Tecniche di Natural Language Processing per il riconoscimento dei discorsi d'odio sui social network

Daniel Scalena 844608

Relatore: Prof.ssa Elisabetta Fersini
Correlatore: Prof. Antonio Candelieri

Laurea triennale in Informatica
Università degli studi di Milano-Bicocca
Anno Accademico 2020 - 2021

Introduzione e stato dell’arte

- Definizione di discorsi d’odio
- Necessità del loro riconoscimento
- Strumenti utilizzati per il riconoscimento

Introduzione e stato dell’arte

- **Definizione** di Hate Speech
- Perché il **riconoscimento automatico** dei discorsi d’odio?
- Metodi usati in **letteratura**
  - Dizionari
  - Bag of words, N-grams
  - SVM, Naïve Bayes
  - Deep learning
- **Stato dell’arte** nel Natural Language Processing

Introduzione e stato dell’arte

- **Definizione** di Hate Speech
- Perché il **riconoscimento automatico** dei discorsi d’odio?
- Metodi usati in **letteratura**
  - Dizionari
  - Bag of words, N-grams
  - SVM, Naïve Bayes
  - Deep learning
- **Stato dell’arte** nel Natural Language Processing

Introduzione e stato dell’arte

- **Definizione** di Hate Speech
- Perché il **riconoscimento automatico** dei discorsi d’odio?
- Metodi usati in **letteratura**
  - Dizionari
  - Bag of words, N-grams
  - SVM, Naïve Bayes
  - Deep learning
- **Stato dell’arte** nel Natural Language Processing

Introduzione e stato dell’arte

- **Definizione** di Hate Speech
- Perché il **riconoscimento automatico** dei discorsi d’odio?
- Metodi usati in **letteratura**
  - Dizionari
  - Bag of words, N-grams
  - SVM, Naïve Bayes
  - Deep learning
- **Stato dell’arte** nel Natural Language Processing

Framework proposto

- Raccolta di dati
- Analisi e classificazione

Framework proposto

- Raccolta e pulizia dei dati
- Classificazione manuale e linee guida
- Classificazione con il lessico **Hurtlex**
- Fine tuning di **BERT base**
  - Selezione dei migliori parametri
- **Ottimizzazioni** su BERT base
  - Modello BERT fine tuned + **Bi LSTM**
  - Modello BERT fine tuned + **Linear** + **Dropout**

Framework proposto

- Raccolta e pulizia dei dati
- Classificazione manuale e linee guida
- Classificazione con il lessico **Hurtlex**
- Fine tuning di **BERT base**
  - Selezione dei migliori parametri
- **Ottimizzazioni** su BERT base
  - Modello BERT fine tuned + **Bi LSTM**
  - Modello BERT fine tuned + **Linear** + **Dropout**

Framework proposto

- Raccolta e pulizia dei dati
- Classificazione manuale e linee guida
- Classificazione con il lessico **Hurtlex**
- Fine tuning di **BERT base**
  - Selezione dei migliori parametri
- **Ottimizzazioni** su BERT base
  - Modello BERT fine tuned + **Bi LSTM**
  - Modello BERT fine tuned + **Linear** + **Dropout**

Framework proposto

- Raccolta e pulizia dei dati
- Classificazione manuale e linee guida
- Classificazione con il lessico **Hurtlex**
- Fine tuning di **BERT base**
  - Selezione dei migliori parametri
- **Ottimizzazioni** su BERT base
  - Modello BERT fine tuned + **Bi LSTM**
  - Modello BERT fine tuned + **Linear** + **Dropout**

Framework proposto

- Raccolta e pulizia dei dati
- Classificazione manuale e linee guida
- Classificazione con il lessico **Hurtlex**
- Fine tuning di **BERT base**
  - Selezione dei migliori parametri
- **Ottimizzazioni** su BERT base
  - Modello BERT fine tuned + **Bi LSTM**
  - Modello BERT fine tuned + **Linear** + **Dropout**

Risultati sperimentali

Come si comportano i metodi di classificazione introdotti?

- Lessico Hurtlex
- BERT base
- Ottimizzazioni di BERT

Risultati sperimentali

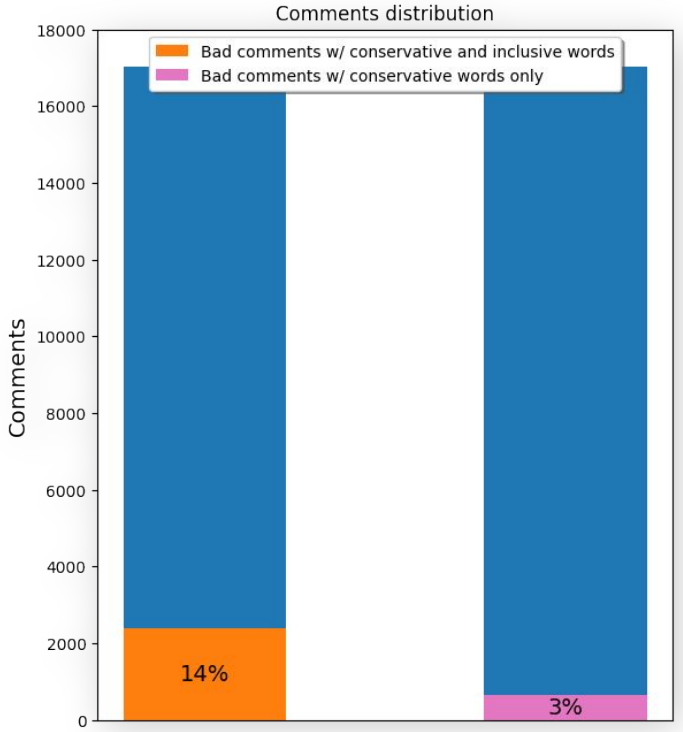
Hurtlex

Commenti classificati come **offensivi**:

- Conservativi e inclusivi: **14%**
- Solo conservativi: **3%**

F1 score sul dataset etichettato:

- Conservativi e inclusivi: **.121**
- Solo conservativi: **.154**



Risultati sperimentali

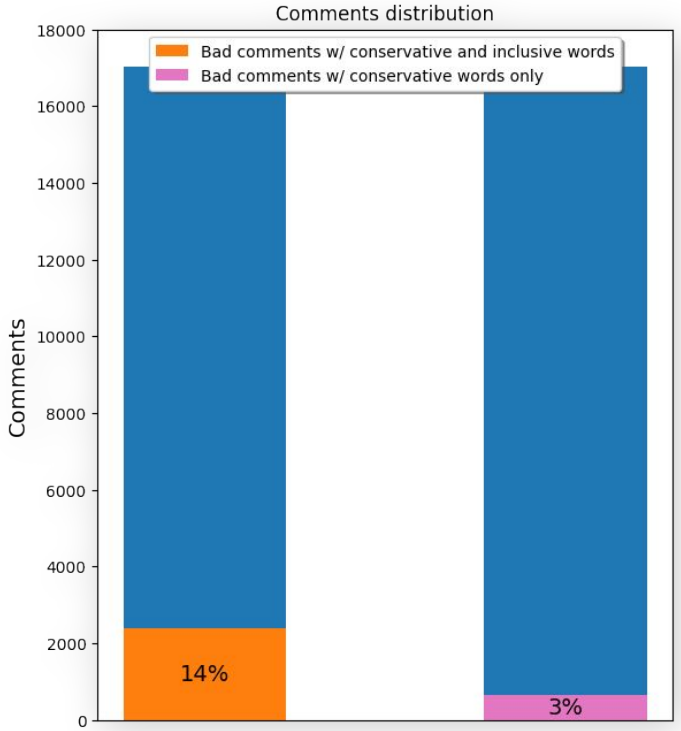
Hurtlex

Commenti classificati come **offensivi**:

- Conservativi e inclusivi: **14%**
- Solo conservativi: **3%**

F1 score sul dataset etichettato:

- Conservativi e inclusivi: **.121**
- Solo conservativi: **.154**



Risultati sperimentali

Training BERT base

*BERT base italian **uncased***

- Miglior batch size: **64**
- Migliori epoche: **< 7**

*BERT base italian **xxl cased***

- Miglior batch size: **16**
- Migliori epoche: **< 4**

Risultati sperimentali

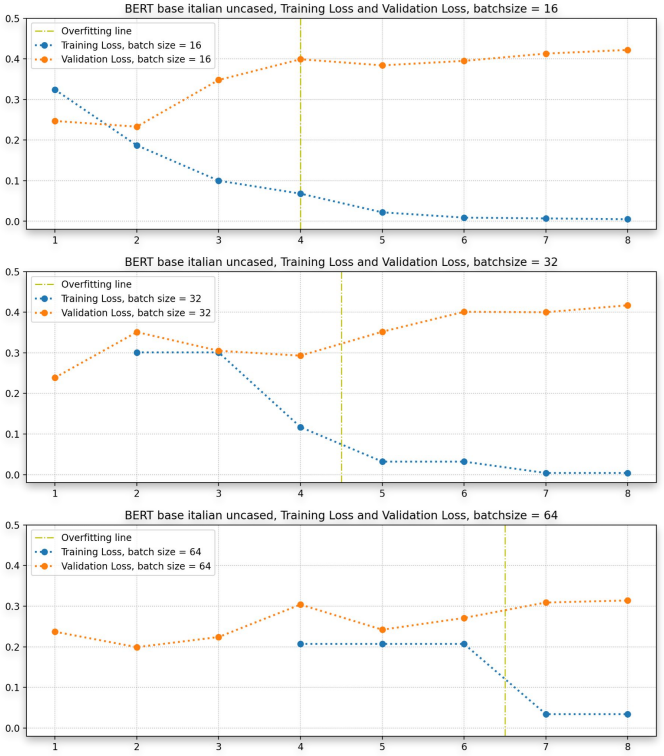
Training BERT base

*BERT base italian **uncased***

- Miglior batch size: **64**
- Migliori epoche: **< 7**

*BERT base italian **xxl cased***

- Miglior batch size: **16**
- Migliori epoche: **< 4**



Risultati sperimentali

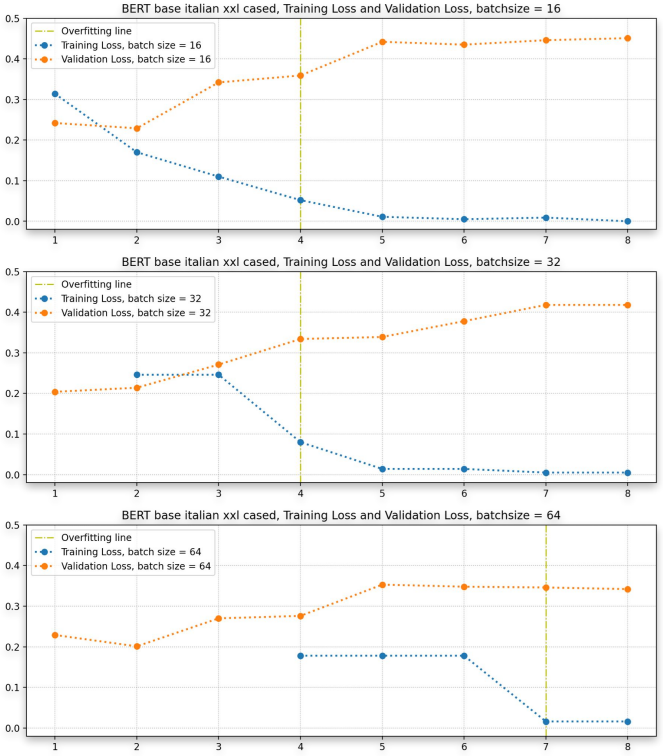
*BERT base italian **uncased***

- Miglior batch size: **64**
- Migliori epoche: **< 7**

*BERT base italian **xxl cased***

- Miglior batch size: **16**
- Migliori epoche: **< 4**

Training BERT base



Risultati sperimentali

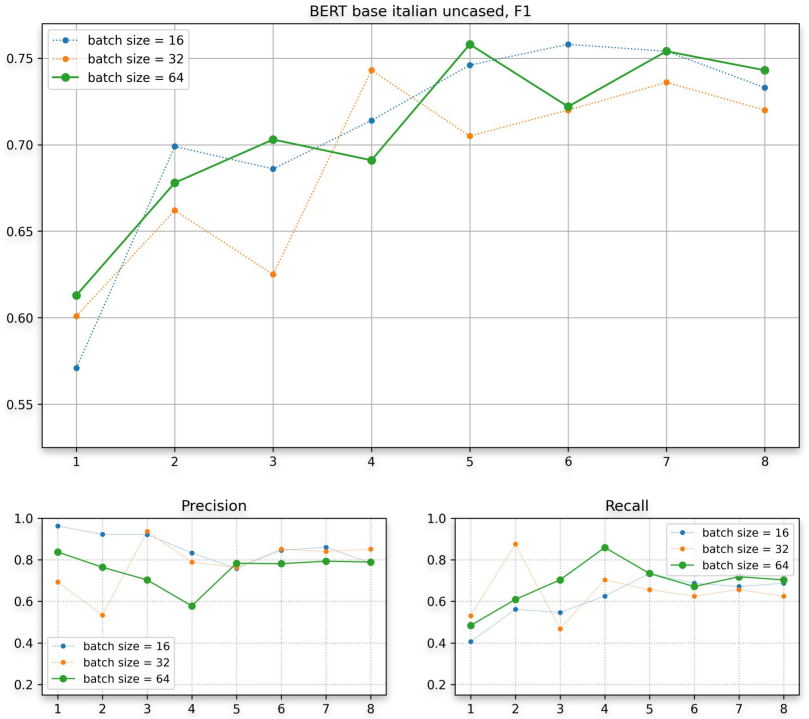
Validating BERT base

*BERT base italian **uncased***

- Miglior punteggio F1: **.758**
  - Precisione: .783
  - Recupero: .734

*BERT base italian **xxl cased***

- Miglior punteggio F1: **.760**
  - Precisione: .807
  - Recupero: .718



Risultati sperimentali

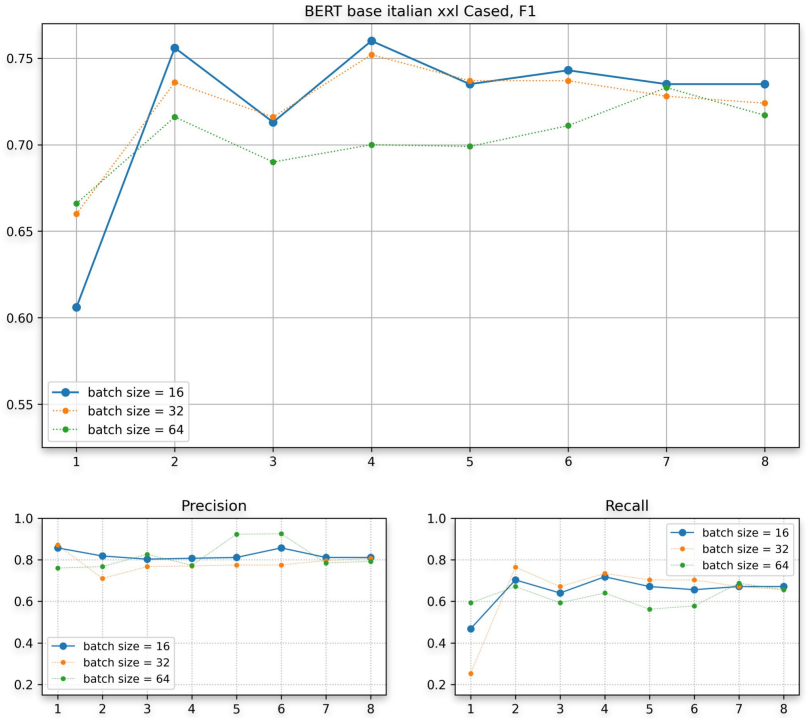
Validating BERT base

*BERT base italian **uncased***

- Miglior punteggio F1: **.758**
  - Precisione: .783
  - Recupero: .734

*BERT base italian **xxl cased***

- Miglior punteggio F1: **.760**
  - Precisione: .807
  - Recupero: .718

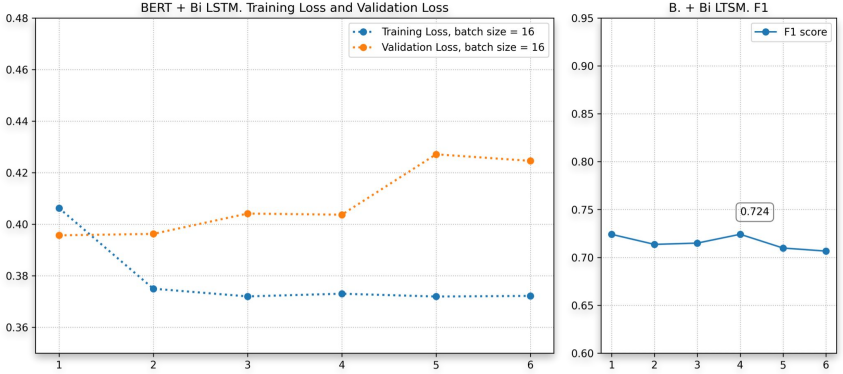


Risultati sperimentali

BERT Optimization

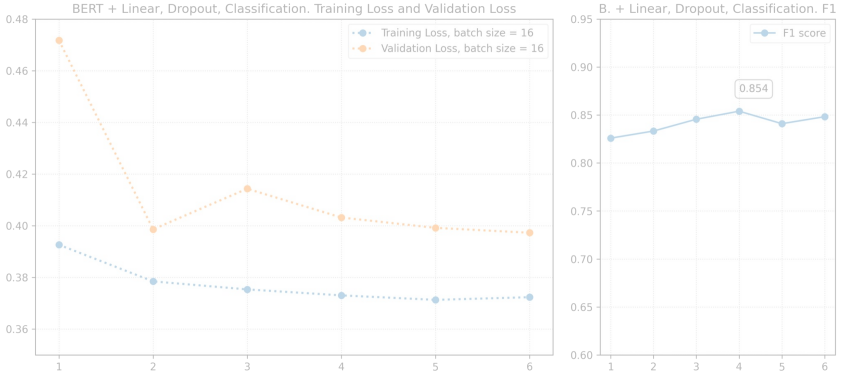
*BERT base + **Bi LSTM***

- Miglior punteggio F1: **.724**
- Numero epoca: **4**



*BERT base + **Linear** + **Dropout***

- Miglior punteggio F1: **.854**
- Numero epoca: **4**

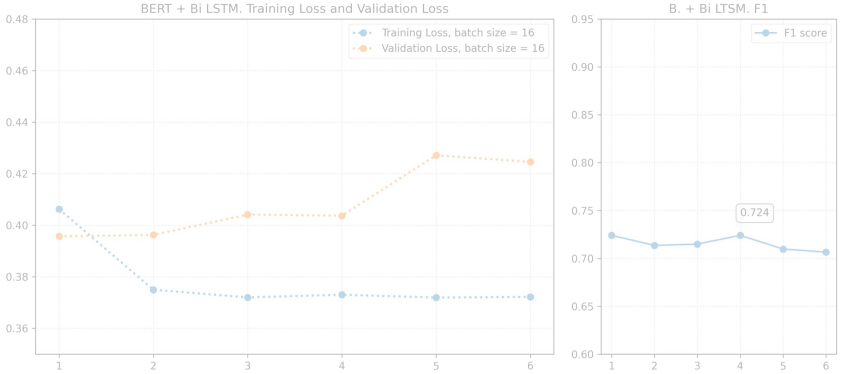


Risultati sperimentali

BERT Optimization

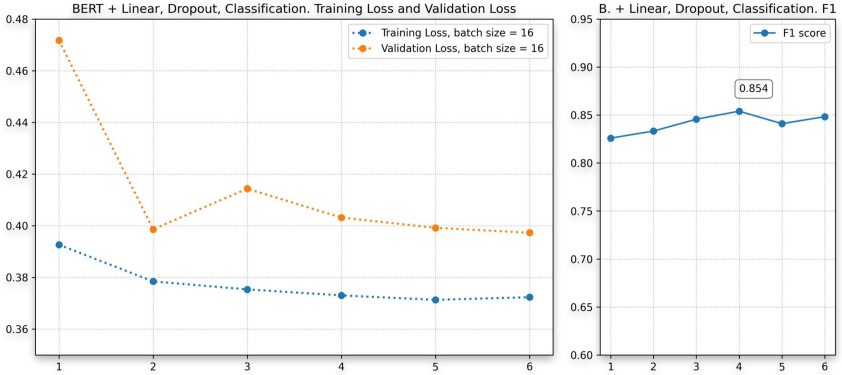
*BERT base + **Bi LSTM***

- Miglior punteggio F1: **.724**
- Numero epoca: **4**



*BERT base + **Linear** + **Dropout***

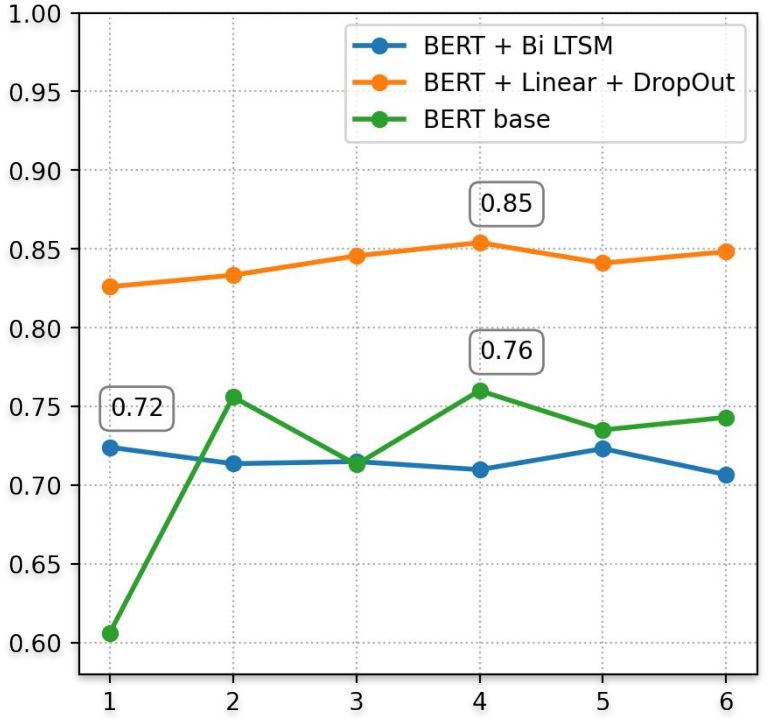
- Miglior punteggio F1: **.854**
- Numero epoca: **4**



Risultati sperimentali

Migliori prestazioni **complessive**:

- Best model: **BERT + Linear + Dropout**
- Best epoch: **4**
- F1 score: **.85**



Conclusioni e sviluppi futuri

- Risultati raggiunti
- Cosa può essere ancora fatto?

Conclusioni e sviluppi futuri

Obiettivi raggiunti:

- Evidenziate le **criticità** di un'analisi esclusivamente lessicale
- Utilizzo di **modelli SOTA** per il riconoscimento dell'Hate Speech
- Ulteriori **miglioramenti dei risultati** aggiungendo layers in coda a BERT

Sviluppi futuri:

- **Aumentare** le dimensioni del dataset
- **Nuove modifiche** alla rete neurale proposta
- Considerare **più features** per la classificazione

Conclusioni e sviluppi futuri

Obiettivi raggiunti:

- Evidenziate le **criticità** di un'analisi esclusivamente lessicale
- Utilizzo di **modelli SOTA** per il riconoscimento dell'Hate Speech
- Ulteriori **miglioramenti dei risultati** aggiungendo layers in coda a BERT

Sviluppi futuri:

- **Aumentare** le dimensioni del dataset
- **Nuove modifiche** alla rete neurale proposta
- Considerare **più features** per la classificazione

Grazie per l’attenzione!