A Twitter Political Corpus of the 2019 10N Spanish Election

Javier Sánchez-Junquera[1][0000−0002−0845−9532], Simone Paolo Ponzetto[2], and Paolo Rosso[1]

[1] Universitat Politècnica de València, Spain
jjsjunquera@gmail.com
[2] Data and Web Science Group, University of Mannheim, Germany

Abstract. We present a corpus of Spanish tweets of 15 Twitter accounts of politicians of the main five parties (PSOE, PP, Cs, UP and VOX) covering the campaign of the Spanish election of 10th November 2019 (10N Spanish Election). We perform a semi-automatic annotation of domain-specific topics using a mixture of keyword-based and supervised techniques. In this preliminary study we extracted the tweets of few politicians of each party with the aim to analyse their official communication strategy. Moreover, we analyse sentiments and emotions employed in the tweets. Although the limited size of the Twitter corpus due to the very short time span, we hope to provide with some first insights on the communication dynamics of social network accounts of these five Spanish political parties.

Keywords: Twitter · political text analysis · topic detection · sentiment and emotion analysis

1 Introduction

In recent years, automated text analysis has become central for work in social and political science that relies on a data-driven perspective . Political scientists, for instance, have used text for a wide range of problems, including inferring policy positions of actors [6], and detecting topics [13], to name a few. At the same time, researchers in Natural Language Processing (NLP) have addressed related tasks such as election prediction [11], stance detection towards legislative proposals [16], predicting roll calls [5], measuring agreement in electoral manifestos [8], and policy preference labelling [1] from a different, yet complementary perspective. Recent attempts to bring these two communities closer have focused on shared evaluation exercises [10] as well as bringing together the body of the scholarly literature of the two communities [4]. The effects of these two strands of research coming together can be seen in political scientists making use and leveraging major advances in NLP from the past years [12].

The contributions of this paper are the following ones: (i) we introduce a corpus of tweets from all major Spanish political parties during the autumn

2019 election; (ii) we present details on the semi-automated topic and sentiment/emotion annotation process; and (iii) we provide a preliminary qualitative analysis of the dataset over different addressed topics of the election campaign. Building this preliminary resource of Spanish political tweets, we aim at providing a first reference corpus of Spanish tweets in order to foster further research in political text analysis and forecasting with Twitter in languages other than English.

In the rest of the paper we will describe how each tweet was annotated with topic information together with sentiments and emotions. Moreover, we will illustrate the preliminary experiments we carried out on topic detection. Finally, we will present some insight about sentiment and emotion topic-related analyses.

2 Related Works

Twitter has been used as a source of texts for different NLP tasks like sentiment analysis [9, 3]. One work that is very related to our study is [7]. They collected a dataset in English for topic identification and sentiment analysis. The authors used distant supervision for training, in which topic-related keywords were used to first obtain a collection of positive examples for the topic identification. Their results show that the obtained examples could serve as a training set for classifying unlabelled instances more effectively than using only the keywords as the topic predictors. However, during our corpus development we noticed that keyword-based retrieval can produce noisy data, maybe because of the content and the topics of our tweets, and we then used a combination of both a keyword-based and a supervised approach.

3 Political Tweets in the 10N Spanish Election

In this paper, we focus on the Spanish election of November 10th, 2019 (10N Spanish Election, hereafter). For this, we analyse tweets between the short time span of October 10, 2019, and November 12, 2019. We focus on the tweets from 15 representative profiles of the five most important political parties (Table 1)[3]: i.e., Unidas Podemos (UP); Ciudadanos (Cs); Partido Socialista Obrero Español (PSOE); Partido Popular (PP); and VOX.

3.1 Topic Identification

Topic categories. We first describe how we detect the topic of the tweets on the basis of a keyword-based and supervised approach. In the context of the 10N Spanish Election, we focused on the following topics that were mentioned in the political manifestos of the five main Spanish parties: *Immigration, Catalonia, Economy (and Employment), Education (together with Culture and Research),*

[3] The dataset is available at https://github.com/jjsjunquera/10N-Spanish-Election.

Table 1: Number of tweets of the five political parties. For each party, we use its official Twitter account, its leader, and the female politician that took part in the 7N TV debate.

Parties	**The main profiles**	**Tweets**
UP	@ahorapodemos, @Irene_Montero_, @Pablo_Iglesias_	671
Cs	@CiudadanosCs, @InesArrimadas, @Albert_Rivera	789
PSOE	@PSOE, @mjmonteroc, @sanchezcastejon	527
PP	@populares, @anapastorjulian, @pablocasado_	684
Vox	@vox_es, @monasterior, @santi_abascal	749
Total		3582

Table 2: Total number of labelled tweets: the training set (i.e., manually annotated, and using keywords), and using automatic annotation. The last column has the total number of labelled tweets considering the training set and the classifier results.

Topic	**Manual annotated**	**Keyword annotated**	**Automatically annotated**	**Total annotated**
Catalonia	115	130	370	615
Economy	71	39	506	616
Education	2	19	23	44
Feminism	10	52	82	144
Healthcare	4	12	7	23
Historical Memory	12	16	30	58
Immigration	9	16	36	61
Other	541	153	1037	1731
Pensions	1	24	55	80
Total	765	461	2146	3372

Feminism, Historical Memory, and Healthcare. We additionally include a category label *Other* for the tweets that talk about any other topic.

Manual topic annotation. We first manually annotate 1,000 randomly sampled tweets using our topic labels.

Table 2 summarizes the label distribution across all parties. After removing the noisy tweets, we are left with only 765 posts. Many tweets in our corpus are not related to any of the topics of interest, and were assigned to the *Other* category. Moreover, during the annotation, we noticed in the manifestos of the five parties little information about topics such as research, corruption, renewable energy, and climate change.

Keyword-based topic detection. Due to the manual annotation is time consuming, we complement it by using topic-related keywords to collect tweets about each topic. We ranked the words appearing in the sections corresponding to the topics of interest with the highest Pointwise Mutual Information (PMI). PMI makes it possible to select the most relevant words for each topic, and is computed as: $PMI(T, w) = \log \frac{p(T,w)}{p(T)p(w)}$. Where $p(T, w)$ is the probability of a word to appear in a topic, $p(T)$ is the probability of a topic (we assume the topic distribution to be uniform), and $p(w)$ is the probability of w. For each topic, we collect the top-10 highest ranked keywords and manually filter incorrect ones (Table 3).

Table 3: Keywords used for collecting training data for topic identification.

Topic	Keywords
Catalonia	*autonómica; cataluña; civil*
Economy	*bienestar; discapacidad; energía; fiscalidad; impuesto; innovación; inversión; tecnológico*
Education	*cultura; cultural; educación; lenguas; mecenazo*
Feminism	*conciliación; familia; machismo; madres; discriminación; mujeres; sexual; violencia*
Healthcare	*infantil; sanitario; salud; sanidad; sanitaria; universal*
Historical Memory	*historia; memoria; reparación; víctimas*
Immigration	*ceuta; extranjeros; inmigrantes; ilegalmente*
Pension	*pensiones; toledo*

Supervised learning of topics. For each topic, we collect all tweets in our corpus in which at least one of its keywords appears. All retrieved tweets are then manually checked to ensure that the annotated tweets have a ground-truth.

Inspired by the work of [7], we use the topic-related keywords to obtain a collection of “positive” examples to be used as a training set for a supervised classifier. However, in our dataset, we noticed that keyword-based retrieval can produce much noisy data. Therefore, the keyword-based collected tweets are manually checked before training the classifier.

While our solution still requires the mentioned manual checking, the advantage of using keywords is that the labelling is more focused on tweets that are likely to be in one of the topics of interest, thus reducing the annotation effort associated with tweets from the *Other* category.

Table 2 summarizes in the second and third columns the number of tweets that we used as a training set. The second column represents the results after manually evaluating the tweets labelled by using the keywords. It is interesting that the annotated data reveal most attention towards some topics such as *Catalonia*, *Feminism* and *Economy.* Finally, the dataset used for training is composed of all the labelled tweets. To avoid bias towards the most populated categories we reduce their number of examples to 100 for training, for which we balance the presence of manually annotated and keyword-based annotated tweets.

We employ a SVM [4] to classifiy the still unlabeled tweets and leave-one-out cross-validation because of the small size of the corpus. We represent the tweets with unigrams, bigrams and trigrams, and use the *tf-idf* weighting scheme after removing the n-grams occurring only once.

Evaluation of topic detection. Table 4a shows the standard precision, recall, and F_1 scores. Table 2 shows in the fourth column the number of tweets annotated using our supervised model. The last column shows instead the total of labelled tweets for each of the topics – i.e., the overall number of labelled tweets obtained by combining manual, keyword-based annotations with the SVM classifier. We break down the numbers of these overall annotated tweets per party in Table 4b. The topic distributions seem to suggest that each party is biased towards specific topics. For instance, *Immigration* seems to be almost only men-

[4] We used the implementation from *sklearn* using default parameter values for with a linear kernel.

Table 4: Results on topic classification the total number of labelled tweets.

(a) Results on topic classification.

Topic	Precision	Recall	F1-score
Catalonia	0.72	0.86	0.78
Economy	0.56	0.7	0.62
Education	0.83	0.48	0.61
Feminism	0.8	0.73	0.77
Healthcare	1	0.38	0.55
Historical Memory	0.82	0.5	0.62
Immigration	0.92	0.44	0.59
Other	0.56	0.6	0.58
Pensions	0.85	0.68	0.76
macro avg.	0.78	0.6	0.65

(b) Total of labelled tweets.

Topic	UP	Cs	PP	PSOE	VOX
Catalonia	40	198	110	50	72
Economy	114	117	203	84	88
Education	12	12	11	5	4
Feminism	44	30	8	29	31
Healthcare	10	2	3	6	2
Historical Memory	25	7	2	8	16
Immigration	4	1	-	7	49
Other	258	262	200	174	243
Pensions	17	2	14	37	10

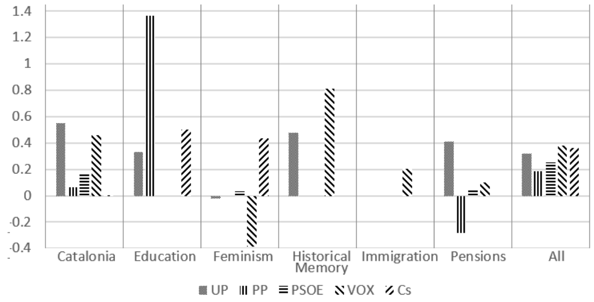

Fig. 1: Expressed sentiment for each topic and party.

tioned by VOX, whereas parties like PP and Cs are mainly focused on *Catalonia* and *Economy*.

3.2 Sentiment analysis

We next analyse the sentiment expressed by the parties about each topic. For this, we use SentiStrength to estimate the sentiment in tweets since it has been effectively used in short informal texts [15]. We compute a single scale with values from -4 (extremely negative) to 4 (extremely positive).

In order to compare for each topic the sentiment expressed by a party, we compute the average of the scores for the party on that topic. Only the topics with a precision greater than 0.6 (Table 4a), and the parties that wrote more than 10 tweets on the corresponding topic, were considered in this comparison. It means that we ignore, for instance, the sentiment showed towards *Economy* (precision lower than 0.6), and *Healthcare* (only UP wrote 10 tweets, see Table 4b, and the sentiment that Cs showed towards *Pensions* (only two tweets, see Table 4b).

Figure 1 shows the expressed sentiment for the parties for each topic. Sentiment scores seem to reveal some common dynamics of political communication from political parties in social networks in that generally, even when the party is known to be negative or have a critical stance with respect to a certain topic (e.g., a populist party on immigration), tweets receive a positive score. Specifically, we see that VOX was the only party addressing the *Immigration* topic, and we observe that in general, its sentiment is positive (i.e., solutions were commented). Also, just two parties show mainly negative sentiments, they are VOX and PP towards *Feminism* and *Pensions* respectively.

3.3 Emotion analysis

We finally analyse the emotions expressed by the parties for different topics using the Spanish Emotion Lexicon (SEL) [14]. SEL has 2,036 words associated with the measure of Probability Factor of Affective (PFA) concerning to at least one Ekman's emotions [2]: joy, anger, fear, sadness, surprise, and disgust. For each tweet, we compute the final measure for each of the five emotions by summing the PFA and dividing by the length of the tweet. We then compute the average PFA of all the emotions for each party and each topic.

Figure 2 (top image on the left) shows the emotions that the parties present in their tweets when talking about different topics. We analyse the emotions of the same pairs of parties and topics we analysed before in Section 3.2. Differently to the case of sentiment, there is a general trend shared in that joy and sadness are very much present across all parties. This could be due to several reasons. First, there is a bias in SEL towards joy (668 words related to joy vs. 391 for sadness, 382 for anger, 211 for fear, 209 for disgust, and 175 for surprise), and second, the terms that help to compute the SentiStrength score are not necessarily the same that are in SEL . Another interesting thing is the presence of joy and sadness in the same topic by the same parties. We attribute this behaviour to the fact that there are tweets describing the current problems and feelings present in the context of the election - e.g., using words like *sufrir* (to suffer), *muerte* (death), *triste* (sad), *grave* (grave), but also there are others with a propositive discourse about the problems - e.g., using words like *esperanza* (hope), *ánimo* (encouragement), *unión* (union), *fiesta* (party).

In Figure 2 we also highlight that PSOE shows contrasting emotions about Catalonia; and Cs shows high score of joy about topics related to feminism. The distribution of the emotions from VOX towards *Immigration* was omitted due to the space. However, despite the positive sentiment that VOX showed in this topic, the predominant expressed emotion was sadness.

4 Conclusions

In this paper we presented a first study about the most relevant topics that have been addressed in Twitter in the context of the 10N Spanish election for the five main political parties, together with their sentiments and emotions.

On the basis of the above analysis, we noticed that each party focused more on specific topics, expressing different sentiments and emotions. Our analysis, although preliminary, indicates potentially interesting dimensions of political

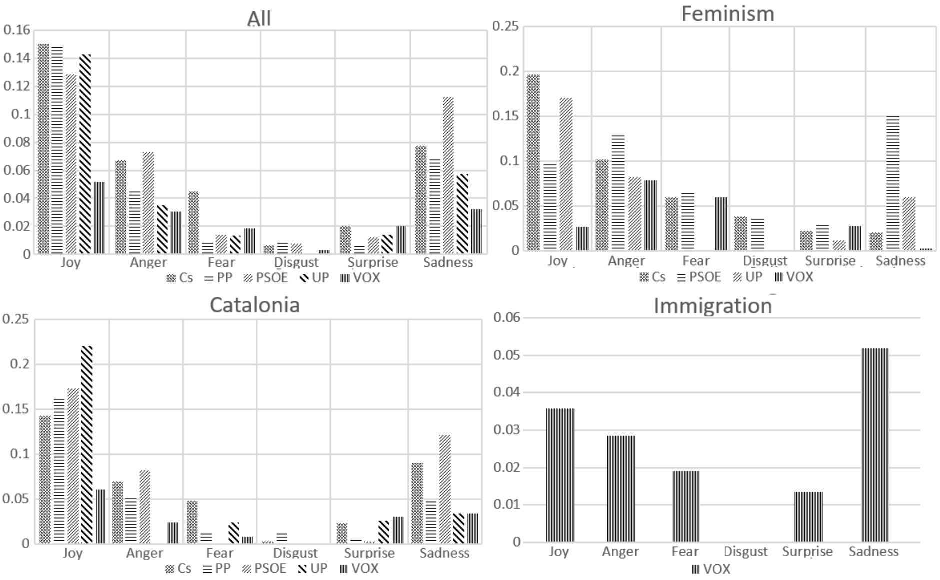

Fig. 2: Emotions distribution across topics.

communications on social networks such as the tendency towards positive tweets, as well the contrasted presence of problems vs. solutions. This work provides a first attempt towards analysing the political communication by the five main political parties in Spain on social networks using NLP techniques. Although we are aware of the limitations of this preliminary study due to the very short time span and the size of the corpus, we hope that this first analysis could contribute to understand how sentiments and emotions were expressed in Twitter by the politicians of the main five parties with respect to the topics mentioned in their manifestos during the political campaign of the 10N Election in Spain.

As future work we plan also to consider additional parties and languages (e.g. Catalan, Basque and Galician) to provide a more comprehensive resource as well as a comparative analysis.

Acknowledgements The work of the authors from the Universitat Politècnica de València was funded by the Spanish MICINN under the research project MISMIS-FAKEnHATE on Misinformation and Miscommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31).

References

1. Abercrombie, G., Nanni, F., Batista-Navarro, R., Ponzetto, S.P.: Policy preference detection in parliamentary debate motions. In: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL). pp. 249–259. Association for Computational Linguistics, Hong Kong, China (Nov 2019)

2. Ekman, P., Friesen, W.V., O'sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W.A., Pitcairn, T., Ricci-Bitti, P.E., et al.: Universals and cultural differences in the judgments of facial expressions of emotion. Journal of personality and social psychology **53**(4), 712 (1987)
3. Gao, W., Sebastiani, F.: Tweet sentiment: From classification to quantification. In: 2015 IEEE/ACM International Conference on ASONAM. pp. 97–104. IEEE (2015)
4. Glavaš, G., Nanni, F., Ponzetto, S.P.: Computational analysis of political texts: Bridging research efforts across communities. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts. pp. 18–23. Association for Computational Linguistics, Florence, Italy (Jul 2019)
5. Kornilova, A., Argyle, D., Eidelman, V.: Party matters: Enhancing legislative embeddings with author attributes for vote prediction. pp. 510–515. Association for Computational Linguistics (Jul 2018)
6. Lowe, W., Benoit, K., Mikhaylov, S., Laver, M.: Scaling policy preferences from coded political texts. Legislative Studies Quarterly **36**(1), 123–155 (2 2011)
7. Marchetti-Bowick, M., Chambers, N.: Learning for microblogs with distant supervision: Political forecasting with twitter. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 603–612. Association for Computational Linguistics, Avignon, France (Apr 2012)
8. Menini, S., Nanni, F., Ponzetto, S.P., Tonelli, S.: Topic-based agreement and disagreement in US electoral manifestos. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2938–2944. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017)
9. Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., Stoyanov, V.: Semeval-2016 task 4: Sentiment analysis in twitter. arXiv preprint arXiv:1912.01973 (2019)
10. Nanni, F., Glavas, G., Ponzetto, S., Tonelli, S., Conti, N., Aker, A., Palmero Aprosio, A., Bleier, A., Carlotti, B., Gessler, T., Henrichsen, T., Hovy, D., Kahmann, C., Karan, M., Matsuo, A., Menini, S., Nguyen, D., Niekler, A., Posch, L., Vegetti, F., Waseem, Z., Whyte, T., Yordanova, N.: Findings from the hackathon on understanding euroscepticism through the lens of textual data. European Language Resources Association (ELRA) (may 2018)
11. O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls: Linking text sentiment to public opinion time series. In: Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010 (2010)
12. Rheault, L., Cochrane, C.: Word embeddings for the analysis of ideological placement in parliamentary corpora. Political Analysis pp. 1–22 (2019)
13. Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B., Rand, D.G.: Structural topic models for open-ended survey responses. American Journal of Political Science **58**(4), 1064–1082 (2014)
14. Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., Díaz-Rangel, I., Suárez-Guerra, S., Trevino, A., Gordon, J.: Empirical study of machine learning based approach for opinion mining in tweets. In: Mexican international conference on Artificial intelligence. pp. 1–14. Springer (2012)
15. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology **61**(12), 2544–2558 (2010)

16. Thomas, M., Pang, B., Lee, L.: Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. pp. 327–335. Association for Computational Linguistics (Jul 2006)