Analysing TV Audience Engagement via Twitter: Incremental Segment-Level Opinion Mining of Second Screen Tweets

Gavin Katz, Bradford Heap, Wayne Wobcke(✉), Michael Bain, and Sandeepa Kannangara

School of Computer Science and Engineering,
University of New South Wales, Sydney, NSW 2052, Australia
gavinkatz@gmail.com, {b.heap,w.wobcke,m.bain,s.kannangara}@unsw.edu.au

Abstract. To attract and retain a new demographic of viewers, television producers have aimed to engage audiences through the "second screen" via social media. This paper concerns the use of Twitter during live television broadcasts of a panel show, the Australian Broadcasting Corporation's political and current affairs show Q&A, where the TV audience can post tweets, some of which appear in a tickertape on the TV screen and are broadcast to all viewers. We present a method for aggregating audience opinions expressed via Twitter that could be used for live feedback after each segment of the show. We investigate segment classification models in the incremental setting, and use a combination of domain-specific and general training data for sentiment analysis. The aggregated analysis can be used to determine polarizing and volatile panellists, controversial topics and bias in the selection of tweets for on-screen display.

Keywords: Social TV · Opinion mining · Machine learning
Social media

1 Introduction

In order to attract a new demographic of viewers, traditional broadcast media have explored ways to embrace new technologies to provide a more engaging and interactive TV viewing experience. A common approach is to augment a broadcast television show through the use of social media channels. This means of audience participation is known as the "second screen" [7]. Common uses of the second screen include voting for contestants on game shows or reality TV, posting messages about a show or the current episode, and communicating with cast members as the show airs. Previous computational work on "second screening" [3,8] has studied sentiment in human annotated datasets [1], time series analysis over the frequency of tweets and human coded content analysis [3], and the interaction between Twitter users [5].

X. Geng and B.-H. Kang (Eds.): PRICAI 2018, LNAI 11013, pp. 300–308, 2018.
https://doi.org/10.1007/978-3-319-97310-4_34

This paper presents a method to analyse audience opinion via the second screen during live television broadcasts of a panel show, the Australian Broadcasting Corporation's weekly political and current affairs show Q&A. During a live broadcast of the program, viewers are encouraged to engage with what is being discussed on the show by posting tweets containing the hashtag `#QandA`, some of which are chosen for display on a tickertape on the bottom of the screen. This mechanism enables feedback, directed towards both topical segments and panellists, to be given in real time, potentially enabling the host to comment on the feedback during the show, or even adjust the content of the show in response to viewer feedback.

In this paper, we show how the model could be used for aggregation of opinions over entities and across segments, to address the questions of identifying polarizing panellists, controversial topics and bias in the selection of tweets for display. We investigate *incremental* opinion mining methods that compute aggregated audience opinions at the end of each segment of an episode. The model classifies audience tweets into two components: the *segment* (part of the show the tweet is about) and the *target* (a panellist or person or organization discussed in the show). Each segment is defined by a question from the studio audience, usually on a topical political issue, which is then discussed by the panellists. There are around 7–8 segments in a typical episode, which runs for 65 min. Issues of computational efficiency and training data are paramount. Importantly, the show is not pre-scripted and questions are not announced in advance, and moreover, there is little overlap between questions from week to week, so it is impossible to use previous episodes as training data for segment classification. Our key idea is to use the episode transcript itself – which is available in real time – as training data for incremental models. We investigate classification of tweets into segments using Multinomial Naïve Bayes (MNB) and Support Vector Machines (SVM) in an incremental setting where models are retrained and run at the end of each segment, using the transcript up to that point, to classify the tweets posted during that segment.

For evaluation, we selected three episodes of Q&A – Episode 1: August 1, 2016 (8 segments); Episode 2: September 12, 2016 (8 segments); Episode 3: September 19, 2016 (7 segments). The dataset consists of all tweets containing the `#QandA` hashtag posted across the live broadcasts from 9.30pm (the start of the show) until 11.00pm AEST, i.e. until around 25 min after the end of the episode, and contains 17,044 tweets from 3,696 distinct authors for Episode 1, 17,274 tweets from 3,029 distinct authors for Episode 2, and 15,983 tweets from 3,765 distinct authors for Episode 3. A ground truth dataset was defined by manually labelling a random sample of 1,659 distinct tweets posted during the live broadcasts of the three episodes. A tweet is categorized into a segment if it references any part of the discussion aired, or is specifically targeting any of the themes raised in the discussion. Otherwise, if the tweet is relevant to the episode but not to any specific segment, it is labelled under a *General* category (300 tweets in the ground truth dataset).

2 Incremental Segment Classification

To explain the idea behind the incremental segment classification models, let the segments of an episode be $S_1, S_2, \cdots$. Then, at the end of each segment S_n (starting with S_3, when there is sufficient training data), a new model M_n is built by training on all the transcript text (ignoring annotations) up to the end of S_n. Thus models $M_3, M_4, \cdots$ are created up to the end of the episode. The models are used incrementally to classify the tweets posted in each new segment: more precisely, M_n is always used to classify the tweets posted during S_n, but note that these tweets could be classified as relating to any of the segments $S_1, \cdots, S_n$ (though not subsequent segments). Equivalently, we can think of the "incremental" model I_n, computed at the end of segment S_n, as a model that uses the time a tweet was posted to "select" the model M_i $(i \leq n)$ to classify the tweet into one of the segments $S_1, \cdots, S_n$.

Use of an incremental model requires a base model for the classification of tweets posted during S_n into the classes $S_1, \cdots, S_n$. We evaluate two commonly used supervised classification models in this setting: Multinomial Naïve Bayes (MNB) [6] and Support Vector Machines (SVM) [4], using Weka's[1] implementation of the Sequential Minimal Optimization (SMO) training algorithm. These algorithms have previously been shown to train quickly and perform well on limited training datasets [2], thus are suitable for use in the context of the show to train and run the classifiers at the end of each segment. To maximize classification accuracy, each algorithm has a tuned text pre-processing stage, and to maximize each classifier's F1 measure, threshold functions are defined for each method, acting as a "confidence level" for a given tweet to be classified: any tweet whose confidence is below the threshold is assigned the *General* class.

Pre-processing techniques were developed for the entire system to use, and then a pre-processing pipeline was chosen for each algorithm that was tuned to maximize the F1 measure on training data. For the training data, all punctuation and annotations were removed and text was reduced to lower case. This results in a transcript that contains only the exact words spoken on the show. Tweets were also subject to two additional pre-processing steps: (1) any mention of an account handle (i.e. word starting with "@") is converted to the full name for that account; and (2) hashtags that run together separate words beginning with a capital letter are split into several words (e.g. the hashtag `#KevinRudd` is converted to "Kevin Rudd"). Finally, for MNB (but not SVM), stop words are removed and words lemmatized, as is standard practice for achieving the best results with these models on text classification problems.

2.1 Segment Classification Evaluation

Incremental segment classification is evaluated using the three episode ground truth dataset described above. MNB and SVM classification models are used as the base models and trained incrementally on the transcript data for these

[1] http://www.cs.waikato.ac.nz/ml/weka/.

episodes. To emphasize that the segment-based incremental model is being used, we call the incremental models I-MNB and I-SVM. These models are compared to a baseline classifier that is a simple time-based classifier where each tweet posted during a segment is assigned to that same segment class, i.e. simply assumes each tweet is relevant to the currently airing segment, as in Diakopoulos and Shamma [1].

Table 1 shows the precision, recall and F1 measure for all tweets in the ground truth dataset for each of the three episodes. These results show that the I-MNB classifier is consistently the best model for classifying tweets into segments. The simple baseline works surprisingly well (though of course fails to capture the "lag" between Twitter stream and show), and even outperforms I-SVM on Episodes 1 and 2. Hence all further analysis in this paper is done with I-MNB.

Table 1. Episode segment classification

	Episode 1			Episode 2			Episode 3		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Baseline	0.69	**0.82**	0.75	0.63	0.77	0.69	0.49	0.59	0.54
I-MNB	**0.80**	0.79	**0.79**	**0.78**	**0.79**	**0.79**	**0.68**	**0.70**	**0.69**
I-SVM	0.59	0.61	0.60	0.60	0.65	0.63	0.52	0.59	0.55

Table 2 shows the per-segment precision, recall and F1 for each of the segment classes in Episode 3, as calculated at the end of the episode, i.e. using I_7, the model that uses M_i to classify the tweets posted during the segment S_i into $S_1, \cdots, S_i$ (and M_3 to classify tweets posted during S_1, S_2 and S_3). Segment 2 (*Plebiscite and Suicide*) has an outlying (low) F1 measure. In this case, the only other question discussed previously in this episode is the first segment, *Plebiscite or Wait*. As there is a drop in F1 measures (from 0.63 to 0.53) between Segment 1 and Segment 2, this suggests that the M_3 classifier, which has to decide between the first three segments, heavily favours the *Plebiscite or Wait* segment: the words overlap between the two segments, and for MNB the length of the training text has a large effect on the behaviour of the classifier. In particular, *Plebiscite or Wait* has a length of 1,136 words, significantly higher than both the average length and the *Plebiscite and Suicide* length of 700 words.

Table 2. Episode 3 - end of episode analysis, MNB

	S_1 Plebiscite or wait	S_2 Plebiscite and suicide	S_3 Hanson speech	S_4 Asylum	S_5 Migrants today	S_6 Alcohol abuse	S_7 Creative copyright
Precision	0.52	0.64	0.80	0.81	0.61	0.88	0.87
Recall	0.79	0.46	0.75	0.79	0.75	0.59	0.72
F1	0.63	0.53	0.78	0.80	0.67	0.71	0.79

Table 3. Episode 3 - end of segment analysis, F1 measure

	S_3 Hanson speech	S_4 Asylum	S_5 Migrants today	S_6 Alcohol abuse	S_7 Creative copyright
Baseline	0.46	0.43	0.46	0.60	0.63
I-MNB	0.88	0.77	0.77	0.73	0.79

Table 3 shows the F1 measure for each segment of Episode 3, calculated at the end of the segment, comparing the score for only that class on tweets posted during that segment for the two classifiers, the MNB model constructed at the end of the segment and the baseline time-based classifier. This is the scenario that arises if those scores were to be calculated during the episode for the most recently aired segment. Immediately noticeable is the very poor performance of the baseline, indicating that in this episode much of the Twitter discussion relating to a segment spills over into following segments. The results for I-MNB are consistent with Table 2, showing that the method can be used incrementally in the desired fashion. F1 is expected to be higher in the per-segment calculations because only F1 for the current class is reported. Taken as a whole, this indicates that I-MNB could be used during the episode for classification of tweets into segments during the show.

3 Aggregated Audience Opinion

We now consider the main purpose of the "second screen" analysis, to address *aggregated* audience opinion, in a mode that, again, could be presented during a show at the end of each segment in a live broadcast. Sentiment analysis was done using SVM trained on a combined general and Q&A specific corpus (for each episode, trained using a general Twitter sentiment analysis corpus[2] and the sentiment-tagged tweets from the other two episodes), which, consistent with prior research, provided the best results, with precision 59%, recall 56% and an F1 measure of 58%.

Specific questions of interest are: (i) which panellists are polarizing, (ii) which panellists are "volatile" in the sense that sentiment towards them varies according to the segment, (iii) which topics are most controversial, and (iv) whether there is any bias in the selection of tweets for on-screen display. This relies on a way to aggregate audience tweets – this is done by identifying the explicit target of each tweet and then clustering tweets with the same target, as determined using single-linkage clustering based on the Sørensen-Dice coefficient with a similarity threshold of 0.15. An aggregated sentiment for an entity cluster or a segment is calculated by taking the mean sentiment over a collection of tweets, where *positive* is 1, *neutral* is 0, and *negative* is −1.

Entity recognition uses two standard tools to identify entities within tweets: Basis Technology's Rosette Entity Extractor (REX) and Stanford CoreNLP

[2] http://www.mpi-inf.mpg.de/~smukherjee/data/twitter-data.tar.gz.

Named Entity Recognition (NER), run with their default settings. If a tweet includes multiple entities, the most important entity is assumed to be the target, and is determined heuristically using a hierarchy of entity types: (i) person, organization, (ii) location, nationality, religion, (iii) title, and (iv) all other types.

3.1 Polarizing and Volatile Panellists

Table 4 shows the number of tweets in the whole Q&A dataset labelled by each of the three methods, and the total number of complete tuples, where a complete tuple is of the form $\langle e_i, c_{ij}, s_{ij} \rangle$ with an entity, segment and sentiment. The methods used are incremental MNB for segment classification and SVM for sentiment classification.

Table 4. Tweets labelled for segment, sentiment and entities

Episode	Tweets	Segment classified	Sentiment classified	Entities recognized	Complete tuples
1	17,044	13,933	16,286	11,586	9,195
2	17,274	13,018	16,529	12,588	9,107
3	15,983	13,343	15,238	12,061	9,708

For each episode, more than 50% of tweets are assigned a complete tuple that can be used for opinion aggregation by segment and entity. However, since a tweet is labelled with the *General* category if the confidence for classifying it into any specific segment falls below a threshold, a large number of tweets containing entities are not assigned to any segment. In this case, the tweet is almost certainly directly about an entity, often a panellist but sometimes also the subject of the discussion. We use tweets of this form to assess aggregated opinion towards entities across an entire episode.

We calculate the mean absolute deviation (MAD) of the sentiments, and the standard deviation and 95% confidence interval (CI) around the mean sentiment. These measures are used to determine the degree of polarity in the tweets that contribute to the mean sentiment. A large MAD, standard deviation and confidence interval suggests that the authors of tweets have divergent views, while a small MAD and standard deviation suggests that the majority of tweeters are in agreement on the sentiment.

Table 5 shows some sample results of the aggregated opinion mining on tweets in the *General* category from the three episodes. All four entity clusters shown in this table correspond to panellists on the episodes: Matt Canavan,[3] Jimmy Barnes, Magda Szubanski[4] and Jacqui Lambie. Across all of these entity clusters,

[3] Resources minister in the government.
[4] Lesbian actor/comedian, strong supporter of same sex marriage.

Table 5. Aggregated opinion – general segment classification

Entity cluster	Tweets	Mean sentiment	MAD	Std deviation	95% CI
MATT CANAVAN, ···	124	−0.46	0.73	0.83	−0.61 to −0.31
JIMMY BARNS, JIMMY, ···	130	−0.34	0.76	0.84	−0.48 to −0.20
YAY MAGDA, MAGDA, ···	89	0.17	0.77	0.85	−0.01 to 0.35
JACQUIE LAMBIE, ···	82	−0.07	0.81	0.89	−0.26 to 0.12

Table 6. Aggregated opinion – specific segment classification

Entity	Tweets	Mean sentiment	MAD	Std deviation	95% CI
Matt Canavan	186	−0.40	0.76	0.84	−0.52 to −0.28
Jimmy Barnes	96	−0.77	0.40	0.60	−0.89 to −0.65

there is a similar high level of standard deviation, indicating a wide variety of sentiment.

As Q&A aims to discuss topical controversial issues, it is unsurprising that much of the sentiment is negative; the confidence intervals for the first three panellists in the table are all clearly in the negative. On the other hand, the entity cluster for Magda Szubanski, with a mean sentiment of 0.17, is highly unusual, being the only entity cluster studied with a positive skew. The last entity cluster for Jacqui Lambie is also unusual in having a high MAD and standard deviation and a neutral mean sentiment. From this it can be concluded that Jacqui Lambie is more polarizing than other panellists, provoking a wider range of sentiment from the TV audience.

Our method of opinion mining also allows tweets to be aggregated by entity within segments. Table 6 shows the results for two of the same panellists, Matt Canavan and Jimmy Barnes, during specific (different) segments when they were active participants in the discussion. What is noticeable is that all values for Matt Canavan are virtually identical to those in Table 5, indicating that sentiment towards him during the segment is representative of overall sentiment. On the other hand, sentiment towards Jimmy Barnes is more volatile, in that sentiment is much more negative towards him during this segment than overall: the mean sentiment in this table (−0.77) falls well outside the confidence interval in Table 5, in support of this conclusion.

3.2 Controversial Segments

Our methods can be used to aggregate sentiment over all tweets classified as belonging to a segment, and to determine controversial topics, we focus on outliers on these metrics. Note that for most segments, the values tend to follow the trends for overall sentiment of Q&A tweets discussed above, so are not revealing. One outlier on standard deviation is the segment *Gun Laws and Terrorism* which has standard deviation of 0.88, indicating an even more highly controversial topic than usual.

Table 7. Aggregated opinion – specific segment classification

Entity	Tweets	Mean sentiment	MAD	Std deviation	95% CI
On-screen tweets					
Kevin Rudd	6	−0.67	0.44	0.47	−1 to −0.29
Tony Abbott	5	−0.6	0.64	0.89	−1 to 0.18
Jacqui Lambie	2	1	0	0	1 to 1
Q&A tweets					
KRUDD, RUDDS, ···	600	−0.49	0.67	0.76	−0.55 to −0.43
TONY ABBOT, ···	301	−0.54	0.65	0.77	−0.63 to −0.46
JACQUIE LAMBIE, ···	180	−0.03	0.84	0.91	−0.16 to 0.10

3.3 Bias in Selection of Broadcast Tweets

As part of the broadcast of Q&A, between 12–20 tweets are chosen manually for on-screen display during the panel discussion. Table 7 shows three examples of people about whom tweets were displayed (in different segments) and who provoked a large number of audience tweets: Kevin Rudd,[5] Tony Abbott[6] and Jacqui Lambie. Only the last of these was a panellist. The confidence intervals for the on-screen tweets for the first two are quite wide (−1 to −0.29 for Rudd; −1 to 0.18 for Abbott), and the mean sentiment derived from the Q&A tweets falls within these ranges. Moreover, the confidence intervals for the Q&A tweets are completely contained within those for the on-screen tweets. Lambie is an anomaly, with only two tweets shown, both positive, whereas, as noted before, she is a polarizing panellist.

4 Conclusion

In this work, we have presented a method for computing the aggregated opinion of Twitter users towards entities within segments discussed on the popular Australian current affairs panel television show Q&A. The idea is that aggregated opinions from the "second screen" could be presented during the show at the end of each segment, and used to provide further feedback to the TV audience.

The key insight is to develop an incremental, segment-level, opinion mining model that can be trained on the transcript of the episode. We used an incremental version of Multinomial Naïve Bayes for classification of tweets into segments, and Support Vector Machines trained on a combination of a general Twitter sentiment corpus and specific Q&A tweets for sentiment classification. We showed how these techniques can be used to address the questions of which panellists most polarize the audience, which panellist's sentiments fluctuate according to the topical segment, which topics are most controversial, and whether there is bias in the selection of tweets for on-screen display.

[5] Former Labor Australian Prime Minister.
[6] Former Liberal Australian Prime Minister.

Acknowledgement. Thanks to Data to Decisions Cooperative Research Centre for supporting this research and supplying full access to the Twitter data for this paper.

References

1. Diakopoulos, N., Shamma, D.A.: Characterizing debate performance via aggregated Twitter sentiment. In: Proceedings of the 28th ACM Conference on Human Factors in Computing Systems, pp. 1195–1198 (2010)
2. Forman, G., Cohen, I.: Learning from Little: Comparison of classifiers given little training. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.): PKDD 2004. LNCS (LNAI), vol. 3202. Springer, Heidelberg (2004). https://doi.org/10.1007/b100704
3. Giglietto, F., Selva, D.: Second screen and participation: a content analysis on a full season dataset of tweets. J. Commun. **64**, 260–277 (2014)
4. Joachims, T.: Text categorization with Support Vector Machines: learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998). https://doi.org/10.1007/BFb0026683
5. Lin, Y.R., Keegan, B., Margolin, D., Lazer, D.: Rising tides or rising stars?: Dynamics of shared attention on Twitter during media events. PloS ONE **9**(5), e94093 (2014)
6. McCallum, A., Nigam, K.: A comparison of event models for Naive Bayes text classification. In: Sahami, M. (ed.) Learning for Text Categorization: Papers from the AAAI Workshop. AAAI Press, Menlo Park, CA (1998)
7. Proulx, M., Shepatin, S.: Social TV. Wiley, Hoboken, NJ (2012)
8. Vaccari, C., Chadwick, A., O'Loughlin, B.: Dual screening the political: media events, social media, and citizen engagement. J. Commun. **65**, 1041–1061 (2015)