Can We Infer Book Classification by Blurbs?

Valentina Franzoni, Valentina Poggioni, and Fabiana Zollo

Department of Mathematics and Computer Science,
University of Perugia, Perugia, Italy

Abstract. The aim of this work is to study the feasibility of an automated classification of books in the social network Zazie by means of the lexical analysis of book blurbs. A supervised learning approach is used to determine if a correlation between the characteristics of a book blurb and the emotional icons associated to the book by the Zazie's users exists.

1 Introduction

Sentiment Analysis and Opinion Mining are receiving increasing attention in many sectors because knowing and predicting opinions of people is considered a strategic added value. In the last years an increasing attention has also been devoted to Emotion Recognition, often by developing automated systems that can associate user's emotions to texts, music or artworks (e.g. [1], [6]).

In the wide scenario of text mining, several attempts have been made in order to associate emotions and moods to blogs [2], tales [3] and newspapers titles, in several domains and contexts. A particular and interesting domain that has been introduced in [9] is the association of emotions to books [1].

The idea of building a model for classifying books from an emotional point of view was born from *Zazie*[2], an Italian social network for readers that, differently from other similar projects e.g., *aNobii* or *Goodreads*, introduces a new dimension for books description, the emotional icon tagging.

Each book in Zazie, besides user's comments and reviews, can be tagged with two MOODs icons, selected in a set of 25 different climates related to the reader's opinion about the book or to the emotion induced by the book, e.g. *cool*, *culture*, *cry*, *sleep*, *love*, *hoax* or *smile*.

Starting from this context, it would be very innovative to provide Zazie users with an emotion-driven search within the social network. The necessity of such an automated system arises also from the presence of a lot of books that have not been tagged yet by the users, for which there is not any information besides the characteristics stored in the database i.e., title, author or publisher.

The first step of this research was focused on the selection of relevant attributes among those usable and available in Zazie to describe a book.
We decided to analyze the book blurb because it can contain relevant emotional

[1] An extended version of this work has been presented in [9]

[2] www.zazie.it We are grateful to Marco Ghezzi and Zazie developers, Joe and David, for the collaboration to this research.

information, since it is generally written for attracting the reader and it can emphasize and highlight some aspects with the use of emotional terms. On the other hand, a possible drawback is the introduction of a bias caused by the excessive use of words with a high emotional meaning, so the problem is not trivial. Moreover, the blurb represents an information always available on Zazie, regardless of user's opinions, reviews or tags. The main original contribution of this work is to determine if the book blurb reflects the same emotions that the reader can find in the book itself. The emotional model used for MOODs representation is directly provided by Zazie by means of its emotional tags (icons) and can be easily correlated to the well known discrete emotional models, such as the ones defined by Ekman [4] or Plutchik [5].

2 Blurb Analysis

The blurb analysis has been realized in three main phases: preprocessing, extraction of emotions and reduction of emotions.

Preprocessing consisted in *stop words deletion*, *tokenization* and *lemmatization*, realized by means of *Morph-it!* , a morphological resource for the Italian language.

Extraction of Emotions was realised by means of *WordNet-Affect*,[8] retrieving for each lemma the *WordNet* synsets associated to it, using the multilingual lexical database *MultiWordNet* [7]. The affective domain WordNet-Affect was exploited in order to obtain all the emotions associated to each synset, filtering out terms that did not convey affective information and taking into account multiple occurrences of the same emotion.

Reduction of Emotions was necessary because the emotional hierarchy of WordNet-Affect is particularly pronged (296 nodes) and the result of the emotion extraction phase can be excessively detailed. It was implemented in two step following two different approaches, first reducing the set of emotions to the 32 emotions corresponding to the third level of WordNet-Affect hierarchy (i.e., the subtree rooted in `emotion`)

and then, associating these 32 emotions to an extended set of Ekman emotions[4], formed by eight emotional categories *happiness*, *anger*, *disgust*, *fear*, *sadness*, *surprise*, *neutral*, *ambiguous.*

3 Experiments

Experiments were carried on to test if an automated classification of book blurbs based on Zazie emotional tags is possible and can actually be used with a satisfactory accuracy.

The database provided by Zazie authors was constituted of 38374 records, each one representing the association of a tag (one of the 25 Zazie MOODs) to a book by an user. Each record is represented by 7 fields (`user_id`, `book_isbn`, `book_title`, `book_pages`, `book_publisher`, `book_blurb`, `mood`). Those data then underwent a filtering and cleaning process as detailed in [9].

In this group of experiments the classification is limited to a subset of the Zazie MOOD set identified by the selected MOODs $\{$*smile*, *love*, *sad*, *think*, *angry*, *cry*$\}$; this subset was chosen because it contains emotional tags that are clearly related to emotions aroused in the reader by the book.

Among the information characterizing a book which is available in the Zazie database, the author and the emotions extracted by the blurb analysis have been used as the sample features. Publisher and number of pages have been discarded, because they are associated to a particular edition of the book and do not characterize the book as its general literary work. Each record in the dataset represents a book and is characterized by either 34 or 10 features:

- the author (nominal attribute)
- the emotions extracted from the blurb (numerical attributes) valued by their occurrences (32 or 8 depending on which strategy for emotion reduction is applied)
- the MOOD tag (nominal attribute) representing the class attribute

Experiments had been carried on by means of the software *Weka*.

The experimentation has been realized through the *cross validation* technique with ten folds using, in particular, algorithms based on decision trees.

Models have been evaluated by the *accuracy*, *recall* and *precision* measures, defined as follow:

- $Accuracy = TP/N$ where TP is the number of instances correctly classified and N is the total number of instances.
- $Recall = \frac{1}{NC}\sum_{i=1..NC} Recall_i$ where NC is the number of classes, $Recall_i = \frac{TP_i}{TP_i+FN_i}$ and TP_i and FN_i are respectively the instances correctly classified as members of class i and the instances wrongly classified as not belonging to the class i.
- $Precision = \frac{1}{NC}\sum_{i=1..NC} Precision_i$ where $Precision_i = \frac{TP_i}{TP_i+FP_i}$ and FP_i is the number of instances wrongly classified as members of class i.

Results for *accuracy*, *recall* and *precision* are presented in Table 1: best accuracy levels are obtained with *J48* and *BFTree* algorithms that have equivalent performances, while, *LADTree* and *RandomForest* did not perform as well as expected. Details about the tested algorithms and their implementations can be found in the *Weka* documentation. Results show that algorithms having an unpruned version perform better. A satisfactory level of accuracy was obtained reaching more than 70% and considering that other experiments and the design of other techniques to build a reliable dataset are ongoing works.

4 Conclusions and Future Work

The blurb is confirmed to be a good source of emotional information about a book and it actually can be analyzed with the aim of sentiment analysis and emotion recognition.

(a) Emotional model with 32 emotions derived from WordNet-Affect

Attributes	Accuracy	Recall	Precision
J48 -U -M 3	70.31%	0.694	0.703
BFTree -U -M 3	72.92%	0.715	0.729
LADTree -B 20	58.85%	0.589	0.556
RandomForest -depth 5	66.15%	0.639	0.661

(b) Emotional model with 8 emotions derived from extended Ekman model

Attributes	Accuracy	Recall	Precision
J48 -U -M 3	72.02%	0.714	0.720
BFTree -U -M 3	70.46%	0.685	0.683
LADTree -B 20	52.85%	0.509	0.528
RandomForest -depth 5	68.91%	0.696	0.689

Table 1: Classification results with respect to selected emotional MOODs

Further developments are split into different directions, ranging from improving the data set implementing different filtering and preprocessing phases to the implementation of a feedback process from Zazie users.

A further research could be directed to the use, in the preprocessing phase, of the web based proximity measures analysed in [10], which can return the similarity between two or more words and have been already applied to semantics-driven search engines.

References

1. Yang, Y. H. and Chen, H. H.: Machine recognition of music emotion: A review. ACM Transactions on Intelligent Systems and Technology, 3(3), 1–30, (May 2012)
2. Gilad Mishne: Experiments with Mood Classification in Blog Posts. Style2005, Stylistic Analysis of Text for Information Access, (2005)
3. C. Ovesdotter Alm, at al.: Emotions from Text: Machine Learning for Text-based Emotion Prediction. Proc. of HLT and EMNL Conferences, 579–586, (2005)
4. Paul Ekman: Facial Expression and Emotion. American Psychologist, 48(4), 384–392, (1993)
5. J. Suttles and Nancy Ide: Distant Supervision for Emotion Classification with Discrete Binary Values. CICLing (2): 121-136. (2013)
6. F. Bertola and V. Patti: Emotional Responses to Artworks in Online Collections. In Proceedings of PATCH 2013, vol. 997 CEUR Workshop Proceedings, (2013)
7. E. Pianta, L. Bentivogli and C. Girardi: MultiWordNet: Developing an aligned multilingual database. Proc. of the 1st Int. WordNet Conference, 293–302, (2002)
8. C. Strapparava and A. Valitutti, WordNet-Affect: an affective extension of WordNet. In Proc. of 4th International Conference on Language Resources and Evaluation (LREC 2004), 1083 – 1086, 2004
9. V. Franzoni, V. Poggioni and F. Zollo: Automated Classification of Book Blurbs According to the Emotional Tags of the Social Network Zazie. In Proc of ESSEM 2013, pp 84–93, CEUR Workshop proceedings.
10. V. Franzoni, A. Milani: PMING Distance: A Collaborative Semantic Proximity Measure. WI–IAT, vol. 2, 442–449, IEEE/WIC/ACM (2012).