Towards Sentiment and Emotion Analysis of User Feedback for Digital Libraries

Stefano Ferilli[1] Berardina De Carolis[1]
Domenico Redavid[2] Floriana Esposito[1]

[1] Dipartimento di Informatica – Università di Bari
name.*surname*@uniba.it

[2] Artificial Brain S.r.l.
redavid@abrain.it

12th Italian Research Conference on Digital Libraries - IRCDL 2016
Firenze, Italy, 4-5 February, 2016

Summary

- Introduction & Motivation
- Pre-Processing
- Author Identification: Relational Approach
- Author Characterization: Statistical Tool
- Conclusions & Future Work

Introduction & Motivation

- Comments in blogs and forums on the Internet allow to study people's attitude on various topics
  - Sentiment Analysis / Opinion Mining
    - Positive/negative valence
  - Emotion Analysis
    - Specific feelings
- Digital Libraries might understand how their content is perceived by their users
  - Useful to suitably direct their future strategic choices

Problems & Objectives

- Problems:
  - complexity of natural language
  - need of dealing with several languages
  - choice of relevant features and of good approaches to building the models
- Interesting results obtained for Italian by a system based on a Text Categorization approach
- Proposed solution: further experiments to check whether reliable predictions can be obtained, both for opinions and for feelings

Sentiment Classification as a Text Categorization Task

- Categories = polarity or emotions
- Differences with respect to topic-based TC
  - topics are objective, sentiments are subjective;
  - there may be hundreds (or even thousands) of topics, but just a few sentiments (at the extreme, just two polarities, positive and negative);
  - topics are application-dependent, sentiment is general;
  - topics may be independent from each other, sentiments typically are not
    - e.g., in an evaluation based on a number of ‘stars’ the categories are different degrees of a single scale

Text Categorization

- Text Categorization (TC) is the activity aimed at mapping documents in natural language to a pre-defined set of categories
  - Formally, given a set of documents *D* and a set of categories *C*, a text *classifier* implements a function that for each document-category pair says whether the document belongs to the category
    - The ‘hard’ categorization can be replaced by a degree of belonging
  - Often, the target function is unknown, and must be approximated
    - Use of Machine Learning techniques

Proposed Approach: Features

- single, normalized words
- abbreviations, acronyms, and colloquial expressions
- *n*-grams with sufficient frequency in the corpus
- PoS tags
- expressive punctuation
- emoticons

Proposed Approach: Algorithms

- Rocchio
  - Similarity-based
  - Classification by cosine similarity
- Naive Bayes
  - Probabilistic
  - Classification by maximum posterior probability
- Committee
  - Weighted sum

Previous Experiment on Opinion Mining

- Dataset: 2000 reviews in Italian, 558 movies
  - http://filmup.leonardo.it/
  - 1000 positive (6-10 stars); 1000 negative (1-5 stars)
- Experimental setting
  - 17 different feature settings
  - 5-fold cross-validation
  - Equal weight to classifiers in the committee
- Good efficiency on mediocre platform
  - PC, Intel Core 2 Duo E6750 @ 2.66 Ghz, 2 GB RAM, Windows 8)
    - Runtime between 3'25" (5892 features) and 13'08" (9001 features, of which 2784 *n*-grams)
      - *n*-grams significantly increase the number of features, and runtime as a consequence

Previous Experiment on Opinion Mining

- Overall accuracy always > 81%
  - Always above 82% for the committee
    - Very good, compared to the state-of-the-art for English and especially for Italian
    - When Rocchio outperformed Naive Bayes, accuracy of the committee was greater than that of the components; in the other cases, corresponding to settings that used *n*-grams, Naive Bayes alone was the winner
  - Even if balanced between positive and negative cases, accuracy on the former always better than that on the latter
    - This is somehow surprising, because it is commonly believed that negative emotions are stronger, and hence easier to recognize

New Experiment on Opinion Mining

- Evalita Sentipolc 2014 dataset
  - 4513 tweets in Italian mainly about politics
  - Standard benchmark for competitions; tweets are shorter than the movie reviews
  - Neutral items removed: 2091 tweets (1412 negative, 679 positive)
  - Only system configuration that provided the best results in the previous experiments

Normalization	PoS tags	Punct./Abbrev.	*n*-grams
lemmas	nouns, verbs, adjectives, adverbs, emoticons	Yes	–

New Experiment on Opinion Mining

- 10-fold cross validation
- Imbalanced dataset -> Precision/Recall
-

Positive			Negative			Average		
P	R	F1	P	R	F1	P	R	F1
0.752	0.498	0.599	0.750	0.901	0.819	0.751	0.700	0.724

- Figures compare well to the state-of-the-art best system in the competition
  - Recall on positives worse than on negatives
    - Due to the difference in the number of examples in the two classes of the dataset?
    - May be a useful outcome, because library managers (differently from e-business site holders) may be more interested in identifying and analyzing criticisms than on reading positive comments

New Experiment on Emotion Analysis

- Dataset purposely collected
  - 800 comments about movies from filmup
  - Shown at random to 11 human raters
    - Asked to evaluate whether the opinion about the movie expressed one of 3 emotions/classes of interest
      - One positive (happiness)
      - One lightly negative (sadness)
      - One strongly negative (anger)

      and in such a case which one
    - Label given to each comment according to the majority agreement criterion
      - Comments for which majority was not reached were discarded

New Experiment on Emotion Analysis

- Final dataset
  - 752 entries
    - 406 for happyness, 175 for sadness, 171 for anger
  - Features involved an extended set of Pos tags

Normalization	PoS tags	Punct./Abbrev.	n-grams
lemmas	nouns, verbs, adjectives, adverbs, articles, pronouns, emoticons	Yes	–

New Experiment on Emotion Analysis

- 10-fold cross validation
-

Anger			Happiness			Sadness			Average		
P	R	F1	P	R	F1	P	R	F1	P	R	F1
0.698	0.408	0.514	0.742	0.870	0.801	0.630	0.575	0.600	0.690	0.617	0.651

-
- Good performance on Happiness
- Less accurate on the other two classes
  - Due to the imbalanced dataset?
  - However, positive emotions are typically harder to recognize than negative ones
    - Combining our classifier with other state-of-the-art ones might improve the overall results

Conclusions

- Comments in blogs and forums may allow DIs to study users' attitude on various topics
  - Valence (sentiment analysis)
  - specific feelings (emotion analysis)
- Complex problem
  - Proposal: a system based on a Text Categorization
- Experimental results compare well to state-of-the-art tools
  - Application to the DL domani seems feasible

Future Work

- Further experiments
  - Use cases specifically concerning DIs dedicated to art
    - Need for data
  - Extended set of emotions to be recognized
    - Including at least all primary ones