ASPECT-BASED SENTIMENT ANALYSIS SU RECENSIONI DI FILM

Provasi Matteo (782922)

STEP

1. **Introduzione, Web Scraping e Dati**
   Creazione del dataset e pre-processing
2. **Spelling correction e Sarcasm detection**
   Creazione di 4 dataset
3. **Identificazione degli aspetti**
   Pattern rilevanti
4. **Sentiment analysis**
   Sentiment adattato al contesto
5. **Classificazione**
   Modelli di classificazione sui 4 dataset
6. **Conclusioni**
   Valutazione dei risultati e possibili miglioramenti

INTRODUZIONE, WEB SCRAPING E DATI

- Aspect-based Sentiment Analysis cerca di superare i limiti della semplice Sentiment Analysis.
- Sempre maggior importanza nell'ambito degli UGC (User Generated Content).
- Difficoltà nel riconoscere gli aspetti e pochi lavori di letteratura.
- Valutazione di nuove tecniche in questo ambito: Spelling correction e Sarcasm detection.

INTRODUZIONE, WEB SCRAPING E DATI

Acquisizione dati:

- Scraping da IMDB
  - 250 film più popolari per ogni anno (1992-2019)
  - Massimo di 1000 recensioni per film
  - Titolo originale del film
  - Personalità legate ad un film (attori, direttori…).

I dati sono stati aggregati in tre dataset rimuovendo i duplicati e le recensioni non scaricate correttamente.

In tutto sono state raccolte 784 mila recensioni.

INTRODUZIONE, WEB SCRAPING E DATI

Per il pre-processing sono state applicate diverse funzioni:

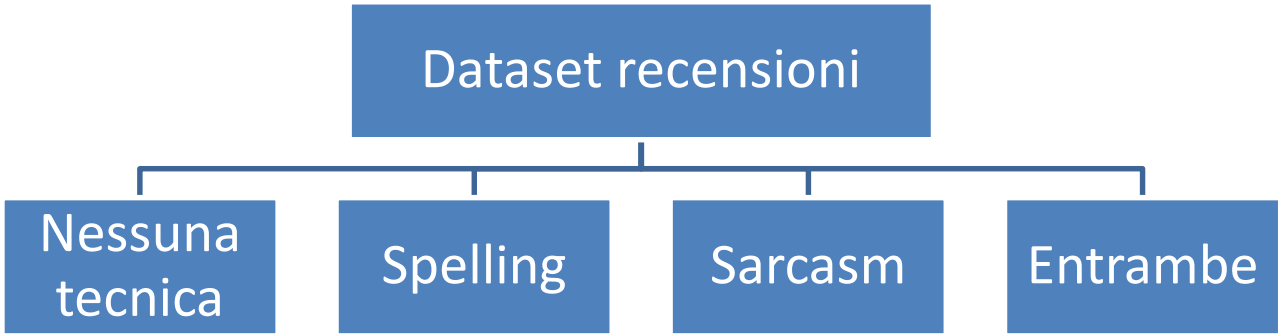
- Rimozione spazi bianchi multipli
- Rimozione numeri e caratteri alfanumerici
- Rimozione dei tag
- Rimozione delle lettere ripetute più di 2 volte
- Estensione delle forme contratte (*I'm = I am*)

Altre funzioni come lower case e la rimozione di stopwords e della punteggiatura sono state applicate successivamente:

- Le parole in maiuscolo sono necessarie per i nomi di attori e la NER.
- I punti, punti di domanda e esclamativi servono per separare le frasi. Le frasi duplicate nella stessa recensione sono rimosse.

SPELLING CORRECTION E SARCASM DETECTION

Spelling correction e sarcasm detection non sono mai state utilizzate in lavori di Aspect-based Sentiment Analysis.



Dal dataset originale si creano delle copie applicando le due tecniche. Le performance dei modelli saranno confrontate tra i 4 dataset ottenuti.

SPELLING CORRECTION E SARCASM DETECTION

I passaggi per la spelling correction sono:

- Creazione vocabolario.
- Filtraggio di termini relativi al film (titolo, nomi di attori).
- NER (Named Entity Recognition), modello Università di Stanford:
  - Parole originali
  - Persone
  - Luoghi
  - Organizzazioni
- Algoritmo SymSpell:
  - Creazione di un nuovo vocabolario.
  - Utilizzo esclusivo della rimozione di caratteri.
  - 100 volte più veloce degli altri metodi.
- Sostituzione dei termini modificati all'interno delle recensioni

IDENTIFICAZIONE DEGLI ASPETTI

- POS Tagging, metodo XPOS con 17 diversi tipi di tag (NN, JJ, VB…). Realizzato sia sul dataset con e senza spelling correction.
- Identificazione di pattern rilevanti delimitati dai tag NN e JJ. Un pattern rilevante inizia con un nome o un aggettivo e si conclude con l'altro tag (3 milioni di sequenze differenti).
- Rimozione pattern non comuni. Lo 0.5% delle sequenze di tag più comuni corrisponde al 67% (14 milioni) dei pattern totali.
- Matrice di similarità sui nomi (30 mila) con il metodo Wu-Palmer. Creazione di una lista di sinonimi.

IDENTIFICAZIONE DEGLI ASPETTI

Utilizzo di tecniche di clustering per ottenere le categorie:

- Clustering agglomerativo.
- HDBSCAN.

Le performance migliorano considerando solo i 1000 nomi più comuni. L'identificazione di sinonimi tramite modelli di word2vec non migliora i risultati.

Performance valutate tramite l'indice di Silhouette. Miglior algoritmo restituisce 12 cluster con un valore dell'indice di 0.3 per entrambi i dataset.

SENTIMENT ANALYSIS

- Libreria VADER in grado di gestire testi UGC (emoji, slang…). Valore di compound preso come riferimento.
- Creazione di un vocabolario adatto al conteso. Esempio: la parola «war» se riferito a film di guerra non ha valenza negativa.
- Utilizzo della metrica *Sentiment Abruptness* per trovare il sarcasmo. Più il sentiment cambia rapidamente in un pattern, più è probabile che sia sarcastico.
- In questi casi i valori di sentiment sono girati. Si creano due dataset uno con questa metrica e uno senza.

CLASSIFICAZIONE

I pattern sono stati classificati in base al numero di parole all'interno dei cluster identificati. Oltre alle 12 classi ne sono state create altre due:

- I pattern senza parole utilizzate per la cluster analysis hanno creato una nuova classe.
- Una nuova classe anche per i nomi delle personalità.

I pattern riguardano gli attori, la produzione, la trama, la colonna sonora, le emozioni e altri meno numerosi.

Per le etichette di sentiment i pattern con valore normalizzato >0.05 positivi, <-0.05 negativi e gli altri neutri.

CLASSIFICAZIONE

Modelli di classificazione utilizzati:

- Naive Bayes
- Complement Naive Bayes
- Classificatore Ridge
- Regressione Logistica (approccio pseudo Newton-Raphson e del gradiente)
- Support Vector Machine (solo kernel lineare)

I modelli sono stati valutati sia su task di classificazione separati (solo pattern e solo sentiment) e sulla classificazione mista.

Le feature date in input al modello sono state ottenute dal metodo TFIDF considerando parole singole e bigrammi.

CLASSIFICAZIONE

Pattern

Modello	Accuratezza (no spelling)	Accuratezza (spelling)
SVM	0,723	0,721
LogReg (sag)	0,721	0,719
Ridge	0,714	0,713
LogReg (lbfgs)	0,709	0,708
CNB	0,679	0,679
NB	0,657	0,659

Sentiment

Modello	Accuratezza (no sarcasm)	Accuratezza (sarcasm)
LogReg (sag)	0,972	0,963
SVM	0,967	0,963
Ridge	0,959	0,948
NB	0,938	0,948
CNB	0,922	0,911

Per tempi di esecuzione il modello Logistico è due volte più veloce (5 minuti) rispetto alla SVM e otto volte sul classificatore Ridge.

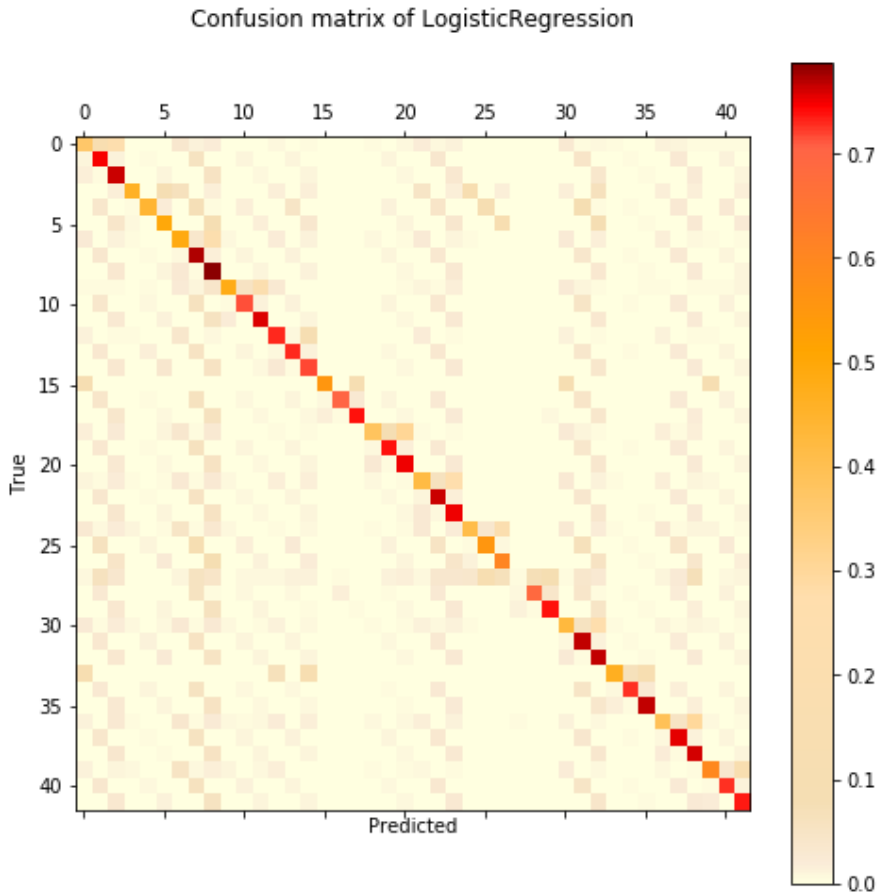
CLASSIFICAZIONE

Pattern + Sentiment

Modello	Accuratezza (no tecniche)	Accuratezza (spelling)	Accuratezza (sarcasm)	Accuratezza (entrambe)
LogReg (sag)	0,687	0,688	0,697	0,687
SVM	0,670	0,664	0,672	0,661
NB	0,602	0,605	0,624	0,600
Ridge	0,548	0,543	0,588	0,544
CNB	0,504	0,505	0,546	0,508

Non ci sono differenze sostanziali tra i dataset. Il modello logistico è sempre il migliore, il dataset con solo sarcasm detection il migliore in assoluto.

CLASSIFICAZIONE



Prestazioni migliori sulle classi con sentiment positivo, in particolare colonna sonora e attori.

Peggiori su nomi di attori ed emozioni negative.

Errori di classificazione sparsi per le classi ma non sistematici.

Le classi erano molto sbilanciate: sono state provate tecniche di resampling, sia casuale sia specifiche come SMOTE, ma non hanno portato a dei miglioramenti.

CONCLUSIONI

Considerazioni finali:

- Il lavoro ha una struttura flessibile che può essere generalizzata ad altri ambiti.
- Non raggiunge i migliori risultati ottenuti in letteratura (0.8 di accuratezza), possibili spiegazioni:
  - Presenza di più classi, 42 contro una media di 15.
  - Assenza di una golden truth, le etichette non sono state assegnate manualmente.
- Spelling correction e sarcasm detection non hanno delle performance significativamente migliori.
- Miglior modello con sarcasm detection
- Possibili miglioramenti:
  - Aspetti impliciti.
  - Eventuale variazione del linguaggio nel tempo.
  - Prove con diversi risultati dalla cluster analysis.

GRAZIE DELL'ATTENZIONE

Provasi Matteo (782922)