

# **Detecting Subjectivity through Lexicon-Grammar Strategies**

## **Databases, Rules and Apps for the Italian Language**

Serena Pelosi  
Matr. 8880300067

### **Abstract English**

The present research handles the detection of linguistic phenomena connected to subjectivity, emotions and opinions from a computational point of view.

The necessity to quickly monitor huge quantity of semi-structured and unstructured data from the web, poses several challenges to Natural Language Processing, that must provide strategies and tools to analyze their structures from a lexical, syntactical and semantic point of views.

The general aim of the Sentiment Analysis, shared with the broader fields of NLP, Data Mining, Information Extraction, etc., is the automatic extraction of value from chaos; its specific focus instead is on opinions rather than on factual information. This is the aspect that differentiates it from other computational linguistics subfields.

The majority of the sentiment lexicons has been manually or automatically created for the English language; therefore, existent Italian lexicons are mostly built through the translation and adaptation of the English lexical databases, e.g. SentiWordNet and WordNet-Affect.

Unlike many other Italian and English sentiment lexicons, our database SentIta, made up on the interaction of electronic dictionaries and lexicon dependent local grammars, is able to manage simple and multiword structures, that can take the shape of distributionally free structures, distributionally restricted structures and frozen structures.

Moreover, differently from other lexicon-based Sentiment Analysis methods, our approach has been grounded on the solidity of the Lexicon-Grammar resources and classifications, that provides fine-grained semantic but also syntactic descriptions of the lexical entries.

According with the major contribution in the Sentiment Analysis literature, we did not consider polar words in isolation. We computed they elementary sentence contexts, with the allowed transformations and, then, their interaction with contextual valence shifters, the linguistic devices that are able to modify the prior polarity of the words from SentIta, when occurring with them in the same sentences. In order to do so, we took advantage of the computational power of the finite-state technology. We formalized a set of rules that work for the intensification, downtoning and negation modeling, the modality detection and the analysis of comparative forms.

With regard to the applicative part of the research, we conducted, with satisfactory results, three experiments on the same number of Sentiment Analysis subtasks: the sentiment classification of documents and sentences, the feature-based Sentiment Analysis and the Semantic Role Labeling based on sentiments.

## Abstract Italiano

Questa ricerca concerne l'analisi computazionale, di fenomeni linguistici relativi alle sfere della soggettività, delle emozioni e delle opinioni espresse nei testi.

La necessità di monitorare, in tempo reale, nel web grandi quantità di dati semi o non strutturati lancia numerose sfide al Trattamento Automatico della Lingua (TAL); disciplina chiamata a fornire strategie e strumenti per l'analisi lessicale, sintattica e semantica dei testi.

Lo scopo più generico dell'Analisi dei Sentimenti, l'estrazione automatica di valore dal caos, è condiviso con quello del TAL, del Data Mining, dell'Information Extraction, ecc...

La differenza più evidente tra questi ambiti della Linguistica Computazionale, invece, riguarda la focalizzazione sulle opinioni piuttosto che sulle informazioni fattuali.

La maggior parte dei lessici di sentimento disponibili per la consultazione o per l'uso, in letteratura e/o in commercio, è stata creata per la lingua inglese; per cui, in linea di massima, le risorse italiane esistenti sono state create a partire dalla traduzione e dall'adattamento dei database inglesi.

Diversamente da molti altri lessici di sentimento, il nostro database italiano *SentIta*, realizzato sulla base dell'interazione tra dizionari elettronici e grammatiche locali lessico-dipendenti, è in grado di gestire strutture, semplici o polirematiche, distribuzionalmente libere, ristrette o idiomatiche.

Inoltre, diversamente da ogni altro metodo di Analisi dei Sentimenti basato su lessici, il nostro approccio si fonda sulla solidità delle classificazioni e della metodologia Lessico-Grammaticale (LG) che fornisce dettagliate descrizioni sintattiche e semantiche di numerose entrate lessicali.

In accordo con i principali contributi in materia di Analisi dei Sentimenti, le parole non sono state considerate in isolamento. Si è tenuto conto, al contrario, dei vari dispositivi linguistici che possono alterare la polarità delle parole di sentimento, quando occorrenti nello stesso contesto testuale. Per fare ciò, ci siamo avvalsi del potere computazionale delle tecnologie a stati finiti. Sono state formalizzate una serie di regole per la gestione di fenomeni come l'intensificazione, la negazione, la modalità e la comparazione.

Per quanto riguarda la parte applicativa della ricerca, sono stati condotti tre esperimenti, con risultati soddisfacenti, su tre delle più importanti sfide dell'Analisi dei Sentimenti: la classificazione automatica di frasi e documenti non strutturati, l'Analisi dei Sentimenti basata sulle caratteristiche di beni e servizi e l'annotazione automatica dei ruoli semantici connessi alla sfera della soggettività.