ISSN 1870-4069

Emotion Classification of Twitter Data Using an Approach Based on Ranking

Cecilia Reyes-Peña, David Pinto-Avendaño, Darnes Vilariño-Ayala

Benemérita Universidad Autónoma de Puebla, Faculty of Computer Science, Puebla, Mexico
reyesp.cecilia@gmail.com, {dpinto,darnes}@cs.buap.mx

Abstract. In this work, a model for textual emotion classification based on Ranking technique is presented. The Ranking technique uses the frequencies of words in order to assign a relevance for each in a tweets (Spanish) after calculating the total relevance of the tweet for each classes. The classes are associated to four emotions: happiness, sadness, anger and fear and the highest relevance indicates to which class the tweet belongs. The training and test corpora are created by manually selected key words as references for a crawling tool, both contain manually tagged tweets extracted from Twitter; the training corpus was validated by K-Fold Cross Validation having a 90% of acceptance. The results are compared with Naïve Bayes and Bigrams Probabilities models using precision, recall and F-measure.

Keywords: Emotion classification, Ranking, Twitter, Crawling.

1 Introduction

Every day, it is common the transfer of information through electronic platforms, which are gradually replacing conventional communication services, this means a huge growth in the amount of information available on the web. In this scope, the social networks have become powerful tools for communication in social topics, generating textual and multimedia contents that open up the studies related to the interpretation of emotions that people express. An example of this is Twitter, which is one of the most important social networks and is a huge database related to public opinions on several topics.

Twitter is characterized for being a platform where each user can create or share own short publications (each publication has a limit of 280 characters) or from other user. These publications are known as tweets and could contain one or more hashtag. Hashtags are textual labels, starting with the symbol '#' and when a hashtag is consulted, all tweets which contain the hashtags are shown. If a hastag is very used in a few time, it is consider relevant and is named Trending Topic.

For data extraction in order to create the training and test corpora, Twitter offers an API (Application Programming Interface) and a crawler, which divide the process into three stages: authentication, crawling and information pre-processing [19].

In this work, we consider four emotion: happiness, sadness, anger and fear, which are opposite to each other (happiness-sadness and angr-fear) [10], and it was necessary to create a corpus that contained tweets associated to each emotion. These corpus is used for training the proposed model and the comparison models: Naïve Bayes and Bigrams Probabilities.

The test results are analyzed with precision, recall and F-measure. Precision is the percentage of the classifier success between all tweets were classified like belong to a class, recall is the percentage of the classifier success between all tweets belong to a class [3], and F-measure is an harmonic between precision and recall and is closer to smaller value of them. The measure most important in this work is Recall because represents the success percentage.

This work is divided as follows: the section 2 contains some related works about emotion classification, the section 3 describes the proposed ranking technique, the Naïve Bayes and Bigrams Probabilities models, in the section 4 the creation of the training and test corpora is introduced, in the section 5 the experimental and measure results are compared among the three models, finally in the section 6 the conclusions and the future works are presented.

2 Related Works

Several projects for natural language processing to identify emotions haven been developed based on different areas as image processing, voice components or textual information. Some related works about textual information retrieval and text classification approach are presented is this section.

There are competition platforms such as SEMEVAL [1] that specializes on semantic similarity systems or tweets classification based on emotion recognition. These systems use manually classified corpus references, which remain static, thus limiting their performance to the new trends of expressions used in Twitter. Therefore, the corpus manual feedback is too late.

Ashequl Qadir et al. [12] propose a tweet tagging system for emotion classification (affect, anger, fear, joy and sadness) using tweets content, patterns and context of the hashtags as base for Bootstrapping technique. Each emotion is associated to five hashtag seeds that represent them. Tweets that cannot be classified in any of the five groups, are tagged by a prefix search process, where the hashtag seeds are used as roots. For feedbacking the system uses the CoTrain technique; in which there are two equal corpora and two models (A and B) with different features, the model A results feedback model B corpus and the model B results feedback model A corpus [11].

Other way for tweets classification is through identification of positive, negative or neutral meaning of each words cointained in a tweet and taking into account negative words that can change the sense or context of possible objects, the combination of what is expressed in each word generates a general result [5, 14].

There are approaches for finding figurative emotions that are more complex to analyze such as sarcasm, metaphor and irony by relying on words and hashtags contained in them for identifying the used words (literally, figuratively, etc.)[4]. The approach proposed by Georgios Paltoglou [8] designs a system for emotion classification related to global events on Twitter based on the analysis of either negative or positive polarity changes from the keywords use, leaving behind the popularity indicator through counters.

3 Ranking Technique

In this paper, we propose a model based on information retrieval Ranking technique, that can be defined as a process that assigns a relevance value to each term within a document belonging to a corpus or collection, with the purpose to satisfy a question-answer task[18]. There are two types of classification: simple ranking and aggregation ranking. The simple ranking consists in creating relevance lists that are based on the features of the documents, while the aggregation ranking takes lists of documents already established to form a new one [6].

In this work, in the simple ranking each emotion is used as a document, thus having a collection of four documents. Every document contains tweets and the frequency of each words on each document ($TF_{t,d}$, Term Frequency in the Document). Then, it is necessary to calculate the relevance of each word in the corpus (IDF_t, Inverse Document Frequency of the Term)that can be found with Eq.1, where N is the total documents of corpus, DF_t (Document Frequency of the Term) are the number of documents where the word appears and a base 10 logarithm is used for smoothing the relevance. To get the final score of each word in a document the Eq. 2 is used:

$$IDF_t = \log_{10}(N/DF_t), \tag{1}$$

$$TF - IDF_{t,d} = TF_{t,d} \times IDF_t. \tag{2}$$

When receiving a tweet to classify it, a similar process to a query is used, first each word in the tweet will have a smoothed weight by the log10 of the $TF_{t,q}$ (Term Frequency in the Query) and will be multiplied by the calculated IDF_t before (Eq. 3), the result is a weighted vector which will apply a dot product over $TF - IDF$ values vector related at the same query words for each document. The document with the greatest score is the classification result (see Fig.1):

$$w_{t,q} = \log_{10}(TF_{t,q} \times IDF_t). \tag{3}$$

3.1 Naive Bayes Classifier

Naive Bayes is a popular supervised probabilistic classifier based in Bayes Theorem. It assumes that some feature in particular of a class is independent to the

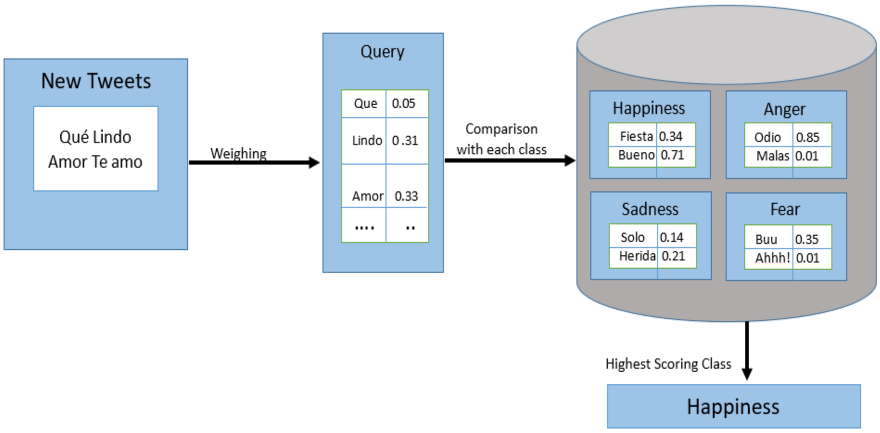

Fig. 1. Ranking Model Diagram.

probability that any other feature belongs or not at the same class [15]. The operation of the classifier lies in calculating the probability of belonging to each feature X in each class C. When the highest value is found, the argument is returned (Eq. 4):

$$\widehat{y} = argmax_{k\in\{1,\ldots,K\}} P(C) \prod_{i=1}^{n} P(x_i \mid C_K). \tag{4}$$

3.2 Probability of Bigrams

The N-gram is a series of N consecutive words that belong to a sentence. In the natural language processing, the most used models based on N-grams are commonly unigrams (one word), bigrams (two words) and trigrams (three words) [13]. The degree of N-grams contains more information than grade N-1. For the calculation of probabilities in models based on N-grams, the Markov assumption is frequently used, in which it is assumed that the probability of a word depends only on the N-1 previous words [2]. Therefore, in a bigram model, the probability of a sentence $P(s)$ made by N words (w_1, w_2, ..., w_n) is given by the multiplication of the probabilities of each word based the previous one (Eq. 5):

$$P(s) = P(w_1)P(w_2 \mid w_1) \ldots P(w_n \mid w_{n-1}). \tag{5}$$

Bigrams have a variety of approaches within the scope of natural language processing. One of them is the detection of spelling errors, in which the frequency of each bigram is calculated in a text corpus and those that are infrequent, possibly contain words with spelling errors, it should be noted that in this type of correctors is not possible identify the types of errors that may occur [16]. Another application of the bigrams is the morphosyntactic labeling from learning having disambiguation benefits based on Markov models [9]. Bigrams have been used in stochastic translation systems by comparing pairs of strings from a source language and a target language [7], and for plagiarism detection [17].

4 Creation of the Corpus

In this work, two sets of tweets manually tagged are used: the training corpus and the test corpus. The training corpus consists of four classes, the tweets were extracted by the Twitter crawler tool, which receives hashtags as keywords associated to each emotion (see Table 1). Each tweet was manually tagged, then the tweets that not express any emotion or expressed contradictory emotions are eliminated. The training corpus size was 105,596 tweets distributed equally for each emotion and for his validation the Naïve Bayes and Bigrams Probabilities classifiers were trained with K-Fold Cross Validation with K=10 (the 10% of the corpus for test and 90% for training in ten different iterations) having 90% of acceptation in average.

Table 1. Crawler keywords.

Emotion	Hashtag Associated
Happiness	#Feliz #Bendecido
Sadness	#Tristeza #Triste #RIP #Depresion #CorazonRoto
Anger	#ALV #Chingada #HDP #TeOdio
Fear	#Miedo #Terror #TengoMiedo #Pavor

For the test set, some common short words were taken as keywords for the crawler (i.e. a, de, un, la) and the tweets were also manually tagged, we obtain is 234 tweets for the four emotions (127 happiness, 31 sadness, 62 angriest and 14 fear).

5 Experimental Results

About the results, the Ranking model has the highest success number respect to Naïve Bayes and Bigrams Probabilities (see Table 2). According the confusion matrices (see Tables 3-5), the emotion with the highest success percentage is happiness in the ranking model, sadness in Naïve Bayes and fear in Bigrams Probabilities. The emotion with the least success percentage is the anger for Ranking model and Bigrams Probabilities, and happiness for Naïve Bayes.

After applying precision, recall and F-measure (see Table 6), it can be seem that the precision the model with the best average of F-measure is Bigrams Probabilities, Ranking model and the worst performance is from Naïve Bayes.

Table 2. General Results.

Model	Success	Errors
Ranking	129	105
Naïve Bayes	92	142
Bigrams Probabilities	124	110

Table 3. Ranking Model Confusion Matrix.

	Happiness	Sadness	Anger	Fear
Happiness	102	8	10	3
Sadness	8	6	15	1
Anger	7	6	11	0
Fear	10	11	26	10
total	127	31	62	14

In the Ranking model, the small precision values for sadness and fear class show the classifier is tagged many tweets that not belonging for these classes. However, the high value in happiness class indicates that the classifier is hitting in the belonging tweets about the total tagged tweets to these class; in recall, the small values for sadness and anger show there are many tweets that ought to belong to these class and are not assigned in these classes, while the high values in happiness and fear indicate that the classifier is hitting in the tagged tweets between the total the belonging tweets to these class. Given these results, F-measure is low because is closer to the lowest values in all classes.

6 Conclusions

Twitter is a great container of social information, so, the manual classification of tweets that is a difficult task because the people has a different ways of expressing. It is vital to know several approaches and techniques that help in the textual emotion classification as well to know the performance of each to select

Table 4. Naïve Bayes Confusion Matrix.

	Happiness	Sadness	Anger	Fear
Happiness	38	0	1	1
Sadness	53	17	16	1
Anger	9	8	27	2
Fear	27	6	18	10
total	127	31	62	14

Table 5. Bigrams Probabilities Confusion Matrix.

	Happiness	Sadness	Anger	Fear
Happiness	73	5	2	1
Sadness	19	19	15	1
Anger	26	5	21	1
Fear	9	2	24	11
total	127	31	62	14

Table 6. Precision, Recall And F-measure Results.

Model	Measure	Happiness	Sadness	Anger	Fear
Ranking	Precision	0.829	0.2	0.458	0.175
	Recall	0.803	0.19	0.177	0.714
	F-measure	0.816	0.196	0.255	0.281
Naïve Bayes	Precision	0.95	0.195	0.586	0.163
	Recall	0.299	0.548	0.435	0.714
	F-measure	0.455	0.288	0.5	0.266
Bigrams Probabilities	Precision	0.901	0.351	0.396	0.239
	Recall	0.574	0.612	0.338	0.785
	F-measure	0.701	0.447	0.365	0.366

the one that provides a better result. This work proposes a new approach about the information retrieval ranking technique, which taking the highest relevant score as a classification result.

Being recall the most important measure in this work the classifier Bigrams Probabilities has the highest average for it despite of the F-measure's average of all classifiers is under the 50% of success. It shows the bigrams approach offers better results.

In future work, it is necessary to modify the Ranking model to get better results, adding other classifiers in order to apply classifiers ensemble for selecting a tweets set and based on it the implementation of automatic corpus feedback technique (Bootstrapping).

References

1. International workshop on semantic evaluation (2016), `http://alt.qcri.org/semeval2017/index.php?id=tasks`
2. Basharin, G.P., Langville, A.N., Naumov, V.A.: The life and work of aa markov. Linear algebra and its applications 386, 3–26 (2004)
3. Feldman, R., Sanger, J.: The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge university press (2007)

4. Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., Reyes, A.: Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). pp. 470–478 (2015)
5. Gonzalez-Marron, D., Mejia-Guzman, D., Gonzalez-Mendoza, M., Enciso-Gonzalez, A.: Explotacion de informacion de redes sociales de microbloggs utilizando analisis de sentimientos y tecnicas propuestas por el big data. Tecnologías de la Información 2(2), 90–103 (2015)
6. Li, H.: Learning to rank for information retrieval and natural language processing. Synthesis Lectures on Human Language Technologies 4(1), 1–113 (2011)
7. Mariño, J.B., Crego, J.M., Lambert, P., Banchs, R., de Gispert, A., R Fonollosa, J.A., Costa-Jussá, M.R.: Modelo estocástico de traducción basado en n-gramas de tuplas bilingües y combinación log-lineal de características. Procesamiento del Lenguaje Natural (35) (2005)
8. Paltoglou, G.: Sentiment-based event detection in t witter. Journal of the Association for Information Science and Technology 67(7), 1576–1587 (2016)
9. Pla, F., Molina, A., Prieto, N.: Evaluación de un etiquetador morfosintáctico basado en bigramas especializados para el castellano. Procesamiento del lenguaje natural 27 (2001)
10. Plutchik, R.: The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. American scientist 89(4), 344–350 (2001)
11. Qadir, A., Riloff, E.: Bootstrapped learning of emotion hashtags# hashtags4you. In: Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis. pp. 2–11 (2013)
12. Qadir, A., Riloff, E.: Learning emotion indicators from tweets: Hashtags, hashtag patterns, and phrases. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1203–1209 (2014)
13. Rodríguez, H.: Lingüística y estadística,¿ incompatibles. Tecnologías del texto y del habla 72, 89–115 (2004)
14. Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S., Ritter, A., Stoyanov, V.: Semeval-2015 task 10: Sentiment analysis in twitter. In: Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015). pp. 451–463 (2015)
15. Russell, S.J., Norvig, P.: Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited, (2016)
16. San Mateo, A.: Un corpus de bigramas utilizado como corrector ortográfico y gramatical destinado a hablantes nativos de español. Revista signos 49(90), 94–118 (2016)
17. Torrejón, D.A.R., Ramos, J.M.M.: Detección de plagio en documentos. sistema externo monolingüe de altas prestaciones basado en n-gramas contextuales. Procesamiento del lenguaje natural 45, 49–57 (2010)
18. Trotman, A.: Learning to rank. Information Retrieval 8(3), 359–381 (2005)
19. Zheng, Z., Alonso-Berrocal, J.L., G Figuerola, C.: Análisis cibermétrico y visual de Twitter. Departamento de Informática y Automática. Facultad de Ciencias. Universidad de Salamanca (2013)