Towards Building User Affect Models from Tweets: a comparative study with various datasets, features and algorithms

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 use machine learning techniques to try to find the best possible solution, including datasets, emotion lexicons, features, and algorithms, for the automatic detection of emotions in tweets. For that purpose, we consider as emotional categories, the six basic emotions from Paul Ekman (*anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*). For the experiments, we rely on five datasets, of which three are regularly found in the literature (news headlines, fairy tales, and blogs sentences), another is directly related with the topic under study (emotional tweets), and the other was composed by us (emotive tweets), specifically for this study and will be primarily used as an evaluation dataset. In addition, we also count with three different emotion lexicons: WordNet-Affect, Roget's Thesaurus, and NRC Emotion Lexicon. From this collection of data, we compare the performance of two different classification algorithms: Naive Bayes and Support Vector Machines (SVM). The experiments evidence the existence of two alternatives that are suitable for the purpose of this study. We evaluate and discuss the results from the point of view of building user affect models, based on the emotional information that can be collected from the tweets.**

***Index Terms*— Machine Learning, Sentiment Analysis, Emotion Detection, Text Classification, Twitter**

I. Introduction

Today, more than ever, text plays a fundamental role in our lives, and its use as a form of communication is assuming more importance, not only due to the use of computers, smartphones and tablets, but also because part of the current social interaction is done online, in a textual form, through the social media (e.g. email, online forums, blogs, and social networks) [1]. In fact, humans make use of these social media technologies to describe their interactions with the environment. Everyday humans access Twitter accounts to say what they are doing, Facebook to tell what they are thinking, Google Places to say where they are, etc.

With the increasing use of the social media, there has been a lot of interest in mining the web for trying to infer the emotions that are directly/indirectly expressed through text [2], [3]. This creates, not only new opportunities for the research itself, but also for the industry [3], [4]. On the one hand, from the point of view of research, there is an increasingly need to improve the human computer interface and the way computers contribute to people's text-based communications, to be more intelligent, and act more naturally and socially [1], [5]. On the other hand, from the point of view of industry, emotion detection can be applied, for example, in the business domain, where it can provide clues on what drives customers toward a product [3].

Recently, there has been a remarkable increase of literature about emotion detection, either based on techniques, systems or devices [1](e.g. [6], [7], [8]). Its relevance results, in part from the fact that emotions are an integral part of the human life, and are heavily involved in the process of decision making [9]. However, in spite automatic emotion detection is a very active research field, understanding the emotional meaning of a text, or sentence, remains an open and quite challenging problem [1]. In the last years, some solutions have been proposed (e.g. [5], [10], [11], [12], [13]). 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 [14] (*anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*), either at sentence, or word level (for an overview about the current state of the art in text-based emotion detection see [3]).

It turns out that this problem is of a multidimensional nature, depending on datasets, emotion lexicons, features, and algorithms, and moreover on their proper combination. In this paper, we try to find the best possible solution to automatically detect emotions in tweets. Currently, Twitter is one of the most popular online social networks, and since its launch year (2006), has been growing at a very fast rate, having recently achieved an important goal: more than 200 million active users, creating over 400 million tweets every day[1]. On Twitter, users post and read messages, called tweets, that are up to 140 characters long. The length of this messages constitute a major limitation [13], however, many of the times tweets include one, or more words, immediately preceded by a hash symbol (#), called hashtags, which serve to indicate additional information, such as the topic, the tone of the message, or even the internal emotions of the users [13]. In this study, we

[1] http://blog.twitter.com/2013/03/celebrating-twitter7.html

focus on this classification domain, with the aim at building user affect models involving the Ekman's six basic emotions. For that purpose, we collected five datasets, of which three are more heterogeneous and regularly found in the literature (news headlines, fairy tales, and blogs sentences), another is directly related with the classification domain (emotional tweets), and the other was composed by us (emotive tweets), specifically for this study and will be primarily used as an evaluation dataset (test set). Additionally, we also collected three different emotion lexicons: WordNet-Affect, Roget's Thesaurus, and NRC Emotion Lexicon. From this collection of data, we compare the performance of two different classification algorithms: Naive Bayes and SVM. The experiments were divided into two parts: (i) evaluation of the classifiers by performing cross-validation on the dataset that was constituted by us, using different features, and different emotion lexicons (ii) evaluation of the classifiers by using a train/test model, in which the training sets correspond to four of the five datasets, the dedicated test set corresponds to our dataset, and similarly to the evaluation (i), we also use different features, and different emotion lexicons.

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

II. Materials & Methods

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

A. Datasets

For this study, we collected five datasets. The information that follows describes each one.

1) Text Affect: This dataset[2] [11] consists of news headlines collected 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 training 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 a given sentence, and 100 represents the maximum emotional value. Besides that, each sentence has also an intensity scale that represents valence annotations. 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]http://www.cse.unt.edu/~rada/affectivetext/#datasets

2) Fairy Tales: This dataset[3] [10], [15] contains 1207 sentences, from 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 using a merged label set: {*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 [15].

3) Blogs: This dataset [12], [16] 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 the categories *mixed emotion*, and *no emotion*. The *mixed emotion* category was thought to include all the sentences that could not be assigned to any basic category, or that would show more than one type of emotion. The *no emotion* category was added to account all the sentences that had no emotional content [12].

4) Twitter Emotion Corpus: This dataset [13] consists of 21,051 emotional tweets, from about 19,000 different people. Many people use hashtags in the tweets to indicate, or notify other users, of the emotions that are associated with the content of the messages. Based on this idea, the authors decided to collect tweets with the hashtags corresponding to the Ekman's six basic emotions: {*#anger*, *#disgust*, *#fear*, *#joy*, *#sadness*, and *#surprise*}. For that purpose, they supplied the online service Tweet Archivist[4] with the six hashtag queries, corresponding to the six basic emotions, and were able to collect about 50,000 tweets. After some data processing operations, to ensure the proper formatting, and to try to remove noisy data, they were left with the 21,051 tweets, which allowed to form this dataset.

5) EmoTweets: This dataset was created by us, specifically for this study and contains 1400 emotive tweets, from random people, which were collected by using the online service Tweet Archivist. Similarly to the previous dataset, we also resorted to that service, but in our case, we just wanted to collect a sample of tweets to build an evaluation dataset. To this end, we used the service to search tweets that were annotated with the hashtags corresponding to the six basic emotions. Additionally, we also collected tweets that were annotated with the hashtag *#noemotion*, just in case if in the future we want to consider the use of messages that do not have any emotional content. The collection process was performed by hand, tweet-by-tweet, in order to filter noisy data, and to confirm if the relevant emotions were, or not expressed in the messages.

B. Emotion Lexicons

The following information describes the three emotion lexicons that were used in our experiments.

1) WordNet-Affect: This resource[5] [17] is an extension of WordNet Domains[6], that includes a subset of synsets suitable to represent affective concepts correlated with affective words.

[3]http://lrc.cornell.edu/swedish/dataset/affectdata/index.html
[4]http://www.tweetarchivist.com/
[5]http://www.cse.unt.edu/~rada/affectivetext/#resources
[6]http://wndomains.fbk.eu/

It is composed by 1128 emotion-related words, which are divided over six lists, corresponding to the six basic emotions.

2) Roget's Thesaurus: This resource[7] [18], [19] is not exactly an emotion lexicon, but rather a lexical knowledge base, which can be used for various natural language processing tasks. Still, from this resource it is possible to derive an emotion lexicon composed by words that are related with the six basic emotions [20]. The semantic relatedness between words is calculated according to the path length between them (nodes), and the similarity measure assigns scores, which can range between 0, for the least related words, and 16, for the most semantically related ones [20], [21]. The words with a score between 12 and 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 [21]. To build an emotion lexicon, initially it is necessary to define a set of words, which will serve as a basis of derivation, and then it is important to decide about the most appropriate similarity score to be considered [20]. For our experiments, initially we defined the following set of derivation: {*angry/anger*, *disgusted/disgust*, *fear*, *happy/happiness*, *sad/sadness*, and *surprised/surprise*}. In relation to the similarity score, due to the experiments reported in [21], we only considered the words that had similarity scores of 12, or more (from intermediate, to the highest level of similarity). Using this implementation, and opting for the choices that were described, we were able to build a lexicon composed by 2678 emotion-related words.

3) NRC Emotion Lexicon: This lexicon [22], [23] consists of 14,182 words that were annotated according to the eight emotions from Robert Plutchik [24] (six basic emotions from Ekman, plus *trust*, and *anticipation*), as well as with positive and negative sentiment. The authors used the Amazon online service called Mechanical Turk[8], to create and compile a manually annotated emotion lexicon. For our experiments, we only considered the annotations corresponding to the six basic emotions.

C. Classification System

The information that follows describes and explains, in detail, the various components that constitute our classification system. Before that, it is important to mention that all the experiments were performed using the *Weka*[9] software.

1) Data pre-processing: In order to detect emotions in text, initially it is important to use processes for ensuring that all data is appropriately formatted [3]. In our experiments, we used stop-words detection, to remove the most commonly used words (e.g. "be", and "for"), with the aid of the *Weka* default stop-words list, based on *Rainbow*[10]. Besides that, we also used the *Porter* stemmer, from the *Snowball*[11] stemmers package, available on *Weka*, to reduce the words to their root forms (e.g. "stemming" and "stemmed" should be reduced to "stem"). Finally, to deal with contractions (short forms), we developed a method to replace the most common positive and negative English short forms, by their corresponding long forms (e.g. "we're" by "we are", and "we couldn't" by "we could not"). This method was designed to deal primarily with negations.

2) Features: Even though the sets of emotional words, collected from the emotion lexicons, may be seen by itself as emotion features, in our experiments we also used unigrams, bigrams, trigrams, and n-grams (ranging between 1 and 3 grams). N-gram models can be visually imagined as a small window that moves through a sentence, or a text, in which only *n* words are visible at the same time [25]. The unigrams are the simplest n-gram models, where only one word can be seen at a time [25]. The bigrams of a sentence can be found by placing the window on the first two words, and by moving it 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]. Therefore, a bigram can be seen as a window that shows two words at a time. The trigrams can have a similar interpretation to bigrams, with the difference that in this case, the window has a length of three words. N-grams can be used to find syntactic patterns in text, and may be useful to deal with negations [26]. The presence of a negation in a sentence can change the emotional category that is really being expressed. For example, the sentence "We will never be happy" should be assigned into the *sadness* category, although, the non-use of n-grams can cause the sentence to be assigned into the *happiness* category.

3) Algorithms: In our experiments, we created two classifiers, to respectively compare two of the top 10 algorithms in data mining [27]: Naive Bayes and Sequential Minimal Optimization (SMO), an implementation of the SVM machine learning model.

4) Evaluation: The experiments of this study are based on two evaluation models. In the first part, the classifiers were evaluated by performing cross-validation on the EmoTweets dataset, using 10-folds, which is a commonly used value [28]. This approach allows us to predict the performance of a classifier that uses our dataset. The second part follows an evaluation based on a train/test model, which is particularly important to analyze how the use of heterogeneous datasets can influence the classification of our dataset. Regarding the evaluation of the results, in our experiments we considered four evaluation metrics that are often found in the literature [3]: accuracy, precision, recall, and f-measure. Accuracy is a simple measure that is used to evaluate the performance of a classifier. Precision and recall are used to evaluate the accuracy of the classification algorithms [3]. F-measure is a simple measure, derived from precision and recall, which indicates the level of performance of a classifier [3]. All the values of precision, recall, and f-measure presented in the Section III are weighted average values, which were automatically calculated by *Weka*, taking into account the respective values of each emotion category.

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

[8]https://www.mturk.com/mturk/

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

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

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

III. Results & Discussion

In this section, we present and discuss the obtained results. All the values in the tables that are in bold text correspond to the highest values, of each evaluation metric, on each dataset. It is worth mentioning that the following tables contain specific abbreviations, which are translated into: *A* (Accuracy), *P* (Precision), *R* (Recall), *F* (F-measure), *TEC* (Twitter Emotion Corpus), *WNA* (WordNet-Affect), *RT* (Roget's Thesaurus), *NEL* (NRC Emotion Lexicon), and *WNA+RT+NEL* (combination between the three emotion lexicons).

A. Cross-validation

In this part of the experiments, we evaluated the classifiers by performing cross-validation on our dataset (EmoTweets), initially using different features (see Table I, and Table II), and then combining it with the different emotion lexicons (see Table III, and Table IV).

TABLE I
Results for cross-validation on the EmoTweets dataset, using different features - Naive Bayes

		A	P	R	F
EmoTweets	*Unigrams*	39,42	**42,40**	39,40	39,90
	Bigrams	28,58	28,60	28,60	27,90
	Trigrams	24,92	31,10	24,90	21,10
	N-grams	**40,42**	41,60	**40,40**	**40,50**

TABLE II
Results for cross-validation on the EmoTweets dataset, using different features - SMO

		A	P	R	F
EmoTweets	*Unigrams*	**40,42**	41,00	**40,40**	**40,20**
	Bigrams	28,50	32,00	28,50	27,60
	Trigrams	21,83	32,20	21,80	16,10
	N-Grams	36,83	**44,00**	36,80	36,70

From a general perspective, and only considering the use of features, the results between both algorithms are relatively similar. We achieved the maximum accuracy rate, 40.42%, in two distinct cases: with the Naive Bayes algorithm, using n-grams, and with the SMO algorithm, using unigrams.

TABLE III
Results for cross-validation on the EmoTweets dataset, using different emotion lexicons - Naive Bayes

		A	P	R	F
EmoTweets	*Baseline*	**40,42**	41,60	**40,40**	**40,50**
	WNA	34,19	39,90	34,20	27,40
	RT	33,99	47,40	34,00	24,00
	NEL	24,54	46,90	24,50	12,90
	WNA+RT+NEL	23,67	**58,40**	23,70	10,40

When we combined the dataset with the different emotion lexicons, the results became more disperse. In the Naive Bayes algorithm, we could not reach the baseline results, however, if we consider the various emotion lexicons, we can see that we achieved the highest accuracy rate, 34.19%, with the introduction of WordNet-Affect. Quite close to this value, with a minimal difference, we obtained 33.99%, by

TABLE IV
Results for cross-validation on the EmoTweets dataset, using different emotion lexicons - SMO

		A	P	R	F
EmoTweets	*Baseline*	40,42	41,00	40,40	40,20
	WNA	**49,96**	**52,60**	**50,00**	**48,50**
	RT	33,24	33,90	33,20	30,00
	NEL	16,46	18,50	16,50	15,80
	WNA+RT+NEL	27,03	29,30	27,00	25,80

combining the dataset with the Roget's Thesaurus lexicon. In the SMO algorithm, the scenario is slightly different, because in one case, we were able to surpass the baseline values, with a difference of almost 10%. That value, 49.96%, which in practice corresponds to the maximum accuracy rate of every test performed through cross-validation, was achieved by combining our dataset with the WordNet-Affect lexicon.

B. Train/Test Set

In this part, we adopted an evaluation based on a train/test model. The training sets correspond to four datasets (Text Affect, Fairy Tales, Blogs, and Twitter Emotion Corpus), and the dedicated test set corresponds to our dataset (EmoTweets). Additionally, we also considered as a training set the use of a global dataset, composed by the combination between the four datasets that were already set for training. This evaluation approach was used in the two types of test of this study, i.e., in the one that involves the use of different features (see Table V, and Table VI), and in the other that involves the combination with different emotion lexicons (see Table VII, and Table VIII).

TABLE V
Results for the various training sets, tested with the EmoTweets dataset, using different features - Naive Bayes

		A	P	R	F
Text Affect	*Unigrams*	**18,33**	**28,10**	**18,30**	**8,70**
	Bigrams	16,67	2,80	16,70	4,80
	Trigrams	16,67	2,80	16,70	4,80
	N-grams	18,25	26,80	**18,30**	8,40
Fairy Tales	*Unigrams*	28,70	**36,40**	28,70	26,80
	Bigrams	20,90	16,20	20,90	14,00
	Trigrams	20,00	4,00	20,00	6,70
	N-grams	**29,30**	35,10	**29,30**	**27,00**
Blogs	*Unigrams*	30,58	**40,70**	30,60	29,30
	Bigrams	19,58	22,50	19,60	13,60
	Trigrams	17,25	24,50	17,30	5,90
	N-grams	**30,75**	37,70	**30,80**	**29,60**
TEC	*Unigrams*	**29,25**	**41,60**	**29,30**	**27,70**
	Bigrams	16,92	26,70	16,90	7,20
	Trigrams	16,92	19,50	16,90	5,30
	N-grams	28,33	38,60	28,30	25,20
Global	*Unigrams*	**28,08**	**40,90**	**28,10**	**26,40**
	Bigrams	16,83	10,00	16,80	7,00
	Trigrams	16,92	19,50	16,90	5,30
	N-grams	27,33	39,80	27,30	24,00

Doing a general analysis of these results, and only considering the use of features, we can immediately mention that in the Naive Bayes algorithm, we could not achieve satisfactory results. The highest accuracy rate that we got was only 30.75%, obtained with the Blogs dataset, using n-grams. In the SMO algorithm, the results were somewhat different, and despite in

TABLE VI
RESULTS FOR THE VARIOUS TRAINING SETS, TESTED WITH THE EMOTWEETS DATASET, USING DIFFERENT FEATURES - SMO

		A	P	R	F
Text Affect	*Unigrams*	**22,08**	**33,10**	**22,10**	**17,30**
	Bigrams	16,92	9,30	16,90	6,10
	Trigrams	16,67	2,80	16,70	4,80
	N-grams	19,17	24,20	19,20	9,80
Fairy Tales	*Unigrams*	**28,50**	36,20	**28,50**	**26,80**
	Bigrams	21,00	38,00	21,00	9,20
	Trigrams	20,40	11,90	20,40	7,90
	N-grams	23,40	**42,00**	23,40	14,50
Blogs	*Unigrams*	**33,33**	44,20	**33,30**	**32,10**
	Bigrams	19,00	38,00	19,00	10,50
	Trigrams	17,08	38,30	17,10	5,80
	N-grams	22,42	**54,90**	22,40	15,50
TEC	*Unigrams*	**41,67**	**51,00**	**41,70**	**40,20**
	Bigrams	28,67	39,00	28,70	26,10
	Trigrams	22,25	45,20	22,30	16,50
	N-grams	40,50	48,30	40,50	38,80
Global	*Unigrams*	41,67	**49,70**	41,70	**40,70**
	Bigrams	30,25	41,40	30,30	27,60
	Trigrams	23,17	38,00	23,20	17,50
	N-grams	**41,83**	48,70	**41,80**	40,60

some cases they were significantly better, there are only two values that should be highlighted. The first, corresponds to the highest accuracy rate, 41.83%, obtained with the Global dataset, using n-grams. The second, with a minimal difference in relation to the first, corresponds to the second highest accuracy rate, 41.67%, and was obtained with the Twitter Emotion Corpus, using unigrams.

TABLE VII
RESULTS FOR THE VARIOUS TRAINING SETS, TESTED WITH THE EMOTWEETS DATASET, USING DIFFERENT EMOTION LEXICONS - NAIVE BAYES

		A	P	R	F
Text Affect	*Baseline*	**18,33**	**28,10**	**18,30**	**8,70**
	WNA	17,42	23,20	17,40	7,10
	RT	17,50	12,50	17,50	7,50
	NEL	16,83	27,00	16,80	7,20
	WNA+RT+NEL	17,17	25,00	17,20	6,00
Fairy Tales	*Baseline*	**29,30**	**35,10**	**29,30**	**27,00**
	WNA	21,00	26,40	21,00	15,50
	RT	18,00	21,80	18,00	11,10
	NEL	18,50	14,90	18,50	12,80
	WNA+RT+NEL	17,00	6,20	17,00	8,40
Blogs	*Baseline*	**30,75**	37,70	**30,80**	**29,60**
	WNA	28,00	37,60	28,00	25,00
	RT	24,17	**46,40**	24,20	21,30
	NEL	21,58	22,60	21,60	16,90
	WNA+RT+NEL	18,00	33,70	18,00	10,10
TEC	*Baseline*	**29,25**	41,60	**29,30**	**27,70**
	WNA	28,33	40,40	28,30	26,70
	RT	27,67	40,80	27,70	25,50
	NEL	24,08	**49,90**	24,10	20,20
	WNA+RT+NEL	21,00	19,80	21,00	15,30
Global	*Baseline*	**28,08**	40,90	**28,10**	**26,40**
	WNA	27,50	41,50	27,50	25,90
	RT	25,75	38,50	25,80	23,40
	NEL	23,17	**45,00**	23,20	19,20
	WNA+RT+NEL	21,00	39,20	21,00	15,70

The combination of the various datasets with the emotion lexicons, generated a greater dispersion of results between the two algorithms. In the Naive Bayes algorithm, we could not

TABLE VIII
RESULTS FOR THE VARIOUS TRAINING SETS, TESTED WITH THE EMOTWEETS DATASET, USING DIFFERENT EMOTION LEXICONS - SMO

		A	P	R	F
Text Affect	*Baseline*	22,08	33,10	22,10	17,30
	WNA	**31,58**	**45,30**	**31,60**	30,10
	RT	26,33	31,20	26,30	24,50
	NEL	30,00	31,70	30,00	28,60
	WNA+RT+NEL	31,17	35,00	31,20	**30,40**
Fairy Tales	*Baseline*	28,50	36,20	28,50	26,80
	WNA	31,42	**43,40**	31,40	29,90
	RT	27,08	33,80	27,10	26,40
	NEL	29,92	33,40	29,90	28,70
	WNA+RT+NEL	**32,25**	36,50	**32,30**	**32,20**
Blogs	*Baseline*	33,33	**44,20**	33,30	32,10
	WNA	34,67	43,50	34,70	33,30
	RT	33,83	39,60	33,80	33,60
	NEL	**34,75**	36,00	**34,80**	**33,70**
	WNA+RT+NEL	34,17	37,30	34,20	**33,70**
TEC	*Baseline*	41,67	51,00	41,70	40,20
	WNA	41,83	52,00	41,80	40,80
	RT	41,58	**52,10**	41,60	40,20
	NEL	42,25	47,50	42,30	40,60
	WNA+RT+NEL	**43,00**	50,80	**43,00**	**42,00**
Global	*Baseline*	41,83	48,70	41,80	40,60
	WNA	41,42	48,60	41,40	40,30
	RT	**42,83**	**50,00**	**42,80**	**41,50**
	NEL	41,67	46,30	41,70	40,20
	WNA+RT+NEL	41,58	48,20	41,60	40,30

reach any of the baseline results. If we only consider the use of the emotion lexicons in separate, in general, we obtained better results through the combinations with the WordNet-Affect. For the SMO algorithm, the scenario is completely different. In all datasets, the introduction of the emotion lexicons allowed us to surpass the baseline results. However, this conclusion is not so simple, because for the various datasets, the best results were achieved by using different lexicons: Text Affect with WordNet-Affect, Fairy Tales and Twitter Emotion Corpus with the combination between the three emotion lexicons, Blogs with the NRC Emotion Lexicon, and the Global dataset with the Roget's Thesaurus lexicon. Even so, there are only two results that deserve a special attention: 43%, obtained with the Twitter Emotion Corpus, when combined with the three emotion lexicons, and 42.83%, with the Global dataset, combined with the Roget's Thesaurus lexicon.

Choosing the best solution is not an easy task. Firstly, and as could be seen, the choice for the best dataset depends very much on the classification domain that is being concerned. For example, in the second evaluation model, the lowest results were obtained from the Text Affect and Fairy Tales datasets. Here, the classification domain involves tweets, which are sentences that are structurally quite different from news headlines, or fairy tales. Furthermore, it is also necessary to choose the classification algorithm, the type of features, and whether to include, or not, an emotion lexicon. In our opinion, our dataset, specially when combined with the WordNet-Affect, has the potential to be a good choice. However, we know that the results were obtained with cross-validation, and not with a train/test model, where we could analyze and study, more carefully, the classification, for example, of another sample

of tweets. On the other hand, there are two other alternatives: Twitter Emotion Corpus, or the Global dataset. In both, we obtained good results, however, the problem here is that a global dataset is computationally more expensive, and for that reason, the Twitter Emotion Corpus has to be our choice. Regarding the features, both unigrams and n-grams allowed us to obtain good results, but specifically for the Twitter Emotion Corpus, unigrams is the feature that best fits. As for the choice of having, or not, a combination with an emotion lexicon, our opinion is that, in most cases, its inclusion is a plus, because we are adding words that directly/indirectly translate emotions. A choice based on the three, i.e., a global lexicon, is an option that is computationally more expensive, and therefore, if we have to choose one of the three, in separate, the choice goes to the NRC Emotion Lexicon, because it is the second best choice for the Twitter Emotion Corpus, and contains the largest amount of emotion-related words.

IV. Conclusion

The results that we achieved in this study allow us to demonstrate that there are several alternatives to automatically detect emotions in tweets. The best solution must include a classifier based on the SMO algorithm. As for the constitution of the data, one alternative is a solution that was confirmed by this study, and is composed by: Twitter Emotion Corpus combined with the NRC Emotion Lexicon, using unigrams. The other alternative is a solution that was not completely confirmed, but has the potential to be practicable. This solution consists of: EmoTweets combined with the WordNet-Affect lexicon, also using unigrams.

Thinking about the future, we are currently developing an application where the use of user affect models, based on social media (e.g. Twitter and Facebook), can really be demonstrated. These models can be applied in different purposes, such as in recommender systems, to try to predict user actions, assuming that an action depends on the emotional state.

References

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

[2] Y. W. Lo and V. Potdar, "A review of opinion mining and sentiment classification framework in social networks," in *Digital Ecosystems and Technologies, 2009. DEST'09. 3rd IEEE International Conference on*. IEEE, 2009, pp. 396–401.

[3] 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.

[4] H. Binali, V. Potdar, and C. Wu, "A state of the art opinion mining and its application domains," in *Proceedings of the 2009 IEEE International Conference on Industrial Technology*, ser. ICIT '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 1–6.

[5] 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.

[6] K. Bloom, N. Garg, and S. Argamon, "Extracting appraisal expressions," in *Proceedings of Human Language Technologies/North American Association of Computational Linguists*, 2007.

[7] J. T. Hancock, C. Landrigan, and C. Silver, "Expressing emotion in text-based communication," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '07. New York, NY, USA: ACM, 2007, pp. 929–932.

[8] L. Zhang, J. A. Barnden, R. J. Hendley, and A. M. Wallington, "Exploitation in affect detection in open-ended improvisational text," in *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, ser. SST '06. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 47–54.

[9] H. Binali, C. Wu, and V. Potdar, "Computational approaches for emotion detection in text," in *Digital Ecosystems and Technologies (DEST), 2010 4th IEEE International Conference on*. IEEE, 2010, pp. 172–177.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

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

[16] 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.

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

[18] 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.

[19] ——, "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.

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

[21] 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.

[22] S. M. Mohammad and P. D. Turney, "Emotions evoked by common words and phrases: using mechanical turk to create an emotion lexicon," in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, ser. CAAGET '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 26–34.

[23] ——, "Crowdsourcing a word–emotion association lexicon," *Computational Intelligence*, 2012.

[24] R. Plutchik, "On emotion: The chicken-and-egg problem revisited," *Motivation and Emotion*, vol. 9, no. 2, pp. 197–200, Jun. 1985.

[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] 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.

[28] 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.