Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1492–1498
Marseille, 11–16 May 2020

EmoEvent: A Multilingual Emotion Corpus based on different Events

Flor Miriam Plaza-del-Arco[1], Carlo Strapparava[2], L. Alfonso Ureña-López[1], M. Teresa Martín-Valdivia[1]

[1]Department of Computer Science, Advanced Studies Center in ICT (CEATIC), Universidad de Jaén, Spain
{fmplaza, laurena, maite}@ujaen.es
[2]Fondazione Bruno Kessler, Trento, Italy
strappa@fbk.eu

Abstract

In recent years emotion detection in text has become more popular due to its potential applications in fields such as psychology, marketing, political science, and artificial intelligence, among others. While opinion mining is a well-established task with many standard datasets and well-defined methodologies, emotion mining has received less attention due to its complexity. In particular, the annotated gold standard resources available are not enough. In order to address this shortage, we present a multilingual emotion dataset based on different events that took place in April 2019. We collected tweets from the Twitter platform. Then one of seven emotions, six Ekman's basic emotions plus the "neutral or other emotions", was labeled on each tweet by 3 Amazon MTurkers. A total of 8,409 in Spanish and 7,303 in English were labeled. In addition, each tweet was also labeled as offensive or non-offensive. We report some linguistic statistics about the dataset in order to observe the difference between English and Spanish speakers when they express emotions related to the same events. Moreover, in order to validate the effectiveness of the dataset, we also propose a machine learning approach for automatically detecting emotions in tweets for both languages, English and Spanish.

Keywords: emotion mining, emotion annotation, social media, affective-computing

1. Introduction

Emotions can be defined as *states that reflect evaluative judgments (appraisal) of the environment, the self and other social agents, in light of the organisms goals and beliefs, which motivate and coordinate adaptive behavior* (Hudlicka, 2011). In psychology, emotions are categorized into basic emotions (those considered universal and innate in human beings such as *joy*, *anger* or *fear*), and complex emotions (those perceived as the result of the combination of the basic ones and that are hard to classify under a single term such as *guilt*, *pride* or *shame*).

In recent decades, research on emotion has become popular in numerous fields including psychology, sociology, neuroscience, endocrinology, medicine, history, and computer science. Emotion analysis in computational linguistics consists of identifying discrete emotion expressed in text and is seen as a natural evolution of sentiment analysis and its more fine-grained model (Seyeditabari et al., 2018). However, as it is a more difficult task than sentiment analysis, this field still has a long way to go.

The automatic detection of emotions in texts is becoming increasingly important due to its vast potential applications in a number of areas, such as marketing to modify or improve business strategies according to the emotion of customers, psychology to detect personal traits, political science to track public emotion on any national, international or political event, education to develop efficient e-learning systems based on student's emotion and so on.

On the other hand, the presence of different languages on the Web is growing every day. However, most of the work and resources developed on emotion analysis have been directed towards English. The emotional expressions of people from other countries and cultures are expressed in different ways since there is a close relationship between the language and the context of its learning, social pressures, cultural influences, and past experience can all help shape the expression of emotion. For this reason it is important to study this field in different languages, since depending on the language there are important cultural differences in the ways emotions are expressed.

In this paper we present a multilingual emotion dataset of tweets based on events related to different domains: entertainment, catastrophes, politics, global commemoration and global strikes. It has been labeled with emotions by three annotators. The selected languages are English and Spanish. This choice of languages intends to show the differences in how people of different language express their emotions in text. Additionally, one of the goals of the dataset is to support further investigations of emotion mining from different languages due to the low availability of datasets annotated in this field.

The remainder of this paper is organized as follows: Section 2 describes the related work presenting some available datasets labelled with emotions. Section 3 describes the dataset creation process. Section 4 presents some statistics on the dataset. Section 5 depicts our baseline evaluation of the dataset based on a machine learning approach. Finally, Section 6 presents conclusions and future work.

2. Related Work

Research efforts in affective computing have focused on classifying text into positive/negative sentiment, while emotion classification models have received relatively less attention.

In recent years, social networks and messaging platforms have attracted the attention of users becoming an important part of our daily lives. For this reason, nowadays we can easily obtain a large amount of data generated by users in

order to obtain a better understanding of emotion.

Corpora are fundamental for training and testing emotion-oriented systems. Currently there is scarce availability of datasets labeled with emotions, and most of them have been generated for English. Some of the most commonly used datasets in recent studies are listed below. EmoBank (Buechel and Hahn, 2017) is a large-scale corpus of English sentences annotated with the dimensional Valence-Arousal-Dominance (VAD) representation format. ISEAR, the International Survey on Emotion Antecedents And Reactions is one of the oldest emotion-labeled datasets and consists of about 76,000 records of emotion provoking text provided by the Swiss Center for Affective Sciences. It contains responses from about 3,000 people around the world who were asked to report situations in which they experienced each of the seven major emotions (*joy*, *fear*, *anger*, *sadness*, *disgust*, *shame*, and *guilt*), and how they reacted to them. The valence and arousal Facebook posts is a dataset of 2,895 Social Media posts rated by two psychologically trained annotators on two separate ordinal nine-point scales. These scales represent valence and arousal. The Affective Text (Strapparava and Mihalcea, 2008) developed for the shared task of affective computing in SemEval 2017 consists of news headlines taken from major newspapers. The annotation was performed manually by six annotators, and the set of labels includes six emotions: *anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise*. In SemEval-2019 Task 3: EmoContext (Chatterjee et al., 2019), the organizers provided a dataset of textual dialogues annotated for four classes: *happy*, *sad*, *anger* and *others*. TEC (Mohammad, 2012) is a large dataset of more than 20,000 emotion-labeled tweets automatically label using hashtags. The set of labels includes six basic emotions: *anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise*. EmoTweet-28 (Liew et al., 2016) is a corpus developed using four different sampling strategies based on random sampling by topic and user. The corpus contains tweets annotated with 28 emotions categories and captures the language used to express an emotion explicitly and implicitly. However, the availability of datasets created specifically for languages other than English is very limited. In SemEval-2018 Task 1: Affect in Tweets, the organizers provided the Affect in Tweets (AIT) dataset for English, Arabic and Spanish tweets (Mohammad et al., 2018). It is composed of a set of tweets annotated for four basic emotions: *anger*, *fear*, *joy*, and *sadness*. A blog emotion corpus was constructed for Chinese emotional expression analysis (Quan and Ren, 2009). This corpus contains manual annotations of eight emotional categories: *expectation*, *joy*, *love*, *surprise*, *anxiety*, *sorrow*, *anger* and *hate*. In particular, we found only a few resources annotated with emotions in Spanish and even most of the English emotion datasets have not been fully annotated manually. For this reason, it is important to focus efforts on creating datasets that are manually labeled and not only in English.

3. Creating the Multilingual Emotion Corpus

Our goal in collecting emotions tweets is to explore great relevant events in a specific time frame on Twitter. In order to accomplish this, we focus on trending topic hashtags. Trending Topics are the most used keywords during a given period of time on Twitter. It is a concept related to fashion trends and topics, what everyone is talking about at any given time. In order to retrieve tweets for each event, we select the trending topic that may contain affective content. In particular, we choose the following events that occurred during April 2019:

1. Notre Dame Cathedral Fire. On 15 April 2019, a structure fire broke out beneath the roof of Notre-Dame Cathedral in Paris.

2. Greta Thunberg. She founded the movement "Fridays for Future". It refers to how she strikes every Friday to protest the lack of effective climate legislation on a governmental level. Students throughout Europe now regularly strike on Fridays.

3. World book day or International Day of the Book, is an annual event organized by the United Nations Educational, Scientific and Cultural Organization (UNESCO) to promote reading, publishing, and copyright. It is marked on April 23, the day of William Shakespeare's birth.

4. Spain Election 2019. The 2019 Spanish general election was held on Sunday, April 2018, to elect the 13th Cortes Generales of the president of Spain.

5. Venezuela's institutional crisis. A crisis concerning who is the legitimate President of Venezuela has been underway since January 10th of 2019, with the nation and the world divided in support for Nicolás Maduro or Juan Guaidó.

6. Game of Thrones. This is an American fantasy drama television series. It is one of the most popular series in the world today. The last season premiered in April 2019.

7. Campeonato Nacional de Liga de Primera Division (La Liga) is the men's top professional football division of the Spanish football league system.

8. The UEFA Champions League (UCL) is an annual club football competition organized by the Union of European Football Associations (UEFA) and contested by top-division European clubs, deciding the best team in Europe.

We find these events very interesting because they belong to different domains such as entertainment (Game of Thrones, La Liga, UCL), catastrophes or incidents (Notre Dame Cathedral Fire), political (Venezuela's institutional crisis, Spain Election), global commemoration (World book day) and global strikes (Fridays for Future). Therefore, we are able to find a variety of emotions in the users who give their opinions on these events.

Event	*Hashtag (SP)*	*# of instances (SP)*	*Hashtag (EN)*	*# of instances (EN)*
Notre Dame	#NotreDameEnLlamas	24,539	#NotreDameCathedralFire	11,319
Greta Thunberg	#GretaThunberg	1,046	#GretaThunberg	1,510
World book day	#diadellibro	8,654	#worldbookday	17,681
Spain Election	#EleccionesGenerales28A	4,283	#SpainElection	493
Venezuela	#Venezuela	5,267	#Venezuela	5,248
Game of Thrones	#JuegoDeTronos	5,646	#GameOfThrones	9,389
La Liga	#Laliga	1,882	#Laliga	1,295
UCL	#ChampionsLeague	6,900	#ChampionsLeague	6,199

Table 1: Hashtags employed to retrieve the tweets for each event and the total number of tweets retrieved in English (EN) and Spanish (SP)

3.1. Hashtag-Based Search on the Twitter Search API

Trending topics are accompanied by hashtags that allow us to easily find all the tweets and conversations by users around that topic.
In order to download the tweets, we used the Twitter Search API[1]. In particular, we used an easy-to-use Python library to access the Twitter API: Tweetpy[2]. It allows us to download messages using a query in a specific language. In our case we chose as a query the trending topic hashtag associated with each event in English and Spanish, as can be seen in Table 1. For each tweet we obtained the following twitter metadata: *id*, *date*, *language*, *location*, *text*, *source*, *followers* and *friends*. We discarded tweets that had less than four words and tweets with very bad spelling. For this, we used a Python spell checker called hunspell[3] which contains a dictionary for English and Spanish. Also, we removed tweets with the prefix "Rt", "RT", and "rt", which indicate that the messages that follow are re-tweets (re-postings of tweets sent earlier by another user).

3.2. Tweet Selection

One of the most commonly used techniques for choosing tweets from a dataset is random selection. However, the problem with this method in our case is that we can obtain many non-affective tweets. Since our goal is to get a dataset mainly labeled with emotions, we followed another strategy for selecting tweets, that of performing a linguistic analysis. It is based on extracting affective features from tweets using the Linguistic Inquiry and Word Count (LIWC) resource (Pennebaker et al., 2001). This is a popular content analysis technique which counts the occurrences of words according to pre-defined psychological and linguistic categories. The LIWC categories are grouped under four main dimensions: Linguistic Dimensions (e.g., word count, pronouns, negations, numbers); Psychological Processes (e.g., positive or negative emotions); the Relativity dimension describes physical or temporal information (e.g., time and space); and Personal Concerns (e.g., occupation, leisure activities). LIWC analysis has been successfully applied to a wide range of data, including determining the linguistic characteristics of emotion, personality, gender and genre (Hancock, et al. 2007; Nowson, et al. 2005). Indeed, this resource is available in English and Spanish. Relying on this resource we focus on the dimension of psychological processes, extracting the following features:

- **Number of affective tweets**. We consider that a tweet is affective if it contains one or more words found in the affective category of LIWC. Otherwise, we assume that the tweet is not affective.
- **Number of positive tweets**. We consider that a tweet is positive if it contains more positive words than negative words. We checked the presence of positive words in the tweets by taking into account the positive category of LIWC.
- **Number of negative tweets**. We consider that a tweet is negative if it contains more negative words than positive words. We checked the presence of negative words in the tweets by considering the negative category of LIWC.

In order to gain a better understanding of the presence of emotion in tweets, we followed a method for calculating a score associated with a given class, as a measure of saliency for the given class inside the tweets collection.
We define the class coverage in the tweets corpus T as the percentage of tweets from T belonging to class C:

$$Coverage_T(C_1) = \frac{\sum_{T_i \in C} Tweets}{Size_T} \quad (1)$$

The prevalence score of class C in the tweets corpus T is then defined as the ratio between the coverage of one class in the corpus T with respect to the coverage of the other class in corpus T.

$$Prevalence_T(C_1) = \frac{Coverage_T(C_1)}{Coverage_T(C_2)} \quad (2)$$

A prevalence score close to 1 indicates a similar distribution of the tweets between class C_1 and class C_2 in corpus. Instead, a score significantly higher than 1 indicates that class C_1 is prevalent in the corpus. Finally, a score significantly lower than 1 indicates that the class C_2 is dominant in the corpus.
In Table 2 we can see the prevalence of the affective and positive classes for the different events in Spanish and English. Interestingly, in both languages the top events where

[1] https://developer.twitter.com/
[2] https://www.tweepy.org/
[3] https://pypi.org/project/hunspell/

the positive class is prevalent are the same: world book day, La Liga and the UCL. However, for the affective class there are more differences between the two languages. For English we find that there is a greater prevalence in the affective class than for Spanish. This means that for these events English speakers express more emotions in tweets than Spanish speakers.

Event	*Prevalence (Affective Class)*		*Prevalence (Positive Class)*	
	SP	EN	SP	EN
Notre Dame	1.37	2.45	0.71	1.43
Greta Thunberg	0.86	1.64	1.31	2.46
World Book Day	1.36	2.25	6.85	12.51
Spain Election	0.92	2.01	1.59	5
Venezuela	1.47	1.44	0.94	1.12
Game of Thrones	0.88	1.53	1.12	1.29
La Liga	0.54	1.27	2.11	10.71
UCL	0.75	1.13	1.93	3.24

Table 2: Prevalent class in the different events

After analyzing the affective and non-affective tweets, we decided to randomly select 1,000 affective tweets and 200 non-affective tweets for each language and event in order to perform the annotation. The final dataset distribution is shown in Table 5.

3.3. Data Annotation

Annotations were obtained via the Amazon Mechanical Turk (MTurk) platform. This is a powerful vehicle for getting tasks done quickly and efficiently.

Customers who complete HITs are called **workers** and customers who publish these tasks are called **requesters**. Requesters can use the MTurk Web user interface to submit the task in small independently solvable units called **HITs** (Human Intelligence Tasks). The annotation provided by a worker for a HIT is called an **assignment**. It is also possible to indicate any additional requirements workers must meet to work on the task. In our case, we selected the location as Spain (ES) to label the Spanish dataset and the United States (US) to label the English dataset. We created HITs for each of the tweets corresponding to the events specified in Table 1. Each HIT had two questions, answered by three different workers. The first question is designed to label the main emotion conveyed by the tweet (*anger*, *fear*, *sadness*, *joy*, *disgust*, *surprise* or *others*), the second one to determine whether the tweet contains offensive language or not.

In order to make the annotation process easier for the workers, we defined some synonyms for each emotion in the case of question 1:

- *anger* (also includes annoyance, rage)
- *disgust* (also includes disinterest, dislike, loathing)
- *fear* (also includes apprehension, anxiety, terror)
- *joy* (also includes serenity, ecstasy)
- *sadness* (also includes pensiveness, grief)
- *surprise* (also includes distraction, amazement)

For the second question, we define the term offensive as: *The text is offensive if it contains some form of unacceptable language (blasphemy). This category includes insults, threats or bad words.*

After the three workers had completed tagging the dataset, we decided the final tweet label based on their labeling in the following way: If two or three annotators agree on the same emotion, we label the tweet with that emotion. Otherwise, we label the tweet as *other*.

3.4. Inter-Annotator Agreement

In order to analyze how often the annotators agreed with each other, we conducted inter-tagger agreement studies for each of the eight emotions. For this we use the Cohen's kappa coefficient and the values are shown in Table 3. In order to measure the level of agreement among the three annotators, we measured the agreement between each annotator and the average of the remaining two annotators.

Emotion	*SP*	*EN*
anger	44.18	19.52
sadness	55.55	38.81
joy	41.10	36.68
disgust	18.61	20.96
fear	29.70	10.08
surprise	17.00	13.22
offensive	54.67	22.15
other	34.78	18.76

Table 3: Kappa coefficient for inter-annotator agreement

As we can see in Table 3 the agreement between the Spanish annotators for each emotion is higher than the one obtained by the English annotators. It can also be noted that the most difficult emotions to label by the annotators are *disgust*, *fear* and *surprise* for both languages.

4. Corpus Statistics

In this section we highlight some statistics regarding the multilingual emotion dataset. These statistics refer to the number of tweets by event, hashtags, emojis and part-of-speech, among others.

Table 4 shows the number of offensive tweets per event in English and Spanish in the dataset. In general, there were few offensive tweets for each event. It is remarkable that in both languages the most offensive tweets were associated with the Venezuelan political incident.

Event	*# of offensive tweets (SP)*	*# of offensive tweets (EN)*
Notre Dame	80	116
Greta Thunberg	6	20
World Book Day	17	24
Spain Election	146	4
Venezuela	184	150
Game of Thrones	165	122
La Liga	17	10
UCL	91	72
Total	706	518

Table 4: Number of offensive tweets in English (EN) and Spanish (SP) in the dataset

Table 5 shows the number of tweets selected by event and language, the average length of the tweets, the number of emojis and the number of unique hashtags found in the

Event	*# of tweets*		*Avg. tweet length*		*# of emojis*		*# of unique hashtags*	
	SP	EN	ES	EN	SP	EN	SP	EN
Notre Dame	1,200	1,200	26.57	26.98	432	242	397	942
Greta Thunberg	630	742	24.91	27.61	279	154	750	1,036
World Book Day	1,200	1,200	23.93	23.83	916	649	827	1,131
Spain Election	1,200	207	20.89	24.67	355	37	373	185
Venezuela	1,200	1,200	24.16	25.16	238	163	681	735
Game of Thrones	1.200	1,200	19.86	21.80	579	565	372	343
La Liga	579	354	19.38	17.70	712	511	372	311
UCL	1,200	1,200	16.77	18.30	782	776	386	641
Total	8,409	7,303	22.06	23.26	4,293	3,097	4,158	5,324

Table 5: Number of tweets by event, average length of tweets, hashtags and emojis in the dataset

Event	*joy*		*anger*		*fear*		*sadness*		*disgust*		*surprise*		*other*	
	SP	EN	SP	EN	SP	EN	SP	EN	SP	EN	SP	EN	SP	EN
Notre Dame	59	148	153	78	2	20	660	234	34	218	27	41	265	461
Greta Thunberg	80	33	33	2	1	9	14	4	3	10	11	5	488	144
World Book Day	465	419	13	39	0	20	32	19	5	74	13	61	672	568
Spain Election	316	190	170	3	44	0	58	6	38	5	38	4	536	146
Venezuela	92	59	283	175	18	57	119	59	55	260	20	20	613	570
Game of Thrones	269	647	107	7	29	3	87	8	9	26	173	12	526	497
La Liga	184	177	23	30	0	28	10	7	1	98	17	6	344	396
UCL	350	366	75	58	2	14	29	79	16	74	45	86	683	523
Total	1,815	2,039	857	392	96	151	1,009	416	161	765	344	235	4,127	3,305

Table 6: Number of tweets by emotion and event in the dataset

dataset. It contains a total of 8,409 tweets for English and 7,303 for Spanish. It should be noted that Spanish users tend to use more emojis than English users to express their opinions on the different events. However, hashtags are more used by English users.

The number of emotion tweets per incident is shown in Table 6, where we can determine which emotions are dominant for each one. World book day was the predominant event for the *joy* emotion. *Anger*, *disgust* and *fear* were more usual for the Venezuela situation. *Sadness* was the most frequent emotion in the case of the Notre Dame Cathedral Fire disaster. *Surprise* was more present at entertainment events such as Game of Thrones and UCL. It is necessary to emphasize that there are some emotions that are difficult to label by human annotators. For example, it can be observed that the number of tweets for *fear*, *disgust* and *surprise* are noticeably lower compared to others (*joy*, *sadness*, *anger*). In particular, *fear* and *surprise* are the most difficult emotions to label. This is because while instances of some emotions tend to be associated with exactly one valence (eg, *joy* is always associated with positive valence), instances of other emotions can be associated with differing valence (sometimes *surprise* or *fear* are associated with positive valence, while other times they are associated with negative valence) (Mohammad, 2016). Therefore, an annotator can be confused to find an opinion that expresses *surprise* but also *joy*. In this case, most of the time the opinion is labelled by the annotator as *joy*.

The grammatical labelling for English and Spanish can be found in Figures 1 and 2. As can be seen, Spanish users tend to use more nouns, verbs and adjectives to express their emotions. However, this is not the case of adverbs, which are more widely used by English users.

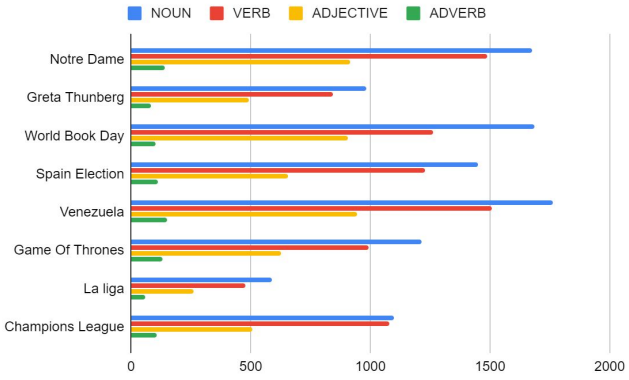

Figure 1: Part-of-speech tagging in the SP dataset

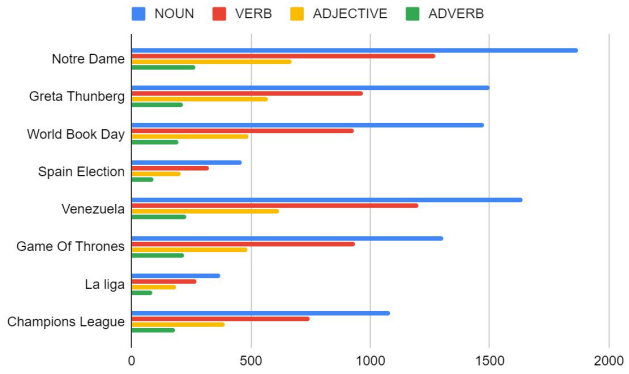

Figure 2: Part-of-speech tagging in the EN dataset

Language	*joy*			*sadness*			*anger*			*fear*			*disgust*			*surprise*			*other*			*macro-avg*			
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	*Acc*
SP	0.60	0.49	0.54	0.79	0.63	0.70	0.55	0.34	0.42	0.63	0.36	0.46	0.21	0.02	0.03	0.39	0.12	0.19	0.64	0.84	0.73	0.54	0.40	0.44	0.64
EN	0.59	0.60	0.6	0.62	0.36	0.46	0.33	0.10	0.16	0.35	0.04	0.07	0.38	0.21	0.27	0.16	0.02	0.04	0.54	0.73	0.62	0.42	0.29	0.32	0.55

Table 7: Results obtained from the multilingual dataset (10-fold cross validation) with SVM

5. Experiments and Results

In this section, we describe the different experiments we carried out to test the validity of the dataset. In particular, we trained a classifier based on machine learning.

5.1. Pre-Processing

Pre-processing the data is the process of cleaning and preparing the text for classification. It is one of the most important steps because it should help improve the performance of the classifier and speed up the classification process. Online texts usually contain a great deal of noise and uninformative parts which increases the dimensionality of the problem and hence makes the classification more difficult. For this reason, we applied pre-processing techniques in order to prepare the data for the text classification. In particular, we preprocessed the tweets following these steps: The tweets were tokenized using NLTK TweetTokenizer[4] and all hashtags were removed.

5.2. Classification

Features in the context of text classification are the words, terms or phrases that express the opinion of the author. These have a greater impact on the orientation of the text. There are several ways to assess the importance of each feature by attaching a certain weight to it in the text. We use the most popular: The Term Frequency Inverse Document Frequency scheme (TF-IDF). Specifically, using this scheme each tweet is represented as a vector of unigrams.
Machine learning techniques are popular in the classification task. For this reason we decide to employ a machine learning algorithm in order to classify the tweets by emotions. In particular, we selected the Support Vector Machine (SVM). It is one of the most well known classifiers since it has been shown to be highly effective and accurate in text categorization.

5.3. Results

In this subsection we report on and discuss the performance of our systems with the multilingual dataset. In order to evaluate and compare the results obtained by our experiments we use the usual metrics in text classification: Precision *(P)*, Recall *(R)*, F-score *(F_1)* and Accuracy *(Acc)*.
We used 10-fold cross validation to evaluate the machine learning classification approach. The results achieved with the SVM algorithm on the multilingual dataset are shown in Table 7. As can be seen, we achieved better results for Spanish (*Acc:* 0.64) than for English (*Acc:* 0.55). For both languages we obtained the best scores on *joy*, *sadness* and *other* labels. However, the other emotions (*anger*, *fear*, *disgust* and *surprise*) are not as easy to detect for our classifier and specifically for English. This may be because we have a lower number of tweets labeled with those emotions and also because they are complementary emotions. It means that, for instance, *anger* and *disgust* may occur at the same time. In fact, the annotators in the labeling process have found it difficult to discern between these two emotions. The same can happen with the *surprise* emotion. Finally, it is important to mention that while in the Spanish dataset we get a good score for the sad emotion (F_1*:* 0.70), this does not occur for the English dataset, where the score is noticeably lower (F_1*:* 0.46).

[4] https://www.nltk.org/api/nltk.tokenize.html

6. Conclusion

In this paper we have described a multilingual dataset of tweets labeled manually with one of seven emotion-categorical labels. In addition, each tweet has also been labeled as offensive or not. The dataset is based on events related to different topics such as entertainment, incidents, politics, global commemoration and global strikes. The selected languages are English and Spanish. EmoEvent contains 8,409 tweets for English and 7,303 tweets for Spanish. We chose to create the dataset in these languages in order to observe the differences in how people of different languages express their emotions in text.
Moreover, in order to validate the effectiveness of the dataset we also propose a machine learning approach for automatically detecting emotions in tweets. Results show that emotion categorization is a complex task and therefore it is important to work on creating resources which will be useful for training and testing algorithms for a number of emotion detection tasks.
As future work we plan to perform more experiments, applying other techniques such as deep learning with the purpose of improving the results regarding those emotions which are more difficult to detect. In the same way, as the dataset is based on different events we will conduct the classification event by event in order to observe the classification behavior of emotions in each of them.

7. Acknowledgements

This work has been partially supported by the Fondo Europeo de Desarrollo Regional (FEDER), LIVING-LANG project (RTI2018-094653-B-C21) and REDES project (TIN2015-65136-C2-1-R) from the Spanish Government.

8. Bibliographical References

Buechel, S. and Hahn, U. (2017). Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585.

Chatterjee, A., Narahari, K. N., Joshi, M., and Agrawal, P. (2019). Semeval-2019 task 3: Emocontext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48.

Hudlicka, E. (2011). Guidelines for designing computational models of emotions. *International Journal of Synthetic Emotions (IJSE)*, 2(1):26–79.

Liew, J. S. Y., Turtle, H. R., and Liddy, E. D. (2016). Emotweet-28: a fine-grained emotion corpus for sentiment analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1149–1156.

Mohammad, S., Bravo-Marquez, F., Salameh, M., and Kiritchenko, S. (2018). Semeval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17.

Mohammad, S. M. (2012). # 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*, pages 246–255. Association for Computational Linguistics.

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement*, pages 201–237. Elsevier.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Quan, C. and Ren, F. (2009). Construction of a blog emotion corpus for chinese emotional expression analysis. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1446–1454. Association for Computational Linguistics.

Seyeditabari, A., Tabari, N., and Zadrozny, W. (2018). Emotion detection in text: a review. *arXiv preprint arXiv:1806.00674*.

Strapparava, C. and Mihalcea, R. (2008). Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560. ACM.