguenti macrostep:

1. Lettura del dataset
2. Shuffle del dataset con lo stesso seed per ogni modello
3. Standardizzare del dataset
4. Creazione di 10 folds in cui viene effettuato lo split del dataset tra training-set e test-set
5. Training dei modelli per ogni folds
6. Valutazione dei classificatori per ogni folds
7. Media delle misure di performance rispetto ogni folds
8. Dump del report così da poter essere successivamente messo a confronto con altri modelli

Essendo scikit-learn un modulo per python, che mette a disposizione delle API, è stato possibile gestire manualmente la politica per il training.

Infatti per questioni di ottimizzazione, fissato un algoritmo e le features per un determinato modello, è stato scelto di creare k processi eseguiti in parallelo, ognuno dei quali si prende in carico il training del modello su un particolare fold dei k considerati. Avendo scelto per gli esperimenti 10 folds, ed utilizzato una macchina con 8 cores, ciò si presta bene sfruttare ognuno degli 8 cores, con il risultato di ridurre il tempo totale impiegato per il training dei diversi modelli, dovendo considerare diverse features per ogni modello e i molteplici algoritmi di classificazione.

4.6.2 Weka

Weka [24] è un software open source per il machine learning. Questo applicativo può essere utilizzato attraverso interfaccia grafica, ed è in generale più immediato rispetto a scrivere del codice in python che utilizza le API fornite dal modulo scikit-learn. Con questo software sono stati creati i modelli utilizzando l'algoritmo:

- Bayesian network

Non è stato utilizzato scikit-learn in quanto questo algoritmo non è implementato nel modulo.

Essendo weka un software che fornisce un'interfaccia grafica, ha il vantaggio di essere più immediato rispetto a scikit-learn, tuttavia non essendo possibile implementare manualmente lo script per il training, in questo caso non vengono sfruttati tutti i cores per il training dei modelli.

Chapter 5

Campagna sperimentale

5.1 Dataset

Il dataset a disposizione è il corpora *TwReyes2013*, composto da 40,000 tweets in lingua inglese. Trattandosi di un approccio supervisionato, è richiesto che il dataset sia etichettato. Per associare una specifica label ai messaggi generati dagli utenti si possono seguire due strade:

- **Self-Tagging:** Twitter mette a disposizione l'utilizzo degli hashtag nei messaggi. Assumendo che un utente utilizzi l'hashtag *#irony* con la volontà di esprimere ironia, è facile collezionare una serie di tweet etichettati come ironici.
- **Crowdsourcing:** I vari tweet vengono etichettati manualmente da alcune persone

Nel caso specifico, viene utilizzato il dataset *TwReyes2013*, composto da 40,000 tweet in lingua inglese, accumulati usando la tecnica self-tagging. Vengono quindi considerati 4 hashtag diversi:

Numero	Hashtag	Label associata
10,000	<i>#irony</i>	ironico
10,000	<i>#education</i>	non ironico
10,000	<i>#humor</i>	non ironico
10,000	<i>#politics</i>	non ironico

Table 5.1: Hashtag e label associate

5.2 10-Folds Cross-Validation

Per la fase di training dei modelli è necessario suddividere il dataset a disposizione in due diversi insiemi: training-set e test-set. Questo è un passaggio fondamentale in quanto non si otterrebbero delle misure di performance valide testando i modelli su dati che sono stati utilizzati per la fase di training. In pratica si vuole testare il modello su dati che non sono mai stati visti in precedenza.

Prima di iniziare la fase di training, i dati in input, che sono composti dalla matrice contenente le features estratte, vengono opportunamente mescolati. Con questo si intende che le righe della matrice vengono casualmente cambiate di posto, chiaramente mantenendo coerenza con le labels

associate. Un aspetto importante è che lo shuffle della matrice è eseguito con lo stesso seed per tutti i modelli, così da ottenere delle misure di performance statisticamente valide.

Le performance ottenute, dipendono anche da come è stato diviso il dataset tra training-set e test-set. Per questo motivo viene utilizzata la tecnica k-Folds Cross-Validation, che permette di sfruttare diverse porzioni di dati per i due insiemi. Il vantaggio di questa tecnica è quello di ottenere valutazioni più affidabile mediando i risultati di ognuno dei k modelli di cui è stato fatto il training, a discapito del tempo necessario, in quanto si effettuerà k volte il training di un modello.

Per gli esperimenti effettuati, è stato scelto $k = 10$, ovvero il dataset è stato diviso 10 volte, utilizzando ad ogni iterazioni una porzione di dati diversa per il training-set e test-set. Avendo scelto questo valore di k , la proporzione tra i due insiemi sarà sempre 90% per il training-set e 10% per il test-set.

Nel seguente diagramma si possono considerare i pallini verdi e rossi come i tweets ironici e non ironici. Quindi, fissata un'istanza del dataset mescolato, si considera una sliding window di dimensione $\#samples \cdot \frac{1}{k}$, dove per ogni k iterazione viene effettuato il training di un nuovo modello.



Figure 5.1: k-Folds Cross-Validation.¹

5.3 Misure di performance

Essendo stati creati diversi modelli, è necessario valutarli. Avendo diviso il dataset a disposizione tra test-set e training-set, una volta completata la fase di training è possibile utilizzare il test-set per la valutazione. Si procede prendendo un tweet e facendolo classificare al modello, il quale verrà quindi identificato come ironico o non ironico. Trattandosi di modelli supervisionati, oltre ai dati si hanno a disposizione le labels associate, è quindi possibile confrontare l'output fornito dal modello con la reale classe di un tweet. Da qui si può capire se per quel particolare individuo, la classificazione è corretta o meno.

Per la valutazione si utilizza tutto il test-set a disposizione e viene contano il numero di:

- **TP** (True positive): Un tweet ironico è classificato come ironico - *corretto*

¹Fonte: [www.wikipedia.org/wiki/Cross-validation_\(statistics\)](http://www.wikipedia.org/wiki/Cross-validation_(statistics)).

- **TN** (True negative): Un tweet non ironico è classificato come non ironico - *corretto*
- **FP** (False positive): Un tweet non ironico è classificato come ironico - *errato*
- **FN** (False negative): Un tweet ironico è classificato come non ironico - *errato*

Si noti che un modello può sbagliare la classificazione in due modi diversi, e c'è differenza tra classificare un tweet ironico come non ironico piuttosto che uno non ironico come ironico. Questi valori verranno poi analizzati sfruttando una matrice di confusione:

		Truth	
		Condition positive	Condition negative
Prediction	Predicted condition positive	True positive	False positive
	Predicted condition negative	False negative	True negative

Figure 5.2: Confusion matrix.²

Basandosi su questi dati vengono definite delle misure di performance per la valutazione:

- **Accuracy:** Il rapporto tra il numero di classificazioni corrette e le classificazioni totali.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

- **Precision:** La porzione di classificazioni positive corrette.

$$Precision = \frac{TP}{TP + FP}$$

- **Recall:** La porzione di sample positivi identificati.

$$Recall = \frac{TP}{TP + FN}$$

- **F-measure:** Metrica per calcolare la valutazione complessiva. Formalmente è definita come la media armonica tra recall e precision.

$$F-measure = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

²Fonte: www.radiopaedia.org/cases/confusion-matrix-3.

5.4 Esperimenti

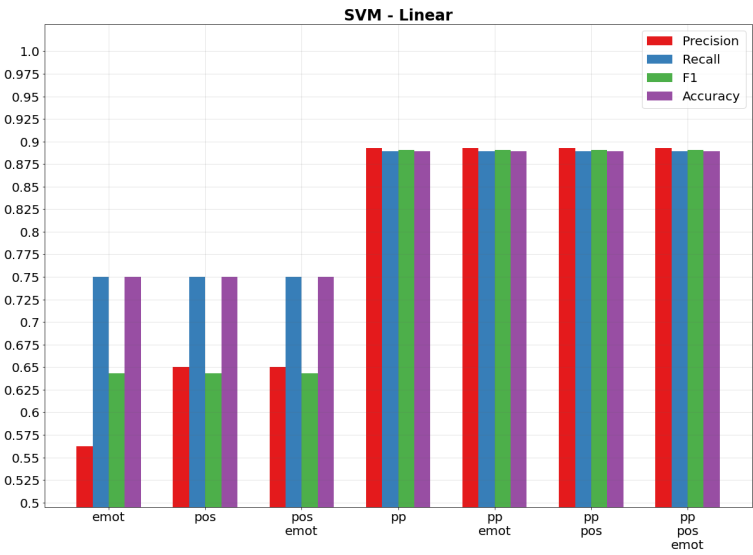
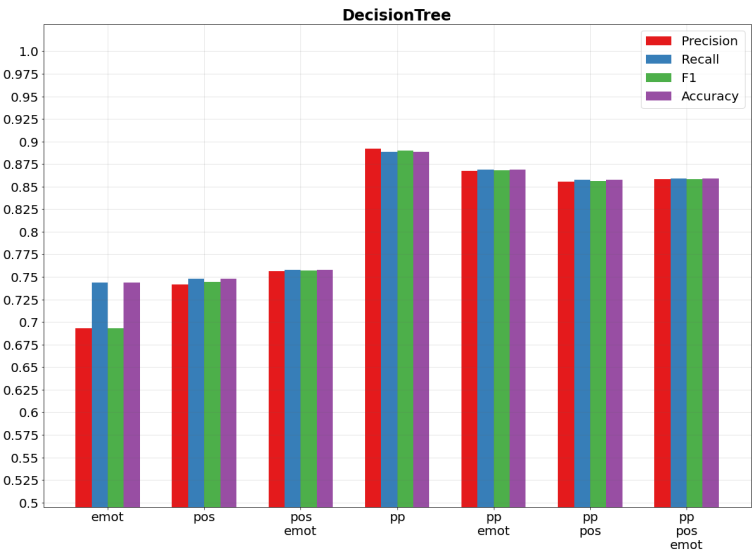
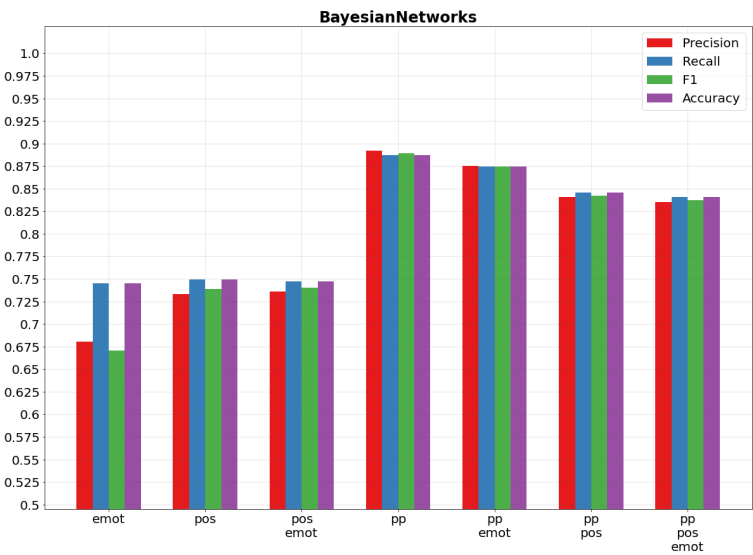
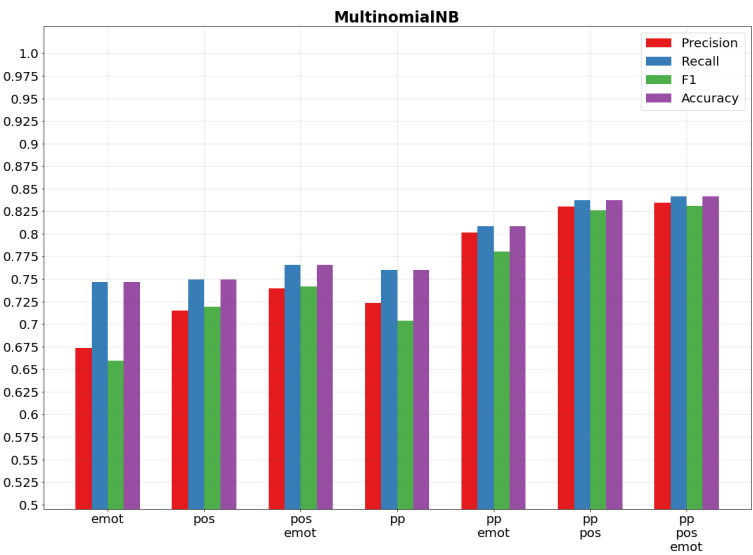
Avendo creato diversi modelli, questi sono stati confrontati per valutare il più performante. Nello specifico sono stati creati i modelli con gli algoritmi di classificazione:

- Decision Tree
- Multinomial Naive Bayes
- Bayesian Network
- Support Vector Machine - Lineare

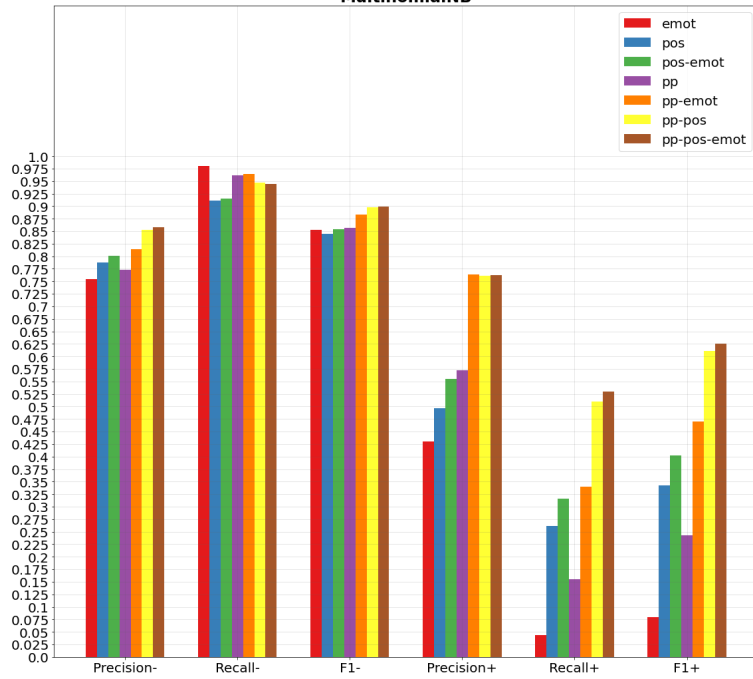
Inoltre sia $F = \{text, pp, pos, emot\}$ l'insieme delle features considerate. Per ogni algoritmo di classificazione, sono stati creati i modelli utilizzando come features $\mathcal{P}(F)$.

5.4.1 Caratteristiche linguistiche

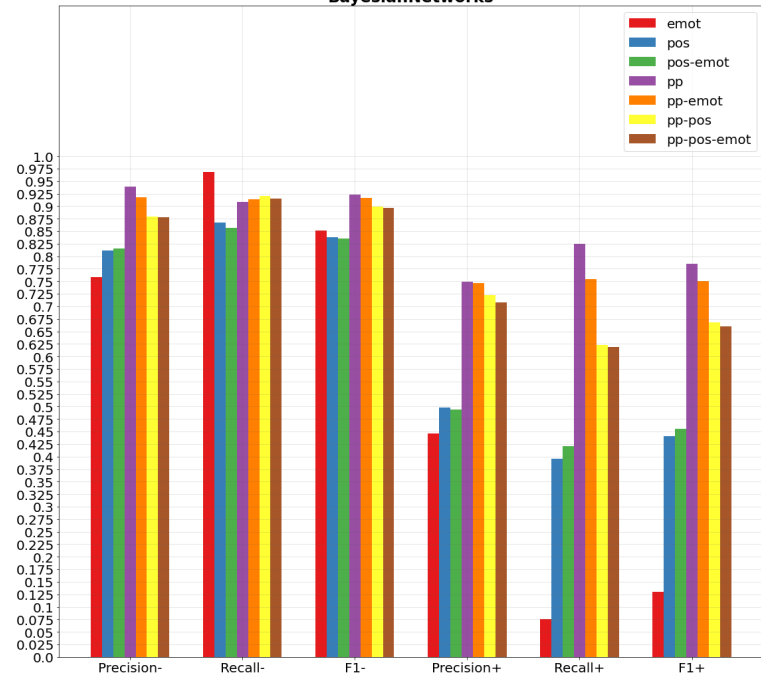
5.4.1.1 Weighted average



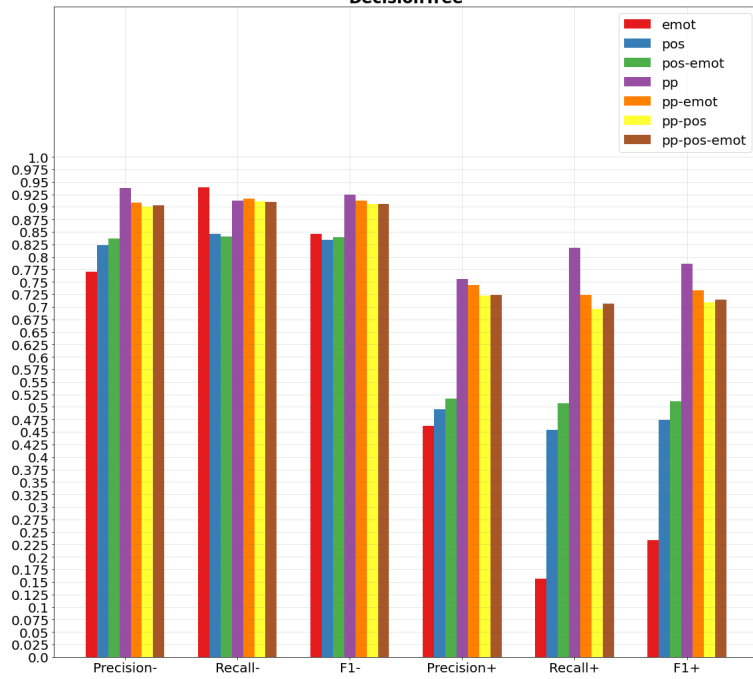
MultinomialNB



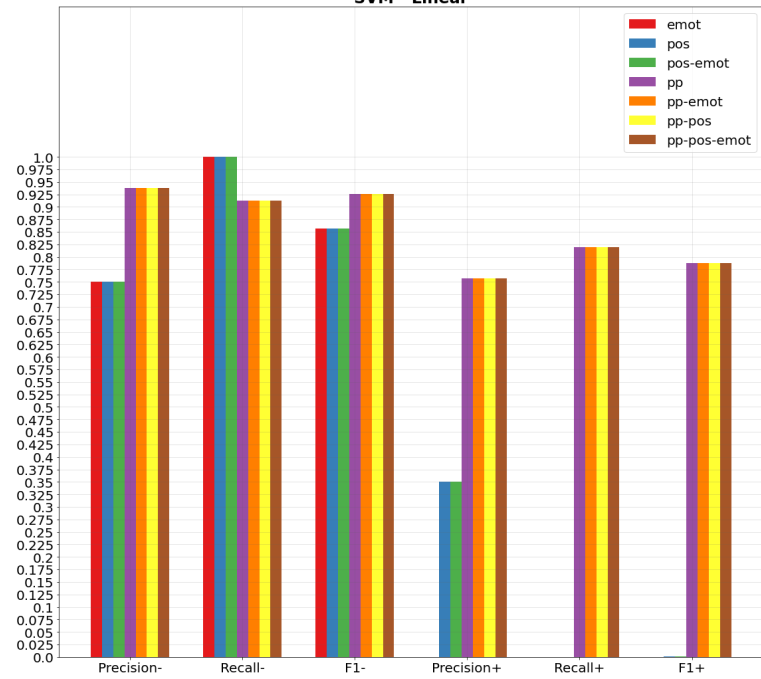
BayesianNetworks



DecisionTree



SVM - Linear

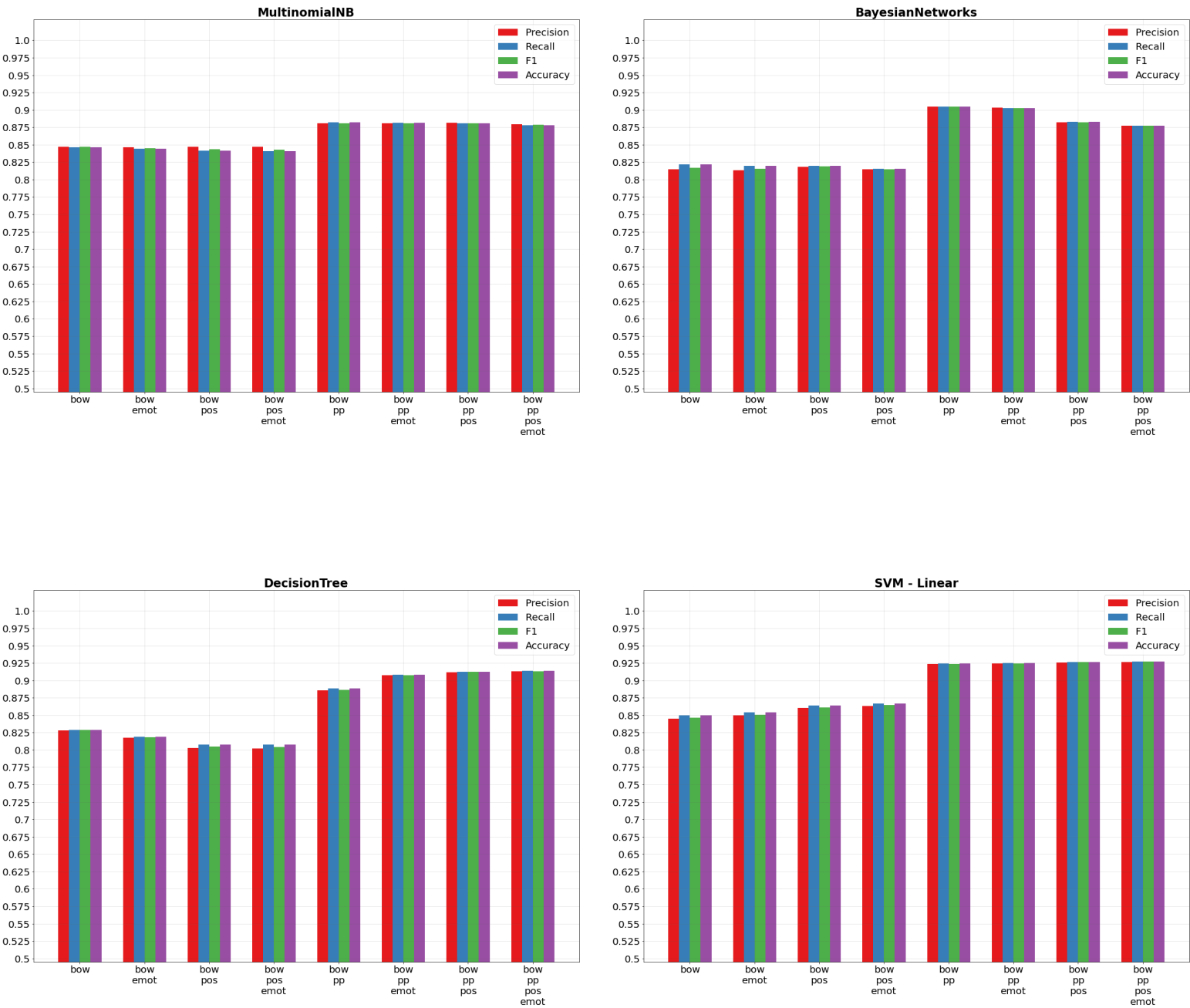


5.4.1.2 Confusion matrix

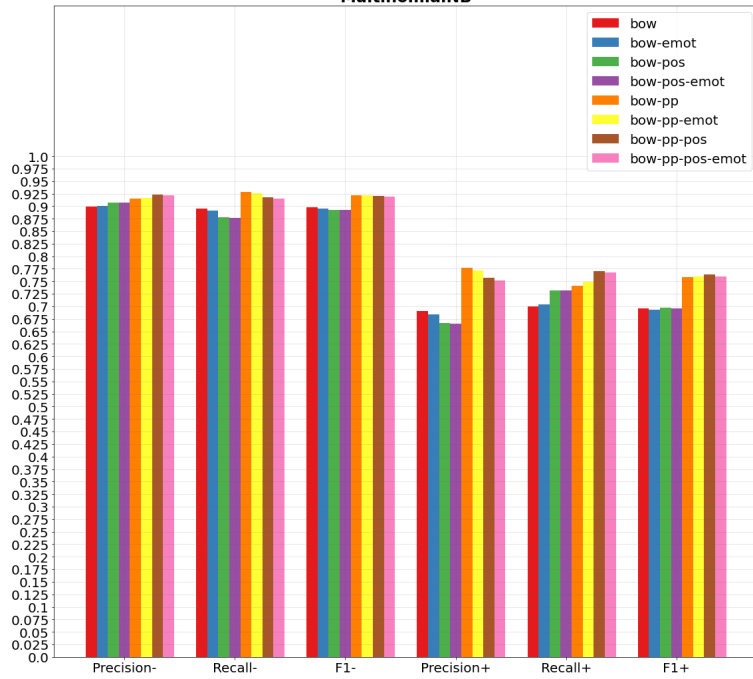


5.4.2 BOW + Caratteristiche linguistiche

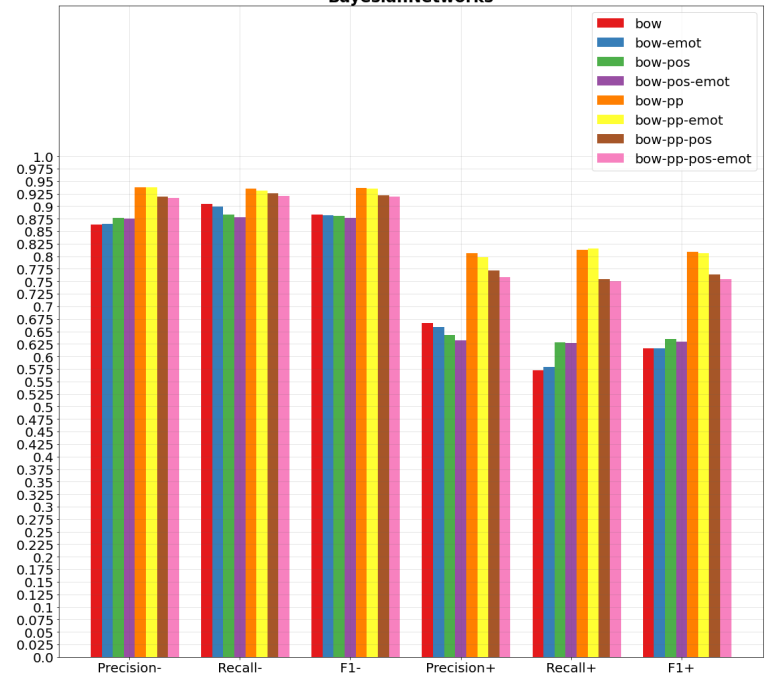
5.4.2.1 Weighted average



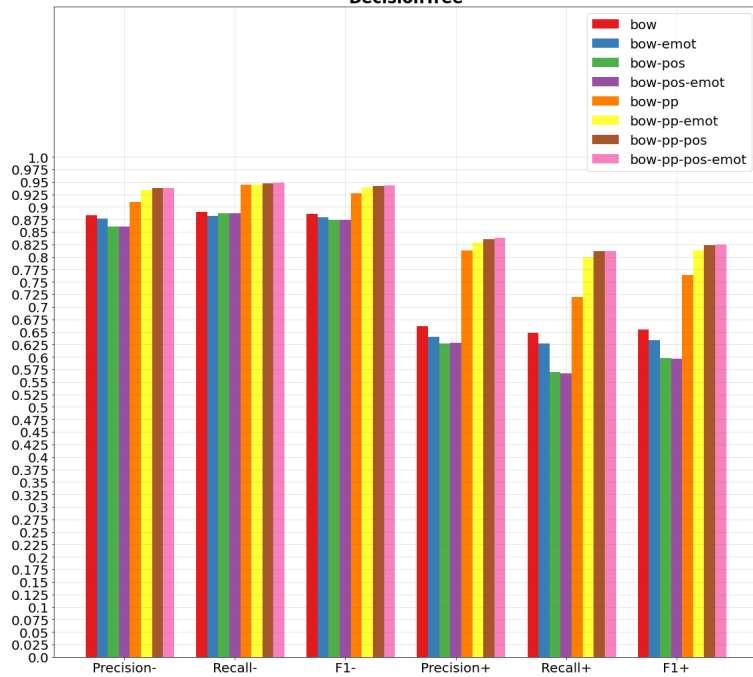
MultinomialNB



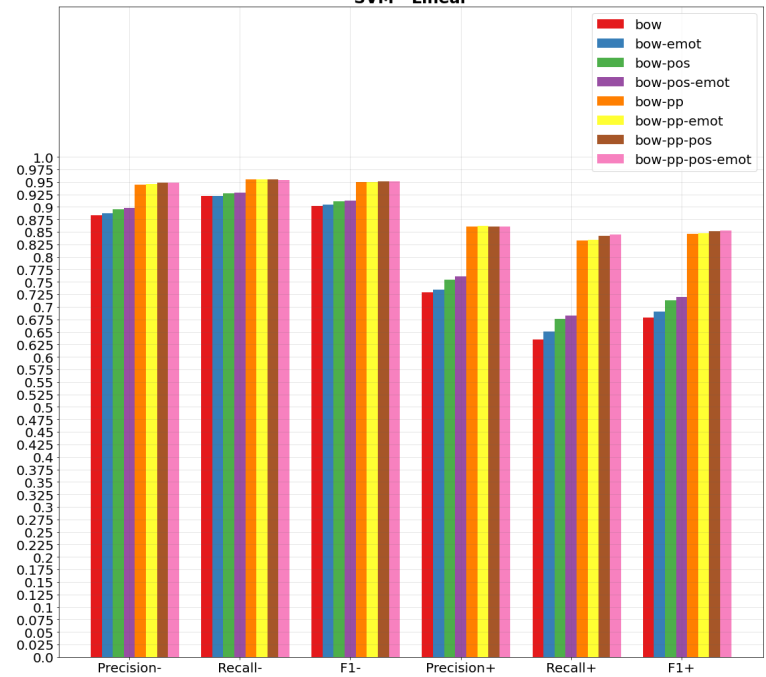
BayesianNetworks



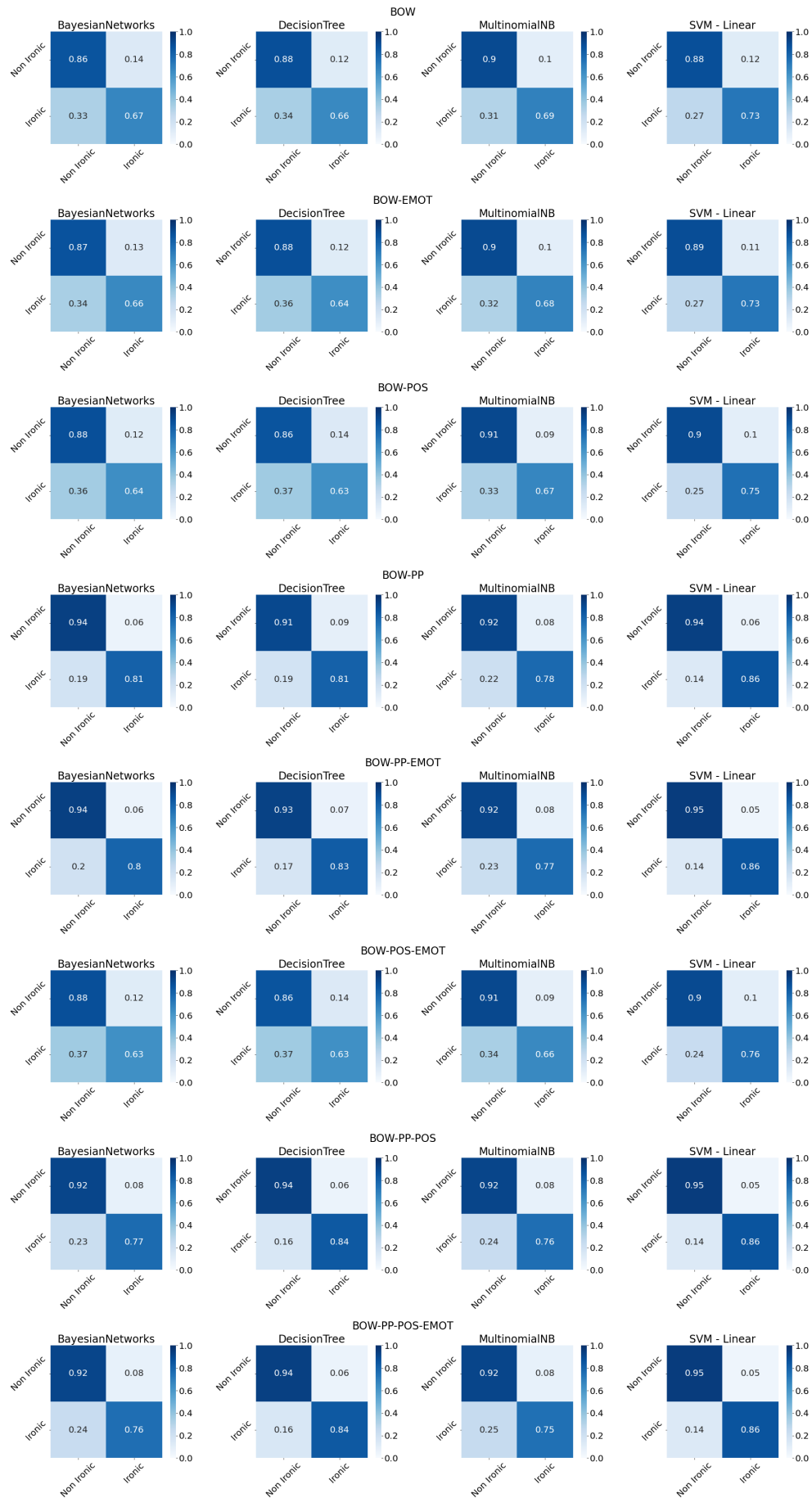
DecisionTree



SVM - Linear

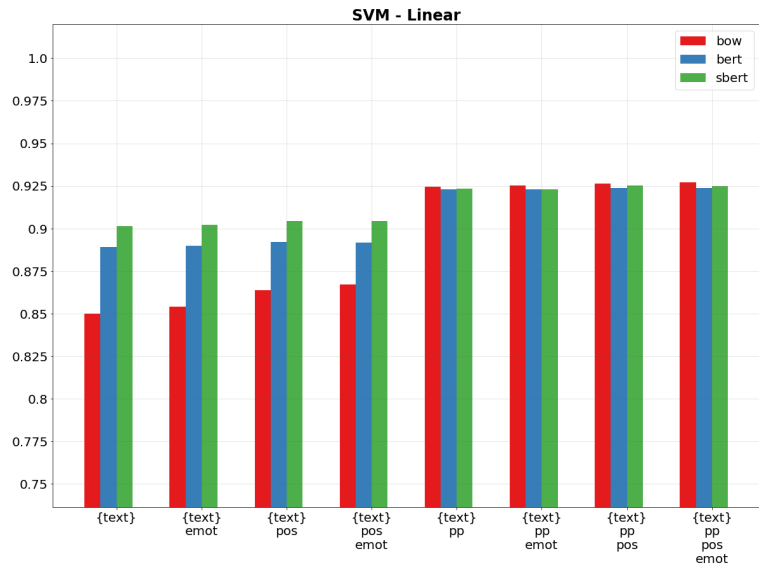
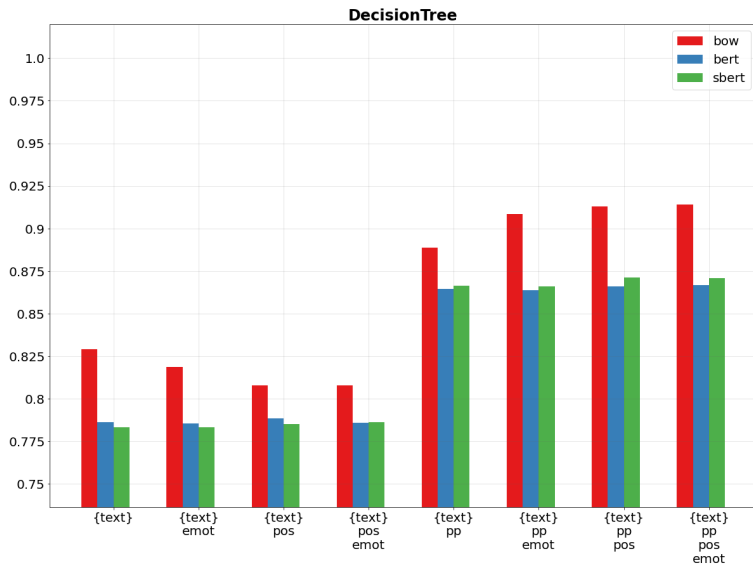


5.4.2.2 Confusion matrix

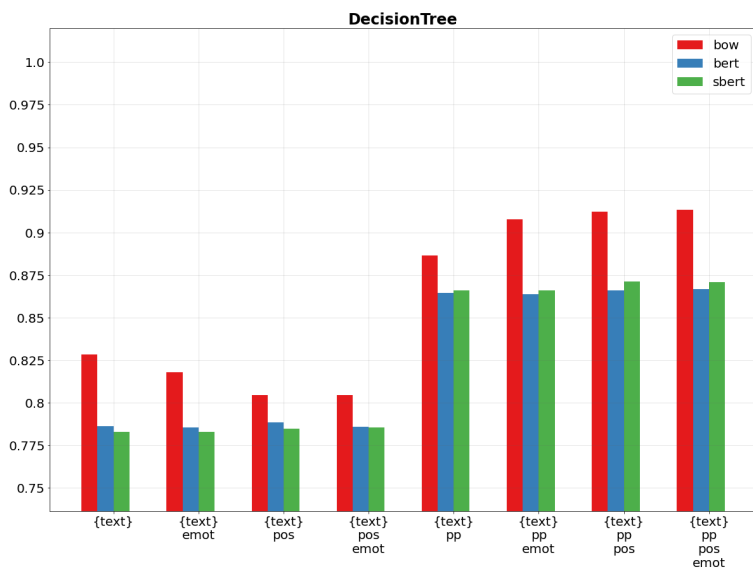


5.4.3 BOW vs BERT vs Sentence-BERT

5.4.3.1 Weighted average - Accuracy



5.4.3.2 Weighted average - F measure



5.4.4 Analisi lessico con PCA

La principal component analysis (PCA) è una tecnica per la semplificazione dei dati, il cui scopo è ridurre il numero di variabili che descrivono il dataset, limitando il più possibile la perdita di informazioni.

Per limitare la perdita di informazioni, si considerano solo quelle variabili con la quantità di varianza spiegata più alta, ovvero quelle che meglio differenziano il dataset, per l'appunto le componenti principali. Di fatto la riduzione delle dimensioni sfrutta questo aspetto, ma nel caso considerato si terrà conto solo del contributo di ogni variabile, senza realmente andare a ridurre le dimensioni del dataset.

A questo scopo si procede standardizzando il dataset, quindi viene calcolata la matrice di covarianza $n \times n$ dove n corrisponde al numero di variabili. Dopodiché si calcolano gli autovalori associati agli autovettori della matrice di covarianza. Come risultato si ottengono le componenti principali, nuove variabili costruite come combinazione lineare delle variabili di partenza. La percentuale di varianza spiegata da ogni componente principale si calcola a partire dagli autovalori:

$$\%PC_k = \frac{\lambda_k}{\sum_{i=1}^n \lambda_i}$$

Questa tecnica viene qui utilizzata per analizzare il lessico e cercare un'eventuale associazione tra parole e ironia. Si considera quindi il dataset ironico e non ironico separatamente, e si procede generando gli embeddings tramite BERT. Si ottiene quindi la matrice M composta dagli embeddings sulle colonne, e le diverse parole che occorrono nel dataset sulle righe. Dal momento che si vuole cercare un'associazione tra le parole e l'ironia, si considera la matrice M^T e si effettua PCA, così da valutare se alcune parole contribuiscono maggiormente relativamente al dataset ironico.

5.4.4.1 Top k componenti principali

Per ogni componente principale, viene calcolata la varianza spiegata. Di seguito viene mostrata la varianza spiegata dalle prime 50 componenti principali (la varianza spiegata totale è 1):

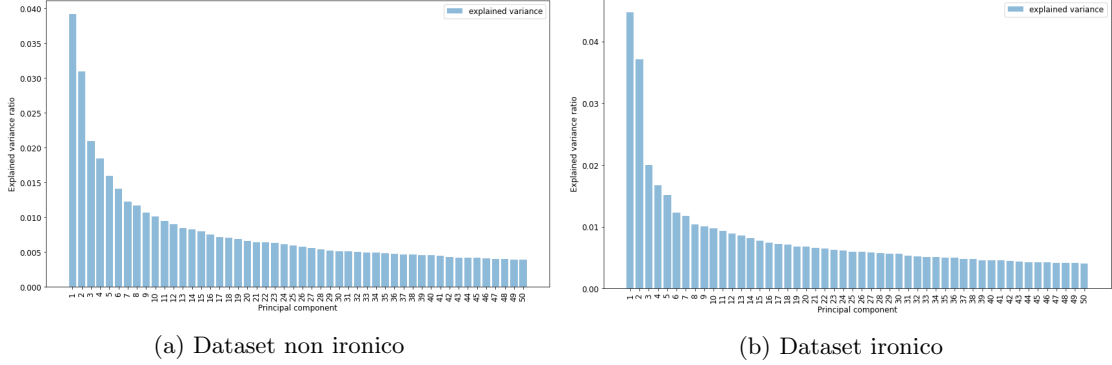


Figure 5.13: Varianza spiegata

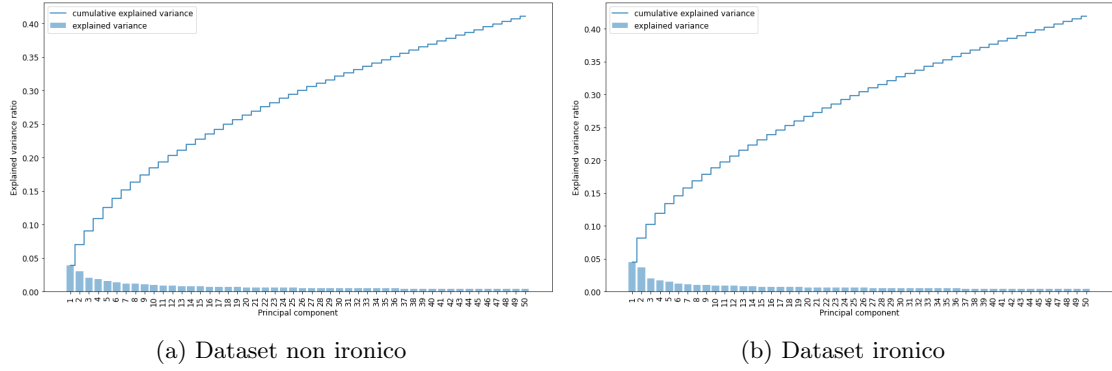


Figure 5.14: Varianza spiegata cumulata

5.4.4.2 Features con contributo maggiore

Per calcolare le features più importanti, si considerano gli autovettori associati agli autovalori, si procede quindi sommando tutti gli elementi del vettore in valore assoluto. Il risultato è il contributo espresso da quello feature rispetto a tutte le componenti principali (maggiore il valore, più alta l'importanza) [25]. Vengono quindi considerate le 50 features con il contributo più alto (il contributo totale è 1):

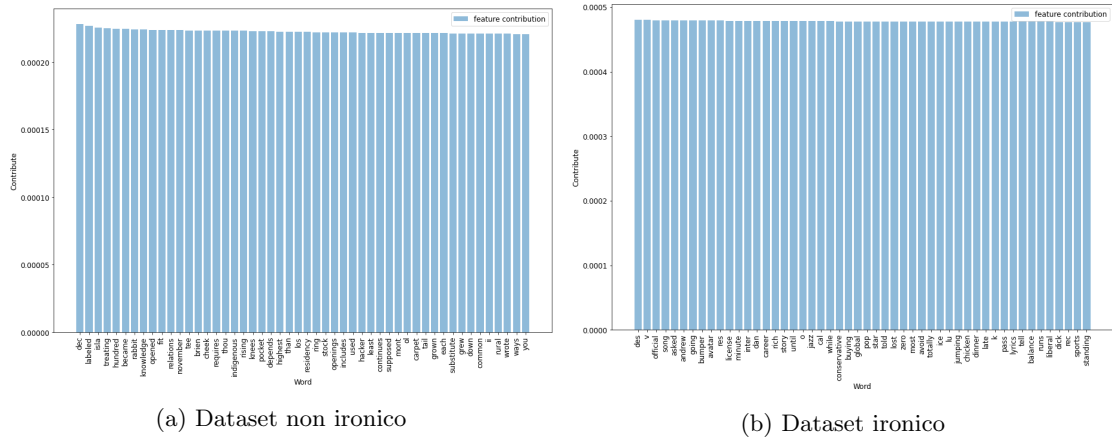


Figure 5.15: Contributo dell features

5.4.4.3 Conclusioni

Come si può notare da queste analisi, pur considerando le prime 50 componenti principali, viene spiegata solo il 40% circa della varianza totale. Inoltre per quanto riguarda il contributo di ogni fetures, nessuna di queste prevale tra le altre. Questi ragionamenti data-driven mostrano come non ci sia una chiara associazione tra lessico e ironia. In effetti il risultato è coerente con ciò che ci si aspetta, per riconoscere l'ironia non è sufficiente basarsi solo sul lessico utilizzato in una frase.

Chapter 6

Conclusioni e sviluppi futuri

I sistemi di sentiment analysis vengono utilizzati per estrarre l'informazione relativa al sentimento e l'opinione espressa in un testo. Nel linguaggio naturale esistono figure retoriche come l'ironia che sono in grado di invertire completamente la polarità di ciò che si esprime, di conseguenza i sistemi di sentiment analysis vengono "ingannati".

In questa tesi sono stati presentati diversi modelli supervisionati di machine learning per il riconoscimento dell'ironia nei tweets. A questo scopo è stato necessario esprimere il linguaggio naturale tramite una rappresentazione matematica, quindi prima con la tecnica BOW e poi mediante l'uso del transformer BERT.

Tra le features utilizzate, è stata prestata particolare attenzione alla componente affettiva. Infatti tramite l'utilizzo di lessici, viene tenuta in considerazione la sfera emotiva di un particolare token presente nel documento. Inoltre nei social network viene fatto largo uso delle emoticons, le quali esprimono una grande componente emotiva. Altri elementi come la punteggiatura, che nei social network non segue convenzioni ortografiche, può avere un carico emotivo.

Trattandosi di modelli supervisionati, è stato utilizzato il dataset *TwReyes2013* in cui tramite la tecnica self-tagging, i vari tweets vengono etichettati dagli stessi utenti facendo uso degli hashtag.

Nella fase di sperimentazione sono stati analizzati e testati diversi algoritmi di apprendimento, così da creare più modelli con i vari algoritmi e considerando per ognuno di questi un insieme di features diverse. I vari modelli sono stati messi a confronto, valutando le performance di ognuno in base a diverse metriche.

Come passo ulteriore è possibile meglio approfondire l'uso di BERT. Negli esperimenti è stato utilizzato il modello pre-trained, mentre si potrebbe utilizzare una versione fine-tuned di BERT ed usare gli embeddings di questo nuovo modello come features testuali.

Bibliography

- [1] B. Wallace, “Computational irony: A survey and new perspectives,” *Artificial Intelligence Review*, vol. 43, 04 2013.
- [2] J. aboobaker and E. Ilavarasan, “A survey on sarcasm detection and challenges,” in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pp. 1234–1240, 2020.
- [3] D. K. Tayal, S. Yadav, K. Gupta, B. Rajput, and K. Kumari, “Polarity detection of sarcastic political tweets,” in *2014 International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 625–628, 2014.
- [4] P. Chaudhari and C. Chandankhede, “Literature survey of sarcasm detection,” in *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pp. 2041–2046, 2017.
- [5] J. Tepperman, D. R. Traum, and S. S. Narayanan, “”yeah right”: Sarcasm recognition for spoken dialogue systems,” in *INTERSPEECH*, 2006.
- [6] A. Joshi, P. Bhattacharyya, and M. J. Carman, “Automatic sarcasm detection: A survey,” *ACM Comput. Surv.*, vol. 50, Sept. 2017.
- [7] A. Rajadesingan, R. Zafarani, and H. Liu, “Sarcasm detection on twitter: A behavioral modeling approach,” in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM ’15*, (New York, NY, USA), p. 97–106, Association for Computing Machinery, 2015.
- [8] R. Bagate and S. Ramadass, *Different Approaches in Sarcasm Detection: A Survey*, pp. 425–433. 01 2020.
- [9] E. Camp, “Sarcasm, pretense, and the semantics/pragmatics distinction*,” *Noûs*, vol. 46, 09 2011.
- [10] F. Barbieri, F. Ronzano, and H. Saggion, “Italian irony detection in twitter: a first approach,” 12 2014.
- [11] S. K. Bharti, K. S. Babu, and S. K. Jena, “Parsing-based sarcasm sentiment recognition in twitter data,” in *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 1373–1380, 2015.
- [12] D. Davidov, O. Tsur, and A. Rappoport, “Semi-supervised recognition of sarcastic sentences in twitter and amazon,” in *Proceedings of the Fourteenth Conference on Computational Natural Language Learning, CoNLL ’10*, (USA), p. 107–116, Association for Computational Linguistics, 2010.

- [13] D. I. H. Fariás, V. Patti, and P. Rosso, “Irony detection in twitter: The role of affective content,” *ACM Trans. Internet Technol.*, vol. 16, July 2016.
- [14] E. Fersini, V. Messina, and F. Pozzi, “Detecting irony and sarcasm in microblogs: The role of expressive signals and ensemble classifiers,” 10 2015.
- [15] A. Joshi, V. Tripathi, K. Patel, P. Bhattacharyya, and M. Carman, “Are word embedding-based features useful for sarcasm detection?,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, (Austin, Texas), pp. 1006–1011, Association for Computational Linguistics, Nov. 2016.
- [16] D. Bharti, K. Babu, and S. Jena, “Parsing-based sarcasm sentiment recognition in twitter data,” 08 2015.
- [17] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O’Reilly Media, 2009.
- [18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 06 2017.
- [19] J. Alamm, “The illustrated transformer.” <http://jalammar.github.io/illustrated-transformer>.
- [20] C. McCormick and N. Ryan, “Bert word embeddings tutorial.” <http://www.mccormickml.com>.
- [21] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” pp. 3973–3983, 01 2019.
- [22] O. Owoputi, B. O’Connor, C. Dyer, K. Gimpel, N. Schneider, and N. Smith, “Improved part-of-speech tagging for online conversational text with word clusters,” *Proceedings of NAACL-HLT*, vol. 2013, pp. 380–390, 01 2013.
- [23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [24] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software: an update,” *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.
- [25] S. Loukas, “Pca clearly explained - when, why, how to use it and feature importance: A guide in python.” <http://towardsdatascience.com/>.