Opinion Mining at Scale: A Case Study of the First Self-driving Car Fatality

Tao Li
Purdue University
taoli@purdue.edu

Minsoo Choi
Purdue University
choi502@purdue.edu

Yuntao Guo
Purdue University
guo187@purdue.edu

Lei Lin
University of Rochester
lei.lin@rochester.edu

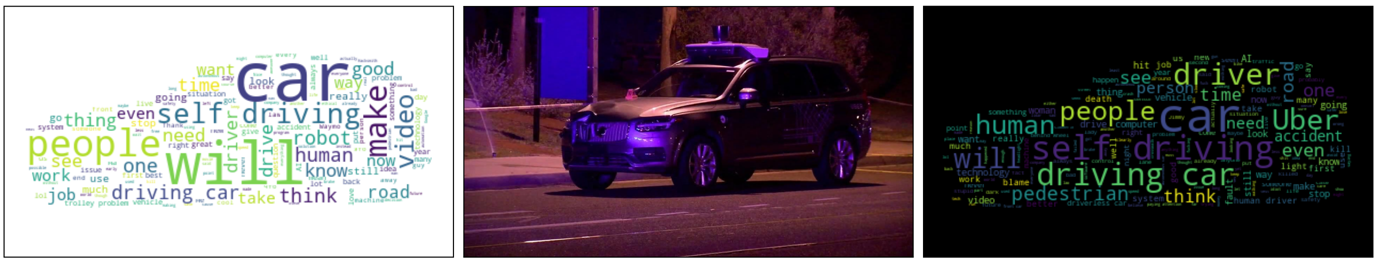

Fig. 1: The first fatal accident of self-driving cars occurred in March 2018: an experimental Uber vehicle in autonomous mode truck and killed a pedestrian in Tempe, Arizona [1]. Word clouds are generated from comments below self-driving-related videos on YouTube (left: one month before the incident; right: one month after; middle: the accident scene, photo by ABC-15).

***Abstract*—We present a comprehensive pipeline for large-scale opinion mining via a case study of the first self-driving car fatality, in an effort to qualitatively and quantitatively evaluate trending techniques in web searching as well as sentiment analysis. We first perform a scalable and fault-resilient web scraping with a partially-stateful data model. We then apply recent advances in deep learning comparing with a commercial software for sentiment detection. Not only do we measure the performances of the models by numerical metrics, we subsequently align the prediction results with amid economic indices and impactful social events. We further discuss trade-offs of above models from perspectives of both performance improvements of computer systems and accuracy enhancements of machine learning models, and provide deeper insights for stakeholders in the autonomous vehicle industry and the computational social science community.**

I. Introduction

With billions of users discussing and sharing opinions online every day, social media (e.g., Facebook, Twitters, and YouTube) is a rich data source for understandings of social and economic semantics. By changing the way we perceive and interact with the world, social media is changing our ways of living profoundly [2], [3], and has attracted tremendous attention from both academia and industry, with concerns ranging from building reliable and scalable systems with high-performance for online data collection, to analyzing such data in a timely and accurate manner [4]–[6].

Recent progresses of autonomous vehicles bring self-driving cars to the forefront of public interest [7]–[10]. Particularily after the first fatal accident of self-driving cars recently happened in Arizona, USA [1], autonomous vehicles have become a popular topic in social media. Multiple attempts have been made to investigate people's safety concerns of autonomous vehicles as well as acceptance levels and willingness to purchase [11]–[13]. However, these studies rely heavily on surveys which usually suffer disadvantages in (i) low response rate ; (ii) uncertainty over the validity of the data and sampling issues; and (iii) concerns surrounding the design, implementation, and evaluation of the survey [14]–[17]. Fortunately, recent techniques (e.g., [18]–[23]) in natural language processing (NLP) provide new possibilities to tackle the disadvantages of traditional surveys. In this paper, we collect textual data from large-scale web scraping and thereby evaluate these quantitative methods in a real-world scenario.

II. Background

A. Large-scale Web Scraping

The first crawler was Wanderer, created by Matthew Gray in 1993 [24], when the World Wide Web (WWW) was created [25]. Since then, many attempts have been made to address the issues of scalability, extensibility, and performance [26]–[28].

B. Sentiment Analysis

Finding out what other people think has always been an important part of information collection and decision making. Gathering opinions of hot, and probably controversial topics in the social media has aroused tremendous attention in the research community. [29] devided the task of sentiment classification into three levels: document-level, sentence-level, and aspect-level. Although the performance of traditional bag-of-words-based methods is good enough for document-level classification, sentence-level sentiment classification is still challenging. Recently, many deep learning-based methods are proposed; to name a few: [30]–[33].

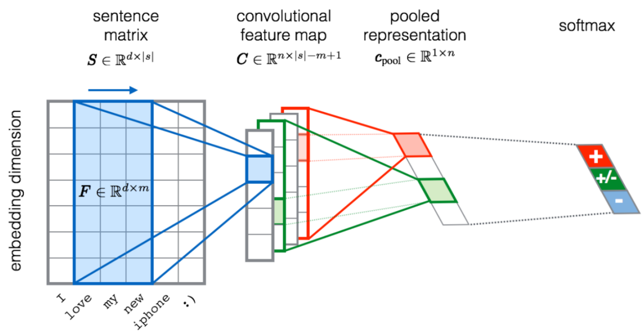

Fig. 2: Architecture of the deep neural network for sentiment classification [33].

III. Methodology

A. Data Collection

One of the key components of the project is YouTube Data APIs, which can be used to download metadata of YouTube videos, including title, description, and comments. The APIs provide Python wrappers which are programmer-friendly, compatible to many existing web scraping frameworks, and powerful in text processing. To maximize the performane of large-scale web scraping with limited computing resrouces, we adopt a partially-stateful data model [34] for high-performance web applications.

B. Analysis Methods

Word cloud is an efficent and appealing visualization method for textual data and has served as a starting point multiple studies across various domains [35]. In Figure 1, we present word clouds generated from YouTube comments one month before and after the fatal incident. It shows a clear distinction of word uses between the two corpuses and provides insights of public opinions toward the accident as well as autonomous vehicles. To quantify the impacts, we adopt a deep neural network composed of a single convolutional layer followed by a non-linearity, max pooling and a soft-max classification layer, as shown in Figure 2. Hyper-parameters and pre-training settings are detailed in [33]. There deep network is trained on the SemEval dataset [36].

IV. Future Works

We discuss high-level ideas in this extended abstract and leave the entire pipeline in the full paper, in which we (i) introduce the data collection process, address its challenges, and adopt a novel data flow model attack the issue; (ii) perform descriptive statistics analysis and data visualization to have a basic sense of the data; (iii) subsequently train a deep neural network for sentiment classification and compare the results with the Google's Cloud Natural Language Processing API; and (iv) finally align the quantitative results with real-world incidents to qualitatively evaluate the prediction models.

References

[1] D. Wakabayashi, "Self-driving uber car kills pedestrian in arizona, where robots roam," *The New York Times*, 2018.

[2] L. Safko, *The social media bible: tactics, tools, and strategies for business success*. John Wiley & Sons, 2010.

[3] W. He, S. Zha, and L. Li, "Social media competitive analysis and text mining: A case study in the pizza industry," *International Journal of Information Management*, vol. 33, no. 3, pp. 464–472, 2013.

[4] A. Ceron, L. Curini, S. M. Iacus, and G. Porro, "Every tweet counts? how sentiment analysis of social media can improve our knowledge of citizens political preferences with an application to italy and france," *New Media & Society*, vol. 16, no. 2, pp. 340–358, 2014.

[5] A. Perrin, "Social media usage," *Pew research center*, pp. 52–68, 2015.

[6] S. Boulianne, "Social media use and participation: A meta-analysis of current research," *Information, Communication & Society*, vol. 18, no. 5, pp. 524–538, 2015.

[7] B. Schoettle and M. Sivak, "A survey of public opinion about autonomous and self-driving vehicles in the us, the uk, and australia," 2014.

[8] S. Gong, A. Zhou, J. Wang, T. Li, and S. Peeta, "Cooperative adaptive cruise control for a platoon of connected and autonomous vehicles considering dynamic information flow topology," *arXiv preprint arXiv:1807.02224*, 2018.

[9] T. Li, K. Kaewtip, J. Feng, and L. Lin, "IVAS: Facilitating safe and comfortable driving with intelligent vehicle audio systems," in *2018 IEEE International Conference on Big Data*, 2018.

[10] C. Wang, S. Gong, A. Zhou, T. Li, and S. Peeta, "Cooperative adaptive cruise control for connected autonomous vehicles by factoring communication-related constraints," in *the 23rd International Symposium on Transportation and Traffic Theory*, 2019.

[11] D. Howard and D. Dai, "Public perceptions of self-driving cars: The case of berkeley, california," in *Transportation Research Board 93rd Annual Meeting*, vol. 14, no. 4502, 2014.

[12] M. Kyriakidis, R. Happee, and J. C. de Winter, "Public opinion on automated driving: Results of an international questionnaire among 5000 respondents," *Transportation research part F: traffic psychology and behaviour*, vol. 32, pp. 127–140, 2015.

[13] B. Schoettle and M. Sivak, "Public opinion about self-driving vehicles in china, india, japan, the us, the uk, and australia," 2014.

[14] K. B. Wright, "Researching internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services," *Journal of computer-mediated communication*, vol. 10, no. 3, p. JCMC1034, 2005.

[15] K. Kelley, B. Clark, V. Brown, and J. Sitzia, "Good practice in the conduct and reporting of survey research," *International Journal for Quality in health care*, vol. 15, no. 3, pp. 261–266, 2003.

[16] F. J. Fowler Jr, *Survey research methods*. Sage publications, 2013.

[17] R. D. Fricker and M. Schonlau, "Advantages and disadvantages of internet research surveys: Evidence from the literature," *Field methods*, vol. 14, no. 4, pp. 347–367, 2002.

[18] B. Pang, L. Lee *et al.*, "Opinion mining and sentiment analysis," *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[19] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining." in *LREc*, vol. 10, no. 2010, 2010, pp. 1320–1326.

[20] A. Esuli and F. Sebastiani, "Sentiwordnet: a high-coverage lexical resource for opinion mining," *Evaluation*, vol. 17, pp. 1–26, 2007.

[21] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining." in *Lrec*, vol. 10, no. 2010, 2010, pp. 2200–2204.

[22] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[23] T. Li, X. Liu, and S. Su, "Semi-supervised text regression with conditional generative adversarial networks," 2018.

[24] M. Gray, "Internet growth and statistics: Credits and background," 1993.

[25] A. Heydon and M. Najork, "Mercator: A scalable, extensible web crawler," *World Wide Web*, vol. 2, no. 4, pp. 219–229, 1999.

[26] J. Edwards, K. McCurley, and J. Tomlin, "An adaptive model for optimizing performance of an incremental web crawler," in *Proceedings of the 10th international conference on World Wide Web*. ACM, 2001, pp. 106–113.

[27] V. Shkapenyuk and T. Suel, "Design and implementation of a high-performance distributed web crawler," in *Data Engineering, 2002. Proceedings. 18th International Conference on*. IEEE, 2002, pp. 357–368.

[28] P. Boldi, B. Codenotti, M. Santini, and S. Vigna, "Ubicrawler: A scalable fully distributed web crawler," *Software: Practice and Experience*, vol. 34, no. 8, pp. 711–726, 2004.

[29] V. Jagtap and K. Pawar, "Analysis of different approaches to sentence-level sentiment classification," *International Journal of Scientific Engineering and Technology*, vol. 2, no. 3, pp. 164–170, 2013.

[30] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[31] T. Li, Z. Liu, and W. Feng, "A distributed search engine instance based on nutch and solr," *GitHub Repository*, 2014. [Online]. Available: https://github.com/Eroica-cpp/DistSearch

[32] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," *arXiv preprint arXiv:1404.2188*, 2014.

[33] A. Severyn and A. Moschitti, "Twitter sentiment analysis with deep convolutional neural networks," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2015, pp. 959–962.

[34] J. Gjengset, M. Schwarzkopf, J. Behrens, L. T. Araújo, M. Ek, E. Kohler, M. F. Kaashoek, and R. Morris, "Noria: dynamic, partially-stateful data-flow for high-performance web applications," in *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*, 2018.

[35] M. P. Anderson, W. W. Woessner, and R. J. Hunt, *Applied groundwater modeling: simulation of flow and advective transport*. Academic press, 2015.

[36] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, A.-S. Mohammad, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq *et al.*, "Semeval-2016 task 5: Aspect based sentiment analysis," in *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, 2016, pp. 19–30.