Learning Comment Controversy Prediction in Web Discussions Using Incidentally Supervised Multi-Task CNNs

Nils Rethmeier, Mark Hübner, Leonhard Hennig
German Research Center for Artificial Intelligence (DFKI), Germany
firstname.lastname@dfki.de

Abstract

Comments on web news contain controversies that manifest as inter-group agreement-conflicts. Tracking such *rapidly evolving controversy* could ease conflict resolution or journalist-user interaction. However, this presupposes controversy online-prediction that scales to diverse domains using incidental supervision signals instead of manual labeling. To more deeply interpret comment-controversy model decisions we frame prediction as binary classification and evaluate baselines and multi-task CNNs that use an auxiliary news-genre-encoder. Finally, we use ablation and interpretability methods to determine the impacts of topic, discourse and sentiment indicators, contextual vs. global word influence, as well as genre-keywords vs. per-genre-controversy keywords – to find that the models learn plausible controversy features using only incidentally supervised signals.

1 Introduction

Online discussion comments are exchanged in parallel, creating redundancy that prohibits discussions from developing beyond a superficial stage of confirming previously held opinions. Instead, Mahyar et al. (2017) recently demonstrated that focusing users on controversial comments – i.e. comments that cause *inter-group agreement-conflicts* (Dori-Hacohen et al., 2015) – helps speed up inter-group consensus finding leading to improved group decisions. However, their system (ConsensUS) uses manual controversy labels which can not capture rapidly evolving comment-controversy at scale or over diverse domains. Hence, to fully automate comment-controversy prediction systems we contribute the following solutions to a number of challenges. **(I)** We extend controversy prediction to *comment-level*, and to *German news discussions*. We evaluate topic, sentiment and discourse importance (Cramer, 2011) and analyze whether models plausibly capture controversy aspects using explainability methods (see Sec. 5.3). **(II)** We use comment vote-agreement to create an incidentally supervised (Roth, 2017) *controversy* signal as seen in Figure 1. Structural (output feature) signals like genre, are predicted by a sub-encoder (see Sec. 4) rather than required as input. **(III)** Sentiment and discourse *input feature* creation work on any tokenizable language (see Sec. 3).

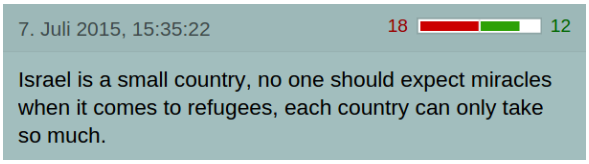

Figure 1: A comment is assumed controversial if its up and down votes show no clear $^2/_3$ majority decision.

2 Related Research

Since predicting user agreement-conflicts *upon web news* comments is a special case of controversy prediction, we list in the following related works that: (a) learn to predict controversy, using (b) incidental supervision, and (c) work on online (news) discussions. Chen et al. (2016) visualized *controversial words* using dissimilarities in pro vs. contra argument embeddings. Garimella et al. (2017) identified *controversial topics* using bipartite Twitter follower-graphs, while Dori-Hacohen and Allan (2015) proposed an incidentally supervised binary classification to predict controversial topics via Wikipedia tags. Jang et al. (2016) used language modeling to predict controversial documents, based on earlier hypotheses by Cramer (2011): "that language in *news discussions* is a good indicator of controversy". Choi et al. (2010) focused on using sentiment polarity indicators and

Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 316–321
Brussels, Belgium, October 31, 2018.
https://doi.org/10.18653/v1/P17

subtopics, i.e. topically related phrases of nouns. *Vote-based learning signals* have been exploited by both Pool and Nissim (2016); Basile et al. (2017) who predict the sentiment distributions of news outlets or find controversial news pieces using Facebook-article emoticon-votes. Instead of predicting controversial topics (articles), we predict *controversial comments*, hence putting the focus on users (commentators) as curators of controversial content.

3 Incidental Supervision Signals

Controversy signal: We use comment vote-agreement ratios and news tags as incidental supervision signals (Roth, 2017) to label comments as controversial and by genre. Comments without a clear $^2/_3$ majority of either agreeing (up) or disagreeing (down) votes are considered controversial – i.e. of conflicted agreement. The ratio is calculated as $r = \min(up, down)/\max(up, down)$. Ratios below 0.5 mark a $2/3$ majority. Ratios above 0.5 mark conflicted agreement. We reduce *labeling noise* via two noise margins: (a) controversial comments must have a vote-ratio $r > 0.6$ and (b) that both the up-votes (group) and down-votes (group) should each have more than 2 votes. **Article Genre signal:** Predicting controversy without context structure is difficult, hence we use article-genre (topic) prediction as an incidental structure signal. The data contains 15 genres – some of which are noisy mixes of others. However, to keep preprocessing general, we use genres "as-is". **Corpus:** We collected comments and the above training signals for every article published by the Austrian newspaper `DerStandard.at` in 2015. Each article has a news genre tag and user comments, that in turn receive up and down votes. The corpus contains 813k comments, from which we extracted 8.9k controversial and 12.6k non-controversial comments after removing duplicates and short comments with less than five words. **Text preprocessing:** is source agnostic without language-specific NLP. We remove noise like low-frequency words. We create special tokens for discourse (repeated punctuation) and reactionary sentiment (emoticons) by categorizing emoticons into four (non-overlapping) types using a Wikipedia emoticon list[1], see Table 1. We keep stop words, as they often overlap with discourse markers (see Sec. 5). Compounds are separated with a $comp$ token. Finally, we pre-trained word2vec (Mikolov et al., 2013) embeddings on 3.35M preprocessed article and comment sentences to cover standard German and mixed (non)dialect.

Pattern	Replacement	Example
URL	url	`web.de`
happy	$happy$	:) :D
sad	sad	:(
skeptical	$skeptical$	:S, :/
unserious	$unserious$	:P ;p
rep. punct.	$.$, $,$, $?$, $!$	... !!!
compounds	word $comp$ word	Go-Fan

Table 1: Text normalization reduces vocabulary noise and creates *input features*.

4 Models

Baselines: As baselines we use Multinomial Naive Bayes (MNB) and Regularized Logistic Regression (LR) trained on TF or TFIDF Bag-of-Ngrams. FastText (FT) (Joulin et al., 2016) is trained on embedding 1-3grams.

Single / Multi-task CNNs : We also use convolutional neural nets (CNN) as they are widely used in text classification. Below, we describe how we modified the single-task model (ST) by Kim (2014) to create a multi-task architecture (MT) as follows. ST: A CNN that predicts comment-controversy only. It uses a deeper classifier, input-token dropout, custom word2vec embeddings and trains on comment, controversy label pairs via a binary cross-entropy – see *Controversy CNN* in Figure 2. MT: This model adds a genre-encoder to the ST. The encoder predicts multi-class genre via categorical cross-entropy and softmax on genre labels. Its penultimate activation map is fed to the ST's controversy classifier, to provide genre plus controversy features – see red downward arrow entitled *genre encoding* in Figure 2. The two losses are trained as a weighted sum. Thus, genre features are not required when predicting on new data.

MT modifications: Since feature extraction module design is central to CNNs, we evaluate a range of different design choices. We separate extraction modules into three categories from left to right: *convolution methods*, *activation schemes*,

[1] `https://en.wikipedia.org/wiki/List_of_emoticons`

and *pooling mechanisms* as seen in the upper and middle parts of Figure 2. White boxes are modules, dashed/dotted lines are module-combination options. Modules are marked by author, or with * for our own modifications. Module details are as follows:

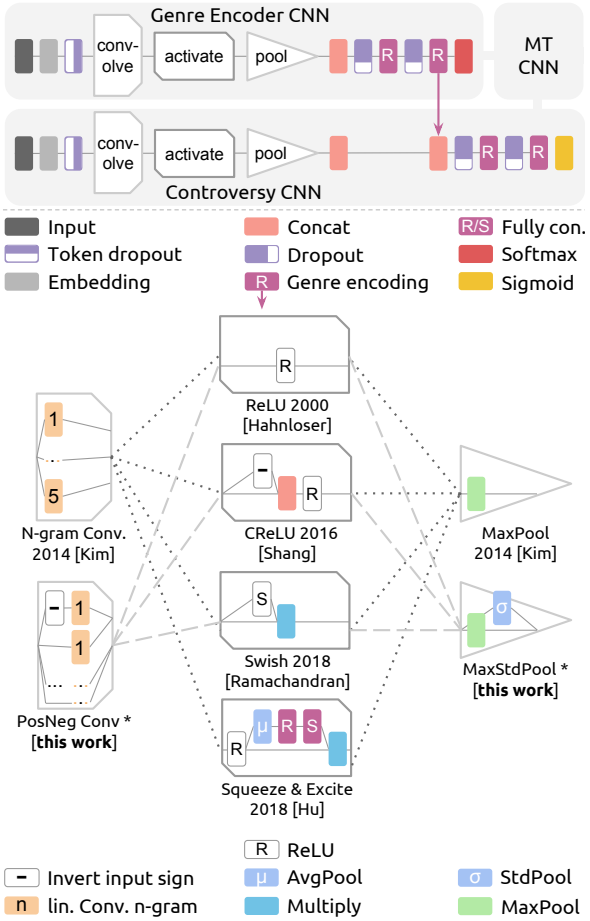

Figure 2: CNN modifications. Upper white box classifies genre to encode it, lower one classifies controversy. Colored rectangles are layers and operations as per the legend.

Conv: Kim (2014). CReLU Appends negated activations before applying ReLU (Shang et al., 2016). PosNeg Conv* (PNC): Learns separate convolutions for negated and positive embedding-activations, to extend CReLU. ReLU: (Hahnloser et al., 2000). Swish: Self-attention multiplying inputs x by their sigmoid $\sigma(x)$ (Ramachandran et al., 2018). Squeeze and Excite (SE): Bottlenecked multi-layer attention that learns convolution filter importances (Hu et al., 2018). MaxPool: (LeCun et al., 1998). Max(SPool)*: Appends per-filter Standard Deviation Pooling (SPool) to MaxPool, to preserve variance info. In the next section we evaluate the most successful combinations.

5 Results and Discussion

We evaluate on 8.9k controversial and 12.6k non-controversial comments that each belong to exactly one genre. We created 5 randomly sampled (stratified) folds – 4 folds for cross validation (CV) and 1 as holdout set. MNB, LR, FT, Conv+ReLU (ST) only predict controversy. The MT models jointly predict controversy + genre and are tested for various modification combos. Finally, we investigate models decision semantics and feature type importances via ablation studies.

5.1 Baselines: MNB, LR, FT

In Table 2 we list F_1, area under the ROC curve (AUC) and accuracy (Acc) controversy prediction results on the holdout test set. We see that FastText is the best baseline[2]. Optimal hyperparameters from 4-fold CV were: word-embedding 1-3gram with 128 dimensional w2v embeddings for FT, and TFIDF 1+2grams with a maximum document-frequency of 100% and a minimum term frequency of 2 for MNB and LR.

5.2 ST, MT CNNs

Stopwords and punctuation are kept as they contain discourse and sentiment features – see sec. 5.3 for details. Low-frequency words are replaced with a pre-trained unknown word token (UNK). Conv+ReLU (ST): The controversy-only CNN outperformed FT at optimal CV parameters of: 1-5gram, global max pooling, 128 filters and 1k classifier widths. More filters or a 4k width decreased CV and test performance. Standard dropout (Hinton et al., 2012) and Batch Normalization (Ioffe and Szegedy, 2015) decreased performance, while 20% token-dropout (Gal and Ghahramani, 2016) led to consistent improvement. Conv+ReLU (MT): Adding a genre-task network to ST improved performances by 2–4 points each, despite working on halved hyper parameters – i.e. MTs performed best using only 64 filters and 512 classifier units, giving less model parameters than the ST, especially since increasing ST's parameters hurt its performance. MT modifications: Since some modifications underperformed we only list combinations that are top-3 in one of the measures. Notice-

[2] An always-controversial predictor gives $F_1 = 58\%$, $Acc = 42\%$ and sample weighted class average $\overline{F_1} = 24\%$. A always-non-controversial predictor gives $F_1 = 42\%$, $Acc = 58\%$ and $\overline{F_1} = 43\%$). Neither is useful in practice.

ably, the MT+PNC+SPool+Swish variant significantly improved AUC_{ROC} and Acc over the simpler Conv+ReLU (MT) model, which produced the best F_1. Overall, we see that adding more incidental supervision signals beats adding advanced network modules.

Model	AUC	F_1	Acc
MNB	59.84	55.72	57.44
LR	62.92	58.14	60.12
FT	65.06	60.57	63.82
Conv+ReLU (ST)	68.25	62.03	66.42
Conv+ReLU (MT↓)	72.12	**64.48**	68.37
PNC+CReLU	*72.06*	63.40	*68.72*
PNC+SPool+Swish	**72.28**	64.21	**68.82**
Conv+SE+ReLU	71.91	*63.93*	68.76

Table 2: Holdout performances for the *controversial class* (y=1). **Baselines:** top 3. **ST:** middle, **MT:** last 4 – only module combinations with top-3 performance in one measure are listed as: **best**, 2nd best, *3rd best*.

5.3 Feature-type ablation

We ablated sentiment, discourse and topical features (Choi et al., 2010; Cramer, 2011). Then, we re-tuned the Conv+ReLU (MT) on the 4, now ablated, CV folds to measured test set performance changes as follows. *Three sentiment ablations*: (1) polarity words (sent ws by Waltinger (2010)), (2) repeated punctuation (punct.), and (3) emoticons (emotes) as mentioned in sec. 3. *Discourse:* Removal of German discourse markers (DiMLex) (Stede and Umbach, 1998). *Topic:* Noun removal as in Choi et al. (2010) to represent topical indicators. Figure 3 shows the relative percentual performance drop per ablation. Thus, for controversy prediction: *topic* was the most important, followed by *discourse* markers[3] and *sentiment* with repeated punctuation and emoticons being impactful style/sentiment features. Polarity words affect prediction, but are not language independent.

[3]Markers overlap with a stop word list in approximately 49% of occurrences in our dataset. Stop words: http://www.ranks.nl/stopwords/german.

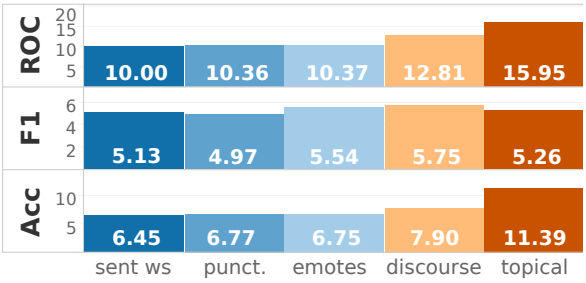

Figure 3: *Relative controversy prediction performance drop in %* for removal of: sentiment (blues), discourse (orange) and topic/nouns (red).

5.4 Per-word impacts

Inspired by explainability methods (Li et al., 2016; Arras et al., 2017) we also measured the controversy prediction-score change when replacing a token with a class neutral UNK token[4].

Discourse or punctuation ($):
Because it not_a UNK country but a dictatorship is
What UNK Putin of human_rights and peace $.$.
Had you the UNK or are you vaccinated $?$?

Emoticons:
They employ the same word_choice :((
Was easily UNK the tradition UNK ? ;)

Context dependent word influence:
Interestingly , if one something negative against ⏎
Windows posts will one instantly_be with UNK ⏎
bombarded .
But 2 years were we by Microsoft marketing ⏎
and by Microsoft fan_boys UNK how cool yet not
Windows 8 and 8 .1 is .

Figure 4: *DE → EN Per-token controversy impacts:* Red is important for controversy. Blue lowers the controversy score. Last paragraph: *context dependent word influence* of the word Windows.

In Figure 4 we colored per-token score drops (red) or increases (blue) for German-to-English word-by-word translations on test set comments. We show examples by ablation types as described in section 5.3. As before, nouns and discourse markers increase controversy, while, expectedly, an (#unserious) ;) emoticon is strongly counter-indicative of controversy. *Repeated punctuation*, like $.$ or $?$, also impacts prediction. Finally, the model learned *context dependent con-*

[4]Removing tokens would create unusual n-grams, and hence wrong results.

domestic politics		international politics		economy		panorama	
kpö	afd	ceasefire	poroschenko	bonds	tsipras	entry	pegida
pühringer	fpö	mariupol	separatists	rbi	troika	battery	prejudices
spö	grünen	rebels	putin	hedge funds	syriza	property dmg.	refugee policy
state elections	parties	hamas	arabs	budget	greece	passage	hate-monger
federal level	faymann	air raid	israelis	credits	varoufakis	tents	antisemitism
genre	+ controversy	genre	+ controversy	genre	+ controversy	genre	+ controversy

Table 3: Token importances in descending order. On the left **genre**: most important genre tokens. On the right (**+ controversy**): most controversial tokens per genre. Tokens are sorted by mean positive impact on genre and genre+controversy predictions.

troversy polarity for the word *Windows*, with has both strong positive and negative polarity.

5.5 Token impacts on genre and controversy

To generate keywords for **controversy** and **genre vs. controversy-per-genre**, we averaged UNK token-replacement prediction-impacts over all occurrences of a token t_i and calculated its impact mean $\mu(impacts(t_i))$ and standard deviation $\sigma(impacts(t_i))$, similar to how Horn et al. (2017) extract topic keywords.

Controversy keywords: In Table 4 we divided tokens into infrequent (top half) and common tokens (lower half). Infrequent tokens have over 10 occurrences, frequent ones at least 200.

(a) 0 con	(b) ↑↓ con	(c) ↑ con	(d) ↓ con
”	pkk	separatists	yet
.	kurds	putin	thx
;	crimea	pegida	has
–	tsipras	israelis	ain’t
possibly	israelis	hamas	yeah
.	eu	eu	have
-	usa	usa	**#happy#**
?	**#unser.#**	country	**#unser.#**
”	**#happy#**	people	anyway
with	**$.$**	austria	from
$\sigma(impacts(token))$		$\mu(impacts(token))$	

[a] No impact := smallest $\sigma(impacts) \approx 0$ top.
[b] Impactful := largest $\sigma(impacts)$ top.
[c] Pro controv. := most positive $\mu(impacts)$ top.
[d] Contra cont. := most negative $\mu(impacts)$ top.

Table 4: Controversy impacts for seldom (upper half) and frequent token (lower half).

The tokens impact controversy either: (a) not at all, (b) positively or negatively, (c) generally increase it or (d) generally decrease it. We see that, standard punctuation has no impact on controversy (a), but repeated punctuation, emotes and political terms do (b). Expectedly, political terms generally increase controversy (c), while colloquialisms and friendly emotes lower it (d).

Genre vs. controversy-per-genre keywords: We examined mean token impacts $\mu(impacts)$ on *genre classification* vs. *per-genre controversy* in Table 3 for the four most interesting genres. The *domestic politics* genre is dominated by established Austrian parties or generic political terms, while right-wing, left-wing and liberal parties characterize domestic controversy. The *international* genre shows mostly war related terms. Its controversy focuses on the 2015 Ukraine and middle east conflicts. Keywords for the *economy* genre are general finance terms, whereas the Greek debt crisis dominates genre controversy. The *panorama* genre focuses on refugee-related terms, where the related right-wing issues caused controversy in 2015.

6 Conclusion

We proposed a fully automated, incidentally supervised, multi-task approach for comment-controversy prediction and showed that it successfully captures contextual controversy semantics despite only using minimal, language independent, preprocessing and feature creation. In the future, we aim to extend data collection to study controversy drift over time.

Acknowledgements

This work was supported by the German Federal Ministry of Education and Research (BMBF) through the project DEEPLEE (01IW17001).

References

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. "What is relevant in a text document?": An interpretable machine learning approach. *PloS one* 12 (2017).

Angelo Basile, Tommaso Caselli, and Malvina Nissim. 2017. Predicting Controversial News Using Facebook Reactions.. In *CLiC-it*.

Wei-Fan Chen, Fang-Yu Lin, and Lun-Wei Ku. 2016. WordForce: Visualizing Controversial Words in Debates. In *COLING*.

Yoonjung Choi, Yuchul Jung, and Sung-Hyon Myaeng. 2010. Identifying Controversial Issues and Their Sub-topics in News Articles. In *PAISI*.

Peter A Cramer. 2011. *Controversy as news discourse*. Vol. 19. Springer Science & Business Media.

Shiri Dori-Hacohen and James Allan. 2015. Automated Controversy Detection on the Web. In *ECIR*.

Shiri Dori-Hacohen, Elad Yom-Tov, and James Allan. 2015. Navigating Controversy as a Complex Search Task. In *SCST@ECIR*.

Yarin Gal and Zoubin Ghahramani. 2016. A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. In *NIPS*.

Venkata Rama Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2017. Reducing Controversy by Connecting Opposing Views. In *WSDM*.

Richard HR Hahnloser, Rahul Sarpeshkar, Misha A Mahowald, Rodney J Douglas, and H Sebastian Seung. 2000. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature* 405, 6789 (2000).

Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR* (2012).

Franziska Horn, Leila Arras, Grgoire Montavon, Klaus-Robert Mller, and Wojciech Samek. 2017. Discovering topics in text datasets by visualizing relevant words. *CoRR* (2017).

Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-Excitation Networks. *IEEE Conference on Computer Vision and Pattern Recognition* (2018).

Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *ICML*.

Myungha Jang, John Foley, Shiri Dori-Hacohen, and James Allan. 2016. Probabilistic Approaches to Controversy Detection. In *CIKM*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759* (2016).

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *EMNLP*.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998).

Jiwei Li, Will Monroe, and Daniel Jurafsky. 2016. Understanding Neural Networks through Representation Erasure. *CoRR* (2016).

Narges Mahyar, Weichen Liu, Sijia Xiao, Jacob Browne, Ming Yang, and Steven Dow. 2017. ConsensUs: Visualizing Points of Disagreement for Multi-Criteria Collaborative Decision Making. In *CSCW Companion*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

Chris Pool 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)*. The COLING Organizing Committee, Osaka, Japan. http://aclweb.org/anthology/W16-4304

Prajit Ramachandran, Barret Zoph, and Quoc V Le. 2018. Searching for activation functions. *ICLR* (2018).

Dan Roth. 2017. Incidental Supervision: Moving beyond Supervised Learning. In *AAAI*.

Wenling Shang, Kihyuk Sohn, Diogo Almeida, and Honglak Lee. 2016. Understanding and improving convolutional neural networks via concatenated rectified linear units. In *International Conference on Machine Learning*.

Manfred Stede and Carla Umbach. 1998. DiMLex: A lexicon of discourse markers for text generation and understanding. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*. Association for Computational Linguistics.

Ulli Waltinger. 2010. GERMANPOLARITYCLUES: A Lexical Resource for German Sentiment Analysis. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*. electronic proceedings, Valletta, Malta.