Overview of the Task on Stance and Gender Detection in Tweets on Catalan Independence at IberEval 2017

Mariona Taulé[1], M. Antònia Martí[1], Francisco Rangel[2,3], Paolo Rosso[2], Cristina Bosco[4], and Viviana Patti[4]

[1] CLiC-UBICS, Universitat de Barcelona, Spain
{mtaule,amarti}@ub.edu,
[2] PRHLT Research Center, Universitat Politécnica de València, Spain
prosso@dsic.upv.es
[3] Autoritas Consulting, S.A., Spain
francisco.rangel@autoritas.es
[4] Università degli Studi di Torino, Italy
{bosco,patti}@di.unito.it

Abstract. Stance and Gender Detection in Tweets on Catalan Independence (StanceCat) is a new shared task proposed for the first time at the IberEval 2017 evaluation campaign. The automatic natural language systems presented must detect the tweeter stance (in favor, against or neutral) towards the target *independence of Catalonia* in Twitter messages written in Spanish or Catalan, as weel as the author's gender if possible. We have received a total of 31 submitted runs from 10 different teams from 5 countries. We present here the datasets, which include annotations for dealing with stance and gender, the evaluation methodology, and discuss results and participating systems.

Keywords: Stance detection, Twitter, Spanish, Catalan, Gender identification

1 Introduction

The aim of the task of Stance and Gender Detection in Tweets on Catalan Independence at IberEval 2017 (StanceCat) is to detect the author's gender and stance with respect to the *independence of Catalonia* in tweets written in Spanish or Catalan. Classical sentiment analysis tasks carried out in recent years in evaluation campaigns for different languages have mostly involved the detection of the subjectivity and polarity of microblogs at the message level, i.e. determining whether a tweet is subjective or not, and, if subjective, determining its positive or negative semantic orientation. However, comments and opinions are usually directed towards a specific target or issue, and therefore give rise to finer-grained tasks such as stance detection, in which the focus is on detecting what particular stance (in favor, against or neutral) a user takes with respect to a specific target.

Stance detection is related to sentiment analysis, but there are significant differences, as is stressed in [9]: in sentiment analysis, the systems detect whether the sentiment polarity of a text is positive, negative or neutral, while in stance detection, the systems detect whether a given text is favorable or unfavorable to a given target, which may or may not be explicitly mentioned in the text. Stance detection is particularly interesting for studying political debates in which the topic is controversial. Therefore, for this task we have chosen to focus on a specific political target: the *independence of Catalonia* [5]. The stance detection task is also related to a textual inference task due to the fact that the position of the tweeter is often expressed implicitly, therefore, the stance has to be inferred in many cases. See, for instance, the following tweet (1).

1. **Language:** Catalan
   Target: Catalan Independence
   Stance: FAVOR
   Tweet: *Avui #27S2015 tot està per fer... Un nou país és possible ‖*‖A les urnes... #27S http://t.co/ls2nkRWt2b*
   Today #27S2015 the future is ours to make... A new country is possible ‖*‖
   Get out and vote... #27S http://t.co/ls2nkRWt2b
   (where ‖*‖stands for the Catalan Independence flag).

Stance detection and author profiling tasks on microblogging texts are currently being carried out in several evaluation forums, including SemEval-2016 (Task-6) [9] and PAN@CLEF [12]. However, these two tasks have never been performed together for Spanish and Catalan as part of one single task. The results obtained will be of interest not only for sentiment analysis but also for author profiling and for socio-political studies.

2 Task description

The StanceCat Task includes two subtasks that are meant to be independent, namely stance detection and the identification of the gender of the author. Moreover, the participation of each team in each subtask can be for one or both languages involved in the contest, i.e. Spanish and Catalan.

As far as the stance detection subtask is concerned, providing that the reference data have been filtered with hashtags and keywords related to a specific topic, i.e. the *independence of Catalonia*, it consists of deciding whether each message is neutral or oriented in favor of or against the given target. The three labels representing the stance of the author in writing the message are mutually exclusive.

The second task consists of identifying the gender of the author of each message and thus labeling it as male or female, as mutually exclusive labels. Section 3.2 provides further explanation and examples about the labels included in the annotation scheme applied to the dataset.

The distribution of the labels (shown in Table 2) for gender in both the training and test sets: half of the data are produced by female authors and the other half by males. In contrast, the distribution of the labels for stance is not balanced. Also the participation varies according to the subtask given that not all the teams took part in the gender classification task but all tackled the stance detection task.

Based on the experience of previous contests, different metrics were adopted for the different subtasks (see section 4) and different rankings of the participants scores were generated for the evaluation of each subtask.

As far as the language is concerned, half of the data are in Spanish and the other half in Catalan and each of the previously described subtasks had to be performed separately for Spanish and Catalan. Each team could decide to perform the task for a single language or for both. Given that most teams performed the selected subtasks in both Spanish and Catalan, an evaluation of performance across the two different languages was done, showing relevant differences in scores.

3 Development and Test Data

3.1 Corpus Description

As usual in the last few years in debates on social and political topics, the discussion on Catalan separatism involved a massive use of social media by users interested in the discussion. In order to draw attention to the related issues, as also happens with commercial products and political elections, users created new hashtags to give greater visibility to information and opinions on the subject.

Among them #Independencia and #27S are two of the hashtags that have been widely accepted with the dialogical and social context growing around the topic, and were widely used within the debate. At the current stage of the development of our project we exploited the hashtag #Independencia and #27S as the first two keywords for filtering data to be included in the *TW-CaSe* corpus. We selected the #27S hashtag because on that date the autonomy elections of Catalonia were celebrated, and were considered as a plebiscite by the pro-independence parties. The hashtag #Independencia and #27S allowed us to select 10,800 original messages -5,400 written in Catalan (*TW-CaSe-ca*) and 5,400 tweets written in Spanish (*TW-CaSe-es*)- collected between the end of September and December 2015 and were also largely retweeted[5]. Half of the tweets in each language were written by female authors and half by male authors.

3.2 Annotation Scheme

This section describes the scheme adopted for the annotation of the *TW-CaSe* corpus with the author's stance and gender.

[5] The dataset was collected with the Cosmos tool by Autoritas (`http://www.autoritas.net`) and it was annotated by the CLiC group at the University of Barcelona (`http://clic.ub.edu`)

In order to annotate the stance, we use the following tags adopting the annotation scheme proposed in [5] and [9]:

- FAVOR: positive stance towards the independence of Catalonia (2).
- AGAINST: negative stance towards the independence of Catalonia (3).
- NONE: neutral stance towards the independence of Catalonia and cases in which the stance cannot be inferred (4).

The possible gender labels are: FEMALE (2) and MALE (3).These tags were automatically extracted from proper nouns dictionaries (INE[6]) and manually reviewed to remove ambiguous names. The following are examples of tweets labelled for both the author's stance and gender in both languages.

2. **Language:** Catalan
   Target: Catalan Independence
   Stance: FAVOR
   Gender: FEMALE
   Tweet: *15 diplomàtics internacional observen les plebiscitàries, serà que interessen a tothom menys a Espanya #27S*
   '15 international diplomats observe the plebiscite, perhaps it is of interest to everybody except to Spain #27S2015'

3. **Language:** Spanish
   Target: Catalan Independence
   Stance: AGAINST
   Gender: MALE
   Tweet: *#27S cuál fue la diferencia en 2012 entre los resultados de la encuesta de TV3 y resultados finales? Nos serviría para hacernos una idea*
   (In 2012, what was the difference between the results of the TV3 poll and the final results? That would give us an idea)

4. **Language:** Catalan
   Target: Catalan Independence
   Stance: NONE
   Gender: MALE
   Tweet: *100% escrutat a Arbúcies #27S http://t.co/avMzng6iyV*
   (100% of votes counted in Arbúcies #27s http://t.co/avMzng6iyV)

[6] `http://www.ine.es`

Although tweets are very short pieces of text, they tend to be complex in their internal structure and often contain considerable informational content. It should be pointed out that for the annotation of stance we took into account all the information appearing in the written text (including emoticons), as well as the information concerning some other user mentioned and hashtags. The mentioned users are identified with the symbol @, and they are also known as mentions; hashtags are semantic labels (introduced with #), which are important for understanding the tweet, and often denote the content highlighted by the author.

It is worth noting that hashtags, like mentions, can appear in any position within the text playing a syntactic-semantic role within a tweet.

We consider that all of these components play a role in the interpretation of the whole tweet and we took them into account in the annotation of stance. However, links -web addresses including photographs, videos and webpages- are also very useful for interpreting the stance, and are especially relevant for the interpretation of ironical tweets, but in this version of the corpus we did not take them into account since the automatic systems do not do so. It is worth noting that we are currently working on a new version of the *TW-CaSe* corpus in which irony and humor are also being annotated, as well as information on the role played by links in the tweet.

3.3 Annotation procedure

In this section, we present the methodology applied in the annotation of tweets, the results of the inter-annotator agreement test carried out and, finally, we analyse the different sources of disagreement.

Three trained annotators, supervised by two senior researchers, carried out the whole manual annotation of *TW-CaSe*. The annotation process was performed in the following way: 1) First, the three trained annotators tagged the stance in 500 tweets in Catalan and 500 tweets in Spanish working in parallel and following the guidelines [5]. 2) We then conducted an inter-annotator agreement test on the 500 tweets tagged in each language in order to test the validity of this annotation (see Table 1), and to detect and solve the disagreements and possible inconsistencies. 3) Finally, the annotators went on to annotate the whole corpus individually. During the annotation process, we met once a week to discuss problematic cases, which were discussed by all the people involved in the annotation process and solved by common consensus.

Table 1 presents the pairwise and average agreement percentages obtained in the inter-annotator agreement test in *TW-CaSe-ca* and *TW-CaSe-es*. In the first four rows (2-5), we show the result of the observed agreement for each pair of annotators (pairwise agreement) and the average agreement (79.26% in *TW-CaSe-ca* and 78.4% in *TW-CaSe-es*). The last row shows the Fleiss' Kappa coefficient (0.60 in both subcorpora). The results obtained show a moderate agreement, demonstrating the complexity of the task. The annotation of the corpus was completed in 16 weeks.

Table 1. Results of the inter-annotator agreement test

Annotator pairs	Pairwise agreement	
	TW-CaSe-ca	TW-CaSe-es
A-B	75.78%	76.40%
A-C	79.54%	77.80%
B-C	82.46%	81%
Average agreement	79.26%	78.40%
Fleiss' Kappa	0.60	0.60

Regarding disagreements, the most problematic cases in the annotation of stance arise when the authors communicative intentions are not clear. For instance, one annotator tagged tweet (5) as being AGAINST independence, probably influenced by the language used in the tweet (Spanish), whereas the other two annotators tagged it as NONE. However, after collectively discussing this case, we agreed to tag the tweet (5) with the NONE stance, because it was not clear enough to which flag (Spanish or Catalan) the writer was referring to.

5. **Language:** Spanish
   Target: Catalan Independence
   Stance: NONE
   Gender: MALE
   Tweet: *#27s voy a denunciar a todo aquel q me siga insultando usando ls red. Yo no soy imbcil, ni mi bandera es n trapo*
   (#27s I'm going to denounce anyone who continues to insult me using the web. Im not stupid, neither my flag is a rag)

6. **Language:** Catalan
   Target: Catalan Independence
   Stance: NONE
   Gender: MALE
   Tweet: *La @cupnacional t la clau de Matrix*
   (The @cupnacional has the key of Matrix

The same problem occurs with tweet (6), in which each annotator assigned a different tag for stance. This is an example of total disagreement. In the end, it was also annotated as NONE since the stance could not be clearly inferred. The cases in which the disagreement was total, we tended to assign the neutral NONE tag.

This is domain dependent information and the annotators knowledge of the domain is therefore crucial. Frequently, the annotators have to infer the stance

and, for doing this inference, they need to know the socio-political context and the social agents involved in the debate, in our case, about Catalan independence, which is not always true for all annotators.

3.4 Format and Distribution

We provided participants with a single development set for training, which consists of a collection of 4,319 tweets in Spanish and 4,319 tweets in Catalan, with annotations concerning the two subtasks: stance detection and identification of gender. For each language, we distributed two files: the first one includes tweets' IDs and textual contents. The data format is as follows: *id* ::: *contents*; the second one includes the truth labels for the two tasks. For the truth files the data format is *id* ::: *stance* ::: *gender* (see Section 3.2 for a description of the possible labels). The language was encoded in the file name.

The test data consist of 1,081 tweets in Spanish and 1,081 tweets in Catalan in the same format: *id* ::: *contents*. Participants therefore did not need to detect the language. Tweets were provided to the participants in two independent files per language, as in the training set. The blind version of the test data did not include the truth files[7].

The distribution in training and testing sets of the data exploited for the stance subtask is balanced in an 80/20 proportion: 80% for training and 20% for testing. The distribution in both training and test data for stance, gender and language is given in Table 2.

Table 2. Distribution of labels for stance, gender and language

	FEMALE			**MALE**			**total**	**dataset**
	FAVOR	**AGAINST**	**NONE**	**FAVOR**	**AGAINST**	**NONE**		
Catalan	1,456	57	646	1,192	74	894	4,319	training
	365	14	162	298	18	224	1,081	test
Spanish	145	693	1,322	190	753	1,216	4,319	training
	36	173	331	48	188	305	1,081	test

4 Evaluation Metrics

The evaluation was performed according to standard metrics. In particular, we used the macro-average of F-score (FAVOR) and F-score (AGAINST) to evaluate stance, in accordance with the metric proposed at Semeval 2016 - Task

[7] Data will be available for downloading at the following address: `http://stel.ub.edu/Stance-IberEval2017/data.html`. In the first stage access has been restricted to participants registered for the task. To acces the dataset, ask for the password by emailing to `stancetask2017@gmail.com`.

6[8]. Gender was evaluated in terms of *accuracy*, in accordance with the metrics proposed at the Author Profiling task at PAN@CLEF[9].

Four different rankings are shown depending on the subtask and language. Concretely, stance ranking for Spanish and Catalan, and gender ranking for Spanish and Catalan. Two baselines are provided for comparison purposes: A random basis approach that returns the majority class, and the Low Dimensionality Representation (LDR) [11] approach. The key concept of LDR is a weight representing the probability of each term to belong to each of the different categories: for stance (in favor vs. against) and gender (female vs. male). The distribution of weights for a given document should be close to the weights of its corresponding category. LDR takes advantage of the whole vocabulary. However, in order to work properly, it needs a sufficient amount of information per author.

5 Overview of the Submitted Approaches

Ten teams from five countries participated in the shared task by sending up to thirty-one runs. Table 3 provides an overview of the teams, their country of origin (C) and the tasks they took part in, i.e. stance (S) and gender (G) for the two languages: Spanish (ES) and Catalan (CA).

Table 3. Teams participating to StanceCat at IberEval 2017

Team	**C**	**Tasks**
ARA1337 [1]	ES	S(ES,CA)
ATeam [14]	ES	S(ES,CA)
atoppe [2]	CH	S(ES,CA)
deepCybErNet [10]	India	S(ES,CA), G(ES,CA)
ELiRF-UPV [7]	ES	S(ES), G(ES)
iTACOS [8]	IT,ES	S(ES,CA), G(ES,CA)
LaSTUS [4]	ES	S(ES,CA), G(ES,CA)
LTL_UNI_DUE [15]	DE	S(ES,CA)
LTRC_IIITH [13]	India	S(ES,CA), G(ES,CA)
LuSer [6]	ES	S(ES,CA)

All the teams participated in the stance subtask in Spanish and nine of them in Catalan. Four teams participated in the gender subtask, both in Catalan and Spanish, whereas only one team participated in the gender subtask in Spanish. Eight teams sent a description of their systems, and used only the training data provided for the task. In what follows, we analyse their approaches from two

[8] `http://alt.qcri.org/semeval2016/task6/index.php?id=data-and-tools`
[9] `http://pan.webis.de/clef16/pan16-web/author-profiling.html`

perspectives: *classification approaches*, and *features* to represent the authors' texts.

Classification approaches. Most participants used SVM: *i*) *ltl_uni_due*, which also applied LSTM and a hybrid system that decides with a decision tree which algorithm to apply; *ii*) *iTACOS*, which also experimented with logistic regression, decision trees, random forest and multinomial NB; *iii*) *ARA1337* and *ELiRF-UPV*, which also used neural networks; and *iv*) *LTRC_IIITH*, which used RBF kernels. Neural networks and deep learning approaches were widely used by participants such as *ltl_uni_due* (LSTM), *ARA1337*, *ELiRF-UPV*, *LuSer* (multilayer perceptron), and *atoppe* (CNN, LSTM, MLP, FASTTEXT, KIM and BI-LSTM).

Features. Both n-grams and embeddings are the most used features. Teams using SVM represented texts with n-gram based approaches, whereas teams using different kinds of deep approaches basically used word embeddings. For instance, *ltl_uni_due* used combinations of word and character n-grams with SVM and word embeddings with LSTM. *LTRC_IIITH* used character and word n-grams with SVM, as well as specific stance and gender indicative tokens. In contrast, teams using deep approaches represented texts with bag-of-words embeddings (*deepCybErNet*), and word and n-gram embeddings (*atoppe*). *ELiRF-UPV* used one-hot vectors to train its networks. Other teams used neural networks as classification algorithms, but with features such as word, tokens and hashtags unigrams (*ARA1337*) or bag of n-grams (*LuSer).* Finally, *iTACOS* combined bag of words with bag of part-of-speech, bag of lemmas, bag of hashtags, bag of words in hashtags and mentions, char n-grams, number of hashtags, number of words starting with capital letter, language, number of words, number of characters, average word length, and bag of words extracted from urls.

6 Evaluation and Discussion of the Submitted Approaches

We evaluated both subtasks (stance and gender) independently. We show results separately for the evaluation of each subtask and for each language. Results are given in F-score in case of stance and accuracy in case of gender.

6.1 Stance Subtask

Ten teams participated in the Spanish subtask, presenting thirty-one runs, and nine teams participated in the Catalan subtask, presenting twenty nine runs. In Table 4, the F-scores achieved by all runs are shown, as well as the two baselines. At the bottom of the table some basic statistics are provided: minimum (min), maximum (max), mean, median, standard deviation (stdev), first quartile (q1) and third quartile (q3).

In the Catalan subtask, the majority of the runs (29 out of 31) obtained worse results than the *majority class* prediction (F-score 0.4882). The only runs that improved majority class prediction belong to the same team (*iTACOS*) with an F-score of 0.4901 and 0.4885. They approached the task with different machine

learning algorithms such as SVM, logistic regression or decision trees, among others, with combinations of different kinds of features (bag of words, bag of parts-of-speech, n-grams) and stylistic features (word length, number of words, number of hashtags, number of words starting with capital letters, and so on). The worst results were obtained with deep learning approaches, with F-scores between 0.2710 (*attope.1*) and 0.3790 (*deepCybErNet.2*).

In the Spanish subtask, twelve runs obtained better results than the *majority class* baseline (0.4479). The best result was also obtained by the *iTACOS* team, with an F-score of 0.4888. The next best results were obtained by different runs of *LTRC_IIITH* (0.4679 and 0.4640) and *ELIRF-UPV* (0.4637). While *LTRC_IIITH* used SVM learning from character and word n-grams besides specific stance features, *ELIRF-UPV* used neural networks and SVM with one-hot vectors and bag-of-words. The worst results were obtained by the *attope* team with word embeddings and combinations of neural networks models (between 0.1906 and 0.2466).

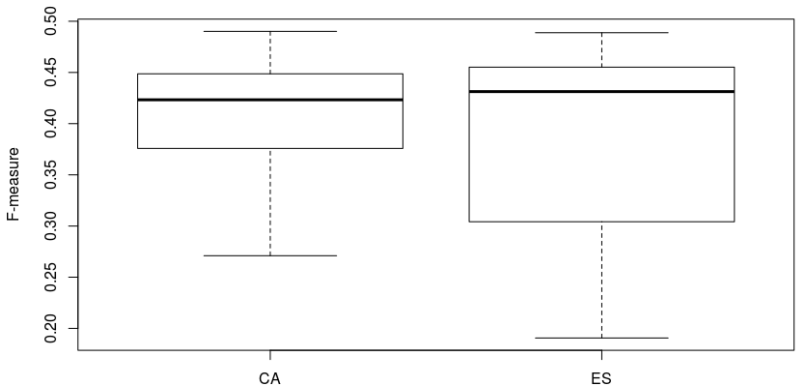

Fig. 1. Distribution of results (F-score) for the stance subtask.

As can be seen in Figure 1, results are similar for mean, max and q3 statistics for both languages, although they are more sparse for Spanish and have lower values for the worst systems. Results for Catalan are between 0.4901 and 0.2710, with an average value of 0.4053. Results for Spanish are between 0.4888 and 0.1906, with an average value of 0.3843.

Table 4. Evaluation results for Stance in Catalan and Spanish (F-score).

Catalan		
Position	Team.Run	F
1	iTACOS.2	0.4901
2	iTACOS.1	0.4885
3	*majority class.baseline*	*0.4882*
4	iTACOS.3	0.4685
5	LTRC_IIITH.system1	0.4675
6	ARA1337.s1	0.4659
7	ARA1337.s2	0.4511
8	iTACOS.4	0.4490
9	iTACOS.5	0.4484
10	ATeam.systemid	0.4439
11	LTRC_IIITH.system3	0.4393
12	LTRC_IIITH.system4	0.4388
13	*LDR.baseline*	*0.4375*
14	LTL_UNI_DUE.hybrid	0.4246
15	LTL_UNI_DUE.svm	0.4233
16	LTRC_IIITH.system2	0.4233
17	LTRC_IIITH.system5	0.4165
18	UPF-LaSTUS.2	0.3955
19	UPF-LaSTUS.1	0.3949
20	UPF-LaSTUS.3	0.3938
21	LuSer.1	0.3909
22	UPF-LaSTUS.4	0.3854
23	deepCybErNet.2	0.3790
24	LTL_UNI_DUE.lstm	0.3726
25	deepCybErNet.1	0.3603
26	attope.2	0.3310
27	deepCybErNet.3	0.3257
28	attope.5	0.3120
29	attope.3	0.2970
30	attope.4	0.2910
31	attope.1	0.2710
32	ELIRF-UPV.1	-
33	ELIRF-UPV.2	-
	min	0.2710
	q1	0.3758
	median	0.4233
	mean	0.4053
	stdev	0.0612
	q3	0.4487
	max	0.4901

Spanish		
Position	Team.Run	F
1	iTACOS.1	0.4888
2	LTRC_IIITH.system1	0.4679
3	LTRC_IIITH.system4	0.4640
4	ELIRF-UPV.1	0.4637
5	ELIRF-UPV.2	0.4637
6	UPF-LaSTUS.1	0.4600
7	iTACOS.2	0.4593
8	LTRC_IIITH.system2	0.4566
9	LTRC_IIITH.system3	0.4552
10	LTRC_IIITH.system5	0.4544
11	ARA1337.s1	0.4530
12	iTACOS.3	0.4528
13	*majority class.baseline*	*0.4479*
14	iTACOS.4	0.4427
15	LTL_UNI_DUE.hybrid	0.4347
16	LTL_UNI_DUE.svm	0.4314
17	ARA1337.s2	0.4313
18	iTACOS.5	0.4293
19	*LDR.baseline*	*0.4135*
20	LuSer.1	0.4060
21	ATeam.systemid	0.3914
22	UPF-LaSTUS.4	0.3812
23	UPF-LaSTUS.2	0.3795
24	deepCybErNet.3	0.3066
25	deepCybErNet.2	0.3042
26	deepCybErNet.1	0.2849
27	LTL_UNI_DUE.lstm	0.2759
28	UPF-LaSTUS.3	0.2505
29	attope.5	0.2466
30	attope.4	0.2438
31	attope.3	0.2426
32	attope.2	0.2074
33	attope.1	0.1906
	min	0.1906
	q1	0.3042
	median	0.4313
	mean	0.3843
	stdev	0.0919
	q3	0.4552
	max	0.4888

LDR obtained worst results than the *majority class* prediction. Since this task was focused on the tweet level instead of the author level, these low results might be expected due to the need of LDR for a large amount of data per author in order to normalise frequency distributions. Something similar might have happened with deep learning approaches that need large amounts of data to learn the models. However, the provided dataset is small and biased towards a majority class.

6.2 Gender Subtask

Five teams participated in the Spanish subtask, presenting nineteen runs, and four teams in the Catalan subtask, presenting seventeen runs. In Table 5 the *accuracies* achieved by all runs are shown, together with the two baselines. At the bottom of the table some basic statistics are also provided: minimum (min), maximum (max), mean, median, standard deviation (stdev), first quartile (q1) and third quartile (q3).

In the Catalan subtask, all the runs (19) obtained worse results than the *majority class* (0.5005) and *LDR* predictions (0.6068). The best results were obtained by *deepCybErNet* (0.4857, 0.4829 and 0.4653) and *LTRC_IIITH* (0.4459 and 0.4440). They used SVM with combinations of char and word n-grams together with specific gender indicators, and deep learning methods respectively. The worst results were obtained by *UPF-LaSTUS* (0.3571 and 0.4043) and *iTACOS* (0.3996 and 0.3987). iTACOS used different machine learning algorithms with a combination of different bags of features, and *UPF-LaSTUS* did not provided a description of their system.

In the Spanish subtask, most runs obtained better results than the *majory class* prediction, although they were below *LDR*. The best results were obtained by *LTRC_IIITH* (between 0.6485 and 0.6401) and *iTACOS* (between 0.6161 and 0.6124). The worst results were obtained by *deepCybErNet* (0.4764, 0.4903 and 0.5014). It is noteworthy that the latter team obtained the best results in Catalan but the worst in Spanish. However, the obtained accuracies were similar (0.4857, 0.4829, 0.4656 vs. 0.5014, 0.4903, 0.4764) for both languages. This demonstrates the stability of this system when applied to different datasets.

As can be seen in Figure 2, results for Catalan are less sparse than for Spanish, though all of them are below the *majority class* and have an average accuracy of 0.4459. There are three outliers corresponding from above to *LDR* (0.6068) and *majority class* (0.5050), and from below to *UPF-LaSTUS* (0.3571). Most results for Spanish are between 0.5495 and 0.6448, with an average accuracy of 0.5935. The maximum value of 0.6855 was obtained by *ELIRF-UPV* and the minimum of 0.4764 by *deepCybErNet*.

LDR obtained the best result for Catalan and the second best result for Spanish, despite the low amount of data per author. The *majority class* prediction coincides with a random classification since the dataset is balanced in terms of gender. Deep learning approaches such as *deepCybErNet* maintained their stability, though with values below those of the *majority class*.

Table 5. Evaluation results for Gender in Catalan and Spanish (accuracy).

Catalan			Spanish		
Position	Team.Run	Accuracy	Position	Team.Run	Accuracy
1	*LDR.baseline*	*0.6068*	1	ELIRF-UPV.1	0.6855
2	*majority class.baseline*	*0.5005*	*2*	*LDR.baseline*	*0.6550*
3	deepCybErNet.3	0.4857	3	LTRC_IIITH.system1	0.6485
4	deepCybErNet.2	0.4829	4	LTRC_IIITH.system5	0.6457
5	deepCybErNet.1	0.4653	5	LTRC_IIITH.system3	0.6448
6	LTRC_IIITH.system3	0.4459	6	LTRC_IIITH.system4	0.6448
7	LTRC_IIITH.system1	0.4440	7	LTRC_IIITH.system2	0.6401
8	LTRC_IIITH.system4	0.4440	8	iTACOS.4	0.6161
9	UPF-LaSTUS.1	0.4431	9	iTACOS.2	0.6142
10	UPF-LaSTUS.2	0.4422	10	iTACOS.5	0.6124
11	iTACOS.5	0.4329	11	UPF-LaSTUS.1	0.6115
12	LTRC_IIITH.system2	0.4320	12	iTACOS.1	0.6115
13	LTRC_IIITH.system5	0.4311	13	iTACOS.3	0.6096
14	iTACOS.2	0.4292	14	ELIRF-UPV.2	0.5874
15	iTACOS.1	0.4274	15	UPF-LaSTUS.4	0.5865
16	UPF-LaSTUS.3	0.4043	16	UPF-LaSTUS.3	0.5495
17	iTACOS.4	0.3996	17	UPF-LaSTUS.2	0.5310
18	iTACOS.3	0.3987	18	deepCybErNet.3	0.5014
19	UPF-LaSTUS.4	0.3571	*19*	*majority class.baseline*	*0.5005*
20	ELIRF-UPV.1	-	20	deepCybErNet.2	0.4903
21	ELIRF-UPV.2	-	21	deepCybErNet.1	0.4764
22	LTL_UNI_DUE.svm	-	22	LTL_UNI_DUE.svm	-
23	LTL_UNI_DUE.lstm	-	23	LTL_UNI_DUE.lstm	-
24	LTL_UNI_DUE.hybrid	-	24	LTL_UNI_DUE.hybrid	-
25	ARA1337.s1	-	25	ARA1337.s1	-
26	ARA1337.s2	-	26	ARA1337.s2	-
27	ATeam.systemid	-	27	ATeam.systemid	-
28	LuSer.1	-	28	LuSer.1	-
29	attope.1	-	29	attope.1	-
30	attope.2	-	30	attope.2	-
31	attope.3	-	31	attope.3	-
32	attope.4	-	32	attope.4	-
33	attope.5	-	33	attope.5	-
	min	0.3571		min	0.4764
	q1	0.4283		q1	0.5495
	median	0.4422		median	0.6115
	mean	0.4459		mean	0.5935
	stdev	0.0513		stdev	0.0613
	q3	0.4556		q3	0.6448
	max	0.6068		max	0.6855

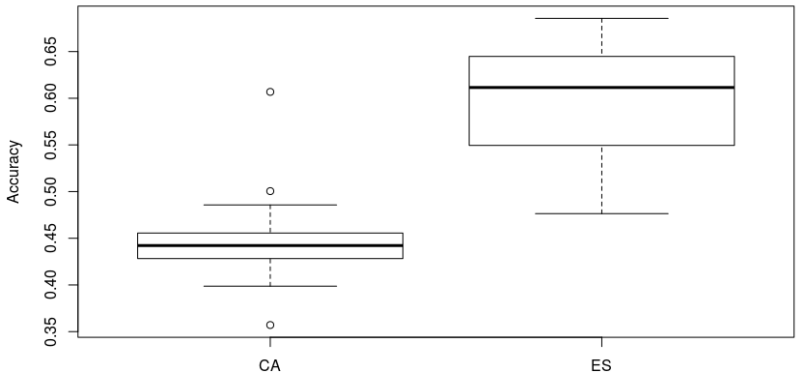

Fig. 2. Distribution of results (accuracy) for the gender subtask.

6.3 Stance vs. Gender

In this section the performance of the systems with respect to both subtasks is analysed together. The aim is to know whether systems performing properly in one subtask, do the same in the other one. The analysis is carried out separately per language.

The results for Catalan are shown in Figure 3. In this language, results for gender were below the *majority class* and *LDR*. *DeepCybErNet* achieved the best results in gender identification, and the worst in stance. This team approached the task with deep learning techniques. On the other hand, systems that obtained some of the best results for stance (*iTACOS.1, iTACOS.2* and *iTACOS.3*), obtained some of the worst results for gender. Systems such as *UPF-LaSTUS.3* and *UPF-LaSTUS.4* obtained some of the worst results both for gender and stance. In this case, they did not provide a description of their system.

The results for Spanish are shown in Figure 4. In this language, results for gender are higher than in Catalan, with most systems over the *majority class* baseline. There is a clearly observable trend for the systems that obtained better results for gender to do the same for stance. For example, *ELIRF-UPV.1* obtained the best result for gender and the third position for stance. In this case, the authors approached the task with one-hot vectors and neural networks. Similarly, *iTACOS.1* obtained the best result for stance, with a value on the median for gender, by using combinations of features and SVM. And finally, the group of results obtained by *LTRC_IIITH* are some of the bests for both subtasks. They learned RBF kernels for SVM with combinations of character and word n-grams with indicative tokens per subtask. On the other hand, *deepCybErNet* and *UPF-LasTUS* obtained the worst results in both subtasks. There is no infor-

mation for *UPF-LasTUS* but *deepCybErNet* used different deep learning-based approaches.

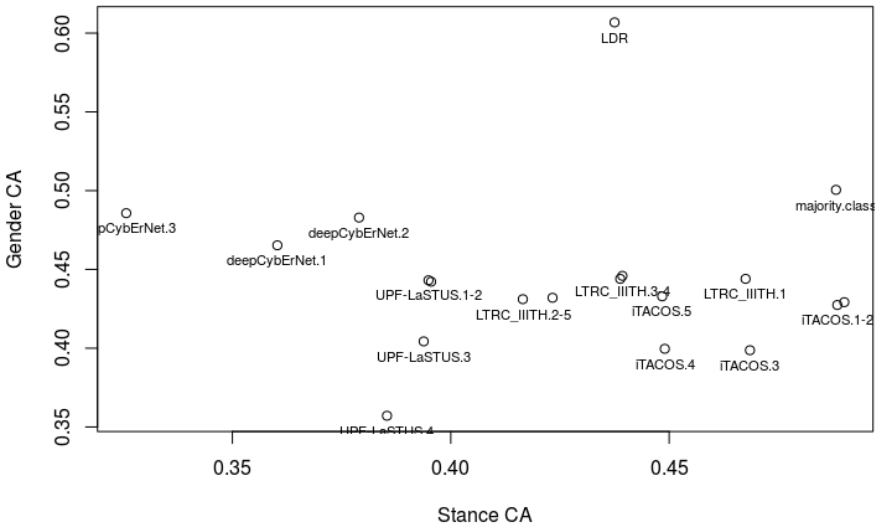

Fig. 3. Stance vs. Gender performances for Catalan.

6.4 Analysis of Error

In this section we analyse errors in stance detection based on the author's gender. We observed two kinds of errors: *i*) the participants interpreted a stance as being "in favor" when the real value was "against" (F ->A); and *ii*) the participants interpreted "against" when it was actually "in favor" (A ->F). We analyse the error rate for these two kinds of error depending on the gender of the author who wrote the tweet. As can be seen in Table 6, in both kinds of errors the rate is higher when the tweets were written by males. The greatest difference occurs with error A ->F in Catalan with a difference of more than 8%. In the case of Catalan, such differences are highly significant (p-value equal to 4.24 and 5.33 respectively). In the case of Spanish, they are only significant when the type of error is F ->A (p-value equal to 2.16). In the case of error type A ->F, the results are only statistically different at level 0.05 (p-value equal to 1.38).

In the case of Catalan, the A ->F error rate is higher, than in Spanish, where it is close to 2%. This may be due to a bias resulting from the difference in the number of tweets classified according to the sentiment expressed: there is

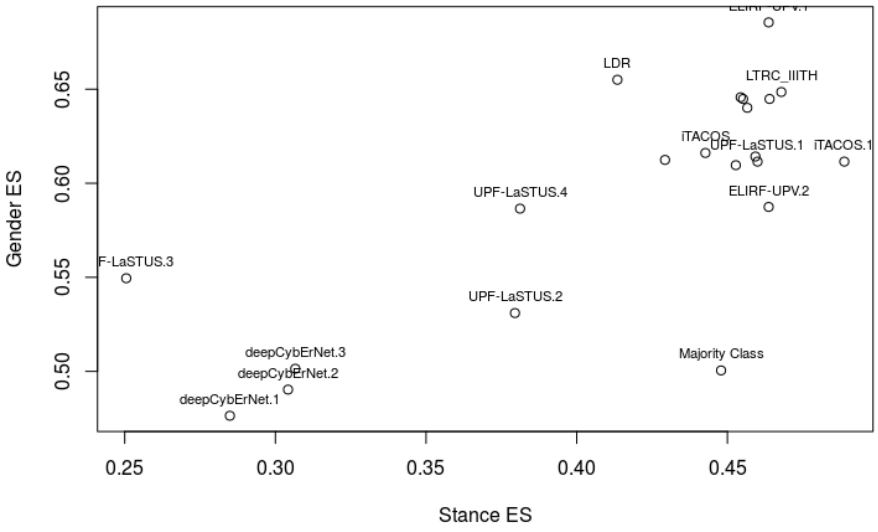

Fig. 4. Stance vs. Gender performances for Spanish.

a higher number of tweets in favor of independence written in Catalan, whereas there is a higher number of tweets against independence written in Spanish.

Table 6. Percentage of error types depending on the gender.

	Catalan		**Spanish**	
Gender	F ->A	A ->F	F ->A	A ->F
Female	12.34%	41.38%	39.07%	1.66%
Male	14.48%	59.00%	43.28%	2.01%

Tables 7 and 8 show tweets that were wrongly classified more often. The tables show five examples per gender, with females examples at the top, and males at the bottom. Taking into account the results shown in Table 6, we can say that it seems more difficult to detect stance for male tweets.

Considering that the average agreement percentage obtained in the inter-annotator agreement test is moderate (around 79%), probably there exists a percentage of inconsistency in the training sets, which could explain the moderate-low results obtained by the systems. Moreover, the analysis of the 40 tweets in Tables 7 and 8, namely those that were wrongly classified more often, does not

Table 7. Tweets more frequently misclassified in Catalan for both Females (top) and Males (bottom).

%	**Favor ->Against**
34.48%	Bastanta por em fa l'actitut de @InesArrimadas @CiudadanosCs De què criden #libertat? #27S #Eleccions27S
31.03%	"@FinancialTimes: Independence parties win in Catalonia http://t.co/pOmcTAG70b" @InesArrimadas Prou de mentides. Ha guanyat el si. #27STV3
27.59%	En quina nit electoral parlen els números 1, 4 i 5? I la Carme Forcadell i la Muriel Casals? De floreros? #JuntsXsiLlistaCiutadana #27STV3
27.59%	Els polítics diuen "catalanes i catalans", per què? No em sento exclosa en el masculí... Eufòria pels resultats! #llengua #27S
27.59%	El Sí no ha estat aclaparador. Em sap greu de debó perquè ho desitjava, però la victòria que celebra @JuntsPelSi no és tal... I ara? #27S
27.59%	Bon dia Catalunya! ll*ll #27s #votar #araeslhora de @srta_borrat. Opina a: http://t.co/8WK4JrTOqj http://t.co/Siqnqvz01G
24.14%	@Bioleg @JuntsPelSi @cupnacional #hovolemtot
20.69%	Bon article d'@eduardvoltas resumint el #27S: Gran victòria independentista http://t.co/e5vlcc8W9z
20.69%	Bon dia, #catalunya. Com ho duis? #27S #27SCatRàdio #27S2015
20.69%	Bufen nous vents!! #catalunya #27S #muntanya #montaña #mountain #trekking #ig_catalonia... https://t.co/XEVU11L1ae
%	**Against ->Favor**
79.31%	#27S ???????? No volem independència. Visca Catalunya i visca Espanya ????
68.97%	#27S Unió té un problema, i es diu 3%. Au va!!!
62.07%	#Eleccions27S ERC + CiU perden 9 diputats i amb tot el suport mediàtic i el bombo i plateret d'aquests dies #QuinExit!
55.17%	Escoltar els crits "Cataluña es España" de Ciutadans i que se'm posi la pell de gallina #NO #independència
55.17%	Gràcies @JuntsPelSi pel resultat de @CiudadanosCs . Sou uns cracks! #eleccionescatalanas #27S
82.76%	Avui més que mai, Catalunya és Espanya. #27S
82.76%	BON DIA A TOTS ELS TONTOS DEL CUL QUE EM VOTARÀN. UN PETONET, IMBÈCILS!! #27S #GuanyemJunts http://t.co/YABQAUzdX1
82.76%	Catalans!!! Heu de follar més i votar menys!! #FollemJunts #27S #GuanyemJunts http://t.co/RZM3cUIsCU
82.76%	avui és el primer dia de la meva vida que he de dir amb tristessa que m'avergonyo de ser del meu poble. #elprat #27s @CiudadanosCs
79.31%	Avui votaré per les valencianes que porten anys de lluita perquè la nostra llengua i cultura seguisquen ben vives. #27S #somdelSud #SomPPCC

allow us to infer the reasons for the low performance of the systems. These facts

highlight the difficulty of this task, in which there is an important subjective component and the linguistic content of the tweets is very scarce.

In order to improve the results, we should probably work with a higher number of tweets, to take into account the information included in the links –to see whether they contribute to detect the stance of the tweet–, and to take into consideration other aspects such as the presence of irony and humor in the tweets. For instance, in our current research about stance and irony, we observed that tweets against independence tend to be more ironic than those that are in favor of independence, and that irony is more common in men than in women.

7 Conclusion

We described a new shared task on detecting the stance towards Catalan Independence and the author's gender in tweets written in Spanish and Catalan, the two languages used by users directly involved in the political debate. Unlike previous evaluation campaigns, we decided to perform stance and gender detection together as part of one single shared task. We encouraged participants to address both sub-tasks, but participation was also allowed only in stance detection, which constitutes the main focus of the shared task. Interestingly, we observed a clear trend showing that systems that participated in both sub-tasks and obtained better results for gender also did so for stance.

StanceCat was proposed for the first time at the IberEval evaluation campaign and was one of the tasks with highest participation in the 2017 edition. We received submissions from ten teams from five countries, collecting more than thirty runs, with systems utilizing a wide range of methods, features and resources. Overall, results confirm that stance detection of micro-blogging texts is challenging, with large room for improvement, as was also observed in the shared task organized at Semeval 2016 for English. We hope that the dataset made available as part of the StanceCat task will foster further research on this topic, also in the context of under resourced languages such as Catalan.

Acknowledgements

The work has been carried out in the framework of the SOMEMBED project (TIN2015-71147), funded by Ministerio de Economía y Competitividad, Spain. The work of the third author has been partially funded by Autoritas Consulting. The work of Cristina Bosco and Viviana Patti was partially funded by Progetto di Ateneo/CSP 2016 "Immigrants, Hate and Prejudice in Social Media" (S1618_L2_BOSC_01).

We would like to thank Enrique Amigó and Jorge Carrillo de Albornoz from UNED[10] for their help during the evaluation with the EVALL platform [3].

[10] `http://portal.uned.es`

References

1. Aineto-García, D., Larriba-Flor, A.M.: Stance detection at ibereval 2017: A biased representation for a biased problem. In: Notebook Papers of 2nd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval), Murcia, Spain, September 19, CEUR Workshop Proceedings. CEUR-WS.org, 2017 (2017)
2. Ambrosini, L., Nicolò, G.: Neural models for stancecat shared task at ibereval 2017. In: Notebook Papers of 2nd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval), Murcia, Spain, September 19, CEUR Workshop Proceedings. CEUR-WS.org, 2017 (2017)
3. Amigó Cabrera, E., Carrillo-de Albornoz, J., Gonzalo Arroyo, J., Verdejo Maillo, M.F.: Evall: A framework for information systems evaluation (2016)
4. Barbieri, F.: Shared task on stance and gender detection in tweets on catalan independence - lastus system description. In: Notebook Papers of 2nd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval), Murcia, Spain, September 19, CEUR Workshop Proceedings. CEUR-WS.org, 2017 (2017)
5. Bosco, C., Lai, M., Patti, V., Rangel, F., Rosso, P.: Tweeting in the debate about catalan elections. In: Proceedings of the International Workshop on Emotion and Sentiment Analysis (co-located with LREC 2016). ELSA, Portoroz, Slovenia (2016)
6. Chuliá, L.C.: Submission to the 1st task on stance and gender detection in tweets on catalan independence at ibereval 2017. http://stel.ub.edu/stance-ibereval2017/ team: Luser. spain.
7. González, J.A., Pla, F., Hurtado, L.F.: Elirf-upv at ibereval 2017: Stance and gender detection in tweets. In: Notebook Papers of 2nd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval), Murcia, Spain, September 19, CEUR Workshop Proceedings. CEUR-WS.org, 2017 (2017)
8. Lai, M., Cignarella, A.T., Hernandez-Farias, D.I.: itacos at ibereval2017: Detecting stance in catalan and spanish tweets. In: Notebook Papers of 2nd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval), Murcia, Spain, September 19, CEUR Workshop Proceedings. CEUR-WS.org, 2017 (2017)
9. Mohammad, S.M., Kiritchenko, S., Sobhani, P., Zhu, X., Cherry, C.: Semeval-2016 task 6: Detecting stance in tweets. In: Proceedings of the International Workshop on Semantic Evaluation. pp. 31–41. SemEval '16, ACL, San Diego, California (June 2016), http://aclweb.org/anthology/S/S16/S16-1003.pdf
10. R, V.: Submission to the 1st task on stance and gender detection in tweets on catalan independence at ibereval 2017. http://stel.ub.edu/stance-ibereval2017/ team: Deepcybernet. india.
11. Rangel, F., Rosso, P., Franco-Salvador, M.: A low dimensionality representation for language variety identification. In: 17th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing. Springer-Verlag, LNCS, arXiv:1705.10754 (2016)
12. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4th author profiling task at PAN 2016: Cross-genre evaluations. In: Balog, K., Cappellato, L., Ferro, N., Macdonald, C. (eds.) Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016. CEUR Workshop Proceedings, vol. 1609, pp. 750–784. CEUR-WS.org (2016), http://ceur-ws.org/Vol-1609/16090750.pdf

13. Swami, S., Khandelwal, A., Shrivastava, M., Sarfaraz-Akhtar, S.: Ltrciiith at ibereval 2017: Stance and gender detection in tweets on catalan independence. In: Notebook Papers of 2nd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval), Murcia, Spain, September 19, CEUR Workshop Proceedings. CEUR-WS.org, 2017 (2017)
14. Verdú, C.: Submission to the 1st task on stance and gender detection in tweets on catalan independence at ibereval 2017. http://stel.ub.edu/stance-ibereval2017/ team: Ateam. spain.
15. Wojatzki, M., Zesch, T.: Neural, non-neural and hybrid stance detection in tweets on catalan independence. In: Notebook Papers of 2nd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval), Murcia, Spain, September 19, CEUR Workshop Proceedings. CEUR-WS.org, 2017 (2017)

Table 8. Tweets more frequently misclassified in Spanish for both Females (top) and Males (bottom).

%	**Favor ->Against**
83.33%	Si como dijo @PSOE no era un plebiscito, porque ahora @sanchezcastejon dice que Mas ha perdido el plebiscito?? Mi no entender #marxem #27s
66.67%	Señora @InesArrimadas que dimisión pide si todavia no hay presidente?! #27S #CatalunyaIndependent #27STV3 ???????? ??????????????????????????????????????
61.11%	Ho acabes de dir, @Albert_Rivera: "Empieza una nueva política para España". #independència #27S #27STV3
61.11%	@_anapastor_ @InesArrimadas . No le han pasado bien los apuntes. Ganan #JuntsPelSi# con un doble apoteosico
55.56%	@InesArrimadas te equivoques nena. Donde ves la mayoría??? Bocazas #JuntesPelSi
62.50%	@Albert_Rivera @CiudadanosCs ha sido quien ha votado la ruptura de España y no la vieja política" #eleccionescatalanas
54.17%	#27STV3 en serio @Albert_Rivera @InesArrimadas @CiudadanosCs alguien os ha enseñado los resultados? Sabéis contar? http://t.co/ccajELgsE4
54.17%	Ahora @CiutadansCs pide nuevas elecciones que sean verdaderamente autonómicas. Al final sí eran un plebiscito? Decídanse #27S
54.17%	Que alguien le diga a Rivera Arrimadas que los reyes son los padres. #27S
52.08%	A los que decían que esto no era un plebiscito lo utilizan ahora al saber los resultados. Me encanta esa lógica. #27S

%	**Against ->Favor**
4.05%	#27S #L6cat. Es evidente que desde Madrid se sigue sin entender nada de nada. Que sordera, que ceguera...es surrealista
2.89%	#27STV3 CUP dice, no se costara un catalan sin comer 3 platos al dia, señor Mas yo no he comido! Pues NO te acuestes!
2.31%	Campeón: @Albiol_XG "Llevo en política muchos años. No he perdido nunca" 2012 471.681 2015 337.645 97% escrut #27STV3 http://t.co/PRSQ2QIA5F
2.31%	Hola @InesArrimadas Soy una más de las orgullosas personas simpatizantes de @CsTorredembarra y con este #Ciutadans25, http://t.co/tNby9XL6zV
2.31%	#27STV3 Pero la Cup no decia que no apoyaria un proceso sin mayoria de votos??????
5.85%	Pues yo quería una independencia de Cataluña,que así puedo decir que tengo familia en el extranjero. #YloqueMolaDecirEsoQue #democracia #27S
5.85%	Puedo entender el deseo de muchos independentistas pero el discurso de Romeva es el nuevo Alicia en el país de las maravillas. #27S
4.79%	CUP rechaza la Unión Europea (Prog #27S pág 13) Romeva: JxSí negociará reingreso con Unión Europea "desde dentro" #Catalunya #InesPresidenta
2.66%	@catsiqueespot no perdamos el rumbo. (Aunque una encuesta no es un referéndum) #CSQEP http://t.co/g3bfHdDtpX
2.66%	Ciutadans gritando: "España unida jamás será vencida" véase la regeneración política. #27Stv3 #CataloniaVotes