Noname manuscript No.
(will be inserted by the editor)

Insert your title here

Do you have a subtitle?
If so, write it here

Bobak Farzin · Second Author

Received: 15 June 2019 / Accepted: date

Abstract Insert your abstract here. Include keywords, PACS and mathematical subject classification numbers as needed.

Keywords First keyword · Second keyword · More

1 Introduction

Included citations as necessary: [Jr(2006)]

2 Task and Dataset Description

The *Humor Analysis based on Humor Annotation (HAHA)* competition asked for analysis of two tasks in the Spanish language based on a coprpus of publicly collected data [Castro et al.(2018)Castro, Chiruzzo, Rosá, Garat, and Moncecchi]:

- **Task1: Humor Detection**:Detemine if a tweet is humorous. System ranking is based on F1 score which will balance precision and accuracy.
- **Task2: Funniness Score**:If humorous, what is the average humor rating of the tweet. System ranking is based on RMSE.

HAHA dataset includes labeled data for 24,000 tweets and a test set of 6,000 tweets (80%/20% train/test split.) Each record includes the raw tweet text

F. Author
first address
Tel.: +123-45-678910
Fax: +123-45-678910
E-mail: bfarzin@gmail.com

S. Author
second address

(including accents and emoticons) and a True/False labels as well as a "Funninness Score" that is the average of the 1 to 5 start votes cast. Examples and data can be found on the CodaLab competition webpage[1].

3 System Description

We follow the method of ULMFiT [Howard and Ruder(2018)] and train a language model based on a larger corpus.

Since we are working with Twitter data, we collected 475,143 tweets in the spanish lanugage using tweepy [?]. Text with citations [?] and [?].

3.1 Data and Cleanup

Using sentencepiece [Kudo and Richardson(2018)] to parte into sub-word units,

Clean by removing the new-line character `\n` in each of the tweets and replacing with `xxnl`.

Use standard pre-processing rules (show example of cleaned up text):

Original:

```
Saber, entender y estar convencides que la frase
#LaESILaDefendemosEntreTodes es nuestra linea es nuestro eje.
#AlertaESI!!!!
Vamos por mas!!! e invitamos a todas aquellas personas que quieran
se parte.
```

Cleaned up:

```
xxbos saber , entender y estar convencides que la frase
# laesiladefendemosentretodes es nuestra linea es nuestro eje.
xxnl  # alertaesi xxrep 4 ! xxnl vamos por mas ! ! ! e invitamos a
todas aquellas personas que quieran se parte.
```

oversample minority class to balance for better training using SMOTE [Chawla et al.(2002)Chawla, Bowyer, Hall, and Kegelmeyer]

Paragraph headings Use paragraph headings as needed.

$$a^2 + b^2 = c^2 \tag{1}$$

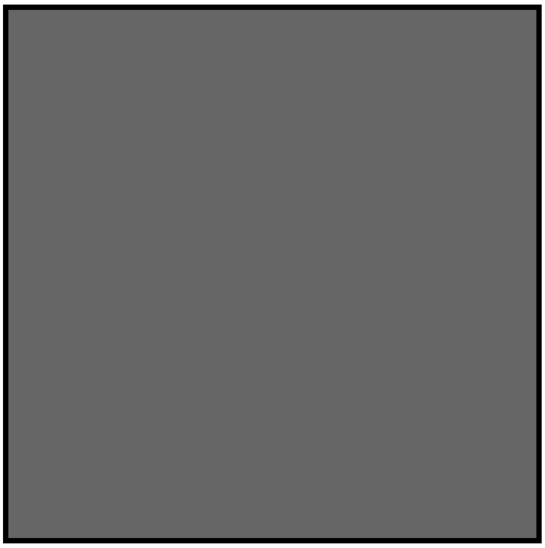

Fig. 1 Please write your figure caption here

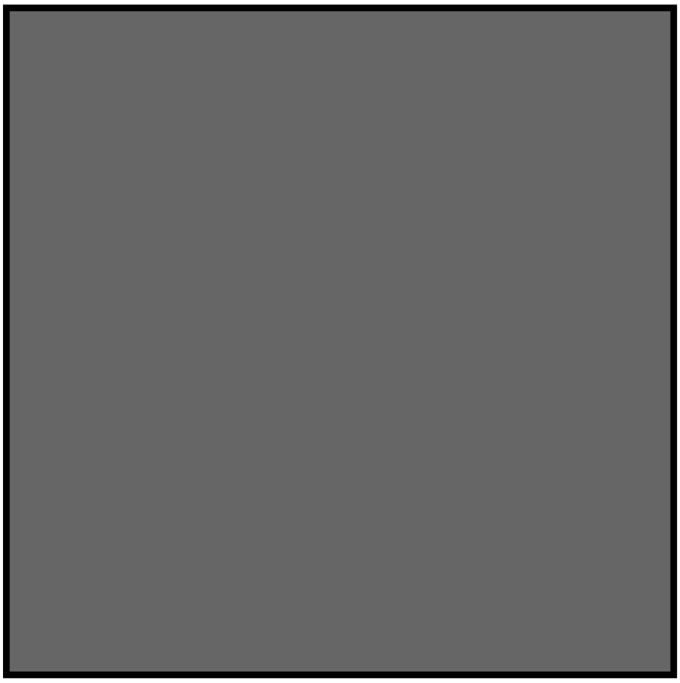

Fig. 2 Please write your figure caption here

Table 1 Please write your table caption here

first	second	third	fourth
number	number	number	
number	number	number	

4 Experiments and Results

5 Conclusion

Acknowlegements

References

[Castro et al.(2018)Castro, Chiruzzo, Rosá, Garat, and Moncecchi] Castro S, Chiruzzo L, Rosá A, Garat D, Moncecchi G (2018) A crowd-annotated spanish corpus for humor analysis. In: Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media, pp 7–11

[Chawla et al.(2002)Chawla, Bowyer, Hall, and Kegelmeyer] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: Synthetic minority over-sampling technique. J Artif Int Res 16(1):321–357, URL http://dl.acm.org/citation.cfm?id=1622407.1622416

[Howard and Ruder(2018)] Howard J, Ruder S (2018) Fine-tuned language models for text classification. CoRR abs/1801.06146, URL http://arxiv.org/abs/1801.06146, 1801.06146

[Jr(2006)] Jr N (2006) My article

[Kudo and Richardson(2018)] Kudo T, Richardson J (2018) Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. CoRR abs/1808.06226, URL http://arxiv.org/abs/1808.06226, 1808.06226

[1] http://competitions.codalab.org/competitions/22194/