

# INFORMAZIONE DAI DATI TESTUALI E RICONOSCIMENTO DI EVENTI DI VIOLENZA: Tecniche di Natural Language Processing per l'analisi degli accessi in Pronto Soccorso

Quarta D<sup>1</sup>, Radicioni DP<sup>2</sup>, Mensa E<sup>2</sup>, Baffigi S<sup>3</sup>, Pitidis A<sup>4</sup>, Dalmasso M<sup>1</sup>, Bianco S<sup>1</sup>, Mamo C<sup>1</sup>

<sup>1</sup> SC a DU Servizio di Epidemiologia, ASL TO3 Grugliasco - Regione Piemonte

<sup>2</sup> Università degli Studi di Torino - Dipartimento di Informatica

<sup>3</sup> Università degli Studi di Torino - Master in Analisi Dati per la Business Intelligence e Data Science

<sup>4</sup> Dipartimento Ambiente e Salute - Istituto Superiore di Sanità

## BACKGROUND

Selene Bianco

### CASO-CONTROLLO. Accessi PS:

- Nei 24 mesi precedenti la morte, le vittime di femminicidio hanno mostrato una frequenza superiore di accessi in PS rispetto alle donne di pari età e condizione sociale decedute per incidente stradale.
- Alcuni codici di patologia risultano predittori.

PIEMONTE  
42 femminicidi  
2005-2010



### TEXT-MINING. Articoli La Stampa e La Repubblica:

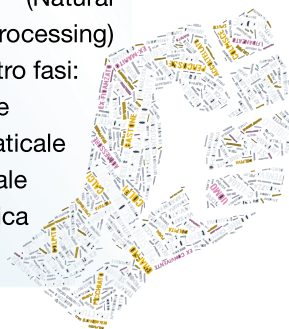
- Parola più frequente: casa, rappresenta la dimensione dell'emergenza domestica.
- Relazioni fra le parole: molto forti quelle fra i rapporti sentimentali (ex -> fidanzata), che sottendono il genere (uomo -> donna).
- Aggettivazioni diverse se la relazione vittima/aggressore comprende immigrati.

## SPERIMENTAZIONE

Tre differenti lavori hanno utilizzato i dati di accesso ai Pronto Soccorso provenienti dal progetto **SINIACA – IDB** (Sistema Informativo Nazionale Incidenti in Ambienti di Civile Abitazione – Injury DataBase), con obiettivi diversi ma sempre volti al miglioramento e all'applicazione di algoritmi su dati non strutturati (testi) per l'estrapolazione di pattern ricorrenti.

Si sono sviluppate quindi tecniche di **NLP** (Natural Language Processing) articolate in quattro fasi:

- analisi lessicale
- analisi grammaticale
- analisi strutturale
- analisi semantica



### FOCUS

- 1) categorizzazione degli accessi in PS in casi di **violenza** o **non violenza**
- 2) estrazione degli schemi ricorrenti nell'ambito della violenza

algoritmo  
NAIVE BAYES

Matteo di Cunzolo

Difficoltà con **sintassi** e **lessico** - gergo medico, abbreviazioni e codici - legate al tempo a disposizione di chi redige la cartella clinica.  
PRECISION: 87,41% • RECALL 95,17% • F-MEASURE 91,12%

### FOCUS

- 1) somministrazione di un questionario a un campione di lavoratori, volontari e studenti arruolati in setting sanitari, per verificare il riconoscimento di casi di **violenza** – **non violenza**.
- 2) evidenza che nella categorizzazione dei documenti gli esseri umani non si basano solo sulla presenza di termini (es. pugno, schiaffo) ma tendono a costruire un modello di evento violento, costituito da elementi come **aggressore** **vittima**, **relazione**, **luogo** della violenza, parte del **corpo** interessata, etc.
- 3) elaborazione di un modello semantico coerente con i risultati del questionario.

Stefano Baffigi

### FOCUS

- 1) miglioramento della categorizzazione degli accessi in PS in casi di **violenza** o **non violenza** sulla base dell'estrazione degli elementi costitutivi dell'evento di **violenza**.
- 2) utilizzo della **Word Sense Disambiguation** al fine di estrarre gli elementi che partecipano all'evento di **violenza**.

SUPERSENSE  
TAGGING

Federico Barbi

Federico: persona è andato a stare al supermercato, location e ha comprato, possession il latte, food

PRECISION: 95,20% • RECALL 92,41% • F-MEASURE 93,78%

## CONCLUSIONI

L'utilizzo degli strumenti di Natural Language Processing, ha dimostrato come anche i dati testuali possano fornire informazione classificabile e analizzabile dai diversi modelli.

Allo stesso tempo però, nonostante un utilizzo più raffinato dello strumento con algoritmi anche personalizzati (**Naive Bayes**, **Word Sense Disambiguation**) e un conseguente miglioramento dei risultati di categorizzazione dei casi di **violenza**, non si è ancora in grado di produrre un algoritmo automatizzato di riconoscimento da stringhe testuali sufficientemente sensibile e specifico.

Dalla sintesi dei lavori fin qui svolti, emergono due visioni:

- 1 – proseguire nel percorso di strutturazione di schemi ricorrenti, al fine di consolidare gli effetti preventivi dei casi di **violenza**;
- 2 – attraverso la formazione degli operatori, abbattere nelle descrizioni testuali a referto le ambiguità/errori che ostacolano sia nell'uomo sia negli algoritmi automatici il riconoscimento dei casi di **violenza**.

## LAVORI

**Bianco S et al.** Are Emergency Department admissions in the past two years predictors of femicide? Results from a case-control study in Italy.

**Bianco S et al.** Come riportano i quotidiani le notizie di femminicidio? L'analisi di lessico e sintassi attraverso tecniche di text-mining.

**Di Cunzolo M.** L'analisi semantica del testo per la categorizzazione di documenti testuali per l'estrazione di informazioni.

**Barbi F.** Il contributo della semantica lessicale nella categorizzazione dei documenti testuali.

**Baffigi S.** Sintassi e semantica: analisi sperimentale dei meccanismi di riconoscimento delle descrizioni di violenza per lo sviluppo di un classificatore automatico.