Facebook Reactions as Controversy Proxies: Predictive Models over Italian News

Angelo Basile*
University of Malta
Rijksuniversiteit Groningen

Tommaso Caselli**
Rijksuniversiteit Groningen

Flavio Merenda†
University of Salerno
Rijksuniversiteit Groningen

Malvina Nissim‡
Rijksuniversiteit Groningen

Discussion on social media over controversial topics can easily escalate to harsh interactions. Being able to predict whether a certain post will be controversial, and what reactions it might give rise to, could help moderators provide a better experience for all users. We develop a battery of distant supervised models that use Facebook reactions as proxies for predicting news controversy, building on the idea that controversy can be modeled via the entropy of the reaction distribution to a post. We create a Facebook-based corpus for the study of controversy in Italian, and test on it the validity of our approach as well as a series of controversy models. Results show that controversy and reactions can be modelled successfully at various degrees of granularity.

1. Introduction

Climate change, abortion, vaccinations are only some of the many topics on which people tend to be very opinionated. Nowadays, such opinions can be openly and widely expressed, even harshly, on social media networks and platforms, such as Facebook, Twitter, Disqus, and Reddit. In this sense, the Web has become a huge *agora* for the interaction between users/consumers/producers (prosumers) of information. Particularly over highly debated topics like the above, such interactions can become lengthy exchanges where opinions of the involved participants not only tend to remain unchanged, but also become more and more polarised towards extreme values. We call these situations *controversies* (Timmermans et al. 2017a).

Modeling and understanding controversies may be useful in many situations. Journalists and news agencies may pay additional attention in the framing of a certain news, government officials and policy makers may be more aware of the issues involved in specific laws, social media managers might be more careful, i.e. monitor controversial content, in order to avoid the spreading of hate speech, and the general public may benefit as well thanks to a reduction of the "filter bubble" effect (Pariser 2011).

* Faculty of ICT, University of Malta, MT & CLCG, Rijksuniversiteit Groningen Groningen, NL, a.basile@student.rug.nl
** CLCG, Rijksuniversiteit Groningen, Groningen, NL t.caselli@{gmail.com|rug.nl}
† University of Salerno, Salerno & CLCG, Rijksuniversiteit Groningen, Groningen, NL f.merenda@rug.nl
‡ CLCG, Rijksuniversiteit Groningen, Groningen NL m.nissim@rug.nl

Recently, computational approaches on controversy detection (Pennacchiotti and Popescu 2010; Awadallah, Ramanath, and Weikum 2012; Mejova et al. 2014; Dori-Hacohen and Allan 2015; Borra et al. 2015; Lourentzou et al. 2015; Garimella et al. 2016; Coletto et al. 2017) have been developed with varying degrees of success and purposes, and, most importantly, on different text genres and domains (i.e. news, social media, and web pages). A commonality of these contributions is the use of a variety of signals to collect the data, such as network structure, topics discussed/debated, edit words and reverses, presence of dispute tags, and specialized lexicons. While this is a reasonable approximation of indicators of controversies, these are also modelling decisions that may not originate directly from the users/producers as direct cues of controversies.

We suggest a more direct approach yet still weakly supervised (so cheap and portable), to modelling controversy. Exploiting the Facebook reaction feature, which lets users express one of six 'emotions' to a post, we build on the assumption that the more varied the reactions to a piece of news are, the more controversial the news is. In other words, rather than manually annotating a corpus, crowdsourcing one, or looking at the structure of the social medium, we use Facebook reactions, as a reasonable signal in the data as a proxy for annotations (Pool and Nissim 2016). We then train a model for predicting the degree of controversy which may be associated to a news, and by extension to an event. On the basis of this model, we predict controversy-related aspects at more or less fine-grained degrees of granularity.

Contributions. This paper offers the following contributions. We provide a corpus for reaction and controversy prediction in Italian, which we make available on request for research purposes. We introduce the idea of *entropy as controversy* building on which we create a variety of models to account for controversy prediction at different levels of analysis and granularity. Specifically, we develop four models:

- a regression model that given a text will predict *the degree of controversy*, based on a measure of entropy; this is also the base model that validates the idea of "entropy as controversy" and on which the other models are built;
- a regression model that given a text will predict *the volume of reactions* to it. This is important as posts with few reactions might not need any attention by moderators or anyone else;
- a multi-label regression model that given a text will predict *the proportion of reactions* (out of six) associated to it;
- a multi-label classification model that given a text will predict *which reaction(s)* (of six) will be associated to it.

While optimally these models would all be dependent and integrated with one another to form a complete processing pipeline, in this paper we explore the (theoretical) feasibility and performance of each separate one, and only suggest their combination in the final discussion towards future work.

We focus on Italian, but since all models are built using *distant supervision*, the methods are completely language independent and can be reproduced for any language for which news are available on Facebook.

The remainder of the paper is structured as follows: Section 2 illustrates the rationale behind this work, namely the idea of using the entropy of Facebook reactions as a proxy for controversy. Section 3 describes the corpus of Italian news that we have

collected, and that we use for our experiments in modelling controversy. These are presented in Section 4. Finally, we discuss related work in Section 5 and draw our conclusions in Section 6. Data and code are made available at `https://github.com/anbasile/predictingcontroversy/tree/multi_class`.

2. Entropy as Controversy

Under the paradigm of distant supervision, so-called silver labels can be obtained for a corpus by exploiting some clues in the data. This has been done for example for relation extraction (Mintz et al. 2009), semantic role labelling (Exner, Klang, and Nugues 2015), sentiment analysis (Go, Bhayani, and Huang 2009), and emotion detection (Pool and Nissim 2016). For the latter, Facebook reactions have proved reliable proxies for annotations. In this work, we extend this principle by taking the *distribution* of Facebook reactions as a signal for the degree of controversy a given post raises (see Section 3 for more details on Facebook reactions).

On the basis of the definition of controversy previously introduced, our working hypothesis is that if users' reactions fall in two or more emotion classes (not necessarily opposed in terms of "polarity") with high frequencies, the controversy of a news item is higher. Building on this, we assume that entropy can be explanatory in modeling news' controversy: the higher the entropy, the more controversial the news. To better clarify this aspect, consider the data in Table 1. Each sample is the text of a Facebook post, for which we report the reaction breakdown (including `LIKE`), and its overall entropy based on reaction counts. Users expressing different reactions suggest that a text is more or less likely to be controversial. For instance, the difference in entropy values between the examples in Table 1 suggest a different level of controversy of a news item. [1]

Table 1
Sample rows from the dataset showing how entropy varies in relation to the reactions.

ID	TEXT	LIKE	ANGRY	HAHA	WOW	SAD	LOVE	entropy
1.)	Come in un reality, Trump nomina Neil Gorsuch in diretta web.	0	25	0	0	0	16	0.9649
2.)	Era la notizia che la famiglia Cucchi attendeva da tempo	5	604	0	0	31	25	0.5632
3.)	In casa c'erano Ornella, l'anziana madre, i nipoti: tutti fuori	8	439	21	109	6	35	1.3384
4.)	#MaxBiaggi lascia l'ospedale entro 2 giorni.	75	0	0	0	0	4	0.2890

To further verify the soundness of entropy as an indicator of controversy, we inspected the top-10 and bottom-10 news of the Italian Controversy Corpus (I-CONTRO v1.0, see Section 3), sorted by entropy (high values on top, high controversy), and manually assigned them to a topic. Table 2 illustrates the results for the top 5 and bottom 5 posts, in terms of entropy score. In addition to identifying a different distribution of topics according to degrees of controversy, we also observed that in some cases, the entities and the specific event mentions interact to generate controversy, supporting findings

1 The translations of the examples in Table 1 are the following: 1.) *Like in a reality show, Trump nominates Neil Gorsuch live on the Web.*; 2.)*It was the news that the Cucchi family was long waiting.*; 3.) *In the house, there were Ornella, the old mother, the nephews: all out.* 4.) *#MaxBiaggi leaves the hospital in 2 days.*

in previous works (Timmermans et al. 2017b). For instance, in the case of the "*25th April*" topic[2], the controversial news involves a political actor (i.e. *ANPI*, the National Association of Italian Partisans), and divisions on the celebration of this day, while the non-controversial news reports on museums being open on that day. The entropy score appears to capture this distinction.

Table 2
Sample of entropy-ranked top-5 and bottom-5 posts.

	TOPIC	TEXT
TOP	Incident	Fuggono dall'aereo in fiamme ma si fermano per scattare un selfie a pochi metri dall'aereo
	25th April	#25Aprile #Anpi: ""Festa di tutti gli italiani"". Roma divisa, due celebrazioni
	Gender/LGBTQ	"Genere: Sconosciuto". E il Canada gli dà' ragione
	Immigration	Emergenza #migranti, nave Rio Segura arrivata a Salerno. A bordo 11 donne incinte, 256 minori e 13 neonati #FOTO
	Animals	#Piacenza, abbattuto il cinghiale #Agostino. Da giorni nel parco urbano di Galleana, avrebbe caricato il personale
BOTTOM	25th April	#25aprile, ecco i musei statali aperti'
	Movies	"La La Land" meritava la statuetta del miglior film, andata poi a "Moonlight"?'
	Sport	Il Presidente della Sampdoria Massimo Ferrero è raggiante per la vittoria nel derby di Genova'
	Arts	Quando Eugenio Corti morì, il 4 febbraio 2014, Sébastien Lapaque, sul quotidiano parigino Le Figaro, lo definì "uno degli immensi scrittori del nostro tempo"'
	Arts	New York New York ricostruisce i legami artistici dal '28 a metà anni '60"

3. The Italian Controversy Corpus

We created the Italian Controversy Corpus v1.0 (I-CONTRO). We used the Facebook Graph API[3] to download news headlines, i.e. Facebook posts, from four major Italian newspapers. Of these, two are slightly politically biased (*Corriere della Sera* and *La Repubblica*, both center/center-left), two openly biased ones (*Il Manifesto*, left-wing and *il Giornale*, right-wing), and one news agency (*ANSA*). Since February 2016, Facebook users can react to a post not only with the well known and standard `LIKE`, but by choosing from a set of 5 emotions: `ANGRY`, `HAHA`, `WOW`, `SAD`, and `LOVE`. Each news is further enriched with all users' reactions.

The corpus currently covers a time period of two months an a half, namely from mid-April to early July 2017. Posts with less than 30 reactions in total were discarded. For each post, we collected: i.) the link to the full article on the source's website, if

2 April 25th is a national holiday in Italy to celebrate the end of World War II.
3 `https://developers.facebook.com/docs/graph-api`

available; ii.) an excerpt of the article (the variable `text`); iii.) additional text commenting the article, when available (the variable `descriptor`); and iv.) the full list of users' reactions. Finally, for a portion of the posts (1024 out of 3595, i.e. 28,48%) we downloaded the entire text of the article (the variable `body`). The full text of the article is not always available or accessible. We made sure that in this version of the dataset, for each source, the same number of posts for which the full body could be downloaded (i.e. 1024) is available. This constraint did not apply to *ANSA*. [4]

Table 3 provides an overview of the data collected, including, for each source, the number of Facebook posts, the number of tokens, the token-post ratio, i.e. the number of tokens per post, the average number of reactions per post, and, finally, the average entropy.

Table 3
Dataset shape and average entropy score (avg H) per source.

SOURCE	# POSTS	# TOKEN	R.TOKEN-POST	AVG. REACTION	AVG H
AgenziaANSA	883	18,635	21.10	392.28	1.0216
corrieredellasera	594	23,811	40.08	572.16	0.9135
ilgiornale	1,022	8,665	8.47	247.19	1.1266
ilmanifesto	752	36,479	48.5	367.56	0.6195
repubblica	344	7,763	22.56	1,451.39	0.9078
total	3,595	95,353	26.52	606.10	0.9386

The average entropy of the dataset is 0.9386, with a standard deviation of 0.4. The average number of reactions is 606.10 with a standard deviation of 486.59. *La Repubblica* is the news source with the highest average of reactions per post (1,451.39), while *Il Giornale* has the lowest one (247.19).

The corpus contains almost 2 millions (1,751,849) reactions. Quite unexpectedly, `ANGRY` is the most frequent reaction rather than `LIKE`, while `HAHA` is lowest one. Table 4 illustrates the distribution of the different reactions in each component of the I-CONTRO corpus, i.e. the news source, as well as the total count in the full corpus.

Table 4
Reaction counts in the I-CONTRO corpus v1.0.

SOURCE	LIKE	LOVE	HAHA	WOW	ANGRY	SAD
AgenziaANSA	251,238	13,907	21,689	10,395	22,261	31,213
corrieredellasera	20,151	18,446	7,411	12,160	248,115	42,735
ilgiornale	5,802	39,458	4,116	23,047	181,440	7,174
ilmanifesto	8,041	12,863	770	4,606	248,236	14,393
repubblica	25,326	23,500	11,175	17,203	398,112	26,866
total	310,558	108,174	45,161	67,411	1,098,164	122,381

The distribution of reactions varies per news source, supporting the intuition that communities on a different political spectrum have different reactions to news. We can

4 We may obtain the full body of the article by scraping the associated URL, however, this operation will open possible issues on copyright infringements related to the distribution of the corpus.

observe that in all newspapers ANGRY is the reaction which is by far preferred by the readers rather than the more neutral LIKE. Such a picture is completely the opposite when looking at the reactions associated with posts from ANSA, where LIKE is the most frequent reaction (251,238). This suggests that different ways of presenting, i.e. framing, the news to the public may affect the reactions of the readers/users: ANSA, being a news agency, seems to use more neutral frames with respect to the other news sources. At the moment, it is not possible to provide a comparison among sources on the same news item due to the fact that the data are not clustered per news topic or event.

Finally, we have identified a total of 49 unique reaction patterns. The most frequent pattern is the one which includes all reactions, i.e. LIKE_LOVE_ANGRY_HAHA_SAD_WOW, representing 18.05% of all reactions' patterns. Patterns containing only one reaction occur but are limited to three cases only: in particular, ANGRY, representing 10.45% of all patterns, LIKE, 2.02%, and HAHA, 0.01% [5]. Looking at the number of reactions used to compose patterns, we can observe that most frequently 3 components are involved (16 out of 49), followed by 4 (14 out 49), 2 (9 out of 49), and finally five (6 out of 49).

4. Experiments

We view the controversy prediction problem from several perspectives, and thus we cast a few separate, but related, tasks. Each of them captures a different aspect of this problem. First, building on the principle described in Section 2, we proxy controversy prediction as entropy prediction over the Facebook reactions, and cast it as a regression task (Section 4.1). We then expand on this problem in two directions. One is a more coarse-grained task, aimed at predicting the *volume* of reactions, without further analysis of their content (Section 4.2). This is also seen a regression problem. The other one is more fine-grained, and is aimed at predicting the actual distribution of the reactions, in proportional terms (Section 4.3). This is cast a multi-label regression problem. A lighter version of this is a multi-label classification task, where we predict which reactions will be given to a post, but not their distribution (Section 4.4). This can be particularly interesting in terms of patterns observed in association to posts. While such distribution can be derived from the multi-label regressor's output, simply obtaining the label distribution is an easier task and it might be more accurate.[6]

We take the core experiment to be the one on predicting controversy as indicated by entropy, as this is also our proof of concept. For this reason, and in the limited context of this contribution, we develop and tune our regression model on this task. Further, we take the best settings and use them also as such in the other tasks. This might be non-optimal, but further dedicated tuning and task integration is left for future work.

4.1 Predicting Entropy as Controversy

For predicting the entropy of the reactions to a given text, we built a system using a sparse feature representation and a linear regressor, with the *scikit-learn* Linear Regression implementation (Buitinck et al. 2013).

5 HAHA alone occurs only 1 time in the full dataset.
6 This problem is further discussed in Section 4.4 and compared results are shown in Table 10.

Settings. We use the *ANSA* dataset for development. The rationale behind this is that, being *ANSA* a news agency, the texts should be more objective and the controversy should depend on the event itself rather than by its framing in a specific, potentially biased, community (it is indeed the case that in the *ANSA* dataset the most frequent reaction is `LIKE` rather than `ANGRY`). The best settings are then used to cross-validate the full dataset (with and without *ANSA*), as well as each Facebook page separately in some of the experiments.

Baseline. As baseline, we use a dummy regressor which always predicts the mean entropy as observed on training data: considering that the values range between 0 and 2.9, with a standard deviation of 0.4, a system that always predicts the mean entropy is already performing reasonably well. Furthermore, this is in line with the average entropy values of each dataset, ranging from 0.6195 (Table 3, *Il Manifesto*) up to 1.1266 (Table 3, *Il Giornale*). As the problem is cast as a regression task, we use mean squared error (MSE) to measure the performance of our system.

Features. We used a tf-idf vectorizer to represent the text as both word and character n-grams. As sentiment might contribute to controversy prediction (Dori-Hacohen and Allan 2015; Timmermans et al. 2017b), we also extended the features with coarse-grained prior polarity information derived from Sentix (Basile and Nissim 2013), a resource for Italian automatically mapped from the English SentiWordNet (Esuli and Sebastiani 2006). We represent each token with the absolute values of its polarity (which in Sentix ranges from -1 to +1). This allows us to ignore the specific positive/negative values, and get a more abstract representation on the subjectivity relevance of a token: high values indicate that the text is rich of subjectivity relevant tokens; 0 means that the text is merely objective. For each post we then compute the average polarity and encoded it into a separate vector. Missing words in the lexicon are simply skipped.

Model development. For development, as mentioned, we only used *ANSA*. We experimented with different features and different sizes of texts. In particular, we ran experiments using: i.) only the `text` variable; ii.) a combination of the `text` and the `descriptor` variables; and iii.) a combination of the `text`, the `descriptor`, and the `body` variables. Furthermore, these three basic settings have been extended with the polarity values from Sentix. To fine tune the parameters, a grid-search of the model using a 10-fold cross-validation was conducted. Table 5 reports the results of the different models as well as of the baselines.

Table 5
Results for the cross-validated *ANSA* dataset.

DATA	BASELINE	MODEL	+ SENTIX
text	0.24	0.154	0.155
text+descriptor	0.24	0.146	0.148
text+descriptor+body	0.24	0.146	0.148

The best model shows an improvement of 0.094 MSE with respect to the baseline when extending the variable `text` with `descriptor` and `body`. The use of the variable `text` alone still beats the baseline, but obtains a lower score than the models which include both the `descriptor` and the `body` variables. The extensions with the polarity

scores from Sentix decrease the model's performance (though still outperforming the baselines). These results are in line with the findings reported in (Mejova et al. 2014), who show that SentiWordNet is one of the sentiment resources with the lowest distribution of polarity oriented words in controversial topics. This calls for better and more context-oriented sentiment lexicons in Italian. Table 6 summarizes the features of the best model, which is based on a combination of the three text variables only: `text`, `descriptor`, and `body` (whenever available), represented as word and character n-grams, ignoring the polarity vectors. This model was used on the remainder of the datasets, even though the variable `body` does not seem to yield any improvements over the `text+descriptor` model. Future work will also explore whether the contribution is instead visible for the subset of cases where body is available.

Table 6
Best model's settings and features.

PARAMETER	VALUE
character ngrams	(2,4)
character binary features	True
character normalization	l2
character sublinear tf	False
word ngrams	(1,3)
word binary features	False
word normalization	l2
word sublinear tf	True

Table 7
Cross-validated results for controversy prediction on all datasets.

	BASELINE	STD	MODEL	STD
ilgiornale	**0.21**	0.03	0.22	0.04
ilgiornale+ansa	0.23	0.04	**0.19**	0.03
ilmanifesto	0.15	0.04	**0.11**	0.04
ilmanifesto+ansa	0.24	0.04	**0.14**	0.03
repubblica	0.22	0.07	**0.18**	0.07
repubblica+ansa	0.24	0.04	**0.15**	0.04
corrieredellasera	0.24	0.06	**0.16**	0.06
corrieredellasera+ansa	0.24	0.03	**0.14**	0.04
full_dataset	0.24	0.02	**0.17**	0.03
full_dataset-ansa	0.24	0.03	**0.17**	0.04

Results on the test set. Table 7 illustrates cross-validated results for each separate newspaper source, as well as for the full dataset (newspapers + *ANSA*). With the exception of *Il Giornale*, our model always beats the baseline, confirming the feasibility of our approach. Extending the newspaper dataset with the data from *ANSA*, we can observe

a reinforcement of the predicting power of the model, with a range between 0.04 to 0.1 points with respect to the corresponding baselines. The positive effect on *Il Giornale* dataset can be due to an extension of the number of tokens, since *Il Giornale* is the dataset with the lowest token-post ratio (8,47 tokens per post), which clearly affects our model.

While we are able to capture entropy with reasonable accuracy, this model has however the following limitations:

- it does not account at all for the actual volume of reactions. Thus, a post with 1000 reactions and a post with 40 reactions which exhibit a similar entropy will be modelled in the same way;
- it does not account for the actual distribution of reactions, though all patterns are obviously not the same. A post with a 70/30 `LOVE/HATE` distribution will be modelled in exactly the same way as one with a 30/70 `LOVE/HATE` distribution. This is still acceptable in terms of controversy, but it provides partial information;
- it will model in the same way equal distribution of reactions that can be positioned on opposite values. For instance, assuming two posts with an equal distribution of reactions distributed over pairs of opposing value, i.e. `LOVE/HAHA` for the first post, and `LOVE/HATE` for the second post, our approach will model them in the same way, while, intuitively, one might assume that the meaning behind such distributions is not the same.

In order to address such limitations, we developed three additional models. One where we predict the volume of reactions (Section 4.2), one with which we predict the proportion of each reaction to a post (Section 4.3), and one that we use to predict presence of absence of each reaction to a post, ignoring though their relative proportion (Section 4.4). While the first two are cast as regression problems, the last one is a multi-label classification task.

4.2 Predicting Reaction Volume

Volumes of reactions can be seen as a component of controversies. Recalling our definition of controversy, the volume of reactions is a good proxy for the number of participants (more reactions, more readers/users).

We cast this problem as a regression task, too, and as mentioned above, we use the same settings of the best model developed for controversy prediction. In order to get more realistic results, we add to our corpus approximately 50% more posts (from the same pages) which have less than 30 reactions each. This gives rise to a new dataset, used only in this experiment, of 5,223 posts, with an average reaction number per post of approximately 340, a minimum of 1 and a maximum of 30,274 (which is also the maximum of the I-CONTRO corpus).

To account for the wide variation in reaction counts across posts, we normalise the data by taking the natural log of the real values, as this allows for more accurate modelling. The predicted log value is then raised to the log power to obtain the predicted actual counts.

Results obtained via 10-fold cross-validation for our linear regression model show a MSE of 1.48 (with a standard deviation of 0.18), against a MSE of 1.54 (with a standard deviation of 0.24) for the baseline model. In Figure 1, we report the plot for the first ten examples of the full datasets (newspapers plus *ANSA*) of the original (gold) down-

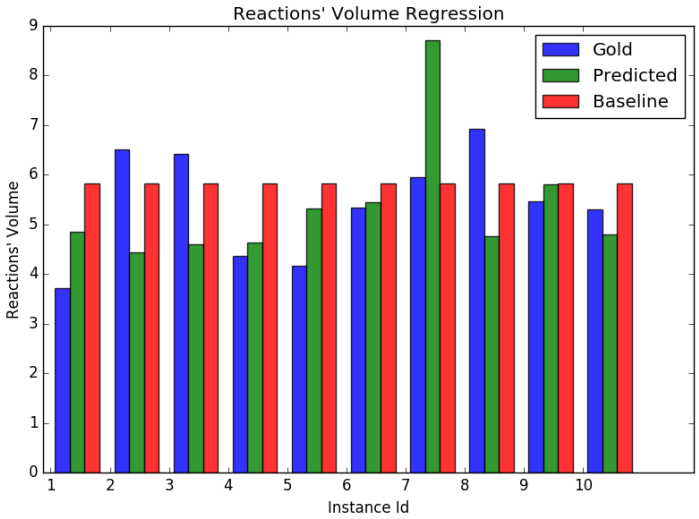

Figure 1
Plot of the reactions' volume, including, the downloaded reactions (gold, blue bar), the predicted (green bars), and the baseline (red bars), for the first 10 posts of the full cross-validated dataset (including ANSA). Volumes are expressed by the natural log.

loaded reactions (blue bar), the predicted volumes with the linear regression model (green bar), and the baseline model (red bar).
As the plot in Figure 1 shows, the model approximates the actual volumes of the reactions more accurately than the baseline. Please, note that the closer the bar to the gold, the more accurate the prediction. In Figure 1, we can see that the baseline performs better than our model only in four cases, namely instances 2, 3, 7 and 8. Predicting the volume of reactions would be valuable information for a moderator who should probably pay more attention to high- rather than low-volume posts, and could also be a valuable feature in further models of controversy. However, this task will only provide us with an indication of the "popularity" of a news item, rather than its controversy, as the two aspects are not necessarily the same. To support this statement we ran a Pearson correlation test between the entropy and the volumes, gold as well as predicted, of these 10 examples. We observe in both cases a weak correlation with entropy (0.4342 with the gold volumes and 0.2497 with the predicted ones), and in both cases the correlation is not statistically significant (p>0.05).

4.3 Predicting Reaction Distribution

One of the limitations listed above is the inability of the controversy prediction model to distinguish *which reactions* are involved, and in *what proportion*. Thus, using again the parameters and features of the controversy predictor, we train a multi-label linear regression model that associates to a post a set of reactions and their relative weight, i.e., the proportion of each reaction.

The reactions' real values are normalised to a 0–1 range, so as to reduce differences in the distributions among the reactions across posts which might have highly varying total volumes. As a baseline, we use again a dummy regressor which always predict

the average value (i.e. proportion) per reaction. We apply 10-fold cross-validation to the whole dataset and obtain an MSE of 0.02, while the baseline's MSE is 0.04.

To better illustrate this task and its results, consider the examples in Table 8, which correspond to the first four posts of the plot in Figure 1 and Table 1 (IDs 1–4). The GOLD columns contains the proportion of the actual reactions, while PREDICTED column those obtained from the model. The order of the reactions is the following: LIKE, ANGRY, HAHA, WOW, SAD and LOVE.[7] For further illustration and more immediate comparison, we report in Figure 2 the gold (G) and predicted (P) plots of the four examples.

Table 8
Sample rows from the dataset showing the label array for multi-label regression.

ID	GOLD	PREDICTED
1.)	[0.0, 0.6097, 0.0, 0.0, 0.0, 0.3902]	[0.02669, 0.8267, 0.0444, 0.0217, 0.0138, 0.0664]
2.)	[0.0075, 0.9082, 0.0, 0.0, 0.04661, 0.0375]	[0.0716, 0.7189, 0.0117, 0.0, 0.1109, 0.0957]
3.)	[0.0129, 0.7103,0.0339, 0.1763, 0.0097, 0.0566]	[0.0302, 0.7179, 0.0259, 0.1419, 0.0663, 0.0176]
4.)	[0.9493, 0.0, 0.0, 0.0, 0.0, 0.0506]	[0.9477, 0.0, 0.0097, 0.0, 0.0, 0.0425]

Example 4.) is an interesting case. First of all, ANGRY is correctly not predicted as present, although this is by far the most frequent reaction, thus showing the robustness of the system. Furthermore, if we look at the reactions which were predicted as present, excluding LIKE, we see that they correspond to positive reactions, namely HAHA and LOVE, which is in line with the content of the post (a famous motorcycle road racer about to leave the hospital after an accident).

4.4 Predicting Reaction Patterns

Specific reaction patterns could be associated to certain event types. For example, happy events would only have certain reactions, while sad ones would have different ones. The pattern of reactions could thus be useful in itself, even independently of the actual proportions of the reactions. We cast this problem as a multi-label classification task where we aim at predicting the presence or absence of each label to a given post. For example, the four posts that we have examined so far, would have the gold labels as reported in Table 9.

Table 9
Sample rows from the dataset showing the label array for multi-label classification.

ID	LIKE	ANGRY	HAHA	WOW	SAD	LOVE	GOLD	CLASSIFIER	BINARIZED REGR
1.)	0	25	0	0	0	16	[0,1,0,0,0,1]	[1,1,1,0,0,0]	[1,1,1,1,1,1]
2.)	5	604	0	0	31	25	[1,1,0,0,1,1]	[1,1,0,0,1,1]	[1,1,1,0,1,1]
3.)	8	439	21	109	6	35	[1,1,1,1,1,1]	[1,1,1,1,1,1]	[1,1,1,1,1,1]
4.)	75	0	0	0	0	4	[1,0,0,0,0,1]	[1,0,1,0,0,1]	[1,0,1,0,0,1]

A dummy baseline will always assign all reactions to all posts (the most frequent pattern), thus yielding 100% recall, but lower precision. Our model, a LinearSVM which

7 Some of the values of the regressor are negative, although the overall sum corresponds to 1. In the table we have reported as 0 the negative values and subtracted a proportion of the negative scores from all reactions with a positive proportion.

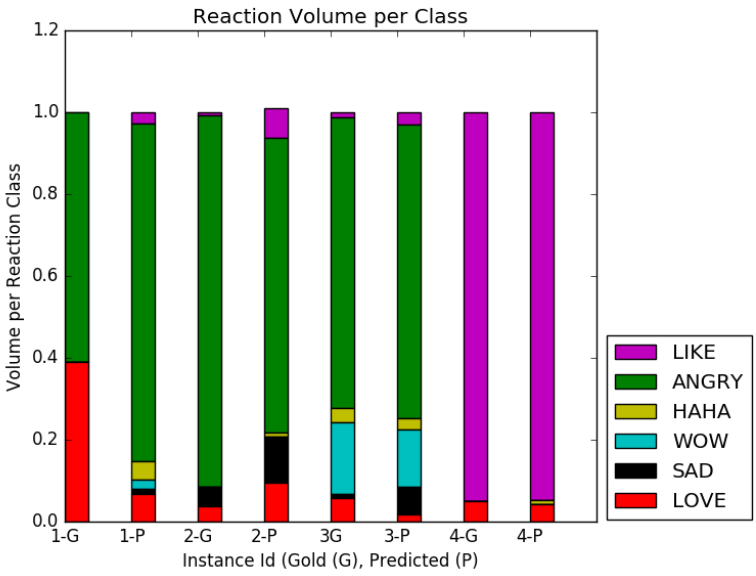

Figure 2
Plot of the reactions' volume per reaction class of the examples in Table 8. Each gold instance (G) is paired with the predicted one (P). Volumes are normalized between 0 and 1.

uses the same features and settings as the controversy regressor, achieves better precision than the baseline on all reactions, as we show in Table 10 [8]. Results are obtained via 10-fold cross validation. Although our model does not always achieve better F-score than the baseline, the improvement in precision is a positive aspect, especially in the perspective of using this information in support to moderators of social media.

It could be observed that this information can be easily obtained by converting the proportions predicted by the multi-label regressor described in Section 4.3. However, a classifier that does not need to predict the weight of each reaction has a simpler task, thus less prone to error, and can provide the desired reaction pattern anyway. For comparison, we took the output of the multi-label regressor, and converted the predicted proportions for each reaction to binary values (1 for all predictions > 0, 0 otherwise), and evaluated them against the gold standard (see last column in Table 9 for a representation example, and the "binarized regr" column in Table 10 for the results).

From the results in Table 10, we can indeed observe that the classifier has a better performance than the regressor on all labels. In terms of F-score, the picture is similar, with the classifier outperforming the regressor on all labels. This indicates that willing to know which reaction pattern will most likely be associated to a post independently on the reactions' relative weight, we are better off with a natural multi-label classification model rather than a conversion of labels from the regressor's output.

8 Note that the evaluation counts as correct only the presence of a reaction, but not the absence.

Table 10
Multi-label classification task for reaction prediction. We show the performances of the multi-label linear SVM classifier, the binarized assignments obtained from the multi-label linear regressor, and the baseline. Predictions are obtained via 10-fold cross-validation.

	CLASSIFIER			BINARIZED REGR			BASELINE		
	prec	rec	f1	prec	rec	f1	prec	rec	f1
LIKE	**0.81**	0.94	0.87	0.83	0.80	0.82	0.80	**1.00**	0.89
ANGRY	**0.95**	0.98	0.97	0.95	0.95	0.95	0.93	**1.00**	0.96
HAHA	**0.72**	0.83	0.77	0.68	0.83	0.75	0.65	**1.00**	0.79
WOW	**0.77**	0.86	0.81	0.70	0.88	0.78	0.67	**1.00**	0.80
SAD	**0.70**	0.78	0.74	0.64	0.85	0.73	0.61	**1.00**	0.76
LOVE	**0.77**	0.88	0.82	0.73	0.89	0.80	0.70	**1.00**	0.82

5. Related Work

Analysis and automatic detection of controversy has received attention in recent years, especially with the explosion of the Web 2.0.

Works specifically targeted to the automatic detection of controversies can be roughly grouped in two areas. The first group includes contributions which use structural features of the source data, like network structure, reply graphs, and patterns of interactions and interconnections (known as *motifs*) (Garimella et al. 2016; Coletto et al. 2017). Further specific structural features have been identified in the analysis of Wikipedia pages, such as edit activities (Borra et al. 2015; Dori-Hacohen and Allan 2015). The second group of works, on the other hand, focuses more on content analysis. In this latter group, we can identify different approaches, ranging from deep parsing of text snippets which expresses an attribution relation between a person, an opinion, and a topic (Awadallah, Ramanath, and Weikum 2012), to the analysis and use of rhetorical structure features (Allen, Carenini, and Ng 2014), or lexical features via specialized lexicons, such as sentiment lexicons or controversial lexicons (Pennacchiotti and Popescu 2010). In this work, we clearly follow a content-oriented approach to the analysis and prediction of controversies. The use of the Facebook reactions' labels as proxies of controversies has not been used before. Furthermore, the method we propose is simple, portable to other languages, robust, and domain independent, thus overcoming some of the criticisms of the language or content analysis approaches (Garimella et al. 2016; Coletto et al. 2017).

Very few are the works which do not assume a predefined set of controversial topics, or a specific domain (e.g. politics), as a starting point. Even works on Wikipedia pages uses as a starting point either a list of known controversial topics (e.g. abortion, gun control, ...) to identify the relevant articles (Dori-Hacohen and Allan 2015) or directly use the pages listed in the "Wikipedia's list of controversial articles" (Borra et al. 2015).[9] Detecting controversy "in the wild", i.e. without any predefined list of known controversial topics or restricted to specific domains, has received limited attention so far, and has been restricted to the social media domain, especially to the analysis of tweets (Garimella et al. 2016; Coletto et al. 2017). The problem we address in this

9 https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues

paper concerns the prediction of controversies in the news without any specific or predetermined domain, an aspect which has been particularly neglected so far.

A further work studied (Mejova et al. 2014) the correlation of distribution of sentiment oriented words, or biased terms, in news. With respect to our work, the authors use a crowdsourced lexicon of controversial terms to identify and aggregate news according to a controversy score, i.e. either as being controversial, somewhat controversial, and non-controversial. Their work, focused on political discussion, showed, for the first time and empirically, a significant correlation between controversial topics and the use of negative polarity-bearing words.

Models of controversies have been proposed, aiming at providing a unitarian framework of the different components identified in literatures, spanning from sociology to linguistics and computer science. One of such models, CAPOTE (Timmermans et al. 2017b), describes controversies (C) as multiple actors (A) involved in a polarized (P), public and open (O) debate persisting over time (T) and involving strong sentiments and emotions (E). Each aspect involved in a controversy is captured by the model, as well as their dependencies and relations. To identify whether a topic is controversial or not the authors ran a crowdsourcing experiment asking the workers to provide a judgment (i.e. yes or no question) on each of the components of the model over the first two paragraphs of 5048 news articles from the *Guardian* together with five comments. Their results shown that all variables are significant for predicting controversy. Furthermore, they also identified that the emotion component is the one which more strongly correlates with controversy with respect to all the other components. On the other hand, actors and openness show the weakest correlation.

Works in the areas of Sentiment Analysis (Zhou et al. 2013; Deng and Wiebe 2015; Deng, Choi, and Wiebe 2013; Chambers et al. 2015; Russo, Caselli, and Strapparava 2015), Emotion Detection (Strapparava and Mihalcea 2007, 2008; Russo et al. 2011; Pool and Nissim 2016), and Stance Detection (Mohammad et al. 2016) are, on the other hand, only partially related, to the controversy issue, as they focus on predicting/classifying the content of a message with respect to specific categories, such as "positive", "negative", "neutral", or "joy", "sadness" (among others), or as "being in favor" or "being against". They may be seen as necessary but not sufficient tools for detecting/predicting controversy (Timmermans et al. 2017a).

Finally, one critical aspect of previous work on controversies concerns evaluation measures and replicability of results. Currently, it is not possible to compare the different approaches proposed as no shared dataset is available. Even for works on Wikipedia pages, dumps of the version used are not reported. With this respect, our contribution makes data and code available, thus facilitating the comparison with other future work and replicability of results.

6. Conclusions and Future Work

We have created I-CONTRO v1.0, the first Facebook-based corpus for studying controversy in Italian. On I-CONTRO, we have developed and tested a battery of models that predict controversy of Facebook posts concerning Italian news with varying degrees of granularity.

Specifically, we have introduced the concept of *entropy as controversy* (the higher the entropy, indicated by highly mixed reactions, the bigger the controversy), and shown that a linear regression model can learn entropy (thus controversy) of posts with accuracy better than baseline. We are also able to approximate the expected volume of reactions to a post, which could come useful in view of directing moderators' attention

to specific posts. We can also successfully predict *which* reactions will be associated to a post, and the proportion of each reaction.

From a feature perspective, we observed that coarse-grained sentiment values are not useful, although this may depend on the quality of the lexicon employed. Further, the approach we have developed is based on discrete linguistically motivated features. This has an impact in the learned model as it is not able to generalise enough when dealing with low-frequency features and unseen data in the test set. To alleviate this issue, we are planning to model the post representations by using word embeddings.

From a more general modelling perspective, we have developed four separate models but they could—and should—be used in a more complex pipeline or at least to inform each other. For example, the predicted volume could be a relevant feature in modelling the degree of controversy. Also, controversy could be derived from the weighted label distribution provided by the multi-label regressors. The ways the models should be combined to optimise performance will be investigated in upcoming work.

Finally, a natural extension of this work is to expand the model to account for perspective bias in different communities (Basile, Caselli, and Nissim 2017). News from different sources may be aggregated per event type, for example via the EventRegistry API[10], allowing to explore entropy (and polarisation of reactions) on exactly the same event instance. A first step in this direction would be to detect and match Named Entities to approximately identify similar events. At the reaction-level, the obvious next step is to explore and experiment with *clusters* of reactions (for instance, positive (`LIKE`, `LOVE`, `AHAH`), negative (`ANGRY`, `SAD`), or ambiguous (`WOW`)), instead of treating them all as single and distinct indicators.

Another natural follow-up is to extend this work to other social media data, such as Twitter. Twitter does not allow for nuances in reactions in the same way that Facebook does, as only one kind of "like" is provided. However, the substantial use of hashtags and emojis might offer alternative proxies to capture a variety of reactions. There is plenty of work on the usefulness of leveraging hashtags as reaction proxies both at a coarse and finer level (Mohammad and Kiritchenko 2015), but this information, to the best of our knowledge, has not been used to predict likelihood of controversy.

References

Allen, Kelsey, Giuseppe Carenini, and Raymond Ng. 2014. Detecting disagreement in conversations using pseudo-monologic rhetorical structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1169–1180, Doha, Qatar, October. Association for Computational Linguistics.

Awadallah, Rawia, Maya Ramanath, and Gerhard Weikum. 2012. Opinions network for politically controversial topics. In *Proceedings of the first edition workshop on Politics, elections and data, PLEAD@CIKM 2012*, pages 15–22, Maui, HI, USA, October-November. ACM.

Basile, Angelo, Tommaso Caselli, and Malvina Nissim. 2017. Predicting Controversial News Using Facebook Reactions. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, Rome, Italy, December.

Basile, Valerio and Malvina Nissim. 2013. Sentiment analysis on Italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@NAACL-HLT 2013*, pages 100–107, Atlanta, Georgia, USA, June.

Borra, Erik, Esther Weltevrede, Paolo Ciuccarelli, Andreas Kaltenbrunner, David Laniado, Giovanni Magni, Michele Mauri, Richard Rogers, and Tommaso Venturini. 2015. Societal controversies in wikipedia articles. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015*, pages 193–196, Seoul, Republic of Korea, April. ACM.

10 `http://eventregistry.org`

Buitinck, Lars, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Chambers, Nathanael, Victor Bowen, Ethan Genco, Xisen Tian, Eric Young, Ganesh Harihara, and Eugene Yang. 2015. Identifying political sentiment between nation states with social media. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 65–75, Lisbon, Portugal, September. Association for Computational Linguistics.

Coletto, Mauro, Kiran Garimella, Aristides Gionis, and Claudio Lucchese. 2017. Automatic controversy detection in social media: A content-independent motif-based approach. *Online Social Networks and Media*, 3:22–31.

Deng, Lingjia, Yoonjung Choi, and Janyce Wiebe. 2013. Benefactive/malefactive event and writer attitude annotation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 120–125, Sofia, Bulgaria, August. Association for Computational Linguistics.

Deng, Lingjia and Janyce Wiebe. 2015. Joint prediction for entity/event-level sentiment analysis using probabilistic soft logic models. In *EMNLP*, pages 179–189.

Dori-Hacohen, Shiri and James Allan. 2015. Automated controversy detection on the web. In *Advances in Information Retrieval - 37th European Conference on IR, Research, ECIR. Proccedings*, pages 423–434, Vienna, Austria, March. Springer.

Esuli, Andrea and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 417–422, Genoa, Italy, May.

Exner, Peter, Marcus Klang, and Pierre Nugues. 2015. A distant supervision approach to semantic role labeling. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics, *SEM 2015*, pages 239–248, Denver, Colorado, USA, June.

Garimella, Kiran, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2016. Quantifying controversy in social media. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 33–42, San Francisco, CA, USA, February. ACM.

Go, Alec, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12).

Lourentzou, Ismini, Graham Dyer, Abhishek Sharma, and ChengXiang Zhai. 2015. Hotspots of news articles: Joint mining of news text & social media to discover controversial points in news. In *2015 IEEE International Conference on Big Data, Big Data 2015*, pages 2948–2950, Santa Clara, CA, USA, November. IEEE.

Mejova, Yelena, Amy X Zhang, Nicholas Diakopoulos, and Carlos Castillo. 2014. Controversy and sentiment in online news. *arXiv preprint arXiv:1409.8152*.

Mintz, Mike, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore, August. Association for Computational Linguistics.

Mohammad, Saif, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California, June. Association for Computational Linguistics.

Mohammad, Saif M and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

Pariser, Eli. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK.

Pennacchiotti, Marco and Ana-Maria Popescu. 2010. Detecting controversies in twitter: a first study. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pages 31–32, Los Angeles, California, USA, June. Association for Computational Linguistics.

Pool, Chris and Malvina Nissim. 2016. Distant supervision for emotion detection using facebook reactions. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 30–39, Osaka, Japan, December. COLING 2016.

Russo, Irene, Tommaso Caselli, Francesco Rubino, Ester Boldrini, and Patricio Martínez-Barco. 2011. Emocause: an easy-adaptable approach to emotion cause contexts. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2011)*, pages 153–160, Portland, Oregon, June. Association for Computational Linguistics.

Russo, Irene, Tommaso Caselli, and Carlo Strapparava. 2015. Semeval-2015 task 9: Clipeval implicit polarity of events. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 443–450, Denver, Colorado, June. Association for Computational Linguistics.

Strapparava, Carlo and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 70–74, Prague, Czech Republic, June. Association for Computational Linguistics.

Strapparava, Carlo and Rada Mihalcea. 2008. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560, Fortaleza, Ceara, Brazil, March. ACM.

Timmermans, Benjamin, Lora Aroyo, Evangelos Kanoulas Tobias Kuhn, Kaspar Beelen, and Gerben van Eerten Bob van de Velde. 2017a. Controcurator: Understanding controversy using collective intelligence. In *Collective Intelligence 2017*.

Timmermans, Benjamin, Tobias Kuhn, Kaspar Beelen, and Lora Aroyo. 2017b. Computational controversy. In Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri, editors, *Social Informatics*, pages 288–300, Oxford, UK, September. Springer International Publishing.

Zhou, Xujuan, Xiaohui Tao, Jianming Yong, and Zhenyu Yang. 2013. Sentiment analysis on tweets for social events. In *Proceedings of the 2013 IEEE 17th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 557–562, Whistler, BC, Canada, June. IEEE.