Combining Machine Learning Techniques and Natural Language Processing to Infer Emotions Using Spanish Twitter Corpus

Gonzalo Blázquez Gil, Antonio Berlanga de Jesús, and José M. Molina Lopéz

Applied Artificial Intelligence Group, Universidad Carlos III de Madrid,
Avd. de la Universidad Carlos III, 22, 28270, Colmenarejo, Madrid, Spain
{gonzalo.blazquez,antonio.berlanga,josemanuel.molina}@uc3m.es
http://www.giaa.inf.uc3m.es

Abstract. In the recent years, microblogging services, as Twitter, have become a popular tool for expressing feelings, opinions, broadcasting news, and communicating with friends. Twitter users produced more than 340 million tweets per day which may be consider a rich source of user information. We take a supervised approach to the problem, but leverage existing hashtags in Twitter for building our training data. Finally, we tested the Spanish emotional corpus applying two different machine learning algorithms for emotion identification reaching about 65% accuracy.

Keywords: Emotion Context, Emotion Recognition, Microblogging, Twitter, Features extraction, Machine Learning.

1 Introduction

Affective Computing (AC) or Emotion-oriented computing is a branch of Artificial Intelligence (AI) that deals with the design of systems and devices that can recognize, interpret, and process human affective states (moods and emotions).

In [1] Picard described three types of affective computing applications: 1) systems which detect user emotions, 2) systems that express what a human would perceive as an emotion (e.g., an avatar, robot, and animated conversational agent), and 3) systems that can actually "feel" an emotion. This paper we will try to create a system which detect user emotions from text using Social Networks Sites.

The question of how humans perceive emotions has become central for the researchers of affective computing [2]. Emotions are fundamental to human experience, perception, and everyday tasks such as learning, communication, and even rational decision-making. Hence, to create a systems able to recognize emotions gives us new clues to understand people behavior.

Although human emotion sensing may be obtained from a wide range of behavioral cues: gestures, facial expression [3], movements [4], speech or physiological signals (heart rate, salivation, ...). In this case, thanks to the rapid growth

J.M. Corchado et al. (Eds.): PAAMS 2013 Workshops, CCIS 365, pp. 149–157, 2013.

of textual content, such as microblog posts, blog posts, forum discussions, and social networks sites (SNS), we propose to develop an automatic tool for identifying and analyzing people's emotions expressed through Computer Mediated Communication (CMC), in concrete using Natural Language Processing Techniques (NLP).

NLP is the application of computational models to tasks involving human language text. NLP research has been active since the dawn of the modern computational age in the early 1950s, but the field has grown in recent years, thanks to the amazing development of the internet and consequent increase in the availability of online text. Nowadays, following the trend, the research in the field of emotion detection from textual data emerged to determine human emotions from another point of view.

Previous works in emotion recognition using NLP methods used small datasets, about thousands of entries, which makes difficult to well-define which emotion is triggered by an events or situations. To overcome the lack of sufficient labeled data is possible to use Social Networks Sites where daily users share their personal information [5]. SNS manage an uncountable gigabytes of useless user information and it is possible to consider SNS's as an emotional sensor [6].

There are different SNS; Twitter, Facebook, Instagram, etc; however not all the SNS are well-fitted to retrieve Emotions from text content. Twitter contains a very large number of short messages (140 characters) created by users. Twitter's audience varies from normal people to celebrities, company representatives, politicians, etc. Relying on the twitter hashtags which are used to mark keywords or topics in a Tweet, we automatically develop an user emotion-annotated training dataset.

The paper deals with the topic of recognizing people's emotion context by analyzing data from Twitter (Microblogging platform).

2 Natural Language Processing

While express emotion through face-to-face channels is easy to recognize, Computer-Mediated Communication (CMC) may be cause confusion. To understand nuances of the expressions, jokes, detecting subjective opinion documents or expressions, non-verbal cues may be an arduous task for humans and an impossible task for computers.

Identifying the expressed emotions in text is very challenging for at least two reasons. The first one is that emotions can be implicit by specific events or situations. In the next sentence *When I see a cop, no matter where I am or what I'm doing, I always feel like every law I've ever broken is stamped all over my body*, it is possible to infer that the person is scared or fear. Second one, gathering distinction between different emotions purely on the basis of keywords can be very subtle.

Although there is not any standard emotion word hierarchy, focus on the related research about emotion recognition, normally emotion is expressed as joy, sadness, anger, surprise, hate, fear according to the Ekman six basic emotions [7].

In the context of emotion detection NLP is normally based on finding certain predefined keywords as happy, sad, anger, etc. A little overview about NLP features extraction techniques is presented:

- Part-of-Speech (POS): In corpus linguistics, part-of-speech tagging is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on its definition, as well as its context. It is also called word class, a category into which words are placed according to the work they do in a sentence. Commonly, there are 8 parts of speech (or word classes) and they are divided into two groups:
  - Open classes: nouns, verbs, adjectives, and adverbs.
  - Closed classes: pronouns, prepositions, conjunctions, and interjections.

  The most common way to classify using POS features is reduced to calculate the percentage of words belonging to each POS in a tweet.
- LIWC Dictionary[1]: Linguistic Inquiry and Word Count3 (LIWC) is a text analysis software which provides a dictionary covering about 4,500 words and word stems from more than 70 categories. The software is available in 11 languages (Spanish is included).

  In this case, the classification method counted the number of positive/ negative words based on the set of collected emotion words, and used the percentage of words that are positive and that are negative as features.
- Adjectives: In sentiment analysis, adjectives are usually considered as effective features since they can be good indicators of sentiment. Some research [8] shows that using adjectives alone produces competitive results with those obtained by using n-grams in sentiment classification of movie reviews. In order to classify each tweet adjective is included in a feature vector.
- Emoticons: Other way to face NLP is rely on the used emoticons. Some recent work, however, notes that emoticons can provide emotion information and improve CMC [9]. Emoticons are described as graphic representations of facial expressions that are included in electronic messages.
- N-grams: In the fields of computational linguistics and probability, an n-gram is a contiguous sequence of n items from a given sequence of text or speech. An n-gram could be any combination of letters. However, the items in question can be phonemes, syllables and letters, although using words give more information to the developer.

2.1 Emotion Representation

As well as the emotion does not have a commonly agreed theoretical definition, a categorization or representation model there is no consensus. Nowadays, there exist two different ways to depict emotions: Categorical and dimensional.

Categorical model of emotion has its roots in the evolutionary theories which claims that emotions are biologically determined, discrete and belong to one

[1] `http://www.liwc.net/`

of a few groups. These groups are consider fundamental or *basic*. However, the problem is which emotions are considered basic. In this case, according with [7] definition of affective state the basic emotions are normally considered: happiness, sadness, surprise, fear, anger and disgust.

Fig. 1. Emotion representation: Categorical (a) and Dimensional model (b)

This model reduce sharply the number of emotions. Some researchers think that any basic emotion may be decomposed into secondary emotions. This process is very similar to the way that any color is a combination of some basic colors. Emotions are extracted by mixing and matching the basic emotional labels as if in a palette of primary colors *Palette theory* as figure 1(a) shows.

In contrast to categorical model, dimensional models do not fix a finite set of emotions. Alternately, they attempt to find a finite set of underlying features into which emotions can be decomposed, any combination of features give a different affective state. Under this model, emotions are described in terms of three components or dimensions [10]. The first dimension aims to describe the degree of pleasantness underlying the emotional experience. The second one describes the level of activation of the emotion and finally the last one defines the level of attention or rejection.

The three dimension approach is synthesized in figure (b) where a concrete emotion (e) is the result of the intersection between every different dimensions (d) whose values are determined by pattern of signals (s).

3 Building a Data Set for Emotion Analysis in Twitter

Due to Twitter restrictions is not possible to use a previous Twitter emotion dataset [11] to compare machine learning techniques. We had to create our own dataset to test NLP techniques in Spanish Tweets.

In this section, it is described how we automatically created a labeled emotion dataset from Twitter SNS. Selected emotions were 6 (fear, anger, sadness, happiness, surprise, and disgust.) according to the Ekman research [7], also called six universal emotions.

Table 1. Matching between emotion hashtags with six universal emotions

Emotion	Hashtag	Instances
Fear	#Miedo, #terror and #aprension	19.39%
Disgust	#Indignado, #asco and #repulsivo	23.74%
Sadness	#Triste, #sad and #infeliz	18.80%
Happiness	#Feliz, #happy and #contento/a	36.28%
Surprise	#Sorprendido, #sorprendida and #sorpersa	0.90%
Anger	#Furioso/a, #cabreado/a, #mosqueado/a and #enfadado/a	0.85%

We firstly collected at least 3 sets of emotion words for 6 different emotions (e.g., word "feliz" for emotion happiness) from existing psychology literature [12]. Subsequently, we retrieved tweets that have one of these emotion words as a hashtag (e.g, #feliz) using Twitter streaming API. Each collected tweet was automatically labeled with one emotion according to its hashtag (See Table 2).

Full sentiment analysis for a given question or topic requires many stages, including but not limited to:

1. Extraction of tweets using Twitter4J which is an unofficial Java library for Twitter API.
2. Filtering out spam and irrelevant items from those tweets. The main filtering steps the we follow are:
   - Anonymized username: We anonymize the usernames since they do not provide relevant emotional information and also in the way to avoid malicious use of the data.
   - Manual retweets (also known as "RT") are deleted because they do not give us relevant information.
   - Tokenization is difficult in the social media domain, and good tokenization is absolutely crucial for overall system performance. Standard tokenizers, usually designed for newspapers or scientic publications, perform poorly because of the Twitter slang. However, we create a tokenizer which treats hashtags, @-replies, abbreviations, strings of punctuation and emoticons as tokens.
   - Removing stopwords we remove prepositions and conjunctions from the set of words since they do not provide enough meaning to the Tweet.
   - Delete repeated characters: All repeated characters like spaces or repeated vowel are deleted in order to join words with the same meaning and slang differences (e.g. holaaaaaaa -¿ hola).
   - Negation form: "no" word is attached to the word which follows it. For example, th next sentence "No quiero ir" will form two different tokens: "no+quiero", "ir". Such a procedure allows to improve the accuracy of the classication since the negation change completely the meaning of

the sentence since it plays a special role in an opinion and sentiment expression [13] and [14].

3. Identifying subjective tweets. A set of filtering heuristics was developed to select the most valuable tweets:
   - We kept only the tweets with the emotion hashtags at the end. In previous works was proved that the most relevant words are at the end of a Tweet [15].
   - We discarded tweets which have less than five tokens, since they may not provide sufficient context to infer emotions.
   - URL del Tweets which contains URL links since the relevant information is stored in the link (e.g. http://example.com).

After the filtering process was conclude, totally, we collected 21,991 relevant tweets from a period spanning December 28th 2012 until January 8th 2012.

4 Emotion Classification Results

We train classifiers with unigram features for each emotion class using Multinomial Naive Bayes (MNB) for predicting the emotion category of the sentences in our corpus. MNB provides good performance with a large-scale dataset and has previously given good performance in sentiment classification experiments.

Table 2. Machine learning accuracy (ngrams)

Features	**Number of ngrams**	**Accuracy**
ngram(n=1)	2264	65.12%
ngram(n=2)	1381	47.64%
ngram(n=3)	164	36.40%
ngram(n=1,2)	3645	49.72%

According to the described features above, one of the best method to analyze emotions in microblogging context is using N-grams. The most common sizes for n are 2 (bigrams), 3 (trigrams) and 4 (four-grams) because unigrams are too narrow a unit of analysis.

In each experiment, we represent every sentence by a features vector indicating if a ngrams appears in the sentence or not. It is made a Boolean feature for each n-gram, which is set to true if and only if the n-gram is present in the tweet.

Our main goal for these experiments is to compare different features in NLP using Spanish Twitter Corpus. Taking into account microblogging text characteristics which maximum text length is 140 characters, we chose small n values. Hence, we decided to compare results between different values of n: Unigrams (n=1), Bigrams (n=2), Trigrams(n=3) and the Unigrams and Bigrams (n=1, 2) combination.

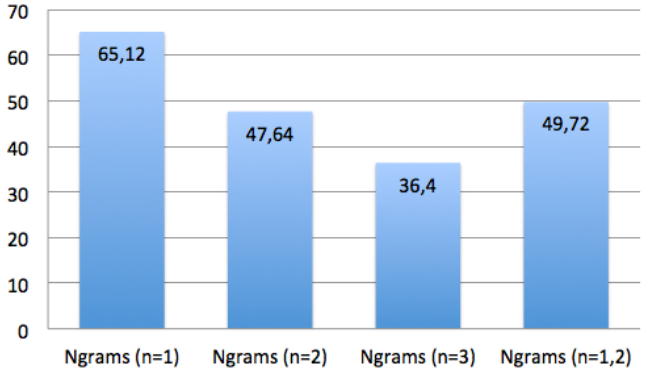

Fig. 2. ngram total accuracy

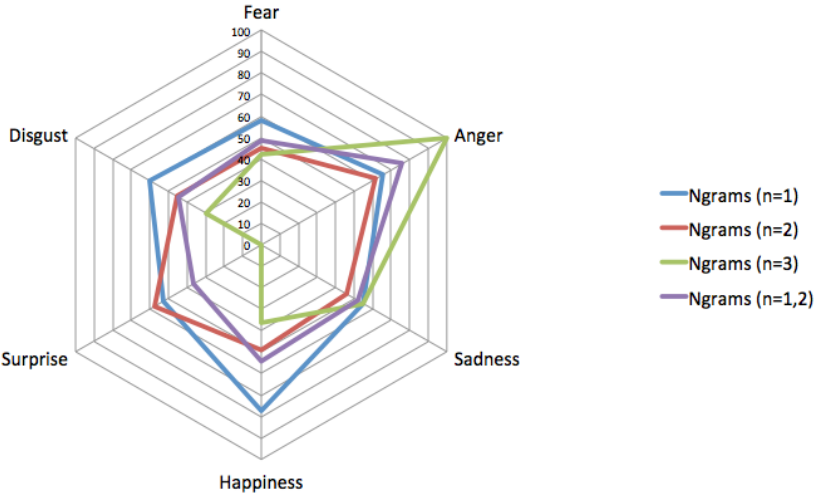

Fig. 3. Accuracy for each emotion

In the first experiment, we use only corpus based unigram features. We obtain high precision values for all emotion classes (as shown in Table 4). Besides, Table 4 shows the overall performance of MNB classifier (trained with all tweets) on each emotion category.

Our experimental results show that unigrams yields better performance than using unigrams alone. While the number of ngrams are increasing the accuracy decreases (from 65.12% to 36.40%). The number of ngrams and the accuracy show that unigrams provides the best performance to infer twitter user emotion. This validates our previous premise since we consider unigrams can help learn lexical distributions well using short sentences (AS Tweets) in order to accurately predict human emotion categories.

It is important to highlight that for the three most popular emotion (joy), which account for 36.28% of all tweets, the classifier achieves precisions of over 75% (Unigrams). On the contrary, performance declines can be seen on less popular emotions (i.e., Surprise and Anger), which consist of 1.75% of all the tweets in our dataset. The precisions of these two emotion categories are relatively high (with lowest precision of 58.1%).

Interestingly, how is decreasing continuously the performance on the evaluation data comes from using bigger n-grams together with the lexicon features and the microblogging features.

Specifically, combining unigrams and bigrams decrease the accuracy to 49.72%. Hence, further incorporation of trigrams was not implemented due to bad result for using one of them alone. As well as existing works on NLP emotion recognition [8] using unigrams alone is better than applying either bigrams, trigrams or a combination of unigrams and bigrams.

5 Conclusions

In this paper, we investigate the utility of linguistic features for detecting the sentiment of Twitter messages in Spanish. Besides, we evaluate whether our training data with labels derived from hashtags is useful for training emotional classifiers.

Moreover, we culled Spanish emotion tweets covering 6 emotion categories for automatic emotion identification. The experimental results show that the feature of unigrams presents better performance than, bigrams, trigrams and the combination of both of them. We achieved the highest accuracy of 65.12% with is more or less the same accuracy that other researchers have obtained in previous works using English Twitter datasets.

Considering future works are to increase the accuracy of the classication, we should discard common n-grams, measured using Chi-squared. For example taking the top 1,000 n-grams. Besides, it is possible to reduce misspellings and grammatical error in order to unify ngrams.

Acknowledgments. This work was supported in part by Projects MEyC TEC2012-37832-C02-01, MEyC TEC2011-28626-C02-02 and CAM CONTEXTS (S2009/TIC-1485).

References

1. Picard, R.: Affective computing (1997)
2. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.: Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine 18(1), 32–80 (2001)
3. Soyel, H., Demirel, H.: Facial expression recognition using 3D facial feature distances. In: Kamel, M., Campilho, A. (eds.) ICIAR 2007. LNCS, vol. 4633, pp. 831–838. Springer, Heidelberg (2007)

4. Reilly, J., Ghent, J., McDonald, J.: Modelling, classification and synthesis of facial expressions. In: Affective Computing: Emotion Modelling, Synthesis and Recognition, pp. 107–132
5. Wang, W., Chen, L., Thirunarayan, K., Sheth, A.: Harnessing Twitter 'Big Data' for Automatic Emotion Identification (2012), knoesis.wright.edu
6. Blázquez Gil, G., Berlanga, A., Molina, J.: Incontexto: Multisensor architecture to obtain people context from smartphones. International Journal of Distributed Sensor Networks 2012 (2012)
7. Ekman, P., Friesen, W.: Facial action coding system: A technique for the measurement of facial movement (1978)
8. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: Sentiment classification using machine learning techniques. In: Proceedings of the ACL 2002 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 79–86. Association for Computational Linguistics (2002)
9. Derks, D., Bos, A., Von Grumbkow, J.: Emoticons and online message interpretation. Social Science Computer Review 26(3), 379–388 (2008)
10. Schlosberg, H.: Three dimensions of emotion. Psychological Review 61(2), 81 (1954)
11. Petrovic, S., Osborne, M., Lavrenko, V.: The Edinburgh twitter corpus. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, pp. 25–26 (2010)
12. Shaver, P., Schwartz, J., Kirson, D., O'Connor, C.: Emotion knowledge: further exploration of a prototype approach. Journal of Personality and Social Psychology 52(6), 1061 (1987)
13. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. Computational Linguistics 35(3), 399–433 (2009)
14. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: Proceedings of LREC, vol. 2010 (2010)
15. De Choudhury, M., Counts, S., Gamon, M.: Not all moods are created equal! exploring human emotional states in social media. In: Sixth International AAAI Conference on Weblogs and Social Media (2012)