Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 41-51 recibido 23-11-2020 revisado 24-01-2021 aceptado 01-02-2021

NECOS: An annotated corpus to identify constructive news comments in Spanish

NECOS: Un corpus anotado para identificar comentarios constructivos de noticias en español

Pilar López-Úbeda, Flor Miriam Plaza-del-Arco, Manuel Carlos Díaz-Galiano, M. Teresa Martín-Valdivia
Department of Computer Science, Advanced Studies Center in ICT (CEATIC)
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
{plubeda, fmplaza, mcdiaz, maite}@ujaen.es

Abstract: In this paper, we present the NEws and COmments in Spanish (NECOS) corpus, a collection of Spanish comments posted in response to newspaper articles. Following a robust annotation scheme, three annotators labeled the comments as constructive and non-constructive. The articles were published in the newspaper *El Mundo* between April 3rd and April 30th, 2018. The corpus is composed of a total of 10 news articles and 1,419 comments. Three annotators manually labeled NECOS with an average Cohen's kappa of 78.97. Our current focus is the study of constructiveness and the evaluation of the Spanish NECOS corpus. In order to address this goal, we propose a benchmark testing different machine learning systems based on Natural Language Processing: a traditional system and the novel Transformer-based models. Specifically, we compare multilingual models with a monolingual model trained on Spanish in order to highlight the need to create resources trained on a specific language. The monolingual model fine-tuning on NECOS obtain the best result by achieving a macro-average F_1 score of 77.24%.
Keywords: Corpora, constructiveness, Natural Language Processing, Transformer-based models.

Resumen: En este artículo presentamos un corpus de noticias y comentarios en español (NECOS). Estas noticias están publicadas en el periódico El Mundo en un período comprendido entre el 3 de abril y el 30 de abril de 2018. El corpus contiene un total de 10 noticias y 1.419 comentarios. Siguiendo un esquema de anotación, tres anotadores etiquetaron manualmente los comentarios como constructivos y no constructivos obteniendo un promedio de 78,97 usando el coeficiente de kappa de Cohen. En este estudio nos centramos en estudiar la constructividad y hacer la evaluación del corpus NECOS. Para abordar este objetivo, proponemos la experimentación con diferentes sistemas basados en Procesamiento del Lenguaje Natural usando aprendizaje automático: un clasificador tradicional y métodos recientes basados en *Transformers*. Concretamente, comparamos modelos multilingües con un modelo monolingüe entrenado para el español. Con ello, pretendemos demostrar la importancia de crear recursos entrenados para un idioma en particular. El modelo monolingüe evaluado en NECOS obtiene el mejor resutlados alcanzando un 77,24% de *macro-average* F_1.
Palabras clave: Corpora, constructividad, Procesamiento del Lenguage Natural, modelos basados en *Transformer*.

ISSN 1989-7553. DOI 10.26342/2021-66-3

1 Introduction

Worldwide, millions of users have the possibility to easily access the web, publishing and sharing content on a variety of topics. For instance, popular newspapers, with the aim of keeping readers up to date, publish daily news from different areas such as politics, technology and socioeconomics, often receiving in a short period of time a large number of comments from users. Sometimes, this leads to inappropriate online content, as it can involve offensive, hateful, fake or non-constructive comments. Since 2007, The New York Times employed a team of full-time moderators to review all comments submitted on its website and to remove inappropriate content (Etim, 2017). However, reviewing these types of comments manually is a very time-consuming process for moderators because of the large number of comments posted. In order to deal with this issue, in the last years a great interest is growing in the Natural Language Processing (NLP) community in the development of systems capable of analyzing comments automatically(Kolhatkar et al., 2019).

In the NLP literature, most of the studies based on the constructiveness of comments identify which ones provide insight, contribute to the dialogue and encourage a healthy discussion (Kolhatkar and Taboada, 2017b; Kolhatkar and Taboada, 2017a). A number of approaches have been proposed for the identification of constructiveness, from traditional supervised systems applying text-based features (Kolhatkar and Taboada, 2017a; Fujita, Kobayashi, and Okumura, 2019) to the use of latest neural network models that are proving to be successful (Kolhatkar et al., 2020). To train these type of systems, linguistic resources such as corpora are essential. The systems use the corpus to generate a language model that allows the prediction of new data. However, there is a great shortage of datasets annotated for constructiveness of individual comments and most of them are in English. To the best of our knowledge, there is no such resource for Spanish which is the second most spoken language in the world and the third most used language on the Web (Instituto Cervantes, 2018).

In this paper, we help bridge this gap by presenting the first Spanish corpus from newspaper comments in order to identify constructive and non-constructive comments. In addition, we propose a benchmark by applying the most advanced approaches in the scope of NLP and machine learning.

In particular, the contributions of this paper are summarized below:

- We release the first Spanish dataset of constructive and non-constructive comments by collecting a number of comments from the newspaper *El Mundo*[1].
- We analyze the constructive and non-constructive comments of the corpus introducing some linguistic statistics.
- We establish a benchmark for our corpus applying Transformer-based language models.
- We conduct an error analysis in order to understand the capabilities and the drawbacks of our system.

The rest of the paper is structured as follows: In Section 2 we present some previous studies related to constructiveness in comments on newspaper articles. Linguistic statistics related to the NECOS corpus are described in Section 3 presenting some statistics. The different machine learning approaches, the results obtained and the error analysis of the best system are shown in Section 4. Finally, conclusions and future work are presented in Section 5.

2 Related work

Constructive comments were initially defined by Napoles et al. (2017) in terms of ERICs (Engaging, Respectful, and/or Informative Conversations). The dataset used was the Yahoo News Annotated Comments Corpus (YNACC) which contains nearly 140k threads posted on Yahoo News articles in April 2016 (Napoles, Pappu, and Tetreault, 2017). In this study, they compared four approaches to classifying ERICs: a pipeline (Conditional Random Fields and binary classification), a linear classifier with linguistic and social features, an augmented pipeline that incorporates features from the linear model, and a Convolutional Neural Network (CNN). The best result obtained was using the augmented pipeline reaching 0.73 in average F1-score.

[1] `https://www.elmundo.es/`

Kolhatkar and Taboada (2017a) provided a new annotated corpus to the scientific community interested in analyzing constructive comments. They crawled 1,121 comments in English from 10 articles of the Globe and Mail news website covering a variety of subjects. In this study they proposed different training sets and their new corpus as a testing set. The two corpora taken into account in the training set were YNACC and the Argument Extraction Corpus (AEC) (Swanson, Ecker, and Walker, 2015). Finally, they carried out the experimentation using Bidirectional Long short-term memory (BiLSTM) architecture achieving 72.59% accuracy.

Following up on their previous research, Kolhatkar and Taboada (2017b) created a new training set in order to identify constructive comments in the news. In this case, the positive examples included in the training set were obtained from the New York Times (NYT) Picks using the NYT API[2], and for the negative examples they used the negative comments from the YNACC. Support Vector Machines (SVM) with features and a BiLSTM were employed for the experimental results. The combination of features included in SVM obtained the best results reaching a 0.84 of average F1-score.

More recently, Kolhatkar et al. (2020) presented a new corpus named Constructive Comments Corpus (C3) including constructive and non-constructive labels in English. They continue to combine training and test sets from different datasets in order to achieve the desired results. The approaches evaluated on this study were more recent including deep learning models (BiLSTM and CNN) and models based on Transformers (BERT) (Vaswani et al., 2017; Devlin et al., 2019). BERT obtained the best result (0.93 of average F1-score) using partitions for training and a test set using C3. In addition, they provided interesting results according to the features that characterize constructive comments.

Regarding languages other than English and following the previous study, Fujita, Kobayashi, and Okumura (2019) created a new dataset including 100K+ Japanese comments with constructiveness scores in collaboration with Yahoo! News. These scores were based on the number of annotators who labeled a comment as constructive. Since the dataset was composed by numerical ranking scores, they used Normalized Discounted Cumulative Gain (NDCG) as the primary evaluation measure. In order to perform a comparison of results they used different methods such as SVM, rankSVM and SVR.

As previous studies show, most of the corpora annotated with constructiveness are in English. However, for Spanish, as far as we know there are no corpora labeled with this phenomenon. For this reason, we contribute to the scientific NLP community by introducing and releasing the first dataset for constructiveness in Spanish named NECOS[3] (News and Comments in Spanish). In addition, we use our dataset to demonstrate the feasibility of implementing systems to automatically detect constructive comments in newspapers.

3 NECOS corpus

In this section, we present NECOS, a corpus of online news comments enriched with constructiveness annotations.

3.1 Data collection

In order to study constructiveness in news comments, we crawled 1,419 comments from 10 articles from the *El Mundo* newspaper. The news articles that were downloaded are dated between April 3rd and April 30th, 2018. These news articles encourage debate and controversy among users and are related to various topics such as politics, assassinations, current affairs, among others.

We selected 10 news items from the newspaper and downloaded up to 150 random comments from these items, although some selected news articles contained less than 150 comments.

3.2 Annotating constructiveness

Based on previous studies based on constructive comments, we followed the same annotation guide to annotate the NECOS corpus (Kolhatkar et al., 2020). Bellow, we point out some characteristics and attributes of constructive and non-constructive comments:

- Constructive:

[2] `https://developer.nytimes.com/`

[3] NECOS corpus: `https://github.com/plubeda/NECOS`

  - Comments which add something substantial to the conversation and encourage dialogue.
  - Comments which are supported by appropriate evidence.
  - Comments which offer solutions, new perspectives and insights.
  - Comments which provide a personal story or experience.

- Non-constructive:
  - Comments which do not have much content or are unsubstantial.
  - Comments which are not civilized.
  - Comments which do not respect the views and beliefs of others.
  - Comments which express opinions without providing evidence.
  - Sarcastic comments.
  - Provocative comments.

The corpus is annotated in a binary form: label 1 has been assigned to constructive comments and label 0 for non-constructive comments. In order to understand the corpus, different comments are shown in Table 1.

As we can see in examples 1 and 2, the comments are constructive because the authors recommend possible solutions. Comment number 3 is not considered constructive because it provides a personal opinion without adding possible solutions or evidence. The author of comment 4 does not provide evidence and the comment does not contain much content.

Annotators should not evaluate comments from a personal point of view. They can agree or disagree with the comment, but this should not affect their labeling of the corpus.

3.3 Inter-annotator agreement

The annotation of the entire corpus was done by three annotators in an average time of 4-5 minutes per comment. These annotators have followed the previously mentioned annotation guidelines.

In binary classification tasks, Cohen's kappa is often used as a quality measure for data annotations because this metric expresses the level of agreement between two annotators on a classification (Cohen, 1960).

In order to measure the level of agreement among the three annotators, we measure the agreement by pairs of annotators using the kappa coefficient and finally we calculate the average to obtain the final value. Table 2 shows the agreement between each pair of annotators and the percentage of coincidence between them.

On the one hand, in this table we can see that the average kappa is 78.97, which corresponds to a moderate level of interpretation (McHugh, 2012). On the other hand, the percentage of agreement ranges from 89.92 to 91.9 and finally obtains an average of 91.03. Both metrics get acceptable values for annotation and agreement in the NECOS corpus.

It is important to emphasize that the most disagreement in the annotation of comments has been on issues of justice, economy and gender violence, where the annotators disagreed on more than 20% of the comments. In contrast, the highest level of agreement reached by annotators has been on political issues. To clarify the disagreement with further analysis, Table 3 shows the number of disagreement annotations for each news item in the NECOS corpus.

3.4 Corpus analysis

In this section we highlight some statistics regarding the NECOS corpus. These statistics refer to the number of comments, number of words and number of sentences in each annotated label.

Table 4 shows some basic linguistic statistics of the corpus. According to the number of comments of each class, we found that the corpus is unbalanced since it contains 985 non-constructive and 434 constructive comments. As the corpus contains more non-constructive comments, the number of total sentences is also higher in this class. On the other hand, although there are more non-constructive comments, the size of the vocabulary is almost the same. Finally, we want to highlight that constructive comments contain more words and sentences than non-constructive comments, which means that people in constructive comments use more words to express their opinion.

Finally, we have analyzed the same linguistic statistics according to the 10 news items downloaded for the NECOS corpus (see the 10 news items in Table 5). These statistics have been shown in Table 6. As we can see, in all the news items there are more constructive than non-constructive comments except in the item number 6. In addition, the constructive comments contain a higher average number of tokens than the

	Comment	Label
1	*Espero que no se aprueben los presupuestos, que se convoquen elecciones y gane ciudadanos y acabe con el cupo y el PP. Asco de partido apoyando a los fascistas nacionalistas vascos por nada.* I hope that the budgets will not be approved, that elections will be called and *Ciudadanos* will win and that the quota and the PP will be ended. A disgusting party that supports the Basque nationalist fascists for nothing.	1
2	*Trump cometió una equivocación porque debería haber dejado que el ejército soltara toda su fuerza para que de una vez el mundo sepa que con los EEUU no se juega.* Trump made a mistake because he should have let the army release its full force so that the world would finally know that the U.S. is not to be trifled with.	1
3	*España como criadero de incompetentes y chorizos. Menos mal que soy catalán.* Spain is a breeding ground for incompetents and crooks. Luckily I am Catalan.	0
4	*Eso es una compra de votos. Es delito. Malversación de fondos públicos.* That's vote buying. It is a crime. Embezzlement of public funds.	0

Table 1: Examples of comments tagged in the NECOS corpus, along with English translations.

	Cohen's kappa	Agreement (%)
Annotator 1 and 2	80.90	91.90
Annotator 2 and 3	76.49	89.92
Annotator 1 and 3	79.53	91.26
Average	78.97	91.03

Table 2: Inter-annotation agreement in the NECOS corpus.

non-constructive ones, which means an increase in the size of the vocabulary in some cases.

4 Experiments and results

In this section we introduce a benchmark for the constructive corpus. In particular, we propose two different approaches: transformer-based methods and a baseline system based on traditional machine learning.

4.1 Constructive classification

First of all, the pre-processing is a fundamental step in NLP. In this process, we prepare and clean up the text before including it in the classification systems. This step is one of the most important because it can help to improve the performance of the classifier and speed up the classification process. The pre-processing addressed in the NECOS corpus was carried out as follows:

- Remove references: references to other comments have been removed in the comments.
- Remove URLs: the existing URLs in the comments have been replaced by the token *URL*.
- Lowercase: the comments have been converted to lowercase.

After performing the pre-processing step, we carry out the experiments on NECOS corpus. In all of our experiments, we use 10-fold cross validation to evaluate the machine learning classification systems.

The models we have chosen to test the effectiveness of NECOS are described below:

News item	# of comments	# of comments disagreeing	Disagreement (%)
1	149	9	6.04
2	148	33	22.30
3	150	13	8.67
4	149	36	24.16
5	149	10	6.71
6	142	12	8.45
7	150	8	5.33
8	128	1	0.78
9	150	26	17.33
10	104	22	21.15
Total	1,419	170	11.98%

Table 3: Analysis of disagreement comments for each news item in the NECOS corpus (see the 10 news items in Table 5).

	Constructive	Non-constructive
Number of comments	434	985
Vocabulary size	6,734	6,367
Avg. of tokens in comments	84.89	32.73
Number of sentences in comments	1,552	1,997
Avg. of sentence in comments	3.57	2.03

Table 4: Dataset analysis.

- **SVM** is a set of supervised learning methods used for classification, regression and outlier detection. (Pedregosa et al., 2011). A number of studies have reported that this classifier is one of the most accurate methods for text classification (Puri and Singh, 2019; Chatterjee, Jose, and Datta, 2019). Therefore, we use this classifier as our baseline for the constructive corpus. For text representation, we use the Term frequency-inverse document frequency (Tf-idf). For the classifier, we use the default configuration provided in the scikit-learn module from Python.
- **Multilingual BERT (aka mBERT)** (Devlin et al., 2019). mBERT is a multilingual model that follows the same model architecture and training procedure as BERT, except with data from Wikipedia in 104 languages. In mBERT, the WordPiece modeling strategy allows the model to share embeddings across languages. In particular, for this study we chose the BERT-Base, Multilingual Cased checkpoint[4].
- **XLM** is a cross-lingual language model pre-trained, which uses a pre-processing technique and a dual language training mechanism with BERT in order to learn the relations between words in several languages. In this study we use the XLM model trained on 100 languages (XLM-100) (Lample and Conneau, 2019).
- **XLM-Roberta** proposed by Conneau et al. (2019). It is based on Facebook's RoBERTa model released in 2019 (Liu et al., 2019). XLM-Roberta is a large multilingual language model, trained from CommonCraws data on 100 different languages.
- **BETO** is a BERT-based language model pre-trained specifically on Spanish data and is similar in size to a BERT model. It has 12 self-attention layers with 16 attention-heads each (Vaswani et al., 2017), using 1024 as hidden size. The model is trained from different sources including Wikipedia and all of the sources of the OPUS Project (Aulamo and Tiedemann, 2019). Specif-

[4] `https://github.com/huggingface/transformers`

	News headline
#1	*Presupuestos 2018: el Gobierno ofrece un 32% más de inversión en el País Vasco tras mejorarle el Cupo.* Budgets 2018: the Government offers 32% more investment in the Basque Country after improving its quota.
#2	*El juez desoye a la Fiscalía y deja libre al informático Falciani, detenido ayer en Madrid a petición de Suiza.* The judge disregards the Prosecutor's Office and releases Falciani, who was arrested yesterday in Madrid at the request of Switzerland.
#3	*El Rey apoya en Barcelona a los jueces como "garantía de los derechos y el respeto a la ley".* The King supports judges in Barcelona as a "guarantee of rights and respect for the law"
#4	*Cristina Cifuentes: "No he cometido ilegalidad, no he mentido y no voy a dimitir".* Cristina Cifuentes: "I have not committed any illegality, I have not lied and I am not going to resign".
#5	Hombres víctimas de la violencia machista, el otro eslabón del maltrato a la mujer. Male victims of male violence, the other link in the chain of abuse of women.
#6	*Trump planeaba un ataque de mayor envergadura pero lo limitó para evitar una confrontación con Rusia.* Trump had planned a larger attack but limited it to avoid a confrontation with Russia.
#7	*El sargento agredido en Alsasua: "Algunos jaleaban y había bastantes móviles grabando".* The sergeant assaulted in Alsasua: "Some people were cheering and there were a lot of mobile phones recording".
#8	*Apagón y silencio absoluto en Podemos ante la gravedad de la crisis interna con Bescansa y Errejón.* Blackout and absolute silence in Podemos in the face of the seriousness of the internal crisis with Bescansa and Errejón.
#9	*La Policía requisa camisetas amarillas a la entrada del Wanda Metropolitano.* Police seize yellow T-shirts at the entrance to the Wanda Metropolitano.
#10	*Detenido un hombre por matar a su ex pareja de una paliza en Burgos.* Man arrested for beating ex-partner to death in Burgos.

Table 5: News headlines from the NECOS corpus.

	Constructive					**Non-constructive**				
	# Comments	Vocab. size	Avg. tokens	# Sents	Avg. sent	# Comments	Vocab size	Avg. tokens	# Sents	Avg. sent
#1	33	1005	87.73	146	4.42	116	1239	35.73	274	2.36
#2	46	1167	76.54	143	3.11	102	958	28.39	196	1.92
#3	33	929	81.03	118	3.58	117	1144	28.22	224	1.91
#4	38	1125	92.61	128	3.37	111	1206	31.15	207	1.86
#5	80	1861	88.31	289	3.61	69	922	35.09	129	1.87
#6	67	2055	103.21	294	4.39	75	1141	39.17	165	2.2
#7	42	1111	77.4	124	2.95	108	1183	31.52	225	2.08
#8	7	276	78.71	28	4	121	1403	36.94	242	2
#9	48	1111	73.08	155	3.23	102	1047	30.14	210	2.06
#10	40	953	73.58	127	3.18	64	731	33.25	125	1.95

Table 6: Dataset analysis by news topic.

ically, in this study we use the BETO cased checkpoint[5].

For all the pre-trained language models we use the same hyperparameters. Specifically, the models are fine-tuning using 2 epochs, a batch size of 16 and a learning rate of 0.0001. We use 256 words as max length.

4.2 Results

In this section we report and discuss the performance of the tested systems on the Spanish constructive classification task. In order to evaluate and compare the results obtained by our systems, we use the usual metrics in

[5] https://github.com/dccuchile/beto

text classification, called precision (P), recall (R), F-score (F_1) and macro-average. The metrics have been computed as follows:

$$P(c) = \frac{TP}{TP + FP} \quad (1)$$

$$R(c) = \frac{TP}{TP + FN} \quad (2)$$

where c is equal to the class (0, 1), TP = True Positive, FP = False Positive and FN = False Negative.

$$F1 = \frac{2 * P * R}{P + R} \quad (3)$$

Table 7 shows the results achieved after experimenting with all the systems. In first place, it can be seen the performance of our baseline model using the SVM classifier. It obtains a macro-avg F1 score of 71.09% which is more than acceptable for a binary classification task. If we focus on the performance of each class, it is worth mentioning that the system is able to detect non-constructive comments more accurately than constructive ones.

Related to the pre-trained language models, XLM-Roberta and XLM, they do not perform as well as SVM. However, mBERT and BETO, which are based on the BERT model, outperform our baseline by a substantial margin. Specifically, BETO is the most accurate system outperforming the rest of the models by achieving a macro-avg F_1 score of 77.24%. We presume that BETO performs significantly better because it was trained on a Spanish corpus. For this reason, we point out the importance of generating resources not only for English but also for other languages such as Spanish.

If we observe the F_1 score of each system, it should be noted that all the systems achieve a better performance in the non-constructive class. On the contrary, there is a significant difference compared to the F_1 score of the constructive class. This could be because there is a higher proportion of non-constructive comments and therefore the system learns better to identify that class.

4.3 Error analysis

The goal of this section is to identify the weaknesses of our systems by conducting an error analysis of the Transformer-based models. For this purpose, we first analyze the errors conducted by the three systems and then focus in more detail on the errors produced by the best system BETO.

The results showing the numbers of wrongly assigned labels for each system are summarized in Table 8. All four models predicted the same wrong labels 91 times. As can be seen, XLM-based models predict more false negatives than BERT-based models. The common errors are highly biased towards false negatives. We have observed that the average number of tokens in the constructive comments mislabelled at the same time by the systems is much lower than the average shown in Table 4, namely 41.12 versus 84.89. Therefore, this may be one reason why systems do not correctly classify this type of comments.

In order to focus on the best system, we performed a manual analysis of some of the mislabelled comments to find the main reasons why BETO does not classify some comments correctly. Table 9 presents some examples of comments incorrectly classified by our system. In particular, there are 4 comments, two false negatives and two false positives. On the one hand, if we look at the false negatives (examples 1 and 2), the reason they are predicted as non-constructive is because BETO was not able to identify the possible solutions that the authors offer in the comments. On the other hand, if we focus on false positives, we consider that examples 3 and 4 are mislabeled by BETO because it identifies an argument, but in this case the author does not add new perspectives or substantial justifications.

Considering the number of false positives and false negatives, and given that we performed 10-fold cross validation, we decided to perform an analysis to determine the number of errors in each partition. Figure 1 shows the percentage of false positives (blue color), false negatives (red color) for each fold of the cross validation. In most of the partitions, we can observe that there are more false negatives than false positives. This problem may be due to the fact that the number of comments that exist in the corpus labeled as constructive is greater than the number of non-constructive. These results are also reflected in Section 4.2 showing a lower value of F_1 in the constructive category.

System	Non-constructive			Constructive			Macro-avg		
	P	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
SVM	80.13	90.05	84.80	68.59	49.31	57.37	74.36	69.68	71.09
XLM-Roberta	75.91	93.58	82.62	34.63	28.04	28.76	55.27	60.81	55.69
XLM	76.62	**93.98**	83.65	43.36	33.85	37.06	59.99	63.62	60.36
mBERT	85.55	83.85	84.27	65.60	66.94	64.98	75.58	75.40	74.61
BETO	**86.70**	85.28	**85.59**	**68.48**	**71.80**	**68.90**	**77.59**	**78.54**	**77.24**

Table 7: Results obtained by different systems on the constructive corpus (10-fold cross validation).

System	Errors	Predicted 1	Predicted 0
XLM-Roberta	378	64	314
XLM	351	57	294
mBERT	294	156	138
BETO	268	143	125
All (in common)	91	6	85

Table 8: Number of instances mislabeled by each system, broken down by wrongly assigned label.

5 Conclusion

We have described the News and Comments in Spanish (NECOS) corpus, a resource for exploring constructiveness in news comments from the newspaper *El Mundo*. Our raw corpus comprises 10 news articles and 1,419 comment threads in response to these articles in April 2018.

Since our resource is manually annotated, we have noticed the difficulty of labeling constructive comments regardless of the annotator personal opinion. Despite this challenge, we have achieved a strong agreement among the annotators, reaching a Cohen's kappa of 78.97. While annotating, on the one hand, we learned that most of the constructive comments are based on offering solutions, new perspectives and insights, and they also provide a personal experience. On the other hand, we noticed that a great amount of non-constructive comments express opinions without providing evidence.

In order to check the effectiveness of our corpus and established a benchmark, we carry out several experiments based on machine learning approaches. Specifically, we use a traditional system (SVM) selecting it as a baseline to compare its results with the latest approaches in NLP including Transformer-based models. Given our results, we conclude that the application of text mining classification systems is a valuable tool for detecting constructiveness in comments from newspapers. Specifically, the best results are obtained using a monolingual Transformer-based model called BETO. With this model we achieved a macro-F_1 score of 77.24 which proves the importance of developing resources for a specific language, in this case Spanish.

A number of research avenues are planned for this corpus. We are interested in studying the relation between constructiveness and toxic language. We believe that this feature could be useful to help the model to easily detect constructive and non-constructive comments. In addition, sentiment analysis could be applied to study the influence of emotions in constructive comments. Finally, we plan to integrate these features into our best system with the purpose of further improving the results.

Acknowledgements

This work has been partially supported by a grant from European Regional Development Fund (ERDF), LIVING-LANG project [RTI2018-094653-B-C21], and the Ministry of Science, Innovation and Universities (scholarship [FPI-PRE2019-089310]) from the Spanish Government.

	Comment	Label	BETO
#1	*La violencia debe ser igualmente reprobable siempre, sea cual sea el sexo del agresor y del agredido. El trato preferente presente es tan aberrante como la discriminación pasada.* Violence must be equally reprehensible at all times, regardless of the sex of the aggressor and of the victim. The present preferential treatment is as bad as past discrimination.	1	0
#2	*Ánimo, pide ayuda médica y verás como dentro de un tiempo verás el suicidio como un disparate, pero lo importante es que cuando uno mismo no es capaz de salir del pozo, pida ayuda y la acepte. Todo es cuestión de paciencia y tiempo.* Come on, ask for medical help and you will see that in a while you will see suicide as nonsense, but the important thing is that when you are not able to get out of the well yourself, ask for help and accept it. It's all a matter of patience and time.	1	0
#3	*Estos sucesos son abominables y su erradicación muy complicada, no vamos a evitar que nazcan criminales en potencia, si al menos este salvaje tuviera el castigo que merece...* These events are abominable and their eradication very complicated, we are not going to prevent potential criminals from being born, if only this savage had the punishment he deserves...	0	1
#4	*En humoristas, titiriteros, raperos, actores, presentadores, activistas, que estén imputados o en la cárcel por expresar sus ideas. Y no se confunda, yo no apoyo lo que dicen, defiendo que puedan decirlo, (casi todo).* In comedians, puppeteers, rappers, actors, presenters, activists, who are accused or in prison for expressing their ideas. And don't be confused, I don't support what they say, I defend their right to say it, (almost everything).	0	1

Table 9: Examples of mislabeled comments using BETO model, along with English translations.

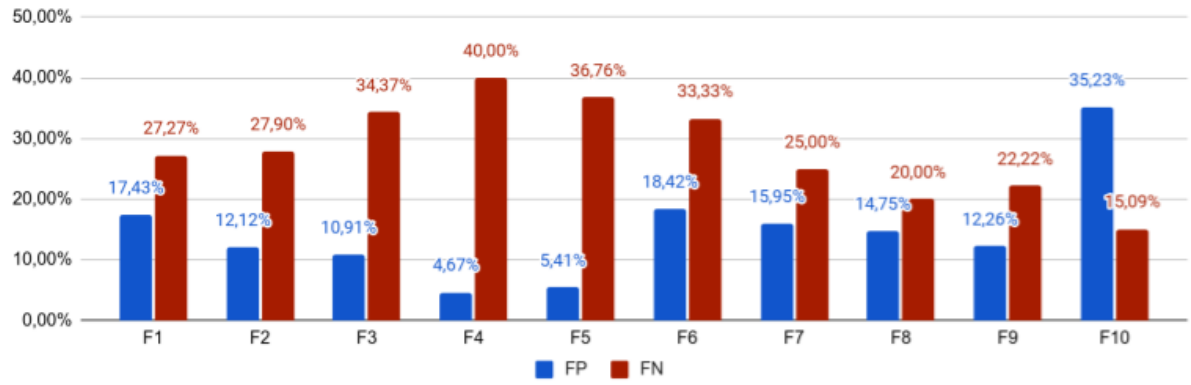

Figure 1: Percentage of false positives and false negatives for each 10-fold cross validation. F: fold.

References

Aulamo, M. and J. Tiedemann. 2019. The OPUS resource repository: An open package for creating parallel corpora and machine translation services. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 389–394, Turku, Finland, September–October. Linköping University Electronic Press.

Chatterjee, S., P. G. Jose, and D. Datta.

2019. Text classification using svm enhanced by multithreading and cuda. *International Journal of Modern Education & Computer Science*, 11(1).

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Etim, B. 2017. The times sharply increases articles open for comments, using google's technology. *The New York Times*, 13.

Fujita, S., H. Kobayashi, and M. Okumura. 2019. Dataset creation for ranking constructive news comments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2619–2626.

Instituto Cervantes. 2018. El español: una lengua viva. `https://cvc.cervantes.es/lengua/espanol_lengua_viva/pdf/espanol_lengua_viva_2018.pdf`.

Kolhatkar, V. and M. Taboada. 2017a. Constructive language in news comments. In *Proceedings of the First Workshop on Abusive Language Online*, pages 11–17.

Kolhatkar, V. and M. Taboada. 2017b. Using new york times picks to identify constructive comments. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 100–105.

Kolhatkar, V., N. Thain, J. Sorensen, L. Dixon, and M. Taboada. 2020. Classifying constructive comments. *arXiv preprint arXiv:2004.05476*.

Kolhatkar, V., H. Wu, L. Cavasso, E. Francis, K. Shukla, and M. Taboada. 2019. The sfu opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, pages 1–36.

Lample, G. and A. Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

McHugh, M. L. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282.

Napoles, C., A. Pappu, and J. Tetreault. 2017. Automatically identifying good conversations online (yes, they do exist!). In *Eleventh International AAAI Conference on Web and Social Media*.

Napoles, C., J. Tetreault, A. Pappu, E. Rosato, and B. Provenzale. 2017. Finding good conversations online: The yahoo news annotated comments corpus. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 13–23.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Puri, S. and S. P. Singh. 2019. An efficient hindi text classification model using svm. In *Computing and Network Sustainability*. Springer, pages 227–237.

Swanson, R., B. Ecker, and M. Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue*, pages 217–226.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.