Predicting Sentiment Evolution via Hidden Markov Models

D. Naskar, M. Rebollo, E. Ondainda
Universitat Politcncia de Valncia
Camino de Vera s/n. 46022 Valencia

1 Introduction

Sentiment analysis in social networks is a current research area that covers many aspects, including sentiment detection, classification and evolution with time. The retweet mechanism in Twitter is a powerful diffusion tool and it makes the original tweets to be quickly propagated through cascades of information. Some studies about sentiment evolution have detected correlations among sentiment variation and some cultural, social, economic or politic events or changes in the weather.

The goal of the present work is to provide a mathematical model that, given a set of tweets related with some event (identified by the usage of a hashtag), determines how those sentiments will be distributed or which one will be the predominant sentiment. The main difference with other approaches is that we consider the person has a different sentiment that the emotion shown in the tweet.

Hidden Markov models (HMM) [1] adapts naturally to this situation. In these models, transitions among a set of non-observable states are defined. From each one of the hidden states, one or more observable emissions can be generated with some probability. Non observable states correspond with the actual, hidden sentiments of the person, whereas the observable emissions are the emotions associated to the written tweets.

2 Sentiment Modelling

A circumflex model has ben chosen to represent the sentiments [2]. They are distributed over a circular space. Any affective experience is the consequence of the linear combination of two factors: the valence (positive/negative) and arousal (activation) (see Figure 1). The set of tweets to be analyzed is formed by all public tweets that contain a determined hashtag. Once they havebeen stored, they are

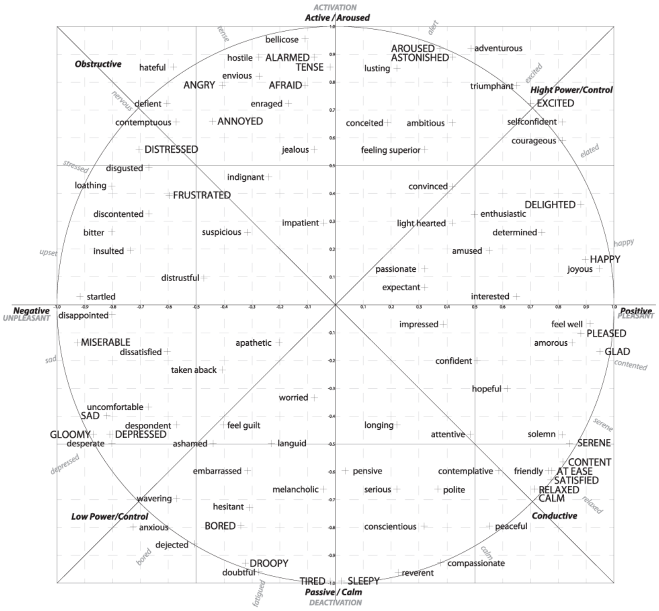

Figure 1: Russell's model of sentiments according the values for valence and arousal

preprocessed to eliminate stop words and the stem words are extracted. To associate the corresponding sentiment, a dictionary that includes the average value for the valence and the arousal and its standard deviation is used. The final value for the valence (arousal) is the weighted average of the valences (arousals) of the the words that appear in the dictionary. From this values, the corresponding sentiment is calculated in the Russell's model.

3 Sentiment Analysis Using HMM

To create the HMM, the probabilities of the transitions among the hidden states and the emissions of the observable ones must be calculated. Let be a user u_i and $t_i = \langle t_i^1, t_i^2, \ldots, t_i^n \rangle$ the sequence of his/her written tweets. To train the model, another sequence with the hidden states $s_i = \langle s_i^1, s_i^2, \ldots, s_i^n \rangle$ is needed. They are calculated from the read tweets, considering that a user has read all tweets in which he is explicitly mentioned. HMM is trained with the 80% of the sequences t_i and s_i of the users, using the Baum-Welch algorithm until it converges. Once the HMM is fitted, it is tested with the remaining 20% of the samples. For each one of the observed sequences t_j for testing, the log-likelihood value $L(M) =$

	Dataset	30-70	50-50	Random	Best	Lin Seq	Markov
Train	Brixit	-4494.3	-4481.7	-4473.6	-4467.0	-4484.5	-4506.8
	Rio	-423.72	-424.55	-417.28	-418.23	-425.57	-418.95
Test	Brixit	-1840.2	-1840.1	-1835.6	-1834.2	-1839.1	-1845.4
	Rio	-85.823	-85.612	-84.776	-84.776	-92.433	-86.517

Table 1: Log-likelihood of the selected models: 30-70 and 50-50 proportion of read/written tweets, uniform random initialization, best of all random, linear sequence and Markov model without hidden sentiments

$\sum_j \log P(M|t_j)$ is calculated, being $P(M|t_j)$ the probability for the sequence t_j to be obtained in the model M. The best model is the one with the highest log-likelihood value $\hat{M} = \arg\max_M L(M)$.

4 Results

To determine the validity of the proposal, different sets of tweets related with different events have been extracted from Twitter. Two representative cases have been chosen: the final of badminton match in the Rio Olympic Games (Aug 2016) and the referendum about United Kingdom's withdrawal from the European Union held on June 23 2016. For *Rio*, we collected over $10\,189$ tweets corresponding to 8496 users and $25\,074$ tweets corresponding to $21\,820$ users for *Brexit*. For each case, the following models have been created:

1. Markov model without hidden states
2. Linear segmentation, assuming that hidden states are arbitrarily ordered and transitions are never backwards. Emissions are uniformly assigned.
3. Proportion linear combination of read and written tweets. Two cases are considered: a 30-70 proportion (read tweets weights 0.3 and written ones 0.7) and 50-50.
4. Random initialization
5. Multiplicative factor to force the hidden sentiment to generate the same emission with a higher probability. The factor varies from 1 (equivalente to a random initialization) to 100. 100 samples have been generated for each factor.

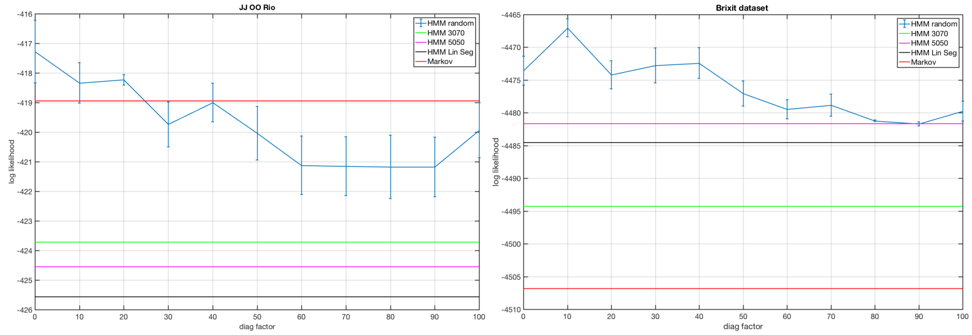

Figure 2: Log-likelihood of the tested methods

5 Discussion

HMM with multiplicative factor obtains the best results in all cases, using a relative small factor. To force a higher correlation between the hidden sentiment and its corresponding expression is not appropriate. We think that this is because (i) the context of the words it is not included, and negation (for example) is not taken into account, and (ii) it is usual that a person writes about meny different things, with different emotions that depends on the topic. Linear segmentation is a good approach for image or natural language recognition, but it is not working for sentiment analysis because any change in the sentiments may be allowed. Finally, in events in which the sentiment is strongly polarized in one direction or the other, HMM does not work well generally and Markov models obtain good enough results. That is the Rio case. But when there is a mixture of sentiments, such as in the Brexit case, then any HMM initialization provides better results.

References

[1] Baum, L.E. . *An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of a Markov Process*. Inequalities. 3: 18 (1972)

[2] James A. Russell. *Core affect and the psychological construction of emotion*. Psychological review, 374 110(1):145172. (2003)