

TwisEnt

A Multi-Stage System for Analyzing Sentiment in Twitter

Subhabrata Mukherjee, Akshat Malu, Balamurali A R, Pushpak Bhattacharyya

Department of Computer Science and Engineering, Indian Institute of Technology Bombay, India

Introduction

- TwisEnt retrieves and categorizes tweets pertaining to an entity searched based on its sentiment content and assigns it an overall sentiment score
- A multi-stage system to classify tweets as **positive**, **negative** or **objective** using sentiment lexicons (rule-based) or SVM's (supervised) for classification
- The sentiment score assigned to the search string reflects the sentiment snapshot of the string in Twitter based on the tweets fetched

Tweet Fetcher

- Obtains tweets pertaining to a search string entered by a user
- The Twitter API retrieves the latest 200 English tweets about the keyword

Spam Filter

- Removes all tweets that are spam, that is, which do not *truly* reflect the author's sentiment. Examples include *bot-generated tweets, promotional tweets, re-tweets etc.*

Extract Dependency

- Extracts the opinion expression specific to the target entity
 - *"The film bombed at the box office despite the actors putting up a good performance"* – is *pos.* w.r.t actors but *neg* w.r.t to film
- *Dependency Parsing* is used to capture associations in a graph which is partitioned to retrieve the opinion pertaining to the *user specified feature*

Spell Checker

- A modified *Levenshtein* dist. based spell checker is used to handle informal language form
- Vowel exchange (*gud*), vowel drop (*btfl*), numeral phonetic change (*gr8*, *fy9*), spelling errors (*redicule*), segmentation (*breathtaking*), normalization (*happyyyy*) etc. are handled

Pragmatics Handler

- Different forms of Pragmatics in Twitter handled by giving them more weightage
- Happiness, joy, excitement, interest are often expressed by – word elongation (*gr888*), *hashtags* (*#worthawatch*), *emoticons* (*😊 😞*) and *capitalization* (*HATED*, *Loving*)

Tweet Sentiment Predictor and Collaborator

- Bing Liu sentiment lexicon is used for *lexicon-based* classification and *SVM's* for supervised classification
- Based on predictions of individual tweets, the Tweet Sentiment Collaborator gives the overall prediction as percentage w.r.t the keyword

- Ablation tests show effect of Pragmatics Handler > Spell Checker > Entity Specificity
- Dependency Parsing fares poorly due to noise

Evaluation

The following phenomena were observed in the course of analyzing human sentiment, a part of which was handled by our system.

Sarcasm

Lack of sense understanding

Lack of Entity Specificity

Spam and Noisy data

Absence of Named Entity Recognition

Requirement of World Knowledge

Mixed Emotion Tweets

Pragmatics

Comparatives