Deusto

Facultad de Ingeniería
Universidad de Deusto

Ingeniaritza Fakultatea
Deustuko Unibertsitatea

Grado en Ingeniería Informática
Informatikako Ingeniaritzako Gradua

Proyecto fin de grado
Gradu amaierako proiektua

Automatic detection of repressed anger from text messages

Jesús María Sesma Solance

Director: Takashi Yukawa

Nagaoka, Mayo de 2017

Abstract

The growth in the usage of social media, microblog and review platforms has resulted in a significant increase of the access to short text messages that reflect individual's opinion and feelings. Automatic detection of people's emotions has a wide range of applications such as producing systems that measure the satisfaction of customers and thus help companies to improve their products or services. This research project focuses on detecting anger, an emotion that is relative difficult to detect compared to other sentiments due to the usage of linguistics figurative language techniques, such as irony, that intends to communicate the opposite of what it is literally said. To this purpose, a review of the state of the art has been made and an experiment using not only traditional machine learning techniques, such as Neural Networks, K-Nearest Neighbor and Support Vector Machines, but also Deep learning algorithms has been conducted in an open social network like Twitter. The proposed methods define the repressed anger detection as a classification problem and to solve it, the task is divided into two subtasks that complement each other. The first focuses on explicit anger detection, while the second subtask's goal is to detect irony. The system make use of selected featured based on characteristics properties of English language and studies emotion and irony in psychology. The model is composed on features such as: frequency of words, style in written and spoken languages, intensity of adverbs and adjectives, structure of the document, use of emoticons, synonymy, ambiguity and the contrast of sentiment and negative situation.

Keywords

Emotion Analysis, Supervised Classification, One-vs-all Classification, Convolutional Neural Networks, Repressed Anger Detection

Contents

List of Figures

Chapter 2

Chapter 3

List of Tables

Chapter 3

Algorithm Index

1. INTRODUCTION

2. PLANNING AND METHODOLOGY

In order to accomplish the Master's thesis during the established period of eight months, the following plan will be followed:

1. **Exploratory phase:** this period of time will consist in searching and studying the literature related to the research field in order to build a solid theoretical framework on which support the rest of the investigation. Even though this process is presented as a isolated task, it is an fundamental activity that will be executed continuously throughout the whole project.
2. **First contact with the technology:** after having reviewed the first phase of the state of art and checked the advantages or limitations of each architecture, the first empirical tests will be done. These experiments will serve to for a better understanding of the literature read in the exploratory phase and to observe how they behave with the gathered sample data.
3. **Specification and design of the solution:** in this phase, based on the knowledge gathered from the empirical tests in the last point and the relevant literature that has been constantly read, the requirements of the solution that best results can achieve for this research topic will be specified.
4. **Design of the functional prototype and evaluation:** after the requirements gathering, during this task, a prototype that justify all the steps followed this far will be developed, and evaluate its performance.
5. **Thesis document writing:** Finally all the relevant knowledge that has been required to complete this research will be gathered, the procedure to develop the system will be described and the conclusion and future work on this topic will be written down altogether, resulting in this master's thesis document, that will be refined until its laster submission and defense.

The figure 2.1 illustrates the methodology used during the investigation process, in which the results obtained from the prototypes might not be satisfactory or could lead to formulate new questions. To solve those doubts a redesign of the system may be needed. In addition, after reporting the results to project supervisor, new relevant information might be provided. This information has to be added to the knowledge obtained from previous state of the art review and reanalyze if the proposed solution requirements are still valid.

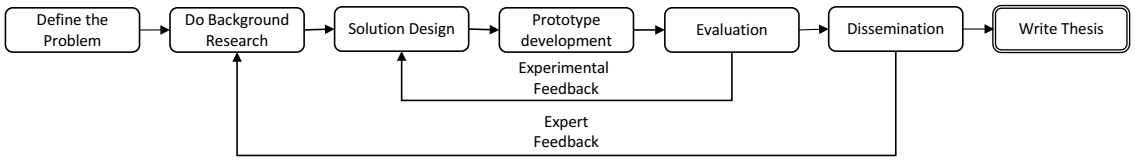

Figure 2.1: Methodology

3. STATE OF ART

TODO

3.1. CLASSIFICATION TECHNIQUES

The aim of this section has two purposes. The first one, is to make an introduction of the basic concepts of classification, which is essential for the detection of repressed anger. The second one, is to explain how the algorithms used in this study work.

3.1.1 Fundamentals of Classification

According to [6], classification can be defined as the task of predicting an outcome from a given input. This outcome is produced by the process of mapping a group of characteristics present in the input to a certain category. In other words, it consists in assigning objects (the input) to one of several predefined classes (the outcome) [5]. Examples of classification can be found in everyday life, such as e-mail spam detection, news classifiers, Optical Character Recognition (OCR), animal kingdom classification (see Figure 3.1), among many others.

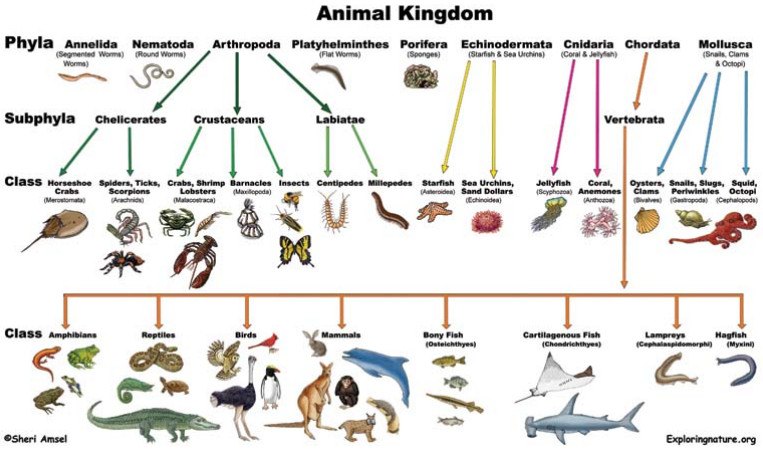

Figure 3.1: Classification of animals. The image is extracted from Exploring Nature.

The input data for a classification task is composed by a collection of records, the dataset. In the same time, each record, also known as an instance, is composed by a set attributes. From all these attributes there is one considered special, which is called the target attribute or the class label. Regular attributes can be both discrete or continuous values. For the values signed for the class label, however, they must be discrete. This characteristic is what distinguishes classification form

regression. Table 3.1 shows a sample dataset for animal classification into the following categories: amphibian, bird, fish or mammal.

Common Name	Hair	Feathers	Eggs	Milk	Aquatic	Legs	Class Label
antelope	Yes	No	No	No	No	4	mammal
catfish	No	No	Yes	No	Yes	0	fish
dolphin	No	No	No	Yes	Yes	0	mammal
dove	No	Yes	Yes	No	No	2	bird
duck	No	Yes	Yes	No	Yes	2	bird
elephant	Yes	Yes	No	Yes	No	4	mammal
flamingo	Yes	Yes	Yes	No	No	2	bird
frog	No	No	Yes	No	Yes	4	amphibian
fruit bat	Yes	No	No	Yes	No	2	mammal
gull	No	Yes	Yes	No	Yes	2	bird
herring	No	No	Yes	No	Yes	0	fish
kiwi	No	No	Yes	No	No	2	bird
lark	No	Yes	Yes	No	No	2	bird
lynx	Yes	No	No	Yes	No	4	mammal
mole	Yes	No	No	Yes	No	4	mammal
mongoose	Yes	No	No	Yes	No	4	mammal
newt	No	No	Yes	No	Yes	4	amphibian

Table 3.1: Animal kingdom dataset.

Tan Pang-Ning et al. propose a more mathematical definition of classification stating that it is the process of learning a target function f, also known as classification model, that maps each attribute set x to one of the predefined class labels y.

Figure 3.2: Classification as a task of mapping a set attributes x into its fitting class label y.

A classification model is useful for the following purposes [5]:

- **Descriptive Modeling:** Since a classification model presents the main features of the data, it can serve as an explanatory tool to distinguish between instances of different categories [4].
- **Predictive Modeling:** A classification model can also be used to predict the class label of an unknown new instance. As shown in Figure 3.2, a classification model can be represented as a black box that automatically assigns a class label to an instance by providing its attribute set.

It is important to remark that classification techniques perform their best when used for predicting or describing datasets which its class label is binary or nominal, Since they no consider properties such ordinality or the implicit order among the categories, they become ineffective with ordinal class labels [1].

3.1.2 General classification problem solving

For general classification problems solving, popular techniques consists on a process that starts with building classification models from a sample dataset [7]. Each technique depends on a learning algorithm witch is in charge of generating the classification model. A good model should define the relationship between the input attribute set and its belonging category that suits the best. Therefore the model should be valid for both, the sample data used to generate the model and also for new unknown instances. Among popular classification techniques Support Vector Machine (SVM), Neural Networks (NNs), Naive Bayes or Decision Trees (DTs) can be found [2].

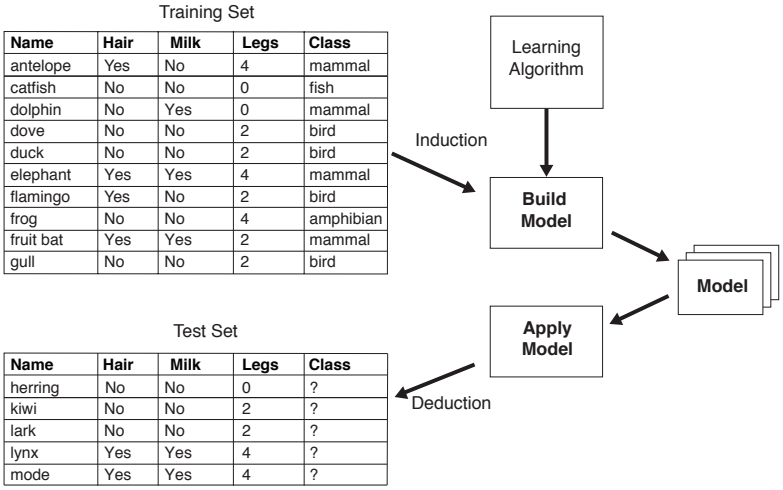

Figure 3.3: General approach for classification model building and new instance category prediction.

As shown in the Figure 3.3, to solve a classification problem a sample dataset must be provided as a training set. This sample is used to build the classification model according to the learning algorithm. After the model is built, it is applied to unlabeled dataset, also called the test set, to predict the categories of each instances of the records. To measure how good the model is, there is only need to count the number of instances have been correctly and incorrectly classified from the test set. Usually, to represent system's performance values, a confusion matrix is used [3].

		Predicted Class	
		$Class = Yes$	$Class = No$
Actual Class	$Class = Yes$	a	b
	$Class = No$	c	d

Table 3.2: Confusion matrix of a binary classification.

As a example, Table 3.2 represents the confusion matrix of a binary classification problem.

4. DEVELOPMENT

5. RESEARCH FRAMEWORK

6. RESULTS

7. CONCLUSION AND FUTURE WORK

8. SELF-ASSESSMENT

Bibliography

[1] Eibe Frank and Mark Hall. 'A simple approach to ordinal classification'. In: *European Conference on Machine Learning*. Springer. 2001, pages 145–156.

[2] GV Garje, Apoorva Inamdar, Apeksha Bhansali, Saif Ali Khan and Harsha Mahajan. 'SENTIMENT ANALYSIS: CLASSIFICATION AND SEARCHING TECHNIQUES'. In: (2016).

[3] Howard Hamilton. *Confusion Matrix*. 2000. URL: `http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matrix/confusion_matrix.html` (visited on 21/10/2016).

[4] David Madigan. *Descriptive Modeling*. 2002.

[5] Tan Pang-Ning, Michael Steinbach, Vipin Kumar et al. 'Introduction to data mining'. In: *Library of congress*. Volume 74. 2006.

[6] Fabricio Voznika and Leonardo Viana. *Data Mining Classification*. 2007.

[7] Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

Acronyms

DT Decision Tree. 7

NN Neural Network. 7

OCR Optical Character Recognition. 5

SVM Support Vector Machine. 7