Early Fusion of Low Level Features for emotion Mining

Fabon Dzogang, Marie-Jeanne Lesot, Maria Rifqi, Bernadette Bouchon-Meunier

Biomedical Informatics Insights

ORIGINAL RESEARCH

Early Fusion of Low Level Features for Emotion Mining

Fabon Dzogang[1], Marie-Jeanne Lesot[1], Maria Rifqi[1,2] and Bernadette Bouchon-Meunier[1]

[1]PhD candidate at University Pierre et Marie Curie, Paris 6, CNRS, UMR7606, LIP6, France.
[2]University Panthéon-Assas, Paris 2, France. Corresponding author email: fabon.dzogang@lip6.fr

Abstract: We study the discrimination of emotions annotated in free texts at the sentence level: a sentence can either be associated with no emotion (neutral) or multiple labels of emotion. The proposed system relies on three characteristics. We implement an early fusion of grams of increasing orders transposing an approach successfully employed in the related task of opinion mining. We apply a filtering process that consists in extracting frequent *n*-grams and making use of the Shannon's entropy measure to respectively maintain dictionaries at balanced sizes and keep emotion specific features. Finally the overall system is implemented as a 2-step decision process: a first classifier discriminates between neutral and emotion bearing sentences, then one classifier per emotion is applied on emotion bearing sentences. The final decision is given by the classifier holding the maximum confidence. Results obtained on the testing set are promising.

Keywords: emotion mining, text analysis, *n*-grams, fusion, entropy

Biomedical Informatics Insights 2012:5 (Suppl. 1) 1–8

doi: xxxxxxxxxx

This article is available from http://www.la-press.com.

Introduction

While opinion mining (the study of opinions as positive, negative or neutral in free texts) has received great attention over the past few years,[1] less work has been performed in the field of emotion mining that aims at identifying emotion labels, as for instance "anger", "love" or "hate". The lack of consensus in emotion models, the difficulty to annotate datasets as well as the complexity of analyzing emotion expressions in free texts strongly participate in this phenomenon. The success of opinion mining can be explained by the availability of Internet user ratings as well as the simplicity of opinion representations: the task of opinion classification is often tackled as a classical binary classification task.

I2B2's challenge track 2 consists in learning to discriminate emotions labels in free texts.[2] To this aim, participants are provided with a training set made of 600 suicide notes annotated at the sentence level according to $M = 15$ predefined emotion labels (see Table 1 for a complete list). Sentences in the learning set are associated with zero or multiple emotion labels, Table 1 gives the distribution of the labels over the whole dataset. We observe that sentences labeled with more than one emotion represent approximately 7% of the whole dataset and up to 5 emotions are labeled at maximum.

Table 1. Number of occurences of each emotion in the training set in decreasing order.

Emotion label	Distribution
No emotion	2460
Instruction	800
Hopelessness	455
Love	296
Information	295
Guilt	208
Blame	107
Thankfulness	94
Anger	69
Sorrow	51
Hopefulness	47
Happiness/Peacefulness	25
Fear	25
Pride	15
Abuse	9
Forgiveness	6

Micro averaged F1 score is employed to evaluate submitted systems over a testing set composed of 300 notes. To our knowledge, it is the first challenge on emotion classification particularly focused on machine learning; SemEval 2007 proposed a track (task 14)[3] consisting in classifying news headlines for several emotions, but due to the small size of the training set, purely linguistic approaches were strongly favored.

We propose a system based on the early fusion of n-grams of increasing orders for representing sentences. Early fusion is the process of merging information from different sources in the input examples. In other words it is the process of taking into account features from different sources at the vector level. Fusion performed at the classifier level is called late fusion, at the similarity function level, intermediate fusion.[4] Here, each order, ie each n value, defines a specific representation of a sentence, a decision surface is then learned in the space made of the concatenation of these representations.

The motivation behind the use of grams of higher orders is to mix features with increasing lengths for representing expressions of emotions. While unigrams are widely employed for representing documents in the classical text classification task, they do not seem to provide enough description in the case of sentiment analysis. By fusing grams of increasing orders, one is able to make use of richer features to describe naturally complex and subtle expressions of emotions. An interesting example is the negation which plays an important role in the detection of emotions' patterns. For instance, given the unigram "bad", the change in polarity held by the expression "not bad" is captured by bigrams. More subtle constructs like "not really bad" are represented by trigrams and higher orders can capture even more complex and subtle expressions.

Given a specific gram's order n, we refer to the set of all unique n-grams in the training set as a dictionary D^n. We must note that the higher the order, the more likely are features to appear uniquely in the dataset and the larger the size of the resulting dictionary. When performing early fusion based on increasing grams' orders, one must therefore consider a feature selection process in order to maintain the different

dictionaries at balanced sizes. In this paper we make use of two criteria: we extract frequent *n*-grams which occur more than a given threshold and we select emotion specific features among these frequent *n*-grams according to their Shannon's entropy measures.

The rest of this paper is organized as follow. Related work is presented in Section 1. We then describe our system: sentences are first lemmatized (to this aim we employ TreeTagger)[5] then represented as binary feature vectors made of the fusion of increasing grams' orders (in the vector, 1 indicates the presence of a feature, 0 indicates its absence). In Section 2 we introduce a method for filtering frequent *n*-grams based on the Shannon's entropy measure, leading to dictionaries specific to each emotion label and each gram's order. The learning of the models is described in Section 3. The decision process is implemented as a 2-step algorithm: a neutral vs. emotion classifier is applied to the pre-processed sentences, sentences recognized as bearing emotions are further ran through M different classifiers, one for each emotion (we adopt the classical one vs. all strategy). Finally, we present the results obtained on the testing set composed of 300 notes in Section 4. Conclusion and perspectives of this work are given in Section 5.

Related Work

Internet user reviews have been extensively studied in the task of opinion mining. Authors tackle the task of sentiment analysis as a binary classification task (positive vs. negative opinions). An early work shows that learning binary vectors of unigrams with linear SVMs produces the most accurate classifiers.[6] In their experiments, the authors find that adding bigrams for representing texts leads to a drop in performance. A second study[7] shows that concatenating bigrams and trigrams to the unigrams vector of representation does improve performance on the condition that the number of unigrams, bigrams and trigrams are maintained at balanced sizes. The authors make use of the weighted likelihood ratio in order to select the best k features from each bigrams and trigrams dictionary. By concatenating the filtered bigrams and trigrams to the original unigrams the authors achieve results competing with state of the art methods in opinion mining. Another study[8] shows that on very large datasets, making use of *n*-grams up to $n = 6$ while keeping dictionaries at balanced sizes does improve performance.

In this paper, we propose to transpose this approach by studying its efficiency in the task of emotion mining. Emotions are more complex and their expressions in text are more subtle than opinion.[9] As it is argued by psychologists,[10] emotions can be segmented in positive and negative emotions, emotion mining may then be regarded as a refinement of opinion mining.

Computing the Dictionaries

In our setting, sentences are represented as binary feature vectors made of the relevant *n*-grams of the training set. In this section, we describe a method for extracting and filtering *n*-grams based on both their frequency in the training set and their discriminative value for each emotion.

Extraction step: frequent *n*-grams

The total number of orders employed in the proposed approach is limited by one factor: above a given n value high orders can become less successful depending on datasets. Indeed an *n*-gram representation with high n may in fact draw full sentences as features for describing the dataset. The resulting features then suffer of a lack of representativity in the whole dataset. An extreme scenario would be a dictionary whose entries correspond to every unique sentence in the dataset. In the experiments described in Section 4 as we compute *n*-grams representations up to trigrams, we observe that from $n = 3$, performance is not improved in a cross-validation setting.

As the size of the dictionary drastically increases with grams' orders we remove every *n*-gram occuring rarely in the whole dataset. The initial 3 dictionaries D^1, D^2 and D^3 composed of respectively all unique unigrams, bigrams and trigrams are therefore cleaned as to keep entries occuring in 3 sentences or more in the training set.

Filtering step: Shannon's entropy

The cleaned dictionaries still contain many entries, among them many correspond to noise (for example the unigrams "the", "a", or "and") and many are simply not good at discriminating the sentences over the emotion labels. With a view to deal with noisy features

one usually employs a "stop word list" whose role is to clean common words out of the dictionaries. While this approach is well suited to unigrams representations, it does not cope with grams of higher orders: defining stop lists for n-grams with $n > 1$ is far from being intuitive.

Instead, we make use of an information measure: while weighted log-likelihood[7] ratios or χ^2 scores[8] have been studied in this context, we propose a method based on Shannon's entropy measure to filter grams of any order while keeping emotion discriminative features. Formally, let P be the frequency of occurence of feature f in sentences labeled with emotion e. Shannon's entropy measures f's ambiguity with respect to e as:

$$H_e(f) = -(1-p)\log_2(1-p) - p\log_2(p)$$

It can be observed that H_e reaches its maximum if f is uniformly distributed, ie $p = (1 - p) = 0.5$ in which case f equally contributes to e and to the other emotions. It reaches its minimum if f is non ambiguous, ie p is close to 0 or 1 in which case f contributes specifically either to the emotion e or to the other emotions.

For each of the 3 cleaned dictionaries, we propose to build one new dictionary per emotion label $D^n(e)$. Taking account of the neutral label, the resulting $3(M + 1)$ new dictionaries are made of the features whose Shannon's entropy measure is higher than a threshold $\varepsilon^n(e)$. Each of them is specialized for one emotion label and one gram's order.

We manually estimate $\varepsilon^n(e)$, based on the performances of the corresponding classifiers as described in Section 4. In our experiments, we find that depending on the dictionary, threshold values comprised between 0.8 and 1 hold the best relevance. It must be noted that dictionaries based on unigrams and on rare emotion labels are associated with threshold values closer to 1. These results are compatible with the intuition that unigrams are less specific than grams of higher orders. They also show that rare emotion labels do not possess a specific set of features.

Final representation

Given a sentence s, we apply the early fusion strategy: we compute $M + 1$ different binary feature vectors $\vec{s}(e)$. Each of them is a representation of s specific to one emotion label:

$$s = \begin{cases} (\underbrace{\quad}_{D^1(no\,emotion)}, \underbrace{\quad}_{D^2(no\,emotion)}, \underbrace{\quad}_{D^3(no\,emotion)})\ \vec{s}(no\,emotion) \\ \dots \\ (\underbrace{\quad}_{D^1(forgiveness)}, \underbrace{\quad}_{D^2(forgiveness)}, \underbrace{\quad}_{D^3(forgiveness)})\ \vec{s}(forgiveness) \end{cases}$$

As presented in Section 3 each classifier, for emotion e, is trained on its associated representation $\vec{s}(e)$.

Learning the Models

As illustrated on Figure 1, the classification of new sentences is viewed as a 2-step process involving $M + 1$ classifiers. Firstly, one binary classifier discriminates between neutral and emotion bearing sentences, sentences bearing emotions are then further processed: adopting the classical one vs. all strategy, M classifiers discriminate one emotion label against all other emotion labels. Finally the classifier holding the highest confidence wins over the others (confidence is measured as the distance to the separating hyperplane). It must be noted that the proposed system only produces one emotion label per sentences even though the training set contains multi-labeled sentences.[a]

We use linear SVMs to learn the classifiers, employing the LIBLINEAR implementation, that solves the L1 regularized SVMs problem in the dual.[11] Linear SVMs compute a separating hyperplane based on the scalar product similarity function, they have been shown to stand for state of the art in traditional text classification and to perform well on the related task of opinion mining. In the neutral vs. emotion setting, sentences associated with no emotion account for the positive examples. In the M other settings, no emotion sentences are removed from the training set and the target emotion label accounts for the positive class, all the other emotion labels standing for the negative class. The soft margin parameter C influencing the trade-off between generalization and accuracy is tuned over a grid search: the consecutive powers of 2 are considered from 0 to 10. Depending on the frequency of the positive class in the training set, a 10-fold cross-validation or a 3-fold cross-validation (for infrequent classes) is performed. Also, to deal

[a]In the training set, 7% of the sentences are multi labeled.

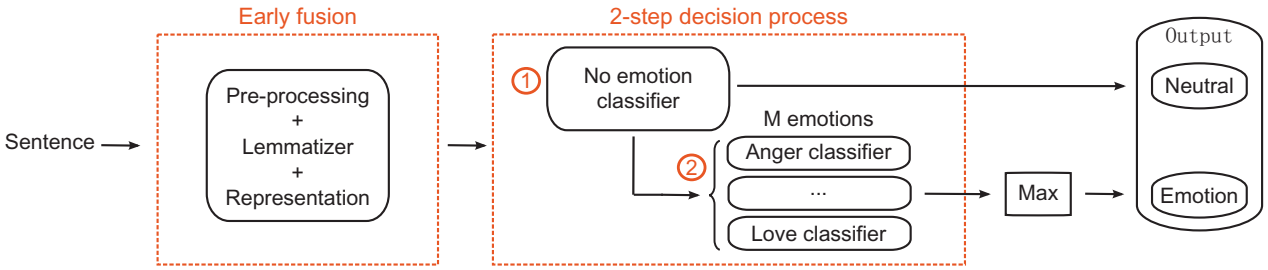

Figure 1. Architecture of the proposed 2-step system: pre-processed sentences are ran through a neutral vs. emotion classifier. Then, emotion bearing sentences are ran through *M* further classifiers, the one holding the most confidence wins over the others.

with imbalanced classes, different costs are introduced for both classes by weighting the supplied C parameter with the corresponding class' frequency. In the end, the $M + 1$ classifiers holding the best averaged F1 score are re-trained on the whole dataset in order to produce the final classifiers.

Experimental Results

In this section we first present and discuss the results obtained by the best performing individual classifiers: we first consider each gram's order independantly, then we consider their fusion. Finally, we give the results achieved by the final system on the testing set composed of 300 notes.

Individual results

Tables 5–7 display the averaged F1 score, precision and recall for the best classifiers (maximizing the F1 score), trained separately on each emotion on respectively unigram, bigram and trigram representations. We must note that in the final system, as a 2-step process is performed, the performances on the emotion labels must be bounded by the performance of the "no emotion" classifier.

We observe that trained separately, grams of lower orders hold better performances than grams of higher orders. It follows our intuition that grams of high orders are more specific and representations relying uniquely on them do not provide enough coverage. Moreover, precision tends to increase on bigrams while recall tends to decrease. However, the gain in precision does not allow the classifiers based on grams of higher orders to discriminate correctly between positive examples and negative examples. This is especially remarkable for the 3 most rare emotions: "Pride" (15 sentences), "Abuse" (9 sentences) and "Forgiveness" (6 sentences). Due to the extreme rarity of these labels in the training set and in spite of the weighting strategy we employed as described in Section 3, the bigrams and trigrams representations alone cannot be exploited to learn an effective classifier (N/A's in the tables indicate that the SVMs learned a majority vote classifier). Nevertheless, we notice that in some cases grams of high orders stand for the best description: for example trigrams provide a representation far better than unigrams and bigrams at describing the emotion "Sorrow" and, to a lesser extent, the emotion "Hopelessness".

Despite a general drop in performance over infrequent emotions, we notice that some emotions seem naturally inclined to separate from the others: for example the emotions "Love" and "Thankfulness" do not occur much in the training set, yet they hold good performances on both unigrams and bigrams.

Table 2. Fusion of unigrams and bigrams: averaged F1 score, precision and recall along with standard deviations (sorted by emotion labels' frequencies).

Emotion label	F1 score	Precision	Recall
No emotion	0.73 ± 0.04	0.8 ± 0.03	0.68 ± 0.05
Instruction	0.85 ± 0.02	0.85 ± 0.03	0.86 ± 0.03
Hopelessness	0.68 ± 0.04	0.68 ± 0.04	0.69 ± 0.06
Love	0.78 ± 0.06	0.78 ± 0.07	0.78 ± 0.08
Information	0.54 ± 0.08	0.52 ± 0.06	0.58 ± 0.12
Guilt	0.53 ± 0.07	0.51 ± 0.08	0.55 ± 0.08
Blame	0.32 ± 0.09	0.34 ± 0.11	0.31 ± 0.08
Thankfulness	0.99 ± 0	0.98 ± 0.01	0.99 ± 0.01
Anger	0.2 ± 0.1	0.18 ± 0.08	0.23 ± 0.14
Sorrow	0.16 ± 0.05	0.1 ± 0.03	0.39 ± 0.12
Hopefulness	0.23 ± 0.07	0.16 ± 0.05	0.38 ± 0.1
Happiness/ Peacefulness	0.16 ± 0.12	0.12 ± 0.07	0.29 ± 0.29
Fear	0.21 ± 0.05	0.23 ± 0.06	0.2 ± 0.07
Pride	0.08 ± 0.02	0.05 ± 0.01	0.27 ± 0.12
Abuse	0.02 ± 0	0.01 ± 0	0.44 ± 0.19
Forgiveness	0.24 ± 0.1	0.14 ± 0.07	0.83 ± 0.29

Table 3. Best ranked SVM features (top 7 unigrams and top 7 bigrams) for final classifiers with F1 scores higher than 0.3. Sorted by classifiers' performance in decreasing order.

Emotion label	Features
Thankfulness	thank/appreciate/than/nice/effort/kindness/under be swell/than you/you dear/appreciate it/too ./have be/for your
Instruction	cremate/call/please/sell/funeral/teach/notify to be/forget me/be good/to have/bury me/dispose of/care of
Love	love/wonderful/bless/watch/beloved/most/loving you ./do ./be wonderful/love you/god bless/your john/me on
Hopelessness	cancer/am/suffer/die/struggle/everybody/tired without you/go on/dear jane/can not/. my/be ./of all
Information	bldg/insurance/key/paper/owe/ticket/in of cincinnati/be pay/ohio ./in this/no ./and my/the key
Guilt	sorry/forgive/excuse/fail/hurt/could/burden have be/forgive me/please forgive/have do/understand ./not to/to help
Blame	sorry/thank/love/please/give/wish/go to be/cause you/of it/you ./you to/in the/to go

This suggests that for some emotions it may exist a specific vocabulary which is easier to identify.

While bigrams capture enriched features at the expense of coverage (higher precision and lower recall), unigrams capture simple and generic features (lower precision and higher recall). The combination of the two representations may therefore lead to a better compromise between precision and recall. Now, because the success of the complex constructs that are captured by trigrams prove to be dependant on emotions, we run further experiments (not reported in this paper) in which trigrams are added to the combination of unigrams and bigrams at the vector level. We observed that on average it did not significantly improve the performance of the classifiers. We therefore decide to only consider the combination of unigrams and bigrams. Table 2 displays the averaged F1 score, precision and recall for the best classifiers trained on the fusion of unigrams and bigrams. Again, in the final system, the performances on the emotion labels must be bounded by the performance of the "no emotion" classifier.

On average, the combination of unigrams and bigrams holds better performances than each representation taken separately. As expected, the resulting classifiers exploit a better compromise between precision and recall than for each of the representations taken separately. A good example is the emotion "Love" for which the fusion strategy of uni-grams and bigrams improves both precision and recall, leading to a better F1 score. We must note that some emotions like "Instruction" do not take benefit from the fusion. For this particular emotion, bigrams prove less successful than unigrams at holding precision. Therefore, the combination of unigrams and bigrams could not benefit from them. Generally speaking, it seems that for the fusion to hold better performances,

Table 4. Results of the final system on the testing set made of 300 notes.

F1 micro avg	Recall	Precision
0.47	0.46	0.49

Table 5. Unigrams: averaged F1 score, precision and recall along with standard deviations (sorted by emotions' frequencies).

Emotion label	F1 score	Precision	Recall
No emotion	0.68 ± 0.02	0.71 ± 0.02	0.66 ± 0.03
Instruction	0.85 ± 0.02	0.86 ± 0.03	0.84 ± 0.03
Hopelessness	0.69 ± 0.04	0.63 ± 0.06	0.76 ± 0.05
Love	0.76 ± 0.04	0.73 ± 0.05	0.8 ± 0.07
Information	0.54 ± 0.04	0.45 ± 0.04	0.68 ± 0.07
Guilt	0.52 ± 0.09	0.42 ± 0.08	0.66 ± 0.1
Blame	0.23 ± 0.09	0.19 ± 0.07	0.31 ± 0.15
Thankfulness	0.98 ± 0	0.99 ± 0.01	0.98 ± 0.01
Anger	0.17 ± 0.06	0.12 ± 0.04	0.29 ± 0.11
Sorrow	0.17 ± 0	0.14 ± 0.01	0.22 ± 0.03
Hopefulness	0.24 ± 0.11	0.18 ± 0.08	0.38 ± 0.16
Happiness/ Peacefulness	0.19 ± 0.13	0.19 ± 0.11	0.2 ± 0.15
Fear	0.19 ± 0.08	0.19 ± 0.04	0.19 ± 0.12
Pride	0.11 ± 0.04	0.06 ± 0.02	0.4 ± 0.2
Abuse	0.02 ± 0	0.01 ± 0	0.44 ± 0.19
Forgiveness	0.26 ± 0.1	0.16 ± 0.06	0.83 ± 0.29

Table 6. Bigrams: averaged F1 score, precision and recall along with standard deviations (sorted by emotions' frequencies).

Emotion label	F1 score	Precision	Recall
No emotion	0.72 ± 0.03	0.84 ± 0.03	0.63 ± 0.04
Instruction	0.82 ± 0.01	0.8 ± 0.02	0.84 ± 0.03
Hopelessness	0.64 ± 0.05	0.66 ± 0.04	0.62 ± 0.08
Love	0.74 ± 0.07	0.76 ± 0.08	0.72 ± 0.08
Information	0.47 ± 0.1	0.43 ± 0.09	0.53 ± 0.14
Guilt	0.5 ± 0.08	0.5 ± 0.08	0.5 ± 0.09
Blame	0.28 ± 0.1	0.27 ± 0.08	0.32 ± 0.14
Thankfulness	0.98 ± 0.01	0.98 ± 0.01	0.99 ± 0.01
Anger	0.14 ± 0.01	0.11 ± 0.01	0.2 ± 0.02
Sorrow	0.05 ± 0.02	0.03 ± 0.01	0.16 ± 0.09
Hopefulness	0.2 ± 0.1	0.2 ± 0.06	0.21 ± 0.13
Happiness/ Peacefulness	0.15 ± 0.04	0.26 ± 0.21	0.12 ± 0.01
Fear	0.13 ± 0.06	0.11 ± 0.04	0.16 ± 0.08
Pride	N/A	0 ± 0	0 ± 0
Abuse	N/A	0 ± 0	0 ± 0
Forgiveness	N/A	0 ± 0	0 ± 0

both representations need to provide different strong points, either in terms of recall or precision.

In order to gain further insight into the final individual classifiers, we observe the best weighted features in the SVMs models. Table 3 gives the best features of classifiers holding F1 scores higher than 0.3. In the table, we reported the 7 top ranked unigrams as well as the 7 top ranked bigrams. While features like "love" and "thank" are naturally high rated by linear SVMs, more complex patterns emerge: for instance, the unigram "." ending a sentence combined with the unigram "too" is more relevant to the emotion Thankfulness than the two of them taken separately. We also notice that while identical unigram features can be shared between different classifiers, bigram features remain specific to each emotions.

Table 7. Trigrams: averaged F1 score, precision and recall along with standard deviations (sorted by emotions' frequencies).

Emotion label	F1 score	Precision	Recall
No emotion	0.6 ± 0.02	0.85 ± 0.02	0.47 ± 0.02
Instruction	0.53 ± 0.07	0.61 ± 0.09	0.47 ± 0.08
Hopelessness	0.8 ± 0.02	0.73 ± 0.03	0.88 ± 0.02
Love	0.22 ± 0.06	0.34 ± 0.15	0.17 ± 0.03
Information	0.53 ± 0.04	0.63 ± 0.09	0.47 ± 0.04
Guilt	0.37 ± 0.06	0.4 ± 0.04	0.35 ± 0.08
Blame	0.1 ± 0.01	0.05 ± 0	0.77 ± 0.02
Thankfulness	0.03 ± 0.01	0.02 ± 0.01	0.51 ± 0.15
Anger	0.4 ± 0.08	0.51 ± 0.07	0.33 ± 0.08
Sorrow	0.98 ± 0.01	0.97 ± 0	0.99 ± 0.01
Hopefulness	N/A	0 ± 0	0 ± 0
Happiness/ Peacefulness	0.04 ± 0.01	0.02 ± 0	0.49 ± 0.08
Fear	N/A	0 ± 0	0 ± 0
Pride	N/A	0 ± 0	0 ± 0
Abuse	N/A	0 ± 0	0 ± 0
Forgiveness	N/A	0 ± 0	0 ± 0

Final results

The final system we prepare for evaluation relies on the fusion of unigrams and bigrams. As described in Section 3, test sentences are first pre-processed then labeled using the 2-step decision process described in Section 3.

It must be noted that the proposed system does not output multiple emotions: for emotion bearing sentences, the classifier with highest confidence wins.

As presented in Table 4, on the testing set composed of 300 notes, it obtains 0.47 on micro averaged F1 score, 0.49 on precision and 0.46 on recall. Among all systems submitted to the I2B2 challenge, the worst micro averaged F1 score is 0.30 while the best is 0.61. The average performance is 0.49 ± 0.07.

Conclusions and Perspectives

In this paper, we presented a system for classifiying sentences' emotional content relying on 3 characteristics: the early fusion of grams of increasing orders, a method for filtering the grams based on Shannon's entropy and a 2-step decision process for dealing with neutral sentences. We showed that unigrams only were not sufficient at describing expressions of emotions, naturally complex and subtle. By adding bigram features at the vector levels, we train classifiers holding better performances on average than on each representation separately. In this setting, unigrams seem to boost the recall while bigrams seem to boost the precision of the resulting classifiers. We also show that, by modeling complex constructs, grams of higher orders like trigrams can provide a better description for discriminating emotions. An interesting developement of this work would be to investigate further types of fusions: we believe that combining low level features with external knowledge

is relevant for discriminating emotions. In this setting, intermediate fusion allows to combine different similarity functions, each specific to one source of information. Another perspective of this work is to study the problem of multi-labeling, for instance considering aggregation functions other than *max*. Finally, grams of high order hold better performance for certain emotion, it can be of interest to adopt emotion dependant representations.

Aknowledgements

This work was supported by the CAP DIGITAL project DOXA funded by DGCIS (N°. DGE 08-2-93-0888).

Disclosures

Author(s) have provided signed confirmations to the publisher of their compliance with all applicable legal and ethical obligations in respect to declaration of conflicts of interest, funding, authorship and contributorship, and compliance with ethical requirements in respect to treatment of human and animal test subjects. If this article contains identifiable human subject(s) author(s) were required to supply signed patient consent prior to publication. Author(s) have confirmed that the published article is unique and not under consideration nor published by any other publication and that they have consent to reproduce any copyrighted material. The peer reviewers declared no conflicts of interest.

References

1. Bo Pang, Lillian Lee. Opinion mining and sentiment analysis. *Found Trends Inf Retr*. 2008;2(1–2):1–135.
2. John PP, Pawel M, Michelle LG, et al. Sentiment analysis of suicide notes: A shared task. 2011.
3. Carlo Strapparava and Rada Mihalcea. Semeval-2007 task 14: affective text. In: *Proc of the 4th Int Workshop on Semantic Evaluations*. 2007.
4. Mehmet Gnen, Ethem Alpaydn. Multiple kernel learning algorithms. *Journal of Machine Learning*. 2011;2211:2268.
5. Helmut Schmid. Probabilistic part-of-speech tagging using decision tree. In: *Proc of the Int Conf on New Methods in Language*. 1994.
6. Bo Pang, Lillian Lee, Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In: *Proc of the Conf on Empirical methods in natural language processing*, Morris-town, NJ, USA. Association for Computational Linguistics; 2002:79–86
7. Vincent NG, Sajib D, Niaz Arifin SM. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In: *Proc of the COLING/ACL*. 2006.
8. Hang Cui, Vibhu M, Mayur D. Comparative experiments on sentiment classification for online product reviews. In: *Proc of the 21st National Conf on Artificial Intelligence*. 2006.
9. Fabon D, Marie JL, Maria R, Bernadette BM. Expressions of graduality for sentiments analysis—a survey. In: *Proc of the Int Conf on Fuzzy Systems*. 2010.
10. Robert Plutchik. *The emotions*. University Press of America; 1990.
11. Rong EF, Kai WC, Cho JH, Xiang RW, Chih JL. Liblinear: A library for large linear classification. *Journal of Machine Learning*. 2008;9:1871–4.