Using Markov Models to Learn the Sentiment of Soccer Fans from Bets and the Result of Matches

Rafael Bomfim
Universidade de Fortaleza - UNIFOR
Av. Washington Soares, 1321, Edson Queiroz
Fortaleza, Brazil
rafaellpontes@gmail.com

Vasco Furtado
Universidade de Fortaleza - UNIFOR
Av. Washington Soares, 1321, Edson Queiroz
Fortaleza, Brazil
vasco@unifor.br

***Abstract*—In this paper we investigate variations of Hidden Markov Models (HMM) as a viable tool for predicting the sentiment of soccer fans based on information regarding the result of matches. The models were constructed from data collected from a social network where fans of a soccer team periodically express feelings towards their team. Our claim is that the change in a fan's sentiment is analogous to a Markovian process of change of state through time. A comparative evaluation performed between variations of the proposed models showed that a second order HMM, considering the match results and fan's gambling information, is the most accurate model.**

Keywords—Hidden Markov Models; Sentiment Analysis; Social Networks

I. Introduction

The interactions performed in digital media, which characterize the era of Web 2.0, allow for studies previously unimaginable of human social behavior. The extraction of patterns from social networks, for instance, provides affordable and reliable means to analyze crowd sentiment. Not surprisingly, researchers explore this subject from multiple fields such as Information Retrieval [1], Natural Language Processing [2][3][4] and Machine Learning [5], among others.

Our work falls into this general context of learning crowd sentiment through data provided by users of a determined social network. Our research goal is to construct a predictive model rather than extract sentiments from interactions of a group of people. People's posts on social networks are not always done immediately after the event affecting their mood. Soccer fans, for instance, participate in social networks with varied intensity according to their team's victories and defeats. To infer their sentiments right after a match, even when knowing the final result, might not be a viable option.

In order to build our model, we will use data from a social network, called *Torcida Virtual*[1] (TV). In this social network, formed by soccer fans, two pieces of information are essential for our studies. The first is the sentiment of soccer team's fans. On TV, fans are periodically invited to express how they are feeling about their teams in a specific moment. The second information is a sport book of official games from Brazilian championships. The information gathered from bets placed on teams will serve as indicators for favorite teams (or not). Such information becomes an equally important feature for the model, because the result of a match and the competence of the opponent team are essential for the change of fans' sentiment.

The data collected from *Torcida Virtual* and our basic research hypothesis have lead us to propose a model to represent fan group's sentiment which evolves over time and is influenced by the results of matches in official championships. This data structure has naturally led us to choose *Hidden Markov Models* (HMM) as a modeling instrument. In the Markov model, the fan's sentiment is represented as latent variables while the results of matches and the characteristics of the teams participating are considered observations, or visible variables. The results obtained with HMM in several variations, both in quantity of states and in the model's order, have shown that second order HMM, considering the match results and fan's bet information, is a more accurate model than other variants.

The rest of this article will be structured as following. First, we will present a brief revision of Hidden Markov Models. Afterward we will describe how HMM was used to model the sentiment of soccer fans. We will describe the different modeling options with a varied degree of complexity depending on the amount of information collected on the matches.

A comparative evaluation will be performed between variations of the proposed models.

II. Background Knowledge

A. Hidden Markov Model

"A HMM is a doubly stochastic process with an underlying stochastic process that is not observable (it is hidden), but can only be observed through another set of stochastic processes that produce the sequence of observed symbols."[6]

Formally, we have a set O of observations o_t for a

[1] http://torcidas.esporte.ig.com.br

978-1-5090-0016-6/15 $31.00 © 2015 IEEE
DOI 10.1109/BRACIS.2015.60

time t and a set of states $S = \{s_1, s_2, s_3, \dots, s_n\}$. The states are what the model attempts to probabilistically infer from the observations.

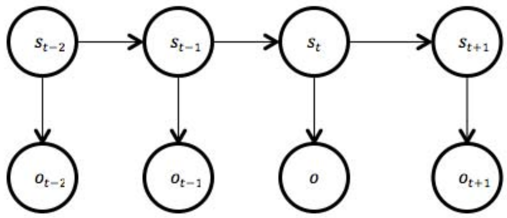

Figure. 1: Markov Chain

Fig. 1 presents a first order Markov chain. In order to estimate a following state, s_{t+1} it is necessary to know only about the current state and the observation o_{t+1}. In greater orders, however, the future state will depend on more than the current state. For instance, in a second order chain, to define s_{t+1} it will be necessary to know of the state s_{t-1} e s_t.

From both variables, state and observation, the Markov model is formed by a transition model, $P(s_{t+1}|s_t)$, which informs the probability of transition between the states, and an observation model, $P(o_t|s_t)$, which offers the probability of an observation given the state, in other words, the probability of observing o_t while being the state s in time t.

In short, a Markov model is composed of the following:

- A set of states $S = \{s_1, s_2, s_3, \dots, s_n\}$
- A set of observations $O = \{o_1, o_2, o_3, \dots, o_n\}$
- A vector π with initial probabilities for each state
- A transition matrix A, with a transition model
- An observation matrix B, representing the observation model

Such model allows for the finding of probabilistic inferences through algorithms such as Baum-Welch [6], which adjusts the model's matrix according to the observation sequence, and the Viterbi algorithm [7], which infers a more probable sequence of states.

B. Related Work

Sentiment extraction from social network logs is an important research area and an important tool for social scientists. Different practical applications are emerging as has been shown by [8]. The authors connect measures of public opinion, which are collected from polls, with sentiment estimated from tweets, and highlight the potential of tweets as a complement for traditional polling. Based on sentiment extraction from twitter, [9] indicate that the accuracy of stock market predictions can be improved by the inclusion of specific public mood dimensions. These works basically identify the polarization of tweets using a dataset of terms with their corresponding polarity. Several methods in Natural Language Processing use a similar strategy [10][11].

Sentiment classification has been used to study the readers' emotions [12]. Emoticon (emotion labels) has also been used for sentiment classification: [13] built a system called MoodLens, in which 95 emoticons are mapped into four categories of sentiments from Chinese tweets in Weibo. [14] designed an approach for classifying headline emotion based on the information collected from the World Wide Web. Also associated with topic models, there are two related works [15][16] that share some similarities with ours. The first studies the change of sentiment over time in written documents using a Topic Sentiment Change Analysis. The second work proposes a probabilistic mixture model, Topic-Sentiment Mixture, to extracts the topics and sentiment from weblogs. It uses Hidden Markov Model to extract topic life cycles and sentiment dynamics from the document.

More related with our approach since it tries to model emotions with respect to time spans, [17] inserted an emotional layer in *Latent Dirichlet Allocation*. It uses topic models for analyses of the correlation between the sentiment of the crowd and topics in News comments.

None of these works studies the relation between people's mood, which continually modifies over time, and events that affect them as we intended to with this research.

C. Torcida Virtual

The social network *Torcida Virtual* (TV) is a digital space for interaction of soccer fans. The fan is weekly consulted about the sentiment he/her is experiencing towards the team at that specific moment. This sentiment can be expressed in six distinct forms: great, good, worry, sad, bad and terrible. Each user has a voting period, which is not necessarily the same as the other users'. TV defines the sentiment of fans for a team at a set time by computing the most voted sentiment registered by fans from the seven previous days.

On TV there are bets, which are made with virtual money and consist of two kinds: bets on the result (victory, tie, defeat) or in the match's exact final score. The social network is composed of more than 200,000 subscribed users with 5% of them considered active. We have analyzed the distribution of the 34,764 votes on sentiment by fans in the past three years. It was possible to observe that most votes are toward positive sentiments (great and good).

III. Applying HMM to Learn About Soccer Fan's Sentiment

A. First Order Basic Models

In order to model the sentiment of fans we initially developed three variations of a first order Markov model, each with more added information. Our hypothesis is the larger the amount of information in the model, the best it will represent reality. The models are the following:

- M1 – It was set up considering the 6 phases that describe fans' sentiment, s, toward a team in *Torcida Virtual*. Each phase was modeled as a state in the Markov model. Thus $s \in$ {great, good, worry, sad, bad and terrible}. The observations (or visible states) represent the result of a match for a determined team. Formally, $o \in$ {victory, defeat, tie}.

- M2 – In this model we added the concept of rout. The expectation is that a defeat by several goals of difference would have a greater effect on fan's mood. A rout is represented by the difference between the goals of both teams during a match, in which that result is equal or greater than three. The model in this case will have five visible states: winning by rout, winning, tying, losing and losing by rout.

- M3 – Due to the number of bets placed it is possible to know which are the favorite and most non-favorite teams for each game. The belief is that if a team, which is considered a favorite to win, ends up losing the game, its fans will be more dissatisfied. In case it does win, the victory does not have a great impact in increasing fan's happiness. Therefore, a model was defined with 15 observations: victory as very favorite, victory as favorite, victory as very dark horse (non-favorite), victory, defeat, tie, defeat as very favorite, defeat as favorite, defeat as very dark horse, defeat as dark horse, tie as very favorite, tie as favorite, tie as very dark horse and tie as dark horse

B. How to Define a Team's Favoritism

The definition of a level of favoritism of a team during a match required a more elaborate strategy to adapt data coming from TV. We began by computing the difference, Δ, between the number of bets favorable to the analyzed team (v_f) and the non-favorable bets (v_{nf}).

$$\Delta = |v_f - v_{nf}| \quad .1$$

It was noticed that the distribution of the differences follows a Power Law. Therefore we have used *Maximum Likelihood Estimator* to estimate the slope (α) of that distribution. The result is α=1.24. From this parameter value a pure Power Law curve was plotted and compared to the curve of data.

We decided to divide the data into four equal parts using the quartile concept, from observing that the data from score differences per number of matches forms a Power Law distribution. Thus when calculating the first, second and third quartile the values obtained were 10, 40 and 90. By assuming the existence of more matches with more balanced votes (difference 10), we defined that the two first quartiles would represent matches set as non-favorite and not-unpopular, hence, any match with a difference between 0 and 40 would have observations with values winning, losing or tying.

Differences over 40 were analyzed separately and divided again into four equal groups, which gave us the new values for the three quartiles of 60, 90 and 120. Subsequently we set the observations favorite and non-favorite as all matches in the first quartile, between 40 and 60, while leaving out observations most non-favorite and most favorite as well as all matches with difference greater than 60.

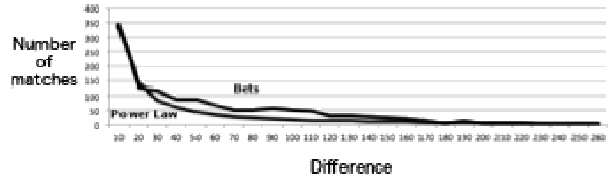

Figure 3: Curve illustrating the number of matches for each difference in comparison to a Power Law

C. Estimating Parameters

The Markov model includes as its parameters a vector of initial probabilities, a transition matrix between states and an observation matrix.

In our initial model the vector of initial probabilities will have a value for each state great, good, worry, bad, sad and terrible.

The transition matrix represents the probability of transition between a specific state in a time *t* into any other state in a subsequent time. To illustrate the transition matrix we initially defined $\gamma(s_i, s_j)$,which represents the frequency of transitions between the state s_i into the state s_j. Therefore, the probability of a transition from s_i to s_j is calculated through the frequency of the transitions between both states divided by the frequency of the transitions of S_i into all other states from the model.

$$P(s_j \mid s_i) = \frac{\gamma(S_i, S_j)}{\sum_{w=1}^{6} \gamma(S_i, S_w)} \quad . \ .$$

The observation matrix stores the probabilities of occurrence of the observations. For instance, in our basic model (*M1*) they represent the possibility of winning, losing or tying given some state of sentiment from the fans, i.e. the probability of an observation given a state $P(o_i|s_j)$.

In our model the probability of an observation O_i being found when a fan is in a state S_j, $P(o_i|s_j)$, is calculated by the frequency with which, in a dataset, a fan is found in the state s_j and observes $O_i, \gamma(o_i|s_j)$, divided by the frequency in which the fan was in the state s_j, $\gamma(s_j)$.

$$P(o_i|s_j) = \frac{\gamma(O_i|S_j)}{\gamma(S_1)} \quad . \ .$$

D. Markov Models of Greater Order

Our hypothesis is that a fan sentiment depends on more than just its state in a short previous moment, but also on what the sentiment was a longer time ago. We decided to represent this trend through a second order Markov model.

We use the same strategy of [18] specifying the second-order Markov chain by a 3-D matrix $\{a_{ijk}\}$.

$$a_{ijk} = P(q_t = s_t / q_{t-1} = s_j, q_{t-2} = s_{i,} q_{t-3} \dots) \quad (4)$$

The probability of the state sequence $Q = (q_1, q_2, q_{3,} \dots q_T)$ is defined as:

$$P(Q) = \Pi_{q1} a_{q1,q2} \prod_{t=3}^{T} a_{qt-2qt-1qt} \quad (5)$$

where Π_i is the probability of state s_i at time *t* and a_{ij} is the probability of the transition $s_i \rightarrow s_j$ at time *t=2*.

Each second-order Markov model has an equivalent first-order model on the twofold product space. The extension of the Viterbi algorithm to HMM2 is straightforward. Instead of referring to a state in the state space S, one must refer to an element of the twofold product space S x S. To adjust the second-order HMM parameters, we have used the extended Baum–Welch algorithm proposed by [18] which is based on modified forward and backward functions. The forward function $\alpha_t(j,k)$ defines the probability of the partial observation sequence $o_1, \ldots, o_t$ and the transition (s_j, s_k) between time *t-1* and *t*. $\alpha_t(j,k)$ can be computed from $\alpha_{t-1}(i,j)$ in which (s_i, s_j) and (s_j, s_k) are two transitions between states s_i and s_k. The backward function $\beta_t(i,j)$ is computed in a similar way.

Considering that $\gamma(s_k s_j, s_i)$ represents the frequency of transitions between the states s_k, s_j and s_i, in which s_i is the state in time $t-2$; s_j is the state in time $t-1$ and s_k is the state in time t, we represented the calculation o the probability for each second order transition by the following equation:

$$P(q_t = s_k \,|q_{t-1} = s_j, q_{t-2} = s_i) = \frac{\gamma(s_k s_i, s_j)}{\sum_{w=1}^{N} \gamma(s_k s_i, s_w)} \; . \; 5$$

Where *N* represents the number of states.

The second order models were generated only for the model with the greatest amount of information (M3).

IV. Empirical Evaluation

The trials were made by using the Baum-Welch algorithm to train the generated Markov models and the Viterbi algorithm to discover which sequence of states was more plausible for an observation sequence.

Our first evaluation compared models with a varied amount of information represented on visible states. We compared these models with a baseline algorithm from an intuition that, whenever a team wins a match, fans' sentiment elevate and, whenever it loses, the sentiment degrades. For instance, if a team is in a good phase and wins, it gets upgraded to a great phase. In case the team loses, fans remove it from a good phase to worry and if the team ties fans let the team continue in its current phase.

To assemble the initial model, which we will refer to as the training model, we estimated the probabilities of the transition and observation matrices from TV data, which captured fan sentiment. This data consists of voting which took place during the seasons of 2012 and 2013 of 8 of the largest Brazilian teams (in terms of number of fans) and also represent the first teams from TV 's fan ranking. The teams consist of Corinthians, Palmeiras, Santos, São Paulo, Botafogo, Vasco, Flamengo and Fluminense. In addition to the sentiment database, we used data from matches from 2012 and 2013. Altogether, 436 match results were utilized. Assuming that the model will attempt to estimate states for soccer matches of teams that do not have sentiments registered, the initial probabilities in our model were the same for all sentiments, $P_{initial}(S_i) = 0.166$.

The results from those teams' matches on season 2014 were used to train the model and infer each week sentiments for 2014 matches. The sentiments inferred were compared with the value from fan sentiment captured by TV. Each of the 8 teams had a average of 45 games played in the analyzed period in 2014.

The inference for fan sentiment was made individually for each team. Table 1 presents the results per team with the success percentage.

TABLE I. Results referred to the 3 first order models and the baseline system

	M1	M2	M3	Baseline
Corinthians	**11%**	**18%**	**36%**	**38%**
Palmeiras	**48%**	**48%**	**60%**	**44%**
São Paulo	**38%**	**21%**	**26%**	**35%**
Santos	**23%**	**31%**	**38%**	**32%**
Flamengo	**28%**	**24%**	**46%**	**46%**
Vasco	**47%**	**13%**	**20%**	**19%**
Fluminense	**20%**	**25%**	**30%**	**41%**
Botafogo	**32%**	**67%**	**60%**	**19%**

It is possible to notice an improvement in results of almost all models, as more information is included in the observations from the Markov model. However they don't have clear advantage compared with the baseline system.

A. Second Order Models

We compared the results from the best model (M3) with its extended version for orders 2 and 3, in order to validate the hypothesis that by increasing the order of models it would be possible to increase accuracy on inferences. Table 2 presents the tests results conducted for second HMM and compares to first order HMM.

There is a significant improvement on inferences made by HMM2 compared to HMM1. Solely the inferences made for the sentiment Palmeira's and Botafogo's fans did not present improvement (we will elaborate more on this matter further). We realized that, by increasing the Markov model order, the amount of examples (with respect to transitions between states and observations) to build the model decrease. In the second order model, some transitions from the transition matrix and some probabilities from the observation matrix remain null. Further tests with more data are necessary to reach definitive conclusions on whether increasing the order will always increase the accuracy of the model.

TABLE II. Results from sentiment inference by First (HMM1) and second (HMM2)

	HMM1	HMM2s
Corinthians	**36%**	**75%**
Palmeiras	**60%**	**36%**
São Paulo	**26%**	**53%**
Santos	**38%**	**55%**

Flamengo	**46%**	**51%**
Vasco	**20%**	**72%**
Fluminense	**30%**	**49%**
Botafogo	**60%**	**17%**

V. Discussion

A. Bias Towards Good Moments

From the data we observed that TV fans use the functionality "Fan Sentiment" more frequently when their team has a positive result. This means that, when a team loses, the fan does not express frustration with the same frequency as he expresses happiness. This is probably due to the fact that TV is a social network, and interacting with friends and acquaintances from the community after a bad result is not particularly pleasant.

We observed that for all 8 considered teams from the database, except for Vasco, fans voted the most when their team attained better results in the matches. Even so, the values for Vasco for victory and defeat were very close. Corinthians, for example, 41% of votes occurred after a victory, 35% after a tie and 23% after a defeat. In such cases, the difference between the amount of votes in victory and defeat is fairly large (e.g. Corinthians, Palmeiras and São Paulo).

These observations show that the models learnt tend to value transitions between positive sentiments, with larger probabilities to positive phases. We have numbered the amount of positive phases (phases great and good) and negative (worry, sad, bad and terrible) and found that more than double the votes are in positive sentiments (27.579) than negative (11.787).

As previously explained, by increasing the order of the Markov model the amount of examples to develop the model decreases and, since the states with more transitions are the ones which involve positive sentiments, it is presumed that the second order model could be even more biased towards positive sentiment. Therefore, predictions over sentiment of fans whose team's data from the test contain more negative sentiments will have lower accuracy.

This assumption made us investigate more closely the relation between positive/negative sentiment and the inference results achieved with HMM2. Palmeiras and Botafogo, the teams with the lowest accuracy in inferences made with second order model, are the only teams, which presented more negative than positive sentiment for each observation in the trial dataset.

For the 42 matches that were tested on Viterbi with HMM2s models, from TV's database for Palmeiras, 20 of them are marked as positive sentiment and 22 as negative sentiment. Botafogo has 14 positive votes and 32 negative ones. Opposed to Corinthians, for which the HMM2 model had the largest accuracy rate with 43 matches tested as Viterbi observations, 38 of them marked as positive sentiment and only 5 as negative sentiment.

B. Tests Performed Removing Data from the Analyzed Team

We have also performed some tests in which we removed the sentiment and results of the matches from the process of constructing the initial model from teams that would have its sentiment inferred from season 2014. This means that, in order to infer Corinthian's fan's sentiment from a determined period, the initial model with data from 2012 and 2013 was created without considering a single sentiment vote made by Corinthians fans. The goal was to verify how many inferences obtained from the model were dependent of sentiment data collected by fans themselves. There was no decline in results. This indicated that the Markov model might be applied in teams not included in the training model in order to infer crowd sentiment from soccer team fans.

VI. Conclusion

This research has investigated a database of sentiments expressed by soccer fans from teams in Brazil. We proposed to create a predictive model based on Hidden Markov Models due to its stochastic characteristics, which describe a procedure that operates over a long period of time. Therefore, the hypothesis that a sentiment is formed over time and not just by a single observation was strengthened. A significant improvement on results was also observed when a second order model was used. This reinforced the hypothesis that current sentiments are dependent on previous states.

Variations of Markov models show that the accuracy rate improves, since observations are capable of expressing winnings or losses as well as the favoritism of the opponents. Another important finding is the fact that sentiment evolution over time also needs to be represented.

The limitations of our approach guide us towards future investigations. More specifically, it will be valuable to study how to create models that are less biased due to data variation, such as the ones occurring due to greater or lesser fan participation when expressing sentiment.

Also we plan to compare the HMM models with other approaches such as classification algorithms.

Finally we are investigating how to represent this problem as a HMM with continuous states. Instead of inferring the sentiment represented by the majority of the votes, we want to infer a histogram, which represents the distribution of fans' sentiment.

References

[1] G. Paltoglou, and M. Thelwall, "A study of Information Retrieval weighting schemes for sentiment analysis", *48th Annu. Meet. Assoc. Comput. Linguist.*, 1386- 1395, 2010.

[2] S. M. Mohammad, S. Kiritchenko, and X. Zhu, "NRC-Canada : building the state-of-the-art in sentiment analysis of tweets", *Seventh International Workshop on Semantic Evaluation, SemEval,* 2013 *321-327*.

[3] A. Vanzo, D. Croce, and R. Basili, "A context-based model for sentiment analysis in twitter", 25th Internacional Confedrence on Computational Linguis, 2014, 2345-2354. Dublin City University and Association for Computational Linguistics, Dublin, Ireland.

[4] D. Gao, F. Wei, W. Li, X. Liu, and M. Zhou,, "Cotraining Based Bilingual Sentiment Lexicon Learning", AAAI Conference, 2013

[5] A. Ortigosa, J. M. Martín, and R. M. Carro, “Sentiment analysis in Facebook and its application to e-learning”, Comput. Human Behav, 2014, 527–541.

[6] L. R. Rabinerand, and B. H. Juang, “An introduction to hidden Markov models”, *IEEE ASSP Mag*., 1986,3: 4-16.

[7] A. J. Viterbi, “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm”, *IEEE Trans, Inform. Theory,* 1967, IT-13:260-269.

[8] B. O’Connor, R. Blasasubramanyan, B, Routledge, and N. Smith, “From Tweets to Polls: Linking Text Sentiment to Public Opinion”, *Fourth International AAAI Conference on Weblogs and Social Media,* 2010.

[9] J. Bollen, H. Mao, and X. Zeng , “Twitter mood predicts the stock market. Journal of Computational Science”, 2011, 2:1-8.

[10] P. Palanisamy, V. Yadav, and H. Elchuri, “Serendio: simple and practical lexicon based approach to sentiment analysis”, *SemEval-2013: Semantic Evaluation Exercises International Workshop on Semantic Evaluation*, 2013, Atlanta.

[11] K. Liu, W. Li, and M. Guo, “Emoticon Smoothed Language Models for Twitter Sentiment Analysis”, *26th AAAI Conference on Artificial Intelligence*, 2012,Toronto, 22-26, pp. 1678-1684.

[12] K. H.-Y. Lin, and H.-H. Chen, “Ranking reader emotions using pairwise loss minimization and emotional distribution regression”, *Conference on Empirical Methods in Natural Language Processing*, 2008,136–144. Association for Computational Linguistics.

[13] J. Zhao, L. Dong, J. Wu, and K. Xu., “Moodlens: an emoticon-based sentiment analysis system for chinese tweets”. *18th ACM SIGKDD international conference on Knowledge discovery and data mining,* 2012, 1528–1531

[14] Z. Kozareva, B. Navarro, S. Vazquez, and A. Montoyo, “Ua-zbsa: a headline emotion classification through web information”. *4th International Workshop on Semantic Evaluations*, 2007,334–337. Association for Computational Linguistics.

[15] Jiang, Yu, Weiyi Meng, and Clement Yu. "Topic sentiment change analysis." Machine Learning and Data Mining in Pattern Recognition. Springer Berlin Heidelberg, 2011. 443-457

[16] Mei, Qiaozhu, et al. "Topic sentiment mixture: modeling facets and opinions in weblogs." Proceedings of the 16th international conference on World Wide Web. ACM, 2007

[17] C. Zhu; H. Zhu; Y. Ge; E. Chen; and Q. Liu, “Tracking the Evolution of Social Emotions: A Time-Aware Topic Modeling Perspective”. *IEEE ICDM,* 2014.

[18] J.-F. Mari and J.-P. Haton, “Automatic word recognition based on second-order hidden markov models”. ICSLP, 1994,pp. 247–250.