Towards Affect-based User Models: a comparative study with various datasets, features and algorithms, for multi-labeled, probabilistic affect detection

Yannick Gaspar
Department of Informatics Engineering
University of Coimbra
Coimbra, Portugal
yannick@student.dei.uc.pt

Luis Macedo
Department of Informatics Engineering
University of Coimbra
Coimbra, Portugal
macedo@dei.uc.pt

***Abstract*—In this paper, we make use of machine learning based techniques to try to find the best possible combination of datasets, affect lexicons, features, and algorithms, for the automatic detection of affect in text. For this purpose, we consider as affective categories, the six basic emotions from Paul Ekman (*anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*). For the experiments, we count with three different datasets, which contain, respectively, news headlines, fairy tales and blogs sentences, and with two affect lexicons: WordNet-Affect and Roget's Thesaurus. From this collection of data, we compare the performance of two different classification algorithms: Naive Bayes and Support Vector Machines (SVM). The experiments allow us to demonstrate that from the various possible combinations, there are two, in particular, that stand out and are more appropriate for the purpose of this study. We evaluate and discuss the results with the aim of, in the future, building user affect models.**

I. Introduction

There is no doubt that from the interplay between humans, and between the humans and the environment results emotions, although, there is much more to be said about emotions [1], [2]. Two functions have been undoubtedly recognized to them: informational and motivational [3]. Emotions are involved in action, and in the perception of the environment by the individuals. Thus, they provide a considerable amount of information about the individuals and their interactions, as well as about the interactions of the individuals with the environment. More information may be gathered, if in addition to emotions, we could also obtain related mental states, such as beliefs and goals/desires, that are on the background of the emotional mental states. Despite all the different kinds of data that can be used for user modeling, emotions and related mental states, such as beliefs and goals/desires, are a valuable aspect to take into account, in order to build better user models, and explore them more effectively. Affective models of the users, and the way they are built, play a very important role in many applications, such as in recommender systems. Hence, affect detection [4] is critical, because an affect-sensitive interface can never respond to the users affective states, if it cannot sense them. More precisely, affect detection is important for increasing the quality of the human-computer interaction, by endowing computers, or better, artificial agents, with the ability to recognize, and react appropriately to the emotions of humans. But how to detect affect? Affect detection does not need be perfect, but it must be an accurate approximation of the real target. However, this is a very challenging problem, because emotions are built considering conceptual quantities, that cannot be directly measured with fuzzy boundaries, and with substantial individual difference variations in expression.

Finding ways to model, infer, detect, or capture emotional data and emotional-related data of humans, associated, for instance, with geographical information, and then visualize that data, has given rise to a whole research industry. For example, [5] stressed the potential political, social, and cultural benefits of visualizing people's biometric and emotional data in a geographic region, through emotional maps, namely in recreation, art, community development, scientific research, architectural planning, or large scale political consultations. In addition to peripheral physiological changes, as used in [5], facial expressions, and oral and written verbal communication are among the main sources of data that can be used to infer emotional states. Specifically, the social media (e.g. email, online forums, blogs, and social networks), constitute an important source of written verbal communication. The increasing use of the social media, mainly caused by the impact of the social networks, has been attracting the interest of researchers, in mining the web for trying to infer the emotions that are directly/indirectly expressed through text [6]. Every day humans are flooded with new social information, access Twitter accounts to say what they are doing, Facebook to tell what they are thinking, Google Places/Foursquare to say where they are, etc. Humans make use of these social media technologies to describe their interactions with the environment. However, the question is how to identify emotions in those texts, or sentences, provided by the social media? In recent years, some approaches have appeared (e.g. [7], [8], [9], [10], [11], [6]). Some only consider the polarity of the texts, and other more elaborated systems analyze texts into several emotions, usually using the list of the six basic emotions from Paul Ekman [12] (*anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*), either at sentence, or word level (for a review in sentiment analysis see [13], and to learn about the current state of the art in text-based emotion detection see [14]).

In this paper, we try to find the best possible combination of datasets, affect lexicons, features, and algorithms, for the automatic detection of affect in text, with the aim at building

user affect models, specifically involving the Ekman's six basic emotions. For that purpose, we collected three different datasets, which contain, respectively, news headlines, fairy tales and blogs sentences, and two affect lexicons: WordNet-Affect and Roget's Thesaurus. From this collection of data, we compare the performance of two of the most common classification algorithms [15]: Naive Bayes and SVM. The experiments were performed using cross-validation, and were divided into two parts: (i) evaluation of the classifiers by performing cross-validation on the three datasets, using different features, such as unigrams, bigrams, trigrams, and n-grams (ii) evaluation of the classifiers by performing cross-validation on the three datasets, when combined with the two affect lexicons.

The rest of the paper is structured as follows. Section II describes the approach of this study, indicating the materials and methodologies that we used in the experiments. In Section III, we present, evaluate and discuss the results. Finally, in Section IV, we present our conclusions and the future work.

II. Materials & Methods

In this section, we describe in detail, the materials and methodologies that we used for our experiments. In the materials, we describe the various datasets and affect lexicons that were used. On the methodologies, we describe and try to explain, in detail, the various components that constitute our classification system.

A. Datasets

For this study, we collected three different datasets (Table I shows the distribution of sentences, of the different datasets, in each affective category). The information that follows describes each one.

1) Text Affect: This dataset [10] consists of news headlines drawn from the most important newspapers, such as New York Times, CNN, and BBC News, as well as from the Google News search engine. It is divided into two subsets, where one is considered a train set, composed by 250 sentences, and the other is the corresponding test set, which contains 1000 sentences. The sentences were annotated in accordance with the six basic emotions by using a vector of scores, where each emotion was assigned with a value between 0 and 100. The value 0 means that an emotion is missing from the given sentence, and 100 represents the maximum emotional value. Besides that, each sentence has also an intensity scale that represents valence annotations. The interval of that scale ranges between the values -100 and 100. The value 0 represents a neutral headline, -100 a highly negative headline, and 100 a highly positive headline. For our experiments, we used this dataset as a whole (1250 sentences), and only considered the emotional annotations, where we assigned to each sentence, the respective most predominant emotion.

2) Fairy Tales: This dataset [9], [16] contains 1207 sentences, from fairy tales by Grimm's, H.C. Andersen, and B. Potter. The sentences with affective high agreements, referred by the author as being sentences with four identical affective labels, were annotated according using a merged label set with six affect classes: (*angry-disgusted*, *fearful*, *happy*, *sad*, and *surprise*). According to the author, the merge between the emotions *anger* and *disgust*, was made due to data sparsity and related semantics.

3) Blogs: This dataset [11], [17] is composed by 4090 emotion-rich sentences, that were collected from blogs. The sentences were manually annotated by four judges, according to eight affective categories. To the Ekman's six basic emotions, the authors added two categories: *mixed emotion* and *no emotion*. The *mixed emotion* category was thought to account all the sentences that could not be assigned to any basic category, or that would show more than one type of emotion, which can happen if a sentence refers to the emotional states of more than one person. The *no emotion* category was designed for all the sentences that had no emotional content.

TABLE I. Distribution of sentences, of the different datasets, in each affective category

Affect	Text Affect	Fairy Tales	Blogs
Anger	87	218	179
Disgust	42		172
Fear	194	166	115
Happiness	441	445	536
Sadness	265	264	173
Surprise	217	114	115
No Emotion	4	-	2800
Total	**1250**	**1207**	**4090**

B. Affect Lexicons

The following information describes the two affect lexicons that we collected for this study. (Table II shows the distribution of words, of the two affect lexicons, in each affective category).

1) WordNet-Affect: This resource [18] is an extension of WordNet Domains[1], that includes a subset of synsets suitable to represent affective concepts correlated with affective words. It is composed by 1128 emotion-related words, which are distributed over six lists, corresponding to the six basic emotions. For our experiments, we used the lists that are publicly available[2].

2) Roget's Thesaurus: This resource[3] [19], [20] is not exactly an affect lexicon, but rather a lexical knowledge-base that can be used for various natural language processing tasks. Nevertheless, from this resource it is possible to find and derive an affect lexicon composed by words that are related with the six basic emotions [21], [22]. This lexicon can be used to calculate a similarity score, which represents a semantic relatedness value between words, based on the path length between them (nodes). The similarity measure assigns scores, which can range between 0, for the least related words, and 16, for the most semantically related ones. The words with a score of 12 to 14 are considered to have an intermediate level of similarity, and those with a score below 10 are considered as having a low level of similarity [23]. To build an emotion lexicon, initially is necessary to define and select a set of words, in order to serve as a basis of derivation, and in a second phase it is important to decide about the most appropriate similarity score to be considered [21]. For our experiments, initially we defined the following set of derivation: {*angry/anger*, *disgusted/disgust*, *fear*, *joy/happy/happiness*, *sad/sadness*, and *surprised/surprise*}. As for the second phase, due to the experiments reported in [23],

[1] http://wndomains.fbk.eu/

[2] http://www.cse.unt.edu/~rada/affectivetext/data/WordNetAffectEmotionLists.tar.gz

[3] For our experiments we used a free implementation of the 1911 Thesaurus that is available at: http://rogets.site.uottawa.ca/

we only considered the words that had similarity scores of 12 or more (from intermediate, to the highest level of similarity). Using this specific implementation, and opting for the choices that were previously described, we were able to build a lexicon composed by 2711 emotion-related words.

TABLE II. DISTRIBUTION OF WORDS, OF THE TWO AFFECT LEXICONS, IN EACH AFFECTIVE CATEGORY

Affect	WordNet-Affect	Roget's Thesaurus
Anger	255	316
Disgust	53	319
Fear	147	318
Happiness	400	579
Sadness	202	961
Surprise	71	218
Total	**1128**	**2711**

C. Classification System

Before starting to specify the classification system, it is important to mention that all the experiments were performed using the *Weka* software[4]. The following information describes and explains, in detail, the various components that constitute our classification system.

1) Data pre-processing: In order to detect affect in text, initially, it is important to use processes for ensuring that all data is appropriately formatted. In our experiments, we used stop-words detection, to remove the most commonly used words (e.g. "are" and "the"), with the aid of the *Weka* default stop-words list, based on *Rainbow*[5]. Besides that, we also used the *Porter* stemmer, from the *Snowball*[6] stemmers package, available on *Weka*, to reduce the words to their root forms, or stems (e.g. "stemming" and "stemmed" should be reduced to "stem"). Most of the languages are full of structural words that provide little, or no meaning to text [24]. For that reason, we used these two pre-processing steps, which help us to differentiate emotion sentences from non-emotion sentences, and to focus the significant linguistic components by removing unimportant features [21], [24]. Finally, to deal with contractions (short forms), we implemented a method to replace the most common positive and negative English short forms, by their corresponding long forms (e.g. "we'll" by "we will", and "we didn't" by "we did not"). This method was created to deal, primarily with negations, and although we do not present comparative results between its use, and non-use, we can state that in most of the tested cases, we were able to get a small performance improvement by using it.

2) Features: Even though the affect lexicons can be seen as affect features, in our experiments we also used other types of features, such as unigrams, bigrams, trigrams, and n-grams (ranging between 1 and 3 grams). A n-gram model can be imagined as a small window that passes over a sentence, and only n words are visible at the same time [25]. The unigrams are the simplest n-gram models, in which just only one word can be seen at a time [25]. The bigrams of a sentence can be found by placing the window on its first two words, and then, by moving that window to the right, one word at a time, in a stepwise manner [25]. This procedure is repeated, until the window covers the last two words of the sentence [25]. Thus, a bigram can be seen as a window that shows to words at a time. The trigrams have a similar interpretation to bigrams, with the difference that in this case, the window has a length of three words. Considering a practical scenario, the sentence "Today was not a happy day" contains the following unigrams: "Today", "was", "not", "a", "happy", "day". Moreover, it also contains the following bigrams: ["Today", "was"], ["was", "not"], ["not", "a"], ["a", "happy"], ["happy", "day"]. As for the trigrams, the representation would be: ["Today", "was", "not"], ["was", "not", "a"], ["not", "a", "happy"], ["a", "happy", "day"]. N-grams are important to find syntactic patterns, and to deal with negations (e.g. "not a happy") [26]. The presence of a negation, in a sentence, can change the emotional category that is really being expressed. For example, the previous sentence, "Today was not a happy day", should be assigned to the *sadness* category, however, the non-use of n-grams can cause the sentence to be assigned to the *happiness* category.

3) Algorithms: In our experiments, we compared two of the top 10 algorithms [15] in data mining: Naive Bayes and Sequential Minimal Optimization (SMO), an implementation of the SVM machine learning model.

4) Evaluation: In this study, the evaluation of the classifiers was performed with cross-validation, using 10-folds, which is a commonly used value [27]. To evaluate the results, we considered four evaluation metrics that are often found in the literature [14]: accuracy, precision, recall, and f-measure. Accuracy is a simple measure that is used to evaluate the performance of a classifier, and consists in counting the proportion of correctly predicted instances in an unseen test dataset [28]. Precision and recall are used to evaluate the accuracy of the classification algorithms [14]. In more detail, precision is the proportion of examples which truly have class x, among all those which were classified as class x, and recall is the proportion of examples which were classified as class x, among all the examples which truly have class x [28]. F-measure is a simple metric, derived from precision and recall, which indicates the level of performance of a classifier [14], [28]. All the values of precision, recall, and f-measure presented in Section III are weighted average values, which were automatically calculated by Weka, taking into account the respective values of each affect category.

III. RESULTS & DISCUSSION

In this section, we present and discuss the results that were obtained for the first and second part of the experiments. All the values within the tables that are in bold text correspond to the highest values, of each evaluation metric, at each dataset. For the experiments, besides the three different datasets that were previously described, we also used and evaluated datasets resulting from the various possible combinations between the three. Therefore, for the experiments, we considered the following datasets: Text Affect, Fairy Tales, Blogs, Text Affect + Fairy Tales, Text Affect + Blogs, Fairy Tales + Blogs, and Global (combination between the three datasets).

A. Datasets with Features

In this part of the experiments, the classifiers were built based on the various datasets, using different features (see Table III, and Table IV).

[4]http://www.cs.waikato.ac.nz/ml/weka/

[5]http://www.cs.cmu.edu/~mccallum/bow/rainbow/

[6]http://snowball.tartarus.org/

TABLE III. RESULTS FOR CROSS-VALIDATION ON THE VARIOUS DATASETS, USING DIFFERENT FEATURES - NAIVE BAYES

		A	P	R	F
Text Affect	*Unigrams*	42,72	47,20	42,70	37,30
	Bigrams	36,80	**55,90**	36,80	22,90
	Trigrams	35,60	33,20	35,60	19,10
	N-grams	**44,16**	48,10	**44,20**	**39,20**
Fairy Tales	*Unigrams*	**58,24**	57,90	**58,20**	**57,70**
	Bigrams	40,93	39,90	40,90	38,70
	Trigrams	39,69	**73,50**	39,70	25,90
	N-grams	55,76	55,30	55,80	55,20
Blogs	*Unigrams*	**70,81**	**68,90**	**70,80**	**68,80**
	Bigrams	62,13	55,60	62,10	55,90
	Trigrams	68,24	49,60	68,20	55,60
	N-grams	63,91	66,20	63,90	64,20
Text Affect + Fairy Tales	*Unigrams*	45,01	47,90	45,00	40,60
	Bigrams	36,14	39,80	36,10	26,40
	Trigrams	36,43	**59,80**	36,40	19,90
	N-grams	**45,10**	48,20	**45,10**	**42,40**
Text Affect + Blogs	*Unigrams*	**56,10**	58,50	**56,10**	**47,60**
	Bigrams	52,21	27,60	52,20	36,10
	Trigrams	52,51	27,60	52,50	36,20
	N-grams	55,88	**59,20**	55,90	47,30
Fairy Tales + Blogs	*Unigrams*	**60,77**	**58,60**	**60,80**	**58,00**
	Bigrams	50,29	42,30	50,30	43,80
	Trigrams	52,86	27,90	52,90	36,60
	N-grams	56,52	56,40	56,50	55,60
Global	*Unigrams*	**50,11**	**54,20**	**50,10**	**43,90**
	Bigrams	43,30	33,80	43,30	30,80
	Trigrams	42,83	18,30	42,80	25,70
	N-grams	48,95	51,30	49,00	**43,90**

TABLE IV. RESULTS FOR CROSS-VALIDATION ON THE VARIOUS DATASETS, USING DIFFERENT FEATURES - SMO

		A	P	R	F
Text Affect	*Unigrams*	48,40	47,10	48,40	**46,10**
	Bigrams	41,60	53,50	41,60	32,30
	Trigrams	37,92	**57,90**	37,90	24,60
	N-grams	**48,56**	55,00	**48,60**	42,50
Fairy Tales	*Unigrams*	**62,47**	62,90	**62,50**	**62,50**
	Bigrams	46,73	56,60	46,70	40,30
	Trigrams	42,09	59,50	42,10	30,70
	N-grams	56,59	**66,60**	56,60	51,10
Blogs	*Unigrams*	**80,81**	**79,60**	**80,80**	**79,00**
	Bigrams	71,37	73,00	71,40	62,50
	Trigrams	69,07	66,20	69,10	57,50
	N-grams	76,19	78,30	76,20	70,70
Text Affect + Fairy Tales	*Unigrams*	**54,33**	53,50	**54,30**	**52,20**
	Bigrams	41,35	53,90	41,40	31,90
	Trigrams	39,40	**60,70**	39,40	26,60
	N-grams	53,15	57,80	53,20	47,60
Text Affect + Blogs	*Unigrams*	**67,68**	65,50	**67,70**	**64,70**
	Bigrams	56,44	62,30	56,40	45,00
	Trigrams	53,97	63,10	54,00	39,60
	N-grams	63,39	**66,40**	63,40	56,90
Fairy Tales + Blogs	*Unigrams*	**74,50**	**73,00**	**74,50**	**72,80**
	Bigrams	61,75	59,60	61,80	55,20
	Trigrams	54,14	60,60	54,10	39,70
	N-grams	72,08	72,20	72,10	68,80
Global	*Unigrams*	**65,59**	**63,90**	**65,60**	**63,40**
	Bigrams	51,58	53,20	51,60	44,50
	Trigrams	44,88	62,50	44,90	30,20
	N-grams	63,33	62,60	63,30	60,80

Analyzing the various results, at least for the Naive Bayes algorithm, there is only one that should be highlighted. This result corresponds to the highest accuracy rate, 70.81%, and was obtained with the Blogs dataset by using unigrams. In the SMO algorithm, the scenario is relatively different. At a first view, we can observe that between the two algorithms, for the different features, we always obtained better results with the SMO algorithm. The highest accuracy rate, 80.81%, was obtained using the same combination that also allowed us to achieve the highest value in the Naive Bayes algorithm, i.e., the Blogs dataset, using unigrams. For this specific combination, the exchange of algorithms, from the Naive Bayes to the SMO, produced an improvement of exactly 10%. There is another case, in which we also obtained an interesting result. The junction between both datasets, Fairy Tales and Blogs, also using unigrams, allowed us to obtain an accuracy rate of 74.5%, which is only a difference of about 6% in relation to the highest value of the SMO algorithm, and an improvement of almost 14% in relation to the same combination when used with the Naive Bayes algorithm.

From a general perspective, the best results were achieved through the SMO algorithm. The best dataset, considering this first part of the experiments, is clearly the Blogs dataset, because it was the one that allowed us to obtain the highest values of accuracy rate, in both algorithms. The worst, or the one that produced the lowest results, was in both algorithms the Text Affect dataset. As regards the features, for most datasets, unigrams has almost always been the option that allowed us to achieve the best results, however, the use of n-grams also allowed us to obtain good results. Inclusive, there were three cases, in which the use of n-grams produced the best results. The worst results are marked, in both algorithms, by the use of trigrams.

B. Datasets with Emotion Lexicons

In this part, we combined the datasets with the affect lexicons, and we tested and evaluated, again, the classifiers (see Table V, and Table VI). It is worth mentioning that the following tables contains specific abbreviations, which are translated into: *WNA* (WordNet-Affect), *RT* (Roget's Thesaurus), and *WNA+RT* (combination between the two affect lexicons).

TABLE V. RESULTS FOR CROSS-VALIDATION ON THE VARIOUS DATASETS, USING DIFFERENT AFFECT LEXICONS - NAIVE BAYES

		A	P	R	F
Text Affect	*Baseline*	**44,16**	48,10	**44,20**	**39,20**
	WNA	38,23	**51,50**	38,20	27,40
	RT	33,88	49,50	33,90	21,60
	WNA+RT	32,97	44,30	33,00	22,70
Fairy Tales	*Baseline*	**58,24**	**57,90**	**58,20**	**57,70**
	WNA	45,48	45,90	45,50	39,70
	RT	38,11	45,50	38,10	29,80
	WNA+RT	34,80	42,90	34,80	26,30
Blogs	*Baseline*	**70,81**	**68,90**	**70,80**	**68,80**
	WNA	57,07	61,00	57,10	47,50
	RT	42,54	47,70	42,50	27,90
	WNA+RT	36,32	37,50	36,30	20,70
Text Affect + Fairy Tales	*Baseline*	**45,10**	48,20	**45,10**	**42,40**
	WNA	39,75	**50,30**	39,70	32,40
	RT	36,30	45,50	36,30	28,30
	WNA+RT	32,70	40,60	32,70	21,40
Text Affect + Blogs	*Baseline*	**56,10**	58,50	**56,10**	**47,60**
	WNA	45,89	**58,80**	45,90	32,50
	RT	35,70	53,80	35,70	20,00
	WNA+RT	32,31	53,20	32,30	18,10
Fairy Tales + Blogs	*Baseline*	**60,77**	**58,60**	**60,80**	**58,00**
	WNA	51,66	52,40	51,70	45,00
	RT	39,54	54,90	39,50	28,60
	WNA+RT	34,95	49,10	34,90	24,40
Global	*Baseline*	**50,11**	**54,20**	**50,10**	**43,90**
	WNA	42,45	51,90	42,40	34,10
	RT	34,93	49,90	34,90	24,60
	WNA+RT	33,70	51,50	33,70	24,30

The introduction of the emotion lexicons, contrary to our expectations, did not cause significant changes, in relation to the results of the first part. In the Naive Bayes algorithm, we could not reach any of the baseline results. However, there are some cases that produced results which are quite close. For example, the combination of the Text Affect dataset with the

TABLE VI. RESULTS FOR CROSS-VALIDATION ON THE VARIOUS DATASETS, USING DIFFERENT AFFECT LEXICONS - SMO

		A	P	R	F
Text Affect	*Baseline*	**48,56**	55,00	**48,60**	**42,50**
	WNA	39,87	54,40	39,90	28,30
	RT	34,69	57,50	34,70	22,90
	WNA+RT	31,05	**58,70**	31,00	19,00
Fairy Tales	*Baseline*	62,47	62,90	62,50	62,50
	WNA	**63,90**	**66,70**	**63,90**	**62,90**
	RT	40,45	41,90	40,50	37,80
	WNA+RT	45,36	49,90	45,40	44,30
Blogs	*Baseline*	**80,81**	**79,60**	**80,80**	**79,00**
	WNA	73,11	74,70	73,10	71,10
	RT	51,26	54,90	51,30	50,30
	WNA+RT	52,13	57,60	52,10	52,10
Text Affect + Fairy Tales	*Baseline*	54,33	53,50	54,30	52,20
	WNA	**57,99**	**60,20**	**58,00**	**56,10**
	RT	42,22	42,50	42,20	39,90
	WNA+RT	45,63	47,70	45,60	44,30
Text Affect + Blogs	*Baseline*	**67,68**	**65,50**	**67,70**	**64,70**
	WNA	63,91	64,60	63,90	61,40
	RT	48,27	50,20	48,30	46,90
	WNA+RT	49,45	52,80	49,40	49,00
Fairy Tales + Blogs	*Baseline*	**74,50**	**73,00**	**74,50**	**72,80**
	WNA	70,05	70,70	70,10	68,50
	RT	52,05	54,80	52,00	51,50
	WNA+RT	51,93	55,70	51,90	51,90
Global	*Baseline*	**65,59**	63,90	**65,60**	**63,40**
	WNA	63,60	**64,30**	63,60	61,90
	RT	49,58	51,00	49,60	48,60
	WNA+RT	50,03	52,40	50,00	49,60

WordNet-Affect lexicon, produced an accuracy rate of 38.23%, which is only a difference of approximately 6%, in relation to the baseline value. In the SMO algorithm, the scenario is not very different, but we were able to surpass the baseline results with two distinct cases. The first one corresponds to the combination between the Fairy Tales dataset and the WordNet-Affect lexicon, from which we obtained an accuracy rate of 63.9%. In other words, this value represents an improvement of about 1.5%, in relation to the baseline value. The second case is relative to the combination of two datasets, Text Affect and Fairy Tales, and with the WordNet-Affect, by which we could achieve 57.99%, which is an increase of about 3.5%, over the baseline value. Considering only the different affect lexicons, in the Naive Bayes algorithm, the best result, with an accuracy rate of 57.07%, was obtained with the Blogs dataset when combined with the WordNet-Affect. In fact, this is a result that is below our expectations, because actually being the highest result, it is very distant from the baseline result. In the SMO algorithm, if we only consider the emotion lexicons, there are two results that should be highlighted: 73.11%, obtained with the Blogs dataset when combined with the WordNet-Affect, and 70.05%, obtained with the dataset defined by the combination between two datasets, Fairy Tales and Blogs, and with the WordNet-Affect.

In general, WordNet-Affect was the emotion lexicon that allowed us to obtain the best results. The Roget's Thesaurus, according to our experiments, was the one that produced the lowest results, and the combination of the two, WordNet-Affect and Roget's Thesaurus, generated, for some cases, even lower results. In this experimental part, we can assume that the best combination, for both algorithms, is composed by the Blogs dataset, combined with the WordNet-Affect.

Choosing the best combination from the results that we have achieved is not difficult. Starting with the algorithm, any combination that is intended to be used, must include the SMO algorithm. Regarding the datasets, in both parts the Blogs was the best dataset that we tested. In the features, using unigrams, or n-grams may depend on the own classification domain, and on the structure of the sentences of each dataset, but specifically for the Blogs dataset, and as we were able to demonstrate, unigrams is the most appropriate feature. The decision to include, or not, an emotion lexicon, is a matter of experimentation, because there are cases in which it is possible to improve the performance, but there are also others, where it can even get worse. It is a decision that may depend on the classification domain, however, we consider the introduction of an emotion lexicon as a plus, because we are adding, to the existing dataset, words that directly/indirectly express emotions, i.e., emotion-related words. From the results of our experiments, we were able to find two alternatives that have potential to be effective, in the construction of user affect models from text. One alternative consists of the Blogs dataset, using unigrams. The other, is based on the first, and includes the WordNet-Affect lexicon.

IV. CONCLUSION

In this study, our goal was to try to find the best combination of datasets, affect lexicons, features, and algorithms, to automatically detect/classify emotions in text. Our experiments allowed to demonstrate that there are several alternatives which are practicable, however, the results point out only two, in particular, that suit the purpose under study. One alternative consists of the Blogs dataset, using unigrams. The other uses the same combination of the first, and includes the WordNet-Affect lexicon. In both, we must include a classifier based on the SMO algorithm.

In the future, we intend to build user affect models based on social media texts (e.g. Twitter and Facebook), and apply them in different purposes, such as in recommender systems, or to forecast user actions assuming that action depends on the emotion, i.e., assuming the motivational function of emotions [29], [3].

REFERENCES

[1] A. Ortony, G. L. Clore, and A. Collins, *The cognitive structure of emotions*. Cambridge university press, 1990.

[2] R. Reisenzein, "Emotional experience in the computational belief–desire theory of emotion," *Emotion Review*, vol. 1, no. 3, pp. 214–222, 2009.

[3] R. Reisenzein and H. Weber, "Personality and emotion," in *The Cambridge Handbook of Personality Psychology*, P. J. Corr and G. Matthews, Eds. Cambridge University Press, 2009, pp. 54–71.

[4] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, Jan. 2010.

[5] C. Nold, Ed., *Emotional cartography : technologies of the self*. s.n., 2009.

[6] S. M. Mohammad, "#Emotional tweets," in *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, ser. SemEval '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 246–255.

[7] H. Liu, H. Lieberman, and T. Selker, "A model of textual affect sensing using real-world knowledge," in *Proceedings of the 8th international conference on Intelligent user interfaces*, ser. IUI '03. New York, NY, USA: ACM, 2003, pp. 125–132.

[8] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," *Language Resources and Evaluation*, vol. 39, no. 2-3, pp. 165–210, 2005.

[9] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: machine learning for text-based emotion prediction," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, ser. HLT '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 579–586.

[10] C. Strapparava and R. Mihalcea, "Semeval-2007 task 14: affective text," in *Proceedings of the 4th International Workshop on Semantic Evaluations*, ser. SemEval '07. Stroudsburg, PA, USA: Association for Computational Linguistics, 2007, pp. 70–74.

[11] S. Aman and S. Szpakowicz, "Identifying expressions of emotion in text," in *Proceedings of the 10th international conference on Text, speech and dialogue*, ser. TSD'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 196–205.

[12] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[13] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, no. 1-2, pp. 1–135, Jan. 2008.

[14] H. Binali and V. Potdar, "Emotion detection state of the art," in *Proceedings of the CUBE International Information Technology Conference*, ser. CUBE '12. New York, NY, USA: ACM, 2012, pp. 501–507.

[15] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, Dec. 2007.

[16] C. O. Alm, "Affect in text and speech," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2008.

[17] S. Aman and S. Szpakowicz, "Using roget's thesaurus for fine-grained emotion recognition," in *Proceedings of the Third International Joint Conference on Natural Language Processing*, 2008, pp. 296–302.

[18] C. Strapparava and A. Valitutti, "Wordnet affect: an affective extension of wordnet," in *LREC*. European Language Resources Association, 2004.

[19] A. Kennedy and S. Szpakowicz, "Evaluation of a sentence ranker for text summarization based on roget's thesaurus," in *Proceedings of the 13th international conference on Text, speech and dialogue*, ser. TSD'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 101–108.

[20] ——, "A supervised method of feature weighting for measuring semantic relatedness," in *Proceedings of the 24th Canadian conference on Advances in artificial intelligence*, ser. Canadian AI'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 222–233.

[21] S. Aman, "Recognizing emotions in text," Master's thesis, University of Ottawa, 2007.

[22] D. Ghazi, D. Inkpen, and S. Szpakowicz, "Hierarchical approach to emotion recognition and classification in texts," in *Advances in Artificial Intelligence*. Springer, 2010, pp. 40–50.

[23] M. Jarmasz and S. Szpakowicz, "Roget's thesaurus and semantic similarity," in *RANLP*, ser. Current Issues in Linguistic Theory (CILT), N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, Eds., vol. 260. John Benjamins, Amsterdam/Philadelphia, 2003, pp. 111–120.

[24] S. M. Kim, "Recognising emotions and sentiments in text," Master's thesis, University of Sidney, 2011.

[25] D. de Kok and H. Brouwer, "Natural language processing for the working programmer," 2011.

[26] S. Chaffar and D. Inkpen, "Using a heterogeneous dataset for emotion analysis in text," in *Proceedings of the 24th Canadian conference on Advances in artificial intelligence*, ser. Canadian AI'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 62–67.

[27] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.

[28] R. R. Bouckaert, E. Frank, M. Hall, R. Kirkby, P. Reutemann, A. Seewald, and D. Scuse, *WEKA Manual for Version 3-6-9*, University of Waikato, 2013.

[29] B. Weiner, *Social motivation, justice, and the moral emotions: An attributional approach*. Psychology Press, 2005.