Experiments with SMS Translation and Stochastic Gradient Descent in Spanish Text Author Profiling

Notebook for PAN at CLEF 2013

Andrés Alfonso Caurcel Díaz[1] and José María Gómez Hidalgo[2]

[1]Universidad Politécnica de Madrid
[2]Optenet
[1]AARcaurcel@gmail.com, [2]jgomez@optenet.com

Abstract Inspired by our ongoing work in the project WENDY, which addresses age detection in social networks by linguistic processing (among other methods), we have built a system that makes use of a number of linguistic resources (a Spanish dictionary, and a SMS-language dictionary) and algorithms (custom text utterances tokenization, SMS to standard Spanish translation, and a number of normalization rules) in order to apply a learning-based approach using a custom Stochastic Gradient Descent algorithm adapted to text, to the Spanish Author Profiling task at PAN'2013. We believe the results obtained in internal testing on a validation set extracted from training dataset do validate our approach in WENDY, while the results obtained in the PAN task are not as good as expected.

1 Introduction

Age Detection has multiple applications, including marketing and security. In Optenet, we face Age Detection as an intermediate step in the overall process of protecting children in the Internet [1]. In fact, we are interested on two particular cases:

- Children posing as adults to be members of a social network. In most European countries, users aged below a certain threshold (typically 14) must not join a social network without explicit parent consent. Most social network operators just react at users claims, but in order to protect minors, a more proactive approach is required. We seek to detect youngsters or children below 14 (either in the case they have omitted their age, or they have entered a false, over 14 one), in order to provide this function as a service for parents and social network operators.
- Adults posing as children in order to establish contact with minors, and sometimes, to perform harassment or sexual predation. In this case, we would be interested on monitor "fake" children (users over 18 but with an underage profile) to check which children they communicate, and to alert their parents or the social networks operators about the contact or when detecting inappropriate behavior.

In order to detect these cases, Optenet research team has been developing a project named "WENDY: WEb-access coNfidence for chilDren and Young"[1] in collaboration

[1] See: http://wendy.optenet.com, in Spanish.

with the Group of Biometrics, Biosignals and Security from the Polytechnical University of Madrid. The project adopts a multi-biometric approach based on monitoring user profiles, communications (posts, chats, etc.) and pictures, and classify the users into three age ranges: $[0-13, 14-17, 17-)$. Classification is performed by a multi-modal classifier which combines several individual classifiers based on the analysis of different kinds of data. For instance, the output of specialized classifiers for the profile information, the posts and the pictures is combined into a single prediction by a second level classifier that has been trained on the outputs for a validation user dataset.

While the age ranges are different from those addressed in the PAN Authorship Profiling task, and we do not address gender recognition in the project WENDY, most of the text-based techniques used in this project can be applied to the task. In fact, we restrict ourselves to the Spanish language case as this is the one addressed in WENDY, and our linguistic resources are language-dependent. In consequence, we take our participation in this task as an opportunity to validate WENDY's text analysis approach in a different domain (from social networks to chats but within the same language).

The general approach followed in several WENDY's mono-modal classifiers is learning a content-based text classifier on a training collection, and it is inspired by previous work by Tam and Martell [3]. This is just the approach we follow in the PAN Author profiling tasks, and it consists of two phases: text analysis and representation, and classifier learning. We discuss these phases in the next sections, along with the results we have obtained in our experiments on the training collections and the lines of future work.

2 Text Analysis

As we face very noisy text, plagued with typos, abbreviations, emoticons and so on, we have designed a pipeline of language analyzers aimed at reduce this noise. The overall process consists of the following steps:

1. Text tokenization via a specific tokenizer designed for noisy Spanish language texts.
2. Token normalization according to a set of linguistic rules designed for noisy Spanish text.
3. SMS word-by-word language translation to standard Spanish by using two language-specific dictionaries.

We discuss each process in the next subsections.

2.1 Text Tokenization

The usage of specific language codes and chat and SMS-like messages is a major trend in electronic communications. This fact makes Natural Language Processing quite hard, even at the simplest step for text message tokenization, due to the widespread usage of non-alphanumeric symbols, frequent typos and non-standard word separators.

Aiming at recognizing this type of language, we have developed a tokenizer that features two steps. The first consists of splitting the original text string into candidate tokens separated by white spaces. For every candidate token, we consider three possibilities:

– It is a sequence of alphanumeric characters, thus most likely making a proper word.
– It is a sequence of punctuation and non-alphanumeric characters, possibly a smiley.
– It is a mixture of both.

In the first two cases, we consider that the character sequences are already words themselves. This implies that non-alphanumeric character sequences, which in particular include punctuation symbols isolated tokens are considered relevant. In the third case, we proceed to a second tokenization step by splitting the candidate token into all sub-sequences of continuous alphanumeric and non-alphanumeric characters. For example, given the Spanish string "Hola:-)ketal"[2], this step would separate it into following tokens: "Hola", ":-)", and "ketal".

As we make use of the learning package WEKA[3], we have implemented this tokenization algorithm in Java, based on the previously existing tokenizer NgramTokenizer[4] in WEKA.

2.2 Token Normalization

The informal Spanish language used in forums and social networks has motivated the development of analyzers able to deal with phenomena such as the following ones:

– Contamination with typical SMS language abbreviations (like e.g. "tkm" – "I love you", "bss" – "kisses", "dsp" – "after", etc.).
– Alternating lower and upper case letters (e.g.: "ThIS iS An ExAMplE").
– Repeating letters (e.g.: "heeeeeellloooooo!").
– Omission of the letter "u" in the syllables "que" or "qui", or replacement og "qu-" by "k" (e.g.: "qiero", "kiero").
– Replacing the syllable "ca" or the letter "c" with the letter "k" (e.g.: "kompra kfe" – "buy kofee").
– Intentional misspellings as "soi" ("I am"), "voi" ("I go"), "i" instead of "y", etc.

These phenomena and others are recognized and analyzed by the Deflogger normalization tool[5], which is a system developed to standardize the language used in the social network Fotolog[6]. We apply this system to the token sequences obtained in the previous step.

2.3 SMS Language Translation

We have developed a system for translating textual elements using a number of existing linguistic resources, namely: a dictionary of SMS-like language for Spanish[7], and a Spanish language dictionary[8].

The translation system operates as follows:

[2] Approximate translation into English: "Hello:-)whatsup".
[3] See: http://www.cs.waikato.ac.nz/ml/weka/.
[4] See: http://weka.sourceforge.net/doc.dev/weka/core/tokenizers/NGramTokenizer.html.
[5] See: http://code.google.com/p/deflog/.
[6] See: http://www.fotolog.com.
[7] See: http://www.diccionariosms.com/contenidos/.
[8] The one included in the linguistic analysis system Freeling: http://nlp.lsi.upc.edu/freeling/.

1. Given a target word (possibly an expression in SMS), we search for it in the standard Spanish dictionary.
2. If the word is present in the dictionary, the process is terminated. If the word is not present in the dictionary, then we search for it the SMS language dictionary.
3. If the word is present in the SMS language dictionary, we select the most common or popular meaning or translation. If the word is not present in the SMS dictionary, we leave the token unchanged.

It should be emphasized that the SMS dictionary includes not only popular expressions as "xq" (the abbreviation of "porque", i.e. "because of"), but it also contains a significant amount of emoticons (like e.g. ":-)" etc.). The previous tokenization system is able to preserve these symbols, allowing us to translate them when possible.

3 Classifier Description

We follow a text learning approach in which we build a classifier using the Stochastic Gradient Descent for Text (SGDtext) algorithm[9] as implemented in WEKA. This method employs the Stochastic Gradient Descent (SGD) [4] algorithm combined with a the WEKA String To Word Vector (STWV) filter in order to build the dictionary and update it in successive iterations. Two particular characteristics of the SGDtext algorithm is that it is specifically designed for working on text problems, and that it is update-able; that is, it builds the learning model incrementally, which allows us to train it in successive instance batches.

The SGD algorithm works as a batch learning method approaching the space defined by feature vectors using a loss function to converge. The vectors used in SGDtext are obtained from filtering plain text instances with the STWV filter, that transforms those text strings into term-weight vectors according to the Vector Space Model [2]. The STWV filter transforms a string attribute in a series of numeric attributes, corresponding with all words in the string and their frequencies.

The options used in the SGDtext classifier are: support vector machines as the loss function, a learning rate of 0.01, a regularization constant of 0.01, 500 iterations without pruning the dictionary, 3 as minimum word frequency, default normalization and no transformation to lower case the input instances. The tokenization of text strings is performed with the tokenizer described previously as a parameter. As the tokenizer is a version of the existing NgramTokenizer, it is possible to build n-grams instead of isolated tokens as text representation terms. We have configured the tokenizer to build unigrams, bigrams and trigrams in order to capture meaningful word sequences.

Regarding the training process, we have divided the original Spanish into four parts. We have accumulated the 30% of longest text files for training an initial model. Then we have randomly built two additional instance batches containing 30% of the instances for incremental learning, and an additional 10% batch kept for testing. In terms of absolut numbers, the original 56,126 have been divided in three batches of 17,176 instances and one batch for testing containing 4,597 instances.

[9] See: http://weka.sourceforge.net/doc.dev/weka/classifiers/functions/SGDText.html.

Due to time constraints and computational limits, we have only been able to train the models on the original batch for both (age and gender prediction) problems, and updated the age prediction classifier with the second learning batch. Thus, the results we have obtained are limited and they do not show the full power of our approach.

4 Results and Analysis

In the table 1 we show the results of the age prediction classifier obtained with our method when evaluated on the test batch we have produced. In the first three columns we show the contingency matrix (with decisions in columns while actual classes in rows), while in the two latest columns we show the precision and recall for each class.

	10	20	30	Prec.	Rec.
10s	78	2	1	0,299	0,963
20s	104	2367	14	0,552	0,552
30s	79	1919	34	0,694	0,017

Table 1. Test results for the age prediction problem.

The results obtained for the age prediction problem are not good in general for the particular setup of the PAN'2013 author profiling task, but they are completely aligned with the kind of setup we face in the WENDY project. While precision for the (very under-represented) 10s class is low, the recall is very high – that is exactly what we have tried to achieve in WENDY, where it is very important to detect all children below 14 because they should not be in a social network without explicit (and legal) parent permission. As in WENDY, we have detected here that an important issue of our approach is that it is not able to discriminate between ages over 18 – which is not a problem, as far as our goal in WENDY is to detect people over 18 posing as minors.

In the table 2 we show the results for the gender prediction classifier on our test batch. The table is similar to the previous one except for the fact that there are only two classes.

	male	female	Prec.	Rec.
male	87	2211	0,509	0,0379
female	84	2215	0,500	0,963

Table 2. Test results for the gender prediction problem.

The results obtained for gender prediction in our internal tests show a major trend of classifying every instance as "female", i.e. the classifier approaches the trivial rejector. As it is a first approach obtained by reusing exactly the same method used for age prediction, and it is clearly under-trained, those results can be expected but anyway they are far from good.

5 Conclusions and Future Work

As an additional validation of the approach we have followed in the WENDY project for a close (but not identical problem), the results we have obtained in our internal experiments fulfill our expectations. In this sense, we must note that PAN Author Profiling and WENDY tasks face different problems. On one side, in PAN we work with individual blog posts while WENDY's goal is to detect age problems on a continuous stream of social network updates and profile information, which can include pictures and other data types as well. On the other, classes and problem statements are clearly different in both tasks; in WENDY, recall is extremely important for children below 14, while precision is a priority for adults (over 18) as they are a majority in the social network Tuenti. However, in PAN all classes are identically important in terms of accuracy metrics.

In the near future, we plan to finish the full experiment on the available training data, and compare it with the same setup without using linguistic resources. We want to extend both experiments to the English collection as well – for this task, specific linguistic resources have to be acquired, and a set of linguistic rules similar to Deflogger ones must be derived for this language.

Acknowledgments

This work has been performed with partial support from the Ministry of Industry, Energy and Tourism and the Center for Industrial Technological Development (CDTI) under the granted project "WENDY: WEb-access coNfidence for chilDren and Young" (TSI-020100-2010-452).

References

1. Gómez Hidalgo, J., Caurcel Díaz, A.: Avances tecnol'ogicos en la protecci'on del menor en redes sociales. Nov'atica, Revista de la ATI (218) (2012)
2. Sebastiani, F.: Machine learning in automated text categorization. Computing Surveys 1(34), 1–47 (2002)
3. Tam, J., Martell, C.H.: Age detection in chat. In: Proceedings of the 2009 IEEE International Conference on Semantic Computing. pp. 33–39. ICSC '09, IEEE Computer Society, Washington, DC, USA (2009)
4. Zhang, T.: Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: Proceedings of the twenty-first international conference on Machine learning. pp. 116–. ICML '04, ACM, New York, NY, USA (2004)