Volume 9, Issue 5
October 2020

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

Controversy Probability Prediction on Twitter Corpus

[1]Harshav Kumar and [2]Gopal Krishan Prajapat

[1, 2] Department of Computer Science & Engineering, BK Birla Institute of Engineering & Technology, Pilani(Raj), India

E-mail: [1]harshavk97@gmail.com, [2]gopal.prajapat@bkbiet.ac.in

ABSTRACT

Nowadays, social media is filled with abundant information on how people communicate and share their opinion with each other's, which is continuously updating to an extent the interest of its users. This leads the researcher to carry out operations on this vast form of data to get useful results which give solution to some problems related to the behaviour and personality of a person. Here, we are evaluating the probability of controversy using tweets. Working on micro blogging sites, like Twitter, where users typically share various types of pieces of evidence to support their opinions over controversial topics on a social network. These types of debate provide the researcher an estimation of argument that have been supported for the cause. So, we have to develop an algorithm to calculate the positive sentiments, negative sentiments, controversy score, user's favs, number of likes, and hashtags. By creating a framework on dataset/corpus of tweet scrapped on the basis of conflicting and are related to some controversy or debate. Our method shows that a classifier trained with additional features is capable of capturing the different forms of representing evidence on Twitter, using the test set we achieve an F1 of 80% score in the detection of controversial versus non-controversial. In general, the feature specific to twitter or social media, are more prevalent in tweets showed by our analysis.

Keywords: *filtering; controversy; tweet; buzzy snapshots*

1. INTRODUCTION

Millions of people post entries and write comments on various topics, such as consumer product and film reviews, news, politics, etc. on social media platforms such as Twitter and Facebook, effectively offering a real-time view of the views, thoughts, behaviours and patterns of individuals and groups across the globe. Latest surveys show that nearly 250 million bloggers approx. 32% worldwide frequently provide opinions on merchandise and makes, active net users approx. 71% read forums, and customers around 70% believe reviews shared on-line by different consumers.

The explosion of social media has enabled researchers to have unprecedented access to data on the points of view wider public on political issues or events in popular culture. [1] Automatic identification of events involving large public social media is an interesting task from both a sociological and a realistic view: For example, display engaging and new events allow web content providers to attract more visitors to their web pages. We present techniques for detecting a specific form of event-controversial event-using twitter as an initial point. Controversial issues raise a public debate in which members from the public, views or concerns they oppose skepticism. Examples involve events that exceed public perceptions of a particular agency or are counter to existing social norms.

Our work is motivated by an interest in observing which events become controversial on social media and understanding why they became controversial. In this project, we are extracting rich features set from tweets and also using features from some external sources to identify the controversy score of an entity.

The micro-blogging platform Twitter provides a forum for people to share their opinions and engage in discussions about a wide variety of topics. There may arise conflicting opinions among users with regard to certain topics and polarization is observed on Twitter around controversial issues [2]. Additionally, it has been noted that user text on Twitter often contains arguments with inappropriate or missing justifications [3]. This leads to the spread of misinformation and rumours and often cause controversies.

Social media, for example, to predict the spread of diseases it has been used in health care. To quantify and monitor actual diseases activity by analyzing frequent changing public opinions on twitter [4] concerning H1N1 or swine flu. The influenza activity predicted one to two weeks before the Centers for Disease Control and Prevention (CDC) by Researchers, CDC and other public health agencies will get assist by these emerging trend identifications in the monitoring of infectious disease as well as public concerns

The new trend is a growing area of interest and value over time on social media platforms such as Twitter, Facebook, forums, etc. The role of emerging trend detection (ETD) is to classify subjects that were previously detected and are increasingly relevant to a broader set of textual data for a limited period of time [5]. Controversial trend is a common topic that invokes contrary opinions or views [6].

2. RELATED WORK

2.1. Detecting Controversial Events

To detect controversial event [7], they are creating snapshots of tweets for different entities, as shown in equation 1.

$$triple\ s = (e, \_t, tweets) \quad (1)$$

They are calculating the controversy scores for all the entities and then ranking them according to the controversy score. For this first, they selected buzzy snapshots out of all by considering the number of tweets they had previously and then set a threshold for them. After this, for all buzzy snapshots, they calculated historical and timely controversy score and based on that calculated overall Controversy score as shown in equation 1, 2 and 3 respectively.

Volume 9, Issue 5
October 2020

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

$hcont(e) = k/|CL|$ (2)

$tcont(s) = a * MixSent(s) + (1 - a) * Controv(s))$ (3)

$cont(s) = B * tcont(s) + (1 - B) * hcont(s)$ (4)

Where k is the number of controversy terms t0 s.t. PMI (e, t0) > A3. PMI is calculated on the basis of the co-occurrence of entities and concepts in Web documents.; they used A = 2. The linear combination of two scores – MixSent(s) and controv(s) will lead out the timely controversy score, where MixSent(s) score applied on snapshot will give out the relative disagreement about the entity in the twitter data and the presence of explicit controversy terms in tweets described by controv(s).

2.2. Telling apart tweets associated with Controversial versus Non-Controversial Topics

The predictability of tweets associated with controversial versus non-controversial topics proposed under this paper [8]. To build their dataset, they developed 8 claim statements covering different topics and then crowdsourced their labelling as controversial or noncontroversial by asking people to rate the topics on a 5-point Likert Scale ranging from 1: non-controversial to 5: controversial. Based on the average rating, the claims/topics were classified as controversial or noncontroversial. Then, they collected all tweets pertaining to a present keyword for each of the claim statements along with related census data using Crimson Hexagon.

From the data, they then extract various features to perform supervised classification of tweets as controversial/ noncontroversial. The feature set includes emphatics features, language-specific features as well as Twitter-specific features. After data pre-processing and feature extraction, they built classification models using different supervised classifiers: Naïve Bayes (NB), Support Vector Machines (SVM), and Decision Trees (DT). They used Weka and an R machine learning package for this task.

To access the accuracy of their models, they used precision, recall, and F-score (β = 1) as metrics. They achieved cross-validation accuracy score of 87% (F1) on the training set and accuracy of 63.4% on the test set.

2.3. Semantics + Filtering + Search = Twitcident exploring information in social web

This paper proposes an incident driven framework [9], where incident detection is a trigger event, and then it starts collecting all the social media and Twitter messages related to that incident. In practice, the program connects Twitcident emergency communications services in the Netherlands to facilitate the collection of relevant information from Twitter sources for incidents identified by these services. We perform large-scale experiments in which we test (i) filtering strategies relevant information for a given event, and (ii) strategies to find specific pieces of information. Twitcident class content of Twitter tweets as victims, injuries or threats, and also categorizes the type of experience recorded in a tweet, e.g. how a tweeter sees, thinks, hears or smells. Classification is accomplished by hand-crafted rules (e.g. when tweeting (X1 AND X2 ...) OR. Then classify as Y) that works on both facet-value pairs and plain words that are listed in a tweet.

2.4. Controversy trend detection in Social Media

In this thesis [10], they concentrate on the early assessment of whether or not the issues-occurring as social media posts, blogs, etc. - are likely to generate significant controversy. They created a corpus consisting of 728 news articles from CNN.com. Twenty annotators from different educational backgrounds classified each news article as controversial or noncontroversial and a voting scheme were used to resolve conflicts. It was observed that there was a fair agreement between the annotators. Pre-processing has been done to delete URLs and stop words from the data as well as articles that consist solely of images. They have developed an algorithm to predict divisive patterns by the opinion conveyed in comments, the explosion of comments, and the substance of the controversy score. The sentiment of the text of the statement was evaluated using SentiStrength to identify the view as positive, negative or neutral [11]. The controversial score was then determined by dividing the total number of negative comments by the total number of comments. Certain features such as number of shares, number of comments, number of users, etc. have also been removed. In order to determine how quickly a controversial article will be found, the comments were separated into various time frames and evaluated separately. All features have been standardized between 0 and 1 and the Decision Tree Classifier has been trained for each time period.

An average 71.3% F-score achieved in the detection of controversial topics across all time period. Their results suggest that early predictions is possible about whether topics are likely to generate controversy on social media.

3. METHODOLOGY

It consists of five major aspects. In the first aspect, the gathering articles and comments from various sources was done to create an annotated corpus. In the second aspect, on extracted tweets to remove URLs and attributes of tweet objects, a pre-processing step is performed. In the third aspect, the snapshot is created for tweet entities using a dictionary for a time period. The fourth aspect, extracting features from tweets and external sources which generalise the controversy identification including sentiments and controversy scores. And the last aspect i.e. the fifth one, a learning model of the machine has been developed to detect controversial trends, including recognition, estimation, calculation and analysis.

3.1. Data Collection

To create a corpus of controversial and non-controversial topics, we explored various online forums, social media websites, and news media websites, and gathered a set of entities. We used the technique of scrapping to extract a large set of entities. Then perform some cleaning on scrapped entities, like removed the ones which has a length less than 3 characters. For all the extracted entities, extracted twitter data

Volume 9, Issue 5
October 2020

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

for a time period. The corpus consists of articles published by 'The Times of India', DNA RSS feed on their online portals. The data collection is done using python script to extract a list of news articles.

3.1.1. Annotations

A python script was used for each article in such a way that annotators categorized whether or not the article was controversial [12]. Where there was a disagreement between annotators in the classification of an article, an arbitration scheme is used in which enforced the majority of Class votes. When the annotator marks the article as disputed, then 1, was stored in the "controversial" tab has been saved otherwise 0.

3.1.2. Data Pre-Processing

While extracting the tweets, the ones which did not meet below criteria, we removed them:

- Tweet does not have its own language ['iso_language_code' attribute of the Tweet object] set as English.
- Tweet contains media files – images, videos, gifs – linked to it.

After this first step of cleaning, we created a snapshot of tweets for each entity, as shown in equation 5.

$$triple\ s = (e, \_t, tweets) \quad (5)$$

To create this triple, we used a dictionary in python, in which key is the entity and value is a list of all the tweets for that entity in a given time period and along with each tweet, its creation date is also associated.

3.2 Controversy detection

Articles are been tested against controversy [13], the text sentiments were analyzed using algorithms, it's a positive tweets fraction (i.e. pol(t)<0) (TW-SENT-POS), a negative tweets fraction(pol(t)<0) (TW-SENT-NEG), neutral tweets fraction (i.e. pol(t)=0) (TW-SENT-NEU). A controversy score was calculated after sentiment classification as follows:

$$TW - CONT - MI = \frac{Min(|Pos|,|Neg|)}{Max(|Pos|,|Neg|)} \cdot \frac{|Pos|+|Neg|}{|Pos|+|Neg|+|Neu|} \quad (6)$$

Where Pos, Neg, Neu are the sets of tweets with positive, negative, and neutral polarity. The contradiction score obtained is as follows:

$$TW - CONT - TSY = \frac{\theta \cdot \sigma^2}{\theta + (\mu)^2} \cdot W \quad (7)$$

Where μ and σ^2 are respectively the mean and the variance of scores pol(t) polarity tweets parameters θ and W are as defined in [9]. Four characteristics, which represents the fraction of the total number of hashtags in the snapshot, the following hashtags "#controv", "#scandal", "#unheard 'and' wft. Percent of tweets with the least controversial word in our lexicon controversy.

Other features also calculated – number of 'Likes', number of posts, retweets, hashtags, URLs, media, user tweets, followers, following, user mention., average comment word count, controversy presence in articles were done by controversy corpus created from various sources. In order to determine how quickly a controversial article would be found [14], the respective articles were divided into time intervals of 6hrs, 8hrs and 12hrs [15]. For each time period, features were extracted and the Decision Tree, Random Forest and Support Vector Machine (SVM) with the Gaussian kernel (RBF) classifier was trained and checked preformation using extracted features from comments belonging to a given time interval. All features have been normalized between 0 and 1 using the algorithm described in figure 1.

1. Let buzzy_enities = the number of buzzy enities
2. let profane_words = the list of bad words
3. for key, value in tweet_triple items:
   a. for items in buzzy_entities:
      i. if key is in items:
         1. let buzzy_entity_features[key] = initialize the empty list
         2. temp_dict = get_entity_feature using key, and profane_words
         3. for k,v in temp_dict items:
            - append v to list buzz_entity_features[key]
4. for key,value in buzz_entity_features items:
   a. controversy_score[key] = sum of value divided by length of value
   b. controversy_score[key] = normalize score of controversy_score[key] obtained previously
5. Done: the value in controversy_score[key] is controversial value for an entity

Figure 1. Pseudocode to calculate controversy score from buzzing entities

3.2.1. Buzzy Snapshot Selection

Given that if an entity is addressed more than in the recent past in a given time span, a controversy regarding an entity is likely to emerge in that time period as shown in equation 8.

$$b(s) = \frac{|tweets_s|}{(\sum_{i \in prev(s,N)} |tweets_i|)/N} \quad (8)$$

Where tweets are a collection of tweets in snapshots and prev(s, N) is a collection of snapshots relating to the same individual of 's' in 'N' time periods prior to 's.' We use 'N=2' in our experiment, i.e. we concentrate on two days before 's.' We hold only those with 'b(s)>30' as buzzy snapshots.

3.2.2. Feature Extraction

Our goal was to extract all those features which generalizes the task of controversy identification. We not only extracted features from tweets, along with that we also extracted some features from external sources as 'The Times of India', DNA news articles, Given the assumption that if an entity is buzzy in news stories at the same time that it is buzzy in a Twitter snapshot, then a snapshot, then the snapshot is likely to apply to a real-world event. So we extracted period-issued news articles (t-1, t+1) for entities obtained after a buzzy snapshot range. And below are the data containing the list of features extracted from twitter data and their explanation:

Volume 9, Issue 5
October 2020

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

- Linguistic: There are different types of features that it contains according to usage, such as the percentage of tokens for the speech component specified in English (TW-LING-NOU), verbs percentage token (TW-LING-VRB), bad words percentage token (TW-LING-BAD), The percentage of tweets that include at least one question (TW-LING-QST), average Levenshtein gap between tweets (TW-LING-LEV), matching every word in the English dictionary percentage token (TW-LING-ENG), percentage of verbs whose corresponding subject is the largest entity (TW-LING-VB), total number of mentions of the target entity in all tweets (TW-LING-ENT-OC), percentage of tweets containing at least the verb whose subject is the larger entity (TW-LING-ENT-TW).
- Structural: As the name suggests, it determines the number of tweets in the snapshot (TW-STRC-TOK), the number of tweets in the snapshot (TW-STRC-TWE), the proportion of tweets that are retweets (TW-STRC-RET), percentage of Tweets that reply(TW-STRC-REP), average number of Tweets per user (TW-STRC-USR), two features reflecting the mean and standard deviation of the Tweet Modeling Distribution Timestamps (TW-STRC-TIM), ratio between number of unique hashtags and the total number of hashtags (TW-STRC-HST).

Below are the data containing the list of features extracted from external source:

- News buzz: number of articles associated with the snapshot (EX-BUZZ-1). Increased number of news reports for the company compared to the recent past:

$$EX - BUZZ - 2 = \frac{|articles| - (\sum_{1<t<N} |articles_t|)/N}{|articles|} \quad (9)$$

where $articles_t$ is the number of articles regarding the target entity in the time preceding (we use N = 7)

- Web-News controversy: Level of controversy the entity in Web data:

$$\text{Ex} - \text{CONT} - \text{HIST} = \frac{k}{|contoversyLexicon|} \quad (10)$$

Where k is the number of words in our contentious lexicon, where the point-wise shared knowledge co-occurs with the target individual on the Internet is greater than 2; and Lexicon is the scale of the contentious lexicon. Overview of overall conflict scores (EX-CONT-HIST) for organizations co-occurring with the target party in the associated news item (EX-CONT-ASS-1). Average of cumulative controversy scores (EX-CONT-HIST) for organizations co-occurring with the target party in the associated news item (EX-CONT-ASS-2). Average number of words per controversial news article (general aligned with the snapshot) (EX-CONT-TRM-1). Total number of controversial terms per news report (overall articles aligned with snapshots) (EX-CONT-TRM-2). Number of articles aligned with snapshots containing controversial terms (EX-CONT-TRM-3).

3.2.3. Supervised Classification

Extracted features will be stored in csv file as shown in figure 2. and that will be used to build the Decision Tree

	A	B	C	D	E	F	G	H
1	controvers	hashtags	media	urls	user_ment	favorite	retweet	verified
2	0	2	0	0	7	0	37	0
3	0	2	1	1	1	0	13	0
4	1	1	0	0	1	0	8	0
5	1	0	0	0	1	0	2457	0
6	1	1	0	0	1	0	37	0
7	1	1	0	1	1	0	5	0
8	0	0	0	0	1	0	25	0
9	0	2	0	0	1	0	35	0
10	0	5	0	0	3	0	8	0
11	1	0	0	0	1	0	180	0
12	1	1	0	0	1	0	142	0
13	1	0	0	0	1	0	434	0

I	J	K	L	M	N	O	P	Q
followers	following	lists	user_twee	user_favs	length	uppercase	exclamatic	profanity
114	293	0	1507	1499	17	2	0	0
314	141	0	19064	1466	10	1	0	0
4694	2704	550	512501	102777	18	2	0	0
74	224	0	552	1237	22	3	0	0
129	523	0	441	2841	14	1	0	0
72	119	46	166	158	17	1	5	0
451	507	12	2956	13821	27	2	0	0
269	470	0	21814	21475	21	2	1	0
9336	245	28	18540	2723	13	2	0	0
1101	207	7	30073	4216	19	1	1	0
223	409	1	10073	1219	20	1	0	0
6258	6750	143	86737	12133	25	3	0	0

Figure. 2. Extracted feature from entities

model, Random Forest model, and Support Vector Machine (SVM) model on top of it after dividing data in training and Testing sets.

4. RESULT

To detect tweet is controversial or not, the features were extracted from the twitter along with external sources-based features, are used to analyze the model's performance. These tweets extracted from entities for a given time period, after selecting a buzzy snapshot (entities), the feature is mapped and implemented on three algorithms Decision Tree classifier, Random Forest classifier and Support Vector Machine by dividing them into training and testing dataset to check model performance.

From the controversy corpus containing over 30000 entities were scarped in eight hours, only 27731 entities were chosen since other entities did not have any activity in a given time period of a tweet posted. Entities contain 13553 controversial entity and 14178 were a non-controversial entity. F-measurement was used to measure the performance of the methodology. Formulating F-score, Precision and recall are shown in equation 11, 12 and 13 respectively.

$$Fscore = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (11)$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (12)$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (13)$$

The extracted features have been stored in a csv format where each line consists of a sample. That sample contained vector features accompanied by a tab-separated description of the sample. The text file has been used to train and evaluate the classifier models. There were two groups – controversial and non-controversial. To evaluate the output of different classifiers, a dataset with features derived from 27731 entities in the time interval was used.

Performance has been contrasted between Decision Tree, Support Vector Machine (SVM) and Random Forest Classifiers. SVM is a supervised learning approach that optimizes the data separation margin. Random Forest works by constructing a multitude of decision trees at the time of training and generating a class is the class fashion production by individual trees. Decision Tree is a rule-based classifier that correlates the characteristics with the acts to be performed; it does not presume that the attributes are separate.

Decision Tree Classifier has been used for training and testing over all time since it received the best results (88.0%) compared to SVM (57.0%) and Random Forest (90.0%). The description of output comparisons between SVM, Random Forest and Decision Tree Classifiers is shown in table 1.

Table 1: Performance comparison between different classifiers

Controversial	0.73	0.88	0.90
Non-Controversial	0.42	0.88	0.91
Avg/Total	0.57	0.88	0.90

5. CONCLUSION

We presented a method for recognize controversy in tweet posted by user, which will give a helping hand in stopping misleading and rumors rise due to the tweet. The experiments were carried out on the twitter corpus, for dataset, rich feature extracted from set of tweets which are controversial in nature and using those features with external sources we determine controversy score of that entity. To get the result, twitter corpus gone through many stages to give probability of an entity or tweet is controversial or not. These stages include data collection and controversy detection. And in controversy detection main step is to find entity being addressed in a given time period. Our result show comparison between three algorithms SVM, Decision Tree, and Random Forest to give probability of entity being controversial or not.

REFERENCES

[1] A. J. Morales, J. Borondo, J. C. Losada and R. M. Benito. “Measuring Political Polarization: Twitter shows the two sides of Venezuela”.

[2] Garimella, K., Morales, G. D., Gionis, A., & Mathioudakis, M. (2016). Quantifying Controversy in Social Media. Proceedings of the Ninth ACM International Conference on Web Search and Data Mining - WSDM '16. doi:10.1145/2835776.2835792.

[3] Addawood, A., & Bashir, M. (2016). "What Is Your Evidence?" A Study of Controversial Topics on Social Media. Proceedings of the Third Workshop on Argument Mining (ArgMining2016). doi:10.18653/v1/w16-2801.

[4] Signorini, Alessio & Segre, Alberto & Polgreen, Philip. (2011). The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. During the Influenza A H1N1 Pandemic. PloS one. 6. e19467. 10.1371/journal.pone.0019467.

[5] April Kontostathis, Leon M. Galitsky, William M. Pottenger, Soma Roy, Daniel J. Phelps. “A Survey of Emerging Trend Detection in Textual Data Mining”.

[6] Choi Y., Jung Y., Myaeng SH. (2010) Identifying Controversial Issues and Their Sub-topics in News Articles. In: Chen H., Chau M., Li S., Urs S., Srinivasa S., Wang G.A. (eds) Intelligence and Security Informatics. PAISI 2010. Lecture Notes in Computer Science, vol 6122. Springer, Berlin, Heidelberg.

[7] Popescu, A., & Pennacchiotti, M. (2010). Detecting controversial events from twitter. Proceedings of the 19th ACM International Conference on Information and Knowledge Management - CIKM '10. doi:10.1145/1871437.1871751.

[8] Addawood, A., Rezapour, R., Abdar, O., & Diesner, J. (2017). Telling Apart Tweets Associated with Controversial versus Non-Controversial Topics. Proceedings of the Second Workshop on NLP and Computational Social Science. doi:10.18653/v1/w17-2905.

[9] Fabian Abel, Claudia Hauff,GeertJan Houben, Ke Tao, Richard Stronkman “Semantics + Filtering + Search = Twitcident Exploring Information in Social Web Streams”.

[10] Chimmalgi, R. V. (2013). Controversy trend detection in social media (Unpublished master's thesis).

[11] Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. Journal of the American Society for Information Science and Technology, 63(1), pages 163-173. doi: 10.1002/asi.21662

[12] Wikipedia. (2012, 9/7/2012). Wikipedia: List of controversial issues Retrieved 9/10/2012, 2012, from http://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues

[13] Anifowose, O. P. (2016). Identifying Controversial Topics in Large-scale Social Media Data (Master’s thesis).

[14] Vuong, B., Lim, E., Sun, A., Le, M., & Lauw, H. (2008). On ranking controversies in wikipedia: models and evaluation. In Proceedings of the international conference on Web search and web data mining, Palo Alto, California, USA, pages 171-182.

[15] Takahashi, T., Tomioka, R., & Yamanishi, K. (2011). Discovering Emerging Topics in Social Streams via Link Anomaly Detection. In 2011 IEEE 11th International Conference on Data Mining (ICDM), pages 1230-1235.

AUTHOR PROFILES

Harshav Kumar received his degree in computer science & Engineering from BK Birla Institute of Engineering & Technology, Pilani(Raj), India. His interest area lies in computer vision, natural language processing, and Scalable Machine learning.

Gopal Krishan Prajapat received his Bachelor of Engineering degree from University of Rajasthan, Jaipur, India, and Master of Engineering from Panjab University, Chandigarh, India in the field of Computer Science & Engineering. His area of interest are Computer Vision, Image Processing and Data Analytics. He is working as an Assistant Professor in Department of Computer Science & Engineering of BK Birla Institute of Engineering & Technology, Pilani(Raj.),India.