Explaining Controversy on Social Media via Stance Summarization

Myungha Jang and James Allan
Center for Intelligent Information Retrieval
College of Information and Computer Sciences
University of Massachusetts
mhjang@cs.umass.eduandallan@cs.umass.edu

ABSTRACT

In an era in which new controversies rapidly emerge and evolve on social media, navigating social media platforms to learn about a new controversy can be an overwhelming task. There has been significant work that studies how to identify and measure controversy online. However, we currently lack a tool for effectively understanding controversy in social media. For example, users have to manually examine postings to find the arguments of conflicting stances that make up the controversy.

In this paper, we study methods to generate a stance-aware summary that explains a given controversy by collecting arguments of two conflicting stances. We focus on Twitter and view this stance summarization task as a ranking problem of finding the top k tweets that best summarize the two conflicting stances of a controversial topic. We formalize the characteristics of a good stance summary and propose a ranking model accordingly. We evaluate our methods on five controversial topics on Twitter. Our user study shows that our methods consistently outperform other baseline techniques in generating a summary that explains the given controversy.

ACM Reference format:
Myungha Jang and James Allan. 2018. Explaining Controversy on Social Media via Stance Summarization. In *Proceedings of The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, Ann Arbor, MI, July 8-12 2018 (SIGIR'18),* 5 pages.
DOI: 10.475/123_4

1 INTRODUCTION

Online controversies often emerge and evolve quickly due to the nature of social media. These platforms force users to be concise and allow them to be casual, requiring less effort to post something on Twitter than other sources, such as Wikipedia or blogs. While existing techniques enable us to identify *whether* a topic is controversial, understanding *why* it is controversial is still left as work for users. For instance, consider a following scenario: A person discovers a new hashtag movement #TakeaKnee[1] on Twitter but does not know what it is about or why it is controversial at all. How would she search for people's opinions to better understand the conflicting stances on this topic?

One straightforward approach to this problem would be for the user to search the topic and manually scan the search results until she has read enough conflicting tweets to understand the controversy. However, current search systems make this navigation difficult due to the filter bubble effect [7]. For example, the top posts are likely to be the ones that the user agrees with because her friends liked the posts or because she or her friends follow the authors.

Another strategy for navigating Twitter is to identify a few key hashtags that indicate stances and then search for posts that contain them. As people are forced to write posts under the strict character limit, certain hashtags are utilized as self-created labels for their opinions (e.g., #imwithher in support of Hillary Clinton or #MAGA in support of Donald Trump during the 2016 US presidential election). However, because the use of hashtags (even the ones that seemingly contain obvious stances) are known to be noisy [11], the user must still carefully read through each tweet. More importantly, she has to go through a large number of noisy tweets that are not useful to understand the controversy while using her own judgment to identify their stance (if they even have one). This process requires substantial effort, critical reasoning, and phenomenal patience. It is clear that users could benefit from automating this process.

We propose a technique that generates a stance-aware summary by selecting the top tweets that best explains a given controversy. Our contributions are as follows:

- This work appears to be the first attempt to automatically summarize controversy on social media.
- We characterize what makes a tweet a good summary of controversy, propose three attributes that should be satisfied (i.e., *stance-indicativeness*, *articulation*, and *topic relevance*), and develop methods to estimate them.
- We propose a novel method to estimate the confidence of stance-indication using automatically-obtained stance hashtags, which have typically been used to filter data during manual annotation.
- We extensively evaluate various methods including a general summarization technique and our methods via user study and demonstrate that the summaries generated by our methods explain controversy better than the ones by other techniques.

2 RELATED WORK

This research is related to a few areas: summarization, stance detection and controversy analysis on social media.

Twitter Summarization: There has been much work on summarizing Twitter postings while most of them focuses on summarizing events [2, 4, 8, 15, 17]. Inouye et al. [13] compare multiple summarization algorithms for Tweet data, and their extensive experiments suggest that the SumBasic algorithm produced the best

[1] This was prevalent during the US national anthem protests that began in 2017.

SIGIR'18, Ann Arbor, MI

DOI: 10.475/123_4

Table 1: An example of good (left) and bad (right) summary tweets on "Abortion" posted on Nov 4, 2016. The good summaries are selected from our method. Examples of stance hashtags are marked in bold.

• We know it's not okay that for 40 yrs politicians have denied a woman coverage of abortion just because she's poor **#BoldTheVote #BeBoldEndHyde** • Read the whole story about #HarvardSoccer before forming idiotic tweets. Don't support **#RapeCulture** by calling it #LockerroomTalk • Hillary Clinton voted no to banning late-term abortions, even though over 80% of Americans support the ban. **#VoteProlife**	• lmaoaoao b**** i would did the abortion myself right there lmaoaoao • before I formed you in the womb I knew you jer 1:5**#prolife #Defundpp** [URL] **#UnbornLivesMatter** • Abortions: the new fall trend in religious circles [URL] • Could you imagine crying over ur uni stopping anti abortion protests, if you're so pro life then go and f***ing get one?

F1-result in human evaluation, which we also adopt as a summarization baseline in this paper.

Stance Detection on Twitter: Stance classification on Twitter has two main tasks: (1) classifying the text's stance (against, favor, or neutral) given a topic, or (2) classifying the twitter users' stances. The former task drew attention when 2016-SemEval Task 6 released a dataset of tweets with stance annotations [11]. The results of various approaches were shared after the competition [12], and later more successful approaches were proposed including one that uses a bi-directional conditional LSTM for classifying the stance and opinion target on Twitter [1]. For the latter type of task, Johnson and Goldwasser developed a method to classify stances of politicians on Twitter using relational representation [10]. While stance detection is closely related to our problem, our goal is not to accurately classify the stances of all tweets. Our problem is also more robust to misclassification errors of stances as we take the tweets with highest stance confidence as part of the summary.

Controversy Analysis on Twitter: Several studies have formally defined a model for controversy detection. Jang et al. propose that controversy should be identified with respect to a given population and argue that contention and topic importance are primary aspects of controversy. [9]. Zielinski et al.'s model determines that a topic is controversial if there is a difference in opinion within a given community of people [18]. Garimella et al. and Fraisier et al. analyze user retweet or follow graphs, which signifies the formation of exclusive communities of like-minded people for controversial topics [5, 6]. Our approach builds on these earlier findings.

3 APPROACH

We first discuss what makes a tweet a good summary. We then develop a ranking model that ranks the tweets by how likely a tweet is part of a good summary. Finally, we propose two methods to select the summary from the ranked tweets.

3.1 Ranking Model

Based on the definition of controversy by previous work, we define a good controversy summary as a description that effectively captures different arguments of two communities that take conflicting stances with each other. After examining many examples (see Table 1), we derive three primary components that characterize a good controversy summary tweet:

- **Stance-indicative (S):** A good tweet strongly indicates its stance and is often followed by some particular stance hashtags that are widely used by users from the same stance community. While both good and bad tweets frequently include stance hashtags, the presence of stance hashtags is a positive reinforcement signal if the the quality of tweet is decent.
- **Articulation (A):** A good tweet is clear, persuasive, and logical. It also written with proper language.
- **Topic Relevance (T):** A good tweet is relevant and self-explanatory in the context of a particular topic.

For any controversial topic $\mathcal{T}$, we assume that there are always two stances that are in conflict with each other. We denote these stances as $\mathcal{S}_A$ and $\mathcal{S}_B$. Let Γ be a summary of a given topic $\mathcal{T}$. We let $\Gamma = [\Gamma_A, \Gamma_B]$ that denotes the summary of $\mathcal{S}_A$ and $\mathcal{S}_B$, respectively. We define a model that computes whether a tweet τ is likely to be in the set Γ_A:

$$P(\Gamma_A|\tau) = f(P_S(\mathcal{S}_A|\tau), P_A(\tau), P_T(\tau|\mathcal{T})) \quad (1)$$

where $P_S(\mathcal{S}_A|\tau)$ computes how likely a tweet indicates $\mathcal{S}_A$, $P_A(\tau)$ computes how articulate the tweet is, and $P_T(\tau|\mathcal{T})$ computes how relevant the tweet is for the topic.

In the next sections, we discuss how to estimate the first two scores. For the topic relevance score, we use the straightforward probability that the tweet sentence was generated from the language model of the given topic, normalized by the tweet length.

3.2 Estimating Stance-indication

To estimate stance-indication, we first identify stance hashtags that statistically characterize the stance community. We use the stance hashtags as a proxy to estimate the tweets that indicate the same stance as follows:

$$P_S(\mathcal{S}_A|\tau) = \sum_{h \in \mathcal{H}} P(h|\tau) \cdot P_S(\mathcal{S}_A|h) \cdot P(h)$$

Then the score boils down to estimating $P(h|\tau)$, a probability that the tweet includes a given hashtag h, and $P_S(\mathcal{S}_A|h)$, a score that indicates how likely h represents $\mathcal{S}_A$. As $\mathcal{S}_A$ and $\mathcal{S}_B$ are mutually exclusive, we penalize ambiguous tweets that are likely to contain stance hashtags of the opposing side by subtracting the score for the opposite stance as follows:

$$P_S(\mathcal{S}_A|\tau) = \sum_{h \in \mathcal{H}_A} \left[P(h|\tau) \cdot P_S(\mathcal{S}_A|h)\right] - \sum_{h \in \mathcal{H}_B} \left[P(h|\tau) \cdot P_S(\mathcal{S}_B|h)\right]$$

where $\mathcal{H}_A$ and $\mathcal{H}_B$ are the set of stance hashtags that represent $\mathcal{S}_A$ and $\mathcal{S}_B$ respectively.

3.2.1 Identifying Stance Hashtags ($\mathcal{H}_A, \mathcal{H}_B$). To obtain a set of stance hashtags, we first identify two communities, C_A and C_B, each of which represents two conflicting stances, $\mathcal{S}_A$ and $\mathcal{S}_B$. As introduced by Garimella et al., we construct a user retweet (RT) graph and partition it into two groups [6]. We use a simple method that produces only two communities so as not to deal with the extra step of classifying several identified communities to two stances. We leave identifying multiple communities and clustering them

into one of the stances of interests to generate the summaries from for the future work.

Once we identify C_A and C_B, we assume that tweets that are written by users from C_A and C_B are likely to indicate $\mathcal{S}_A$ and $\mathcal{S}_B$ respectively. From the two sets of tweets, we compute the information gain [16] that each hashtag gets for the information of the community class when they are present in the tweets: if we know nothing about the tweet but the hashtag presence, which hashtag best indicates its stance community? Finally, we define $\mathcal{H}_A$, the set of stance hashtag of $\mathcal{S}_A$, as follows.

$$\mathcal{H}_A = \{h \in \mathcal{H} | h \in TopN(IG) \wedge freq_A(h) > freq_B(h)\}$$

where IG is a function that returns the information gain value for the two stance classes for a given hashtag, $freq_A$ is the frequency of h in the tweets published from C_A, and $TopN$ is a set of n items sorted by the scores of the given function. In our experiments, we set $n = 30$. We then let $P_S(S_A|h)$ be the normalized score of $IG(h)$ for all hashtags in the set $\mathcal{H}_A$.

3.2.2 Estimating $P(h|\tau)$ via Latent Hashtags. If we think of hashtags as user-generated annotations, hashtags are incomplete annotations. It means that a lack of a certain hashtag does not necessarily mean that it is not a relevant label. To better utilize hashtags as more accurate signals, we make hashtags more complete annotations by estimating $P(h|\tau)$ for all hashtags, the probability that tweet τ generates a hashtag h. Therefore, we adopt a character composition model, Tweet2Vec, which finds a vector space representation of tweets to predict user-annotated hashtags [3]. By finding the embeddings of tweets and hashtags, we estimate $P(h|\tau)$ for hashtags that were not explicitly used in the given tweet.

3.3 Estimating the level of articulation

We build a regression model that predicts how well the tweet is written and generate an annotated set of 150 articulate and 150 non-articulate tweets on arbitrary topics. The annotation criteria between the two classes is whether the given tweet is logical, the grammar is sound, and it is written with proper language.

Similarly, Duan et al. propose a classifier to evaluate the content quality of tweets [4]. In addition to their features, we include a large set of POS tags that are Twitter-specific provided by TweeboParser [14], N-grams of the POS tags sequence to capture the structural flow of the good sentences, and the ratio of offensive words to penalize usage of inappropriate language. This model is generalizable since the features are not content-specific. We trained a logistic regression model and obtained 89.9% classification accuracy using 5-fold cross validation.

Table 2: The features used to train a regression model for predicting the level of tweet articulation.

Feature	Description
Tweet POS Tags	The ratio of Tweet POS tags [14]
OOV words	The ratio of words that are not in the dictionary
Offensive Words	The ratio of offensive/profane words
POS Tags N-grams	N-grams of Tweet POS Tag sequence
Stop words	The ratio of stop words
Tweet length	The number of characters in a tweet
Avg. word length	The avg. number of characters in tweet words

Table 3: The amount of data used to train Tweet2Vec and summary generation. The number in parentheses refers to the number of tweets published by the stance community.

Topic	Tweet2Vec		Summary	
	# Tweets	# Users	# Tweets (# in C)	RT ratio
Election	10.8M	4.3M	10000 (4268)	70.9%
#TakeAKnee	565K	692K	44167 (17217)	71.1%
Abortion	692K	539K	3477 (1262)	57.6%
Feminism	1.7M	1.7M	50323 (20783)	41.3%
Climate Change	546K	360K	10234 (3915)	60.1%

3.4 Summary Selection

We propose two algorithms that aggregate the three probability scores to generate the final k summary, which we set as 10 in our experiments. To produce a final summary to equally cover two stances, both algorithms select $k/2$ tweets from each stance.

SumSAT ranks the tweets by setting the aggregation function f (in Eq. 1) to be a harmonic mean for the three scores described earlier. HashtagSumSAT, on the other hand, while using the same aggregation function, first identifies the top $k/2$ stance hashtags for each stance and selects the top tweet for each hashtag. While we use a harmonic mean as f, any aggregator can be plugged in. The difference of the two algorithms come from whether it globally ranks the tweets or ranks the tweets per each hashtag.

4 EVALUATION

We evaluate our methods by running them on real data and conducting user studies to capture the utility of our algorithms.

4.1 Experiment Setup

We consider five controversial topics including two short-term, event-based controversies (2016 US Presidential Election and 2017 US National Anthem Protests which we refer to as `#TakeAKnee`), and three long-term ethics-related controversies (Abortion, Feminism, and Climate Change).

Our goal is to generate a summary that can explain why the topic is controversial. For each topic, we generate a pair of summaries and used Amazon Mechanical Turk to ask people which summary explains the controversy better in a blind fashion. The participants could also say that the quality of the two summaries is the same. To observe whether a subset of tweets whose author's stance is identified from the community generates a better quality summary, we experiment with two cases for each algorithm: (1) using all tweets as summary candidates or (2) using only tweets whose author belongs to one of two stance communities we identified. We distinguish the second case by adding 'C' (for the community) to the method name. We also generate summaries including the following baseline methods:

- **Random**: A random set of k tweets from a unique set of tweets.
- **MostRT**: The top k most-retweeted tweets in a given day
- **SumBasic [13]**: A general summarization technique. We preprocess the tweets to exclude Twitter-specific stop words.

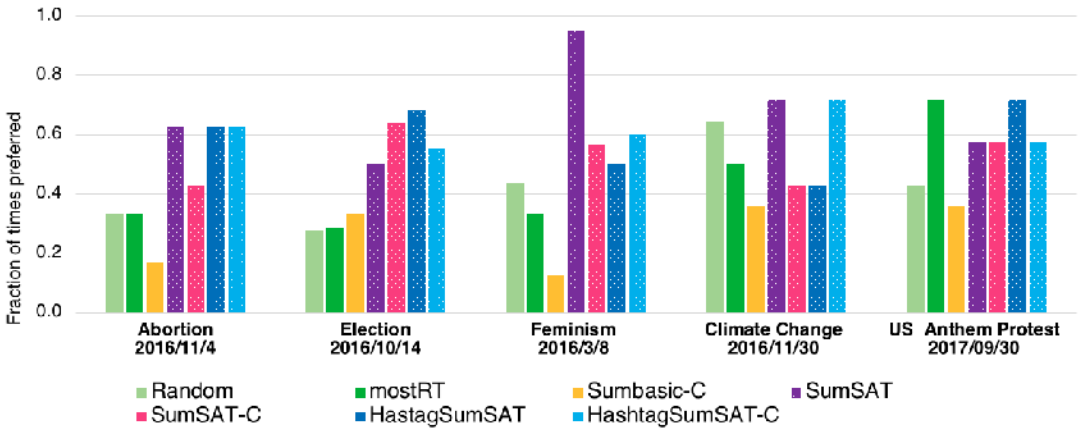

Figure 1: The user study results by the topics. The rightmost four bars indicate our methods. We did not include SumBasic in the graph because it was the worst method for all topics, being preferred only 8% of times overall.

4.2 Results and Discussion

The user study shows that our methods were consistently more effective than other baselines across all five topics (Figure 1). Overall, SumSAT generated the summaries that were preferred the most (68%) followed by HashtagSumSAT-C (61%).

Controversy summarization as a new task: Overall, both Sumbasic (8%) and Sumbasic-C (42%) generated the worse summary than the naive baselines such as mostRT or random. This suggests that controversy summarization is an inherently different task from a general topic summarization.

MostRT is often a strong baseline, but its performance is not reliable: For the topic of #TakeAKnee, the mostRT baseline was as effective as our top approach. The topic also particularly had a high ratio of retweets compared to other topics (Table 3). However, depending on the topic and the day, mostRT can also be the worst feature, even worse than random selection as in the case for the topic of Feminism. For example, the top retweets in Feminism include 'Happy International Women's day!'. Retweets can often be tweets for entertainment and can easily be dominated by people on one side of stances who are more vocal on Twitter.

Social features seem to be more useful than the content itself in stance summarization: We learned that in identifying and finding stance-indicative tweets, social features are far more important than the content itself. For example, mostRT outperforms a general summarization technique that only considers the text content most of the times. This finding aligns with the findings of the previous study on detecting controversy on Twitter [6].

Utility of stance hashtags: While SumSAT was an overall winner, HashtagSumSAT outperformed SumSAT for two topics: US Election and #TakeAKnee. We observe a tendency in the event-based controversies like those topics show more active usage of stance hashtags as there were specific actions people try to promote via stance hashtags. In such type of controversies, stance hashtags were particularly effective to generate a summary around.

5 CONCLUSION AND FUTURE WORK

We introduce and tackle a new task of generating a stance-aware summary to explain controversy on social media. Our goal is to provide a tool that helps people navigate controversy effectively. We propose a ranking model that considers three factors that suggest a tweet be part of a good summary derived from our qualitative observations. We assume that a good summary tweet is clear, articulate, and relevant to the topic. Our algorithm characterizes two conflicting stances by identifying two communities from a retweet graph and retrieving the tweets published by them. We define and identify "stance hashtags" that are distinctively used to indicate their opinions in each community and propose a probability model that computes how a tweet is likely to indicate the stance of the community based on the probability that the tweet is likely to generate those hashtags. Our user study demonstrates that users prefer the summaries from our methods over the ones from other reasonable baselines.

6 ACKNOWLEDGEMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by the Air Force Research Laboratory (AFRL) and IARPA under contract #FA8650-17-C-9118 under subcontract #14775 from Raytheon BBN Technologies Corporation. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

REFERENCES

[1] Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance Detection with Bidirectional Conditional Encoding. In *EMNLP*.
[2] Deepayan Chakrabarti and Kunal Punera. 2011. Event Summarization Using Tweets. In *ICWSM*.
[3] Bhuwan Dhingra, Zhong Zhou, Dylan Fitzpatrick, Michael Muehl, and William W. Cohen. 2016. Tweet2Vec: Character-Based Distributed Representations for Social Media. *CoRR* (2016).
[4] Yajuan Duan, Zhimin Chen, Furu Wei, Ming Zhou, and Harry Shum. 2012. Twitter Topic Summarization by Ranking Tweets using Social Influence and Content Quality. In *COLING*.
[5] Ophélie Fraisier, Guillaume Cabanac, Yoann Pitarch, Romaric Besançon, and Mohand Boughanem. 2017. Uncovering Like-minded Political Communities on Twitter *(ICTIR '17)*. 261–264.

[6] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2016. Quantifying Controversy in Social Media *(WSDM '16)*. 33–42.

[7] Mathew Ingram. 2016. Here's What's Wrong With Algorithmic Filtering on Twitter. (2016). `http://fortune.com/2016/02/08/twitter-algorithm`.

[8] David I. Inouye and Jugal K. Kalita. 2011. Comparing Twitter Summarization Algorithms for Multiple Post Summaries. *PASSAT and SocialCom* (2011), 298–306.

[9] Myungha Jang, Shiri Dori-Hacohen, and James Allan. 2017. Modeling Controversy within Populations. In *ICTIR*.

[10] Kristen A Johnson and Dan Goldwasser. 2016. Identifying Stance by Analyzing Political Discourse on Twitter.

[11] Saif Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2016. Stance and Sentiment in Tweets. *CoRR* (2016).

[12] Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 Task 6: Detecting Stance in Tweets *(SemEval)*.

[13] Ani Nenkova and Lucy Vanderwende. 2005. *The impact of frequency on summarization.* Technical Report. Microsoft Research.

[14] Olutobi Owoputi, Brendan OConnor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *HLT-NAACL*. 380–390.

[15] Beaux Sharifi, Mark-Anthony Hutton, and Jugal Kalita. 2010. Summarizing Microblogs Automatically *(NAACL-HLT)*. Stroudsburg, PA, USA, 685–688.

[16] Yiming Yang and Jan O. Pedersen. 1997. A Comparative Study on Feature Selection in Text Categorization. In *ICML (ICML '97)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 412–420.

[17] Evi Yulianti, Sharin Huspi, and Mark Sanderson. 2016. Tweet-biased Summarization. *J. Assoc. Inf. Sci. Technol.* 67, 6 (June 2016), 1289–1300.

[18] Kazimierz Zielinski, Radoslaw Nielek, Adam Wierzbicki, and Adam Jatowt. 2018. Computing controversy: Formal model and algorithms for detecting controversy on Wikipedia and in search queries. *Information Processing Management* 54, 1 (2018), 14–36.