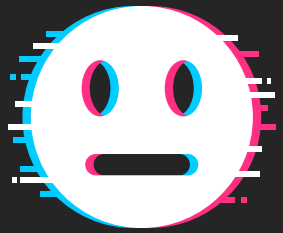
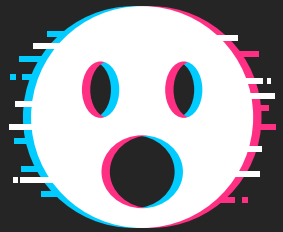
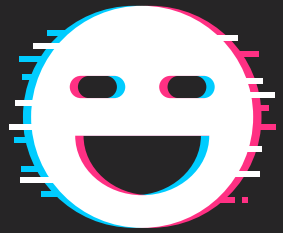







EmoEvent

A Multilingual Emotion Corpus based on different Events

A paper by F.M. Plaza-del-Arco, C. Strapparava,
L.A. Ureña-López, M.T. Martín-Valdivia

Introduction to EmoEvent

EmoEvent is a multilingual **corpus** (English and Spanish) for **emotion detection** sourced from X (formerly Twitter), focusing on tweets about 8 significant **events** from April 2019.

Notre-Dame Cathedral Fire

A structure fire broke out beneath the roof of Notre-Dame Cathedral in Paris.

Fridays for Future

International climate movement started by Greta Thunberg's school strike for climate.

World Book Day

Annual event organized by UNESCO to promote reading, publishing, and copyright.

Spanish General Election

Election of the Spanish parliament.

Venezuela's institutional crisis

A crisis concerning who is the legitimate President of Venezuela.

Game of Thrones

American fantasy drama that aired from April 2011 to May 2019.

La Liga

Spanish men's football championship.

Champions League

Annual club football competition organized by the Union of European Football Associations (UEFA).

Tweets are annotated for seven **emotion-categorical labels** and offensiveness. Emotion types include **Ekman**'s six basic emotions plus the "neutral or other emotions" category.

 anger  sadness  joy  disgust  fear  surprise

Goal: Improving **emotion mining**, a complex area of sentiment analysis that lacks annotated gold standard resources, and providing insight into how English and Spanish speakers express emotions differently.

Related Work

Research in affective computing has primarily focused on classifying text into positive/negative sentiment, with less attention given to **emotion classification**.
There is a lack of datasets fully manually labeled with emotions, especially for languages **other than English**.

EmoBank	Large-scale corpus of English sentences annotated with the dimensional Valence-Arousal-Dominance (VAD) representation format.
ISEAR	76,000 records of emotion provoking text provided by the Swiss Center for Affective Sciences.
Valence and arousal Facebook posts	2,895 Social Media posts rated by two psychologically trained annotators on two separate nine-point scales representing valence and arousal.
Affective Text	News headlines annotated manually by six annotators for six emotions: anger, disgust, fear, joy, sadness, and surprise.
SemEval-2019 EmoContext	Textual dialogues annotated for four classes: happy, sad, anger and others.
Twitter Emotion Corpus	Over 20,000 emotion-labeled tweets automatically labeled using hashtags for six basic emotions: anger, disgust, fear, joy, sadness, and surprise.
EmoTweet-28	Tweets annotated with 28 emotional categories, randomly sampled by topic and user.
SemEval-2018 Affect in Tweets	Dataset for English, Arabic and Spanish tweets annotated for anger, fear, joy, and sadness.
Chinese blog emotion corpus	Manually annotated for eight emotional categories: expectation, joy, love, surprise, anxiety, sorrow, anger and hate.

Corpus Creation

Tweets in English and Spanish were collected via the X (Twitter) API based on trending hashtags for each event, with a focus on capturing a variety of emotions from different event types: entertainment, incidents, politics, global commemoration, global strikes.

A linguistic analysis was performed to select tweets, aiming to create a dataset mainly labeled with emotions.

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

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

The Linguistic Inquiry and Word Count (LIWC) resource was used to extract affective features from tweets. LIWC is a content analysis technique that counts the occurrences of word according to predefined psychological and linguistic categories.

1,000 affective tweets and 200 non-affective tweets for each language and event were selected. The top events where the positive class is prevalent are the same in both languages. English speakers express more emotions in tweets than Spanish speakers for these events.

Tweet Selection

The researchers developed a method to understand the presence of emotions in a collection of tweets. They calculated a score for each class (or category) of emotion, which they referred to as the class's saliency within the collection of tweets.

They defined the coverage of a class in the tweet corpus as the percentage of tweets belonging to that class.

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

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

The researchers then defined the prevalence score of a class in the tweet corpus as the ratio of the coverage of one class to the coverage of another class.

A prevalence score close to 1 indicates a similar distribution of tweets between the two classes in the corpus. A score significantly higher than 1 suggests that the first class is more prevalent in the corpus. Conversely, a score significantly lower than 1 indicates that the second class is more dominant in the corpus.

Data Annotation

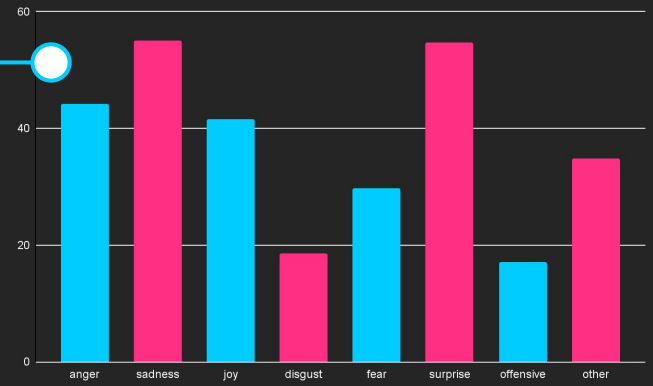
Data was annotated using the **Amazon Mechanical Turk** (MTurk) platform, on which the tasks, known as **Human Intelligence Tasks** (HITs), were published for workers to complete.

Each HIT consisted of two questions: one to label the main **emotion** conveyed by a tweet (anger, fear, sadness, joy, disgust, surprise or others), and the other to determine whether the tweet contains **offensive** language.

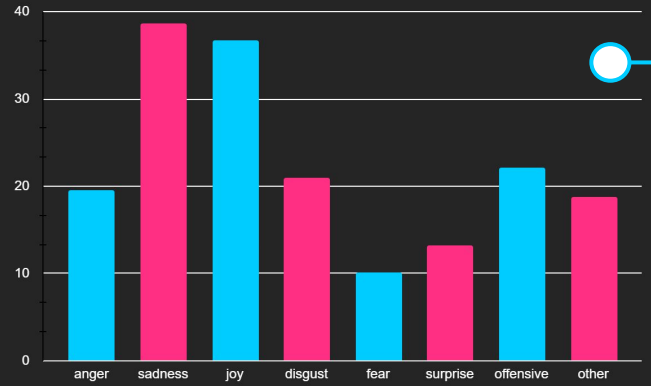
Synonyms were provided for each emotion to aid the workers in the annotation process. Offensive language was defined as containing unacceptable language, including insults, threats, or bad words.

The final label for each tweet was decided based on the agreement of at least **two out of three** annotators. An **inter-annotator agreement** study was conducted to measure the level of agreement among the annotators for each of the eight labels, using the **Cohen's Kappa coefficient**.

Spanish



English



Kappa coefficient for inter-annotator agreement

Corpus Statistics

The dataset contains 8,409 tweets in English and 7,303 in Spanish.

The number of offensive tweets per event in both languages was generally low, with the most offensive tweets 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

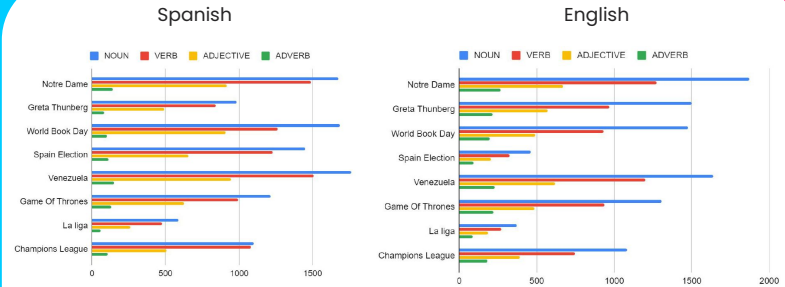
Event	*# of tweets*		*Avg. tweet length*		*# of emojis*		*# of unique hashtags*	
	SP	EN	ES	EN	SP	EN	SP	EN
Notre Dame	1,200	1,200	26.57	26.98	432	242	397	942
Greta Thunberg	630	742	24.91	27.61	279	154	750	1,036
World Book Day	1,200	1,200	23.93	23.83	916	649	827	1,131
Spain Election	1,200	207	20.89	24.67	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

Spanish users tend to use more emojis than English users to express their opinions on different events, while hashtags are more used by English users.

Corpus Statistics

The emotions expressed in the tweets vary by event. For instance, joy is predominant for World Book Day, while anger, disgust, and fear are more common for the Venezuela situation. Sadness was most frequent for the Notre Dame Cathedral Fire disaster, and surprise was more present at entertainment events.

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



Some emotions, such as fear and surprise, are difficult to label due to their association with differing valence (positive or negative depending on the context).

In terms of grammatical labeling, Spanish users tend to use more nouns, verbs, and adjectives to express their emotions, while English users use more adverbs.

Experiments and Results

Pre-processing

The data was **cleaned** and prepared, since online text usually contains **noise** and uninformative parts, which increases the dimensionality of the problem. The tweets were **tokenized** using NLTK TweetTokenizer and all **hashtags** were removed.

Classification

TF-IDF (Term Frequency Inverse Document Frequency) and Support Vector Machine (**SVM**) were used for the classification task.

Evaluation

The performance of the classifier was evaluated using metrics such as **Precision**, **Recall**, **F-score**, and **Accuracy**. **10-fold cross validation** was used to evaluate the machine learning classification approach.

Language	*joy*			*sadness*			*anger*			*fear*			*disgust*			*surprise*			*other*			*macro-avg*			*Acc*
	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	
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

The SVM algorithm performed better in detecting certain emotions like joy, sadness, and others, while it struggled with emotions like anger, fear, disgust, and surprise. This could be due to the complementary nature of these emotions and their lower representation in the tweets.

The model balanced precision and recall differently for different emotions. For instance, Sadness had high precision but moderately high recall. But, for emotions with lower scores, the model struggled both in correctly identifying the emotion and in avoiding false identifications.

The model was more effective on the Spanish dataset compared to the English dataset. This could be due to various factors, including the nature of the dataset, linguistic nuances of each language, or the distribution of training data.

The F1 score showed that the model achieved a reasonable balance between precision and recall for some emotions, but struggled to maintain this balance for others, especially those with lower scores.

The results highlighted the inherent challenges in automated emotion detection, particularly in multilingual contexts. Emotions that are more nuanced or less frequently represented in the training data might be harder for the model to learn and predict accurately.

Critical Insight

During the exploration of the research conducted for EmoEvent, several opportunities were identified to improve the methodologies and results of the conducted research, based on the **conclusions** of the authors and the topics studied during the **course**. Here are some targeted suggestions for refining the work conducted by the authors:

Comparative Model Analysis

As also suggested by the authors, it would be beneficial to conduct experiments comparing the performance of deep architectures (e.g. BERT) against SVM.

Event-Specific Emotion Classification

Classify emotions event-wise using distinct models for each event subset. This allows for understanding of unique emotional expressions in different contexts.Analyze emotional response variance across events to identify patterns and anomalies.

Enhanced Feature Extraction

Improve pre-processing by identifying linguistic nuances like emoticons, character repetitions, and uppercase usage in tweets, assigning them emotional weight.

Custom Lexicon Development

Develop lexicons for specific events by performing frequency analysis of relevant terms and phrases from the dataset, enriched with domain-specific words to effectively capture diverse emotions.

Negation Handling

Implement a negation handling technique, such as appending a "_NEG" suffix to words following a negation, to ensure that the sentiment analysis accurately reflects the tweet's true emotional content.

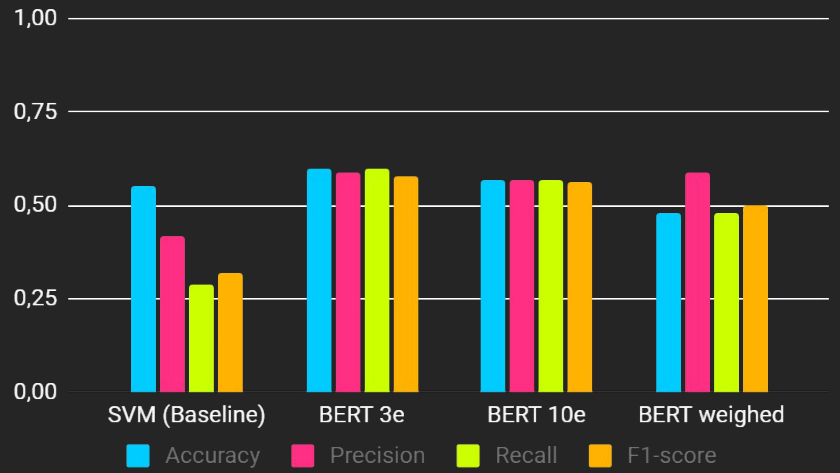
Aspect-Based Sentiment Analysis

Identify key aspects of each event type in the dataset (e.g., characters in TV shows, political figures in elections) and tailor the emotion analysis to these aspects. This approach would involve modifying the annotation guidelines to include aspect-specific emotion labeling.

Comparative Analysis

Building upon the insights and suggestions identified in the previous discussion, since a version of the dataset was made publicly available by the authors, a performance comparison with a deep learning model (**BERT**) was implemented by conducting three experiments with different hyperparameters on the English portion of the dataset.

- **Experiment 1**: Adam optimizer, learning rate of 5e-5, 3 epochs, batch size of 32, Sparse Categorical Cross Entropy loss function
- **Experiment 2**: Same as Experiment 1 but trained for 10 epochs.
- **Experiment 3**: Lower learning rate of 3e-5, 5 epochs and a smaller batch size of 16, class weights and early stopping.



BERT 3e demonstrates the best overall performance with the highest accuracy, precision, recall, and F1-score. It effectively classifies tweets into correct emotional categories and maintains a good balance between precision (minimizing false positives) and recall (minimizing false negatives).

BERT 10e and **BERT weighed** have good precision but slightly lower recall compared to BERT 3e, which might lead to underrepresentation of some emotions in their classifications.

SVM (Baseline) shows the lowest performance across all metrics, indicating it struggles with the complex task of specific emotion classification.