Sentiment analysis e Reddit scores a confronto: trends nel contesto delle presidenziali USA 2020

Tesi di Laurea in

Ingegneria Informatica

Candidato
Gabriele Frassi

Relatori
prof. Marco Avvenuti
ing. Lorenzo Cima

UNIVERSITÀ DI PISA

Introduzione e Problema

reddit   

- **Contesto**:
  - Le elezioni presidenziali USA 2020 si svolgono in un clima estremamente polarizzato, con controversie sull'esito.
  - *Reddit* è uno dei principali social network negli USA. Gli utenti esprimono gradimento ai commenti con uno *score*.
- **Obiettivo**:
  - Vogliamo individuare possibili correlazioni tra *score* e *sentiment analysis*, analizzando il periodo elettorale.
  - Studio su due subreddit: *r/democrats* ed *r/Republican*
  - Analisi di due finestre temporali: [07/10, 03/11], [04/11, 01/12]

Soluzioni

- VADER (*Valence Aware Dictionary and sEntiment Reasoner*) tool per lo svolgimento della sentiment analysis

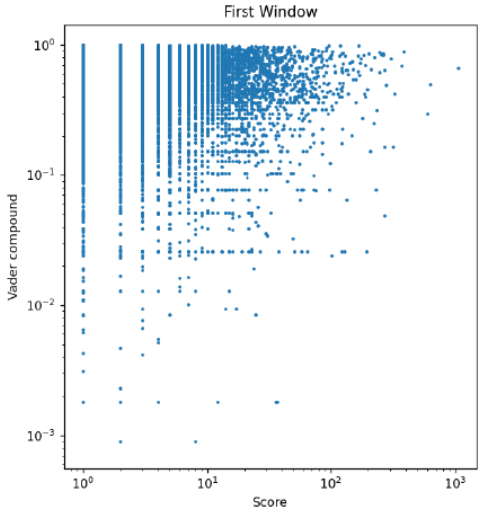
$$Risultato\ sentiment\ analysis = \begin{cases} positive & \text{compound} \geq +0.05 \\ neutral & \text{compound} \in]-0.05, +0.05[\\ negative & \text{compound} \leq +0.05 \end{cases}$$

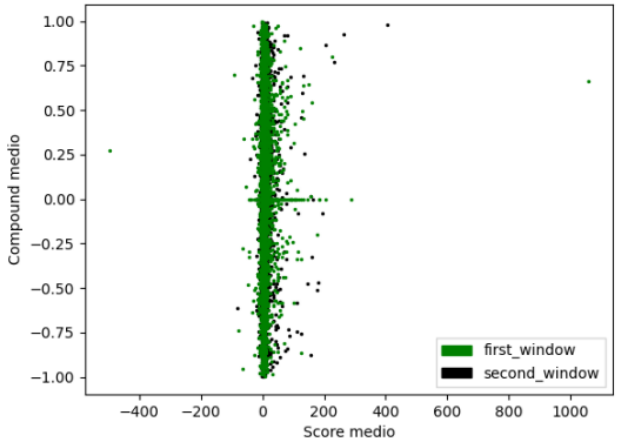
  È stato calcolato il *compound* per ogni commento del dataset.

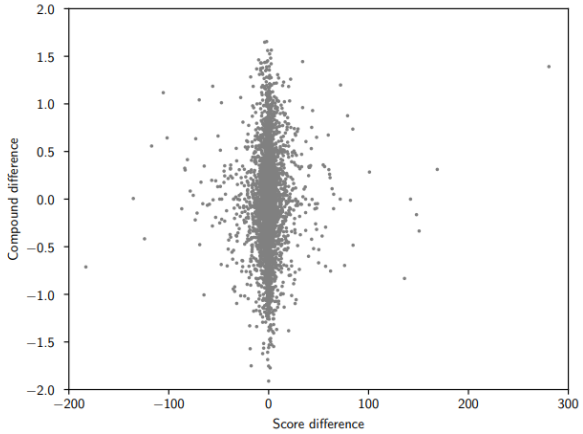
- **Analisi statistica**:
  1. Correlazioni tra *score* e *compound* rispetto ai commenti (*scatterplot*)
  2. Correlazioni tra score e compound rispetto agli utenti
     a) Utenti attivi in entrambe le finestre temporali (*scatterplot* con differenze di medie)
     b) Utenti attivi in una sola delle due finestre (*scatterplot* con medie e distinzione finestre)
  3. Evoluzione giornaliera di *score medio* e *compound medio*
  4. Analisi degli outlier di quanto fatto al punto (4), individuati con algoritmo IsolationForest.

Risultati

- **Assenza di correlazioni dirette tra *score* e *compound***







- **Maggiore polarizzazione**