Advances in Intelligent Systems and Computing 767

Atilla Elçi
Pankaj Kumar Sa
Chirag N. Modi
Gustavo Olague
Manmath N. Sahoo
Sambit Bakshi *Editors*

Smart Computing Paradigms: New Progresses and Challenges

Proceedings of ICACNI 2018, Volume 2

Springer

Better Quality Classifiers for Social Media Content: Crowdsourcing with Decision Trees

Ian McCulloh, Rachel Cohen and Richard Takacs

Abstract As social media use grows and increasingly becomes a forum for social debate in politics, social issues, sports, and brand sentiment; accurately classifying social media sentiment remains an important computational challenge. Social media posts present numerous challenges for text classification. This paper presents an approach to introduce guided decision trees into the design of a crowdsourcing platform to extract additional data features, reduce task cognitive complexity, and improve the quality of the resulting labeled text corpus. We compare the quality of the proposed approach with off-the-shelf sentiment classifiers and a crowdsourced solution without a decision tree using a tweet sample from the social media firestorm #CancelColbert. We find that the proposed crowdsource with decision tree approach results in a training corpus with higher quality, necessary for effective classification of social media content.

Keywords Social media · Sentiment · Classifier · Machine learning · Decision tree · Twitter · Turk

1 Introduction

Developing automated classifiers for social media content is an important problem for data scientists and privacy researchers. People are increasingly using social media to express opinions on a wide range of issues spanning politics, social injustice, corporations, sports teams, and more. User's opinions can vary, ranging from support to opposition. Data scientists may apply automated classifiers to these data to understand public sentiment toward certain brands or social issues that are being debated

I. McCulloh (✉) · R. Cohen · R. Takacs
Johns Hopkins University, 11100 Johns Hopkins Road, 20723 Laurel, MD, USA
e-mail: ian.mcculloh@jhuapl.edu

R. Cohen
e-mail: rachel.cohen@jhuapl.edu

R. Takacs
e-mail: richard.takacs@jhuapl.edu

A. Elçi et al. (eds.), *Smart Computing Paradigms: New Progresses and Challenges*,
Advances in Intelligent Systems and Computing 767,
https://doi.org/10.1007/978-981-13-9680-9_8

online. Privacy researchers may be interested in using these classifiers to identify specific individuals within those online discussions to target key users or communities for influence interventions. Developing automated classifiers to measure support and opposition is complicated by several factors to include short size of text [1–3], sarcasm [3–7], humor [5, 7, 8], political alignment [9], emoticons [10], among other factors. We investigate methods to overcome these issues using crowdsourcing and decision trees.

Twitter is an online news and social networking service that allows users to post and interact with 280-character messages called "tweets". As of the second quarter of 2017, the microblogging service averaged 328 million monthly active users, making it the eighth most popular social network in the world [11]. As the number of users continues to rise, Twitter, as a social platform, demonstrates the ability of social media to influence politics and social debate. As a result, Twitter's rich data can provide insight as to how the general public perceives a topic. With that value realized, a single tweet object can serve as a trove of metadata, including the tweet's content, the location it was sent from, and the time it was sent. Twitter provides a well-documented application programming interface (API) as a way to query Twitter data and fetch results in a standardized format, which can then be parsed for areas of specific interest.

Developing text classifiers for Twitter data using a supervised learning approach requires a gold-standard training dataset to develop and evaluate the veracity of potential classifiers. While there exist several off-the-shelf text classifiers for Twitter data [15–19], they may not be tailored to specific applications. We posit that the nature of an online firestorm (large, negative, online discourse) may be fundamentally different than the nature of discourse comprising the off-the-shelf training corpus. Furthermore, the performance of various off-the-shelf classifiers may differ and exhibit varied performance when applied to newly captured data. Finally, we posit that achieving agreement among humans classifying tweets in emotionally charged firestorms is highly problematic due to personal bias and conflation of sentiment, position, humor, sarcasm, and other challenges of assessing micro-blog data.

In this paper, we propose a method to develop a high-quality, gold-standard training corpus tailored for firestorms. The proposed method utilizes crowdsourcing and guided decision trees to aid people in systematically labeling tweets. We contrast this approach with off-the-shelf classifiers and crowdsourcing without decision trees. We demonstrate that crowdsourcing with guided decision trees improves the quality and feature space of the training corpus for developing automated classifiers.

2 Background

This work contributes to the study of online firestorms. Firestorms are defined as "an event where a person, group, or institution suddenly receives a large amount of negative attention [online]" [12]. A firestorm can be characterized by an instance where sudden negative attention is in response to a recent action and arises without prior discussion. For the purpose of this research, the focus was placed on online protest and social debate—these events are often fast moving and often have part in influencing the public's perception of an issue.

This paper will focus on a specific firestorm that erupted in 2014. The firestorm surrounds comedian, Stephen Colbert, and his show *The Colbert Report* on Comedy Central. The hashtag #CancelColbert was given to the controversy that began after a tweet was sent from a Twitter account associated with Colbert's then show. The tweet, that was offensive to many Asian-Americans, seemed to many as a mockery of Asian speech. A certain individual, activist Suey Park, started the campaign with a single tweet: "The Ching-Chong Ding-Dong Foundation for Sensitivity to Orientals has decided to call for #CancelColbert. Trend it" [13]. The hashtag grew in popularity and ultimately made it on Twitter's list of trending topics for a nontrivial period, but it did not come without response from supporters of Colbert in defense of his comedic style. Because of this polarity, this firestorm shows the importance of detecting sentiment around a given hashtag, as individuals can tweet (using the hashtag) having strayed from the sentiment that the original author had hoped for.

The data used in this paper are selected from a corpus of 80 firestorms presented by Lambda et al. [12]. Their data were obtained using Twitter's decahose, which represents an approximately 12% random sample of the Twitter content associated with the #CancelColbert firestorm. The additional 2% above the 10% of tweets is obtained by extracting retweet and mention messages from the 10% random sample of the corpus. Their #CancelColbert firestorm sample consisted of 10.1 MB of data and included 15,591 unique tweets. A sample of 200 tweets from this corpus was selected at random for use in an Amazon Mechanical Turk (AMT) experiment and in comparison with off-the-shelf classifier performance. While limitations of Twitter's sampling methodology are noted [14], these data represent a sufficient corpus for the purpose of evaluating construction of a gold-standard training set.

Several off-the-shelf classifiers exist for assessing sentiment within Twitter data. AFINN, from Finn Årup Nielsen, is an English wordlist-based approach for sentiment analysis. The AFINN lexicon assigns words with a score that runs between −5 and 5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment [15, 16]. The NRC Emotion Lexicon, from Saif Mohammad and Peter Turney, is a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). This lexicon categorizes words in a binary fashion ("yes", "no") if it fits into one of the emotion categories [17, 18]. The Bing classifier, by Liu et al., categorizes words in a binary fashion ("positive", "negative") [19]. When applied to a data corpus, however, each of the off-the-shelf sentiment classifiers differs somewhat in their assessment of sentiment. There are many potential reasons for this such as how classifiers treat the presence of sarcasm, humor, and colloquial symbols or text. Differences in classifier performance, however, bring into question the veracity of a given classifier for firestorm sentiment analysis. Developing a tailored classifier for firestorms requires the construction of a gold-standard training corpus.

AMT is an on-demand crowdsourcing marketplace that allows individuals to request work from others online. This marketplace allows people to complete "human intelligence tasks" (HIT)—tasks that humans can currently do more intelligently than computers. It allows requesters to crowdsource data from tasks ranging from object detection in photos to text translation. The requester of the work specifies how many workers can complete a task, determines a monetary value to reward them with, and for the case of sentiment analysis of tweets, asking several workers to provide annotations for each tweet will improve the accuracy of the results. By having multiple annotators assess the tweets provided in the task, there will be random overlap of workers annotating the same tweet.

An important measure of quality in training data is the inter-annotator agreement (IAA). IAA is the level of agreement between raters (annotators), which is high if all raters consistently agree when independently labeling data and low when they disagree. Krippendorff's Alpha is best suited for this data because it can be adjusted for a variable number of annotators assessing different tweets, handles missing data, and is uniformly more powerful than competing methods [20–23]. Missing data will be an important consideration when using a decision tree approach, where annotators have the option of choosing "*Not Applicable*" when coding text labels. Alpha (α) is given by:

$$\alpha = 1 - \frac{D_\mathrm{o}}{D_\mathrm{e}}$$

where D_o is the disagreement observed:

$$D_\mathrm{o} = \frac{1}{n} \sum_{c \in R} \sum_{k \in R} \delta(c, k) \sum_{u \in U} m_u \frac{n_{cku}}{P(m_u, 2)}$$

and D_e is the expected disagreement:

$$D_\mathrm{e} = \frac{1}{P(n, 2)} \sum_{c \in R} \sum_{k \in R} \delta(c, k) P_{ck}$$

where P_{ck} is the number of possible pairs that could be made. Here, $\alpha = 1$ indicates 100% reliability and $\alpha = 0$ indicates a complete lack of reliability [22].

The resulting α ranges from -1.0 to 1.0, where a 1.0 indicates perfect agreement and -1.0 indicates perfect disagreement. The benchmark for random chance agreement scales with the number of annotators and converges to 0.0 as the number of annotators becomes large. Beyerl et al. [21] proposed a benchmark of quality using Krippendorff's alpha, where $0.66 < \alpha < 0.80$ is good agreement, corresponding to approximately 80% probability of joint agreement (PJA) and $\alpha > 0.80$ is excellent agreement, corresponding to better than 90% PJA.

3 Methods

Research consisted of three phases. The first phase involved exploring the competing performance of the three off-the-shelf sentiment classifiers. The second phase involved conducting an AMT experiment to compare IAA with the off-the-shelf classifiers. The third phase involved incorporating a guided decision tree within Amazon Mechanical Turk to extract additional data features and structure the responses of the crowdsourced annotators. For each phase of analysis, we used the random sample of 200 #CancelColbert tweets from the Lambda et al. [12] firestorm corpus. The same dataset is used across all phases for more accurate comparison and was limited to a sample of 200 for cost efficiency. IAA is calculated using Krippendorf's Alpha for sentiment ratings of positive/negative/neutral across the four off-the-shelf sentiment classifiers. IAA is calculated for the second phase of the Amazon Mechanical Turk experiment positive/negative/neutral and benchmarked against off-the-shelf classifiers. Finally, IAA is calculated for the third phase of the AMT experiment using guided decision trees.

AMT was used to crowdsource coding of tweets for phases two and three. AMT annotators were paid a reward of $0.07 per HIT. Each tweet was coded by nine independent annotators. Annotators coded between one and 20 tweets. All annotators were "Masters", who are seen as top workers in the AMT marketplace and have been awarded this qualification due to their high degree of success across different AMT tasks.

For phase two, annotators were asked to code tweets on a five-point Likert scale, ranging from strongly positive to strongly negative. They were also provided the option of "I Don't Know". For each option, the AMT worker was provided guidance for their choice. For example, the strongly positive rating was accompanied by the guidance, "Select this if the item embodies emotion that contains extreme support, satisfaction, or general happiness. For example, 'This candidate is going to change our country for the better. I support them all the way'". In the authors' prior work using AMT, providing workers guidance increased the inter-annotator agreement when viewing responses. This guidance is provided to establish the best-case IAA score in comparison with the guided decision tree approach.

For the guided decision tree, annotators were asked to respond to four questions. For two of the four questions, the annotator was asked to respond to a sub-question depending upon their response. The questions for the guided decision tree were:

1. Does the author of the tweet express a certain position regarding canceling The Colbert Show?
    a. If yes, what is the author's position regarding canceling The Colbert Show?
2. Does the author of the tweet convey a clear sentiment toward the subject of the tweet?
    a. If yes, is the sentiment positive or negative?

   b. If no, is the tweet a straightforward reporting of facts or a headline with/without a link?

3. Did the author of the tweet use sarcasm to mock or convey contempt toward the subject of the tweet?
4. Did the author of the tweet use humor to provoke laughter or amusement?

Prior to beginning the AMT task, subjects acknowledged the following instruction: *The purpose of this project is to analyze the sentiment of politically, socially, and/or culturally charged tweets. Please try and identify them in this context. Keep in mind that you are to rate the Twitter user's sentiment or position, and that you are NOT rating the tweet based on your personal feelings on the subject. Please answer the following questions for a given tweet.*

Structuring the AMT crowdsourcing task in this manner allows additional features to be captured for the training corpus and is likely to screen out ambiguous sentiment in the cognition of the annotator. We posit that this will improve IAA.

Ethics approval for this study was approved by the Johns Hopkins University Institutional Review Board (IRB) for use of AMT workers in the crowdsourcing tasks. All AMT workers were presented with the background and purpose of the study and could opt-out at any time. AMT workers acknowledged informed consent prior to participation in the project.

4 Findings

The IAA rating for off-the-shelf sentiment classifiers was $\alpha = 0.525$. This rating falls below acceptable limits according to the Bayerl scale. The IAA rating for the basic AMT experiment was $\alpha = 0.301$. The rating could be improved slightly to $\alpha = 0.308$ by including the "I Don't Know" as a response choice and excluding ratings that were completed in 3 s or less. It is surprising that the IAA rate for the AMT experiment performs worse than the off-the-shelf sentiment classifiers. Consistency among off-the-shelf classifiers with each other, however, does not mean that those classifiers are more accurate at classifying sentiment within this specific Twitter firestorm. It could represent a systematic bias. In any case, the best-case scenario still falls below established quality standards for use.

The results of the IAA for the guided decision tree are displayed in Fig. 1. It can be seen that all questions meet the Bayerl scale for good agreement and questions 1A, 2B, and 4 exceed the benchmark for excellent agreement. These questions require AMT workers to accurately assess tweet position, lack of sentiment, and humor, given that they accurately recognized whether a tweet expressed a position, sentiment, or humor. The guided decision tree allows the data scientist to effectively control for recognition when assessing IAA on coding/label assignment to Twitter data.

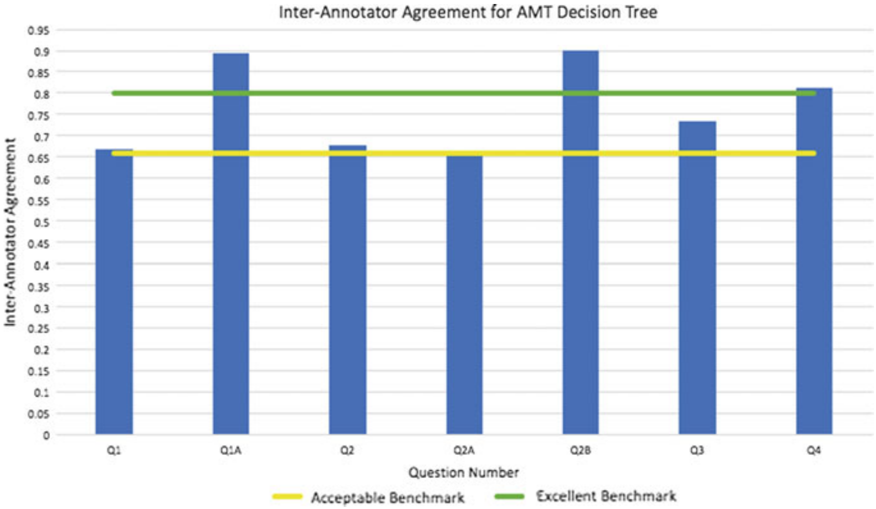

Fig. 1 Inter-annotator agreement (IAA) for guided decision tree crowdsourcing approach. This shows that all IAA ratings exceed the acceptable benchmark of 0.66 and several exceed excellent benchmark of 0.80

5 Conclusion

Constructing a Twitter training corpus using crowdsourcing with a guided decision tree appears to improve the quality of the corpus over off-the-shelf or simple crowdsourcing approaches. The decision tree effectively broke down a complex task into smaller problems that allowed AMT workers to achieve higher rates of agreement. It also provided a means for raters to remove bias when evaluating different types of tweets by first cognitively assessing the position and tone of the tweet prior to rendering a judgment. The ability to detect a positive, negative, or neutral sentiment remains a nuanced process, however, especially when language constructs such as humor and sarcasm are in play. Since these language paradigms are often difficult for a machine or artificial intelligence to detect, it is notable that AMT workers had a higher percentage of agreement when rating tweets for these factors in the guided decision tree approach. Arguably, linguistic factors such as humor and sarcasm are just as important (if not more important) as detecting sentiment, especially in a firestorm context surrounding a social debate. It is also notable that when AMT workers were not guided through a decision tree, the presence of humor and sarcasm degraded their IAA rating.

This work faced several limitations. The AMT experiment was focused on a single firestorm corpus and utilized a random sample of 200 tweets. Future research may investigate additional firestorm corpora and may include random samples with more tweets. The required sample size estimation could be performed by investigating the sensitivity of IAA rating by randomly removing sampled tweets. Similarly, the same decision tree structure could be applied to different corpora to evaluate the sensitivity of the data corpus on the findings.

Despite these limitations, this paper further demonstrates the inherent limitations of off-the-shelf classifiers. It highlights unique challenges for sentiment classification present within online protests. As Internet activism flourishes due to the immediacy of social media, the hasty spread of information, and political engagement; construction of high-quality training corpora becomes more important. These data challenges call for more work in classifying additional data features such as emotion, sarcasm, humor, polarizing positions, and even "hijacked hashtags". The crowdsourced, decision tree approach presented in this paper has proven effective in developing a high-quality gold-standard training dataset that outperforms off-the-shelf solutions.

Acknowledgements This work was supported by the Office of Naval Research, Grant No. N00014-17-1-2981/127025

References

1. E. Kouloumpis, T. Wilson, J.D. Moore, Twitter sentiment analysis: The good the bad and the omg!, in *ICWSM 11* (2011), pp. 11538–541
2. A. Bermingham, A.F. Smeaton, Classifying sentiment in microblogs: Is brevity an advantage?, in *19th ACM International Conference on Information and Knowledge Management, ACM* (2010), pp. 1833–1836
3. H. Yu, V. Hatzivassiloglou, Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences, in *EMNLP* (2003)
4. D. Maynard, M.A. Greenwood, Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis, in *Language Resources and Evaluation Conference* (2014), pp. 4238–4243
5. A. Reyes, P. Rosso, T. Veale, A multidimensional approach for detecting irony in Twitter, in *Language Resources and Evaluation*, vol. 47, no. 1 (2013), pp. 239–268
6. O. Tsur, D. Davidov, A. Rappoport, A great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews, in *Fourth International AAAI Conference on Weblogs and Social Media* (2010)
7. R.W. Gibbs, H.L. Colston, in *Irony in Language and Thought* (Routledge (Taylor and Francis), New York, 2007)
8. C. Bosco, V. Patti, A. Bolioli, Developing corpora for sentiment analysis: The case of irony and senti-TUT. IEEE Intell. Syst. **28**(2), 55–63 (2013)
9. S. Park, M. Ko, J. Kim, Y. Liu, J. Song, The politics of comments: Predicting political orientation of news stories with commenters' sentiment patterns, in *ACM 2011 Conference on Computer Supported Cooperative Work, ACM* (2011), pp. 113–122
10. K.-L. Liu, W.-J. Li, M. Guo, Emoticon smoothed language models for twitter sentiment analysis, in *AAAI* (2012)
11. J. Dunn, Facebook totally dominates the list of most popular social media apps, [online] Business Insider, Available at http://www.businessinsider.com/facebook-dominates-most-popular-social-media-apps-chart-2017-7 (2017)
12. H. Lamba, M.M. Malik, J. Pfeffer, A tempest in a teacup? Analyzing firestorms on twitter, in *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference*, *IEEE* (2015), pp. 17–24

13. J. Kang, Campaign to 'Cancel' Colbert, [online] The New Yorker, Available at https://www.newyorker.com/news/news-desk/the-campaign-to-cancel-colbert (2014)
14. F. Morstatter, J. Pfeffer, H. Liu, When is it biased?: Assessing the representativeness of twitter's streaming API, in *23rd International Conference on World Wide Web ACM* (2014), pp. 555–556
15. M.M. Bradley, P.J. Lang, Affective norms for english words (ANEW) instruction manual and affective ratings, Technical Report C-1, The Center for Research in Psychophysiology University of Florida (2009)
16. F.Å. Nielsen, A new ANEW: Evaluation of a word list for sentiment analysis in microblogs, arXiv preprint arXiv:1103.2903 (2011)
17. X. Zhu, S. Kiritchenko, S. Mohammad, NRC-Canada-2014: Recent improvements in the sentiment analysis of tweets, in *SemEval@ COLING* (2014), pp. 443–447
18. S. Kiritchenko, X. Zhu, S.M. Mohammad, Sentiment analysis of short informal texts. J. Artif. Intell. Res. **50**, 723–762 (2014)
19. B. Liu, Sentiment analysis and opinion mining, in *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1 (2012), pp. 1–167
20. R. Artstein, M. Poesio, Inter-coder agreement for computational linguistics. Comput. Linguist. **34**(4), 555–596 (2008)
21. P.S. Bayerl, K.I. Paul, What determines inter-coder agreement in manual annotations? A meta-analytic investigation. Comput. Linguist. **37**(4), 699–725 (2011)
22. A.F. Hayes, K. Krippendorff, Answering the call for a standard reliability measure for coding data, in *Communication Methods and Measures*, vol. 1, no. 1, pp. 77–89 (2007)
23. K.A. Neuendorf, *The Content Analysis Guidebook* (Sage, 2016)