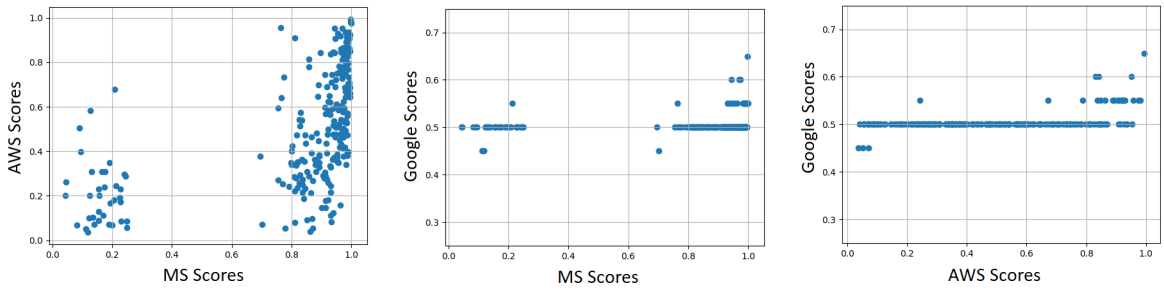

Figure 1: Scatter plots of the different sentiment scores. From top to bottom: Amazon (AWS) scores against Microsoft scores; Google scores against Microsoft scores; Google scores against Amazon (AWS) scores.

Table 2: Correlation coefficients (Kendall's tau) between the LIWC'15 lexical analysis and the cloud service tools

		LIWC'15 scales						
		Affect	Positive affect	Negative affect	Anxiety	Anger	Sadness	Swearing
Amazon	Positive	.284	.410	-.160	*n.s.*	-.158	*n.s.*	-.164
	Negative	*n.s.*	-.265	.416	*n.s.*	.420	.107	.294
	Mixed	-.279	-.171	-.212	*n.s.*	-.238	-.096	-.159
	Neutral	.226	.140	.228	*n.s.*	.228	.110	.156
Google	Feeling	.264	.349	-.186	*n.s.*	-.213	*n.s.*	-.127
	Intensity	.104	.106	*n.s.*	.103	.105	.121	.089
Microsoft	Sentiment score	.106	.305	-.342	*n.s.*	-.308	-.170	-.271

inspection of the data as shown in Figure 2 indicates that the three APIs show very different results even when normalised to the same scale. Most notably, the scores from the Google API are clustered around 0.5 which, in conjunction with mostly high magnitudes, means it has detected very mixed affect. This is in stark contrast with the scores of Amazon (AWS) and Microsoft. Microsoft seems to be applying a possibly non-linear function to the output of its sentiment analysis to force scores either close to 0 or 1. As a result, it primarily classifies the conversations as positive, with a sizeable minority in the negative sentiment. Normalised sentiment scores from Amazon appear to be spread fairly uniformly across the whole range between 0 and 1.

Kendall's τ is used as a measure for correlation due to the non-normal distribution of Microsoft's results. For Amazon and Microsoft scores τ is 0.49 ($p < .001$), which can be considered a large effect [9, 15]. Google's results correlate less well with both Microsoft ($\tau = .26$, $p < .001$, medium effect) and Amazon ($\tau = .33$, $p < .001$, medium effect) as it squashes all results close to 0.5. Figure 1 shows the scatter plots of the three cloud services.

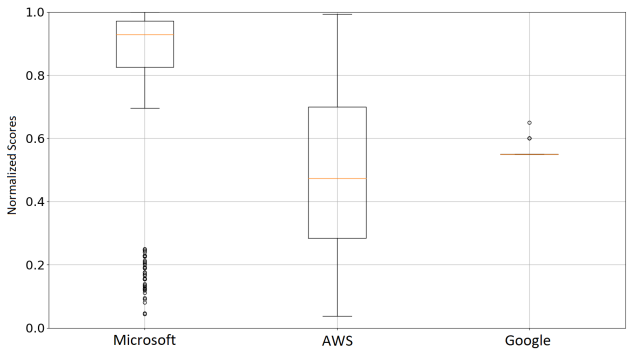

Figure 2: Distribution of sentiments (with 1 being very positive, 0 being very negative, and .5 being neutral) according to the different APIs

3.2 Agreement between cloud service tools and LIWC'15

As can be seen in Table 2, the LIWC'15 and most cloud service analysis results correlated reasonably well.

Amazon. Although most correlations are both present and in the direction that could be expected, (for example, positive affect as identified by Amazon's Comprehend analysis tool correlating positively with Positive affect from LIWC'15) some correlations, or lack of correlations, are quite surprising. One of those is the positive correlation between Amazon's Neutral scale and nearly all of LIWC'15's affect scales, suggesting that as Amazon identified a conversation as more Neutral in tone, LIWC'15 rated it as containing more affect. Conversely, a negative correlation between Mixed affect by Amazon and LIWC'15's rating of Affect indicates that as Amazon identified a conversation as being more mixed in its sentiment, LIWC'15 identified it as containing less affect. Moreover, the lack of significant correlation between the Positive and Negative Amazon scales on one hand and LIWC'15's Anxiety subscale on the

other hand, suggests that Amazon may do well in identifying general positive and negative affect, but does not discern well between the different kinds of negative affect.

Google. The analysis outcomes of Google's Natural Language Processing tool correlated well with the LIWC'15 analysis results. Interestingly, Anxiety as identified by LIWC'15 did not correlate with Google's Feeling scale.

Microsoft. The Sentiment score of Microsoft's Text Analysis tool correlated well with the LIWC'15 results. Like the sentiment analysis scales of Amazon and the Feeling scale of Google, Microsoft's Sentiment score did not correlate with Anxiety. This pattern of non-significant correlations suggests that either the cloud services do not identify anxious sentiment well, or that the LIWC'15 does not detect actual anxiety.

3.3 Agreement between tools and human judgement

3.3.1 *Cloud service tools.*

Amazon. Surprisingly, there was a marginally significant negative correlation between the Positive sentiment scale of Amazon's Comprehend analysis tool and human judgement of how positive a conversation was (τ = -.078, p = .058, small to medium effect). In addition, the human Negative sentiment ratings correlated negatively with Amazon's Mixed sentiment (τ = -.093, p = .025, medium effect) and marginally with Neutral sentiment (τ = .075, p = .071, small to medium effect). This suggests that the Amazon sentiment analysis is not capable of correctly identifying any sentiments in human-machine conversations. In the best case, its results are unrelated to human judgement and in the worst case they are diametrically opposed.

Google. The outcomes of Google's Natural Language Processing service did not correlate very well with human judgement either. The sole correlation that was marginally significant was a curiously positive correlation between Google's Feeling scale and the human judgement of Neutral sentiment (τ = .095, p = .063). These scores suggest that the Google cloud service sentiment analysis tool is not capable of correctly identifying any sentiments in human-machine conversations.

Microsoft. The single sentiment score that was returned by Microsoft's Text Analytics tool correlated positively with the Neutral judgement by the human raters (τ = .094, p = .041, medium effect), suggesting that it is not capable of correctly identifying the accurate sentiments in human-machine conversations.

3.3.2 *LIWC'15.*

As can be seen in the bottom half of Table 3, the different sentiment scales of the LIWC'15 did not correlate with the human judgements on many points. The Affect scale of LIWC'15 correlated positively (τ = .089, p = .036, medium effect) with the Mixed scale of the human judgements. There were negative correlations between the human judgement of Neutrality on one hand and the LIWC'15 scale Negative Affect (τ = -.127, p = .003, medium effect) and its subscales Anxiety (τ = -.105, p = .025, medium effect) and Anger (τ = -.100, p = .019, medium effect). Moreover, a positive correlation was found between the LIWC'15 Anxiety subscale and the human judgement of Negative affect (τ = .103, p = .027, medium effect), as well as a marginally significant correlation between the Anger subscale of the LIWC'15 and the human judgement of Mixed sentiments (τ = .071, p = .097, small effect). Neither the Sadness subscale nor Swearing scale of the LIWC'15 showed any significant correlation with human judgements, so were not included in Table 3.

Table 3: Correlation coefficients (Kendall's τ) between human judgement and the scales of the cloud service tools and the LIWC'15 lexical analysis.

		Human judgements			
		Positive	Negative	Mixed	Neutral
Amazon (AWS)	Positive	$-.078^{\dagger}$	*n.s.*	*n.s*	*n.s.*
	Negative	*n.s.*	*n.s*	*n.s.*	*n.s.*
	Mixed	*n.s.*	-.093	*n.s.*	*n.s.*
	Neutral	*n.s.*	$.075^{\dagger}$	*n.s.*	*n.s.*
Google	Feeling	*n.s.*	*n.s.*	*n.s.*	$.095^{\dagger}$
	Intensity	*n.s.*	*n.s.*	*n.s.*	*n.s.*
Microsoft	Sentiment	*n.s.*	*n.s.*	*n.s.*	.094
LIWC'15	Affect	*n.s.*	*n.s.*	.089	*n.s.*
	Positive	*n.s.*	*n.s.*	*n.s.*	*n.s.*
	Negative	*n.s.*	*n.s.*	*n.s.*	-.127
	Anxiety	*n.s.*	.103	*n.s.*	-.105
	Anger	*n.s.*	*n.s.*	$.071^{\dagger}$	-.100

† indicates a marginally significant (.05 < p <.10) correlation

The combination of the (negative) correlations between the human judgement of a neutral tone in the conversations and LIWC'15's detection of Negative affect, and the (positive) correlation between the Affect and Anger scales and the human judgement of anger, suggest that the tool was able to pick up on some sentiment but couldn't correctly identify any of the affect save for Anxiety.

3.4 Agreement amongst human judgements

As stated in section 2.1.4, each conversation was scored by four (or sometimes three) different human raters; these raters were not identical for all conversations. As some raters completed only a handful of ratings, and others finished many, inter rater agreement cannot be computed. In order to still give some indication of agreement, the standard deviation between ratings on each scale (i.e. Positive, Negative, Neutral and Mixed) within each unique conversation was calculated. This standard deviation statistic indicates how close the ratings on each particular scale for each individual conversation were to one another. For example, if the four raters of a single conversation had rated it 2, 3, 4, and 3 out of 5 on Negativity respectively, the standard deviation of the Negative scale would be .82. The higher the standard deviation, the less the raters agreed with one another.

4 DISCUSSION

The current paper tested the usability of different sentiment analysis tools for conversational data, by measuring to what extent the tools agreed with both human raters and each other on the sentiment in human-bot conversations. The cloud-based sentiment analysis tools

agreed reasonably well with one another and the lexical analysis software. However, the sentiment analysis tools did not correlate with human ratings, suggesting that although the tools picked up on some cues in the text, they either missed other important cues, or misinterpreted the cues they picked up, or both.

The human raters largely agreed among each other on the positive and negative scales, which suggests that the stimuli used in our study are sufficiently unambiguous. The different software tools, however, were unable to detect the sentiment patterns in the conversations. While it is hardly surprising that humans are better at detecting sentiment than software, the size of the disparity casts considerable doubt on the reliability and validity of the sentiment analysis tools currently offered for use in human-agent interaction.

For example, the conversation below was rated as *mostly mixed* by Amazon, *neutral* and *not very intense* by Google, and *very positive* by both Microsoft and LIWC'15.

CLEVERBOT: What do you want to do?
USER: I want to slit your face open with a hot knife.
CLEVERBOT: So do I, dingbat!
USER: What ever.
CLEVERBOT: Do you live in America?
USER: Nope.
CLEVERBOT: Good. I would've asked to move there.
USER: I live in africa and i have ebloa [*sic*].
CLEVERBOT: Is that where your pineapple is?
USER: Yes.
CLEVERBOT: I take it you like it there.
USER: Yupo.
CLEVERBOT: Why not?
USER: Ee your [*sic*] too irrelevant.
CLEVERBOT: No, it's not honey.
USER: Byew bitcgh [*sic*] v.
CLEVERBOT: What are you talking about?
USER: Your arse.

In contrast, human raters judged this conversation as very negative, slightly mixed, not positive, and not neutral.

Since the cloud service tools are not open about how their algorithms were developed, it is hard to determine what caused the discrepancy between human and tool ratings. One possibility would be that the service tools are mainly trained on a different type of data (like tweets and reviews) and as a result cannot deal well with the more volatile nature of human-agent dialogue. It is also possible that the internet slang, swearing and sarcasm in the conversations threw the tools off balance. Of course, these explanations are not mutually exclusive. Supporting such explanations, Ribeiro et al. [31] ranked various sentiment analysis tools yet still concluded that their benchmarks depended considerably on the data set used. Despite rapid progress owing to deep learning, domain adaptation across different data sets remains the panacea for natural language processing algorithms. In short, while sentiment analysis may be tremendously useful on other types of data (e.g. tweets, reviews, Facebook status updates), we would like to caution against the naive use of these services on data types with which they may not be familiar with, such as human-agent interaction and other types of human-machine interaction.

Table 4: Descriptives of the standard deviations between ratings of the same conversation, per scale.

Scale	Median	Mean	67% interval
Positive	.71	.75	[.58; .96]
Negative	.58	.72	[.50; .96]
Neutral	.96	.98	[.82; 1.15]
Mixed	.82	.83	[.58; .96]

4.1 Limitations

The conversations between users and Cleverbot are often strange and jump from one topic to the next. While this makes for quite particular data, we argue that the current analysis is relevant for a broader scope of human-agent interaction as well. As technology progresses and AI becomes a more central part of everyday life, informal conversations between humans and agents will become more ubiquitous - already, people use their Alexa or Siri for a wide range of conversation topics which stretches far beyond mere inquiries about the weather or news headlines [11, 14]. While a recreational bot like Cleverbot may be less sophisticated than an Alexa, we expect there to be at least some overlap in the conversations people will attempt to have with either agent. Thus, the inability of the sentiment analysis tools to correctly identify affect in quirky, unpredictable conversations is an issue not just restricted to Cleverbot data.

Another point to consider is that our analysis used utterances from both users and Cleverbot, who at times could be inconsistent. It could be argued that a more favourable image of the sentiment analysis tools would have emerged if solely the utterances of the user had been taken into account. However, as also shown in the conversation snippet above, humans often change the topic of conversation at will. The sentiment analysis tools discussed in this paper are designed to analyse discussion in online forums and social media commentary. They should therefore be designed to take contributions from multiple, occasionally inconsistent, conversation partners into account. We therefore do not believe that the current study disadvantaged the sentiment analysis tools by feeding them the full conversations.

4.2 Future work

When making sense of what the conversation partner is trying to convey, humans use a wide range of cues, like facial expressions, volume of speech, and body language, but also less obvious cues like personality or how someone dresses. Sentiment analysis therefore could take not only the bare text into account but also consider the voice of the speaker, self-corrections and hesitations in the speech, and all other non-verbal cues. Ideally, the cloud-based services would expand their service to also include non-verbal data in order to improve the quality of their analyses.

REFERENCES

[1] Christoph Bartneck, Andreas Duenser, Elena Moltchanova, and Karolina Zawieska. 2015. Comparing the similarity of responses received from experiments conducted in Amazon's Mechanical Turk to experiments conducted with traditional methods. *PLOS One* 10, 4 (2015), e0121595. https://doi.org/10.1371/journal.pone.0121595

[2] Christoph Bartneck and Michael J. Lyons. 2009. Facial Expression Analysis, Modeling and Synthesis: Overcoming the Limitations of Artificial Intelligence with the Art of the Soluble. In *Handbook of Research on Synthetic Emotions and Sociable Robotics:New Applications in Affective Computing and Artificial Intelligence*, Jordi Vallverdu and David Casacuberta (Eds.). IGI Global, Chapter 3, 33–53.

[3] B. Bazelli, A. Hindle, and E. Stroulia. 2013. On the Personality Traits of StackOverflow Users. In *IEEE International Conference on Software Maintenance*. 460–463. https://doi.org/10.1109/ICSM.2013.72

[4] Cindy L Bethel and Robin R Murphy. 2008. Survey of non-facial/non-verbal affective expressions for appearance-constrained robots. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38, 1 (2008), 83–92. https://doi.org/10.1109/TSMCC.2007.905845

[5] Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of computational science* 2, 1 (2011), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

[6] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*. 4349–4357.

[7] Scott R Braithwaite, Christophe Giraud-Carrier, Josh West, Michael D Barnes, and Carl Lee Hanson. 2016. Validating machine learning algorithms for Twitter data against established measures of suicidality. *JMIR mental health* 3, 2 (2016). https://doi.org/10.2196/mental.4822

[8] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[9] Jacob Cohen. 1992. A power primer. *Psychological bulletin* 112, 1 (1992), 155.

[10] Henriette Cramer, Jorrit Goddijn, Bob Wielinga, and Vanessa Evers. 2010. Effects of (in) accurate empathy and situational valence on attitudes towards robots. In *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*. IEEE, 141–142.

[11] Amanda Cercas Curry and Verena Rieser. 2018. # MeToo Alexa: How Conversational Systems Respond to Sexual Harassment. In *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*. 7–14.

[12] Xing Fang and Justin Zhan. 2015. Sentiment analysis using product review data. *Journal of Big Data* 2, 1 (2015), 5. https://doi.org/10.1186/s40537-015-0015-2

[13] Golnoosh Farnadi, Susana Zoghbi, Marie-Francine Moens, and Martine De Cock. 2013. Recognising personality traits using Facebook status updates. In *Proceedings of the workshop on computational personality recognition (WCPR13) at the 7th international AAAI conference on weblogs and social media*. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6245/6309

[14] Leah Fessler. 2017 (accessed 06-March-2019). We tested bots like Siri and Alexa to see who would stand up to sexual harassment. https://bit.ly/2lLGF8w

[15] Andrew R Gilpin. 1993. Table for conversion of Kendall's Tau to Spearman's Rho within the context of measures of magnitude of effect for meta-analysis. *Educational and psychological measurement* 53, 1 (1993), 87–92. https://doi.org/10.1177/0013164493053001007

[16] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning*. 513–520. http://www.icml-2011.org/papers/342_icmlpaper.pdf

[17] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep learning*. Vol. 1. MIT press Cambridge.

[18] Scott Gray, Alec Radford, and Diederik P Kingma. 2017. *GPU kernels for block-sparse weights*. Technical Report. Technical report, OpenAI.

[19] Emitza Guzman and Walid Maalej. 2014. How do users like this feature? a fine grained sentiment analysis of app reviews. In *Requirements Engineering Conference (RE), 2014 IEEE 22nd International*. IEEE, 153–162. https://doi.org/10.1109/RE.2014.6912257

[20] Louisa Hall. 2017. How We Feel About Robots That Feel. *MIT Technology Review* (2017). https://www.technologyreview.com/s/609074/how-we-feel-about-robots-that-feel/

[21] Eva Hudlicka. 2008. What are we modeling when we model emotion?. In *AAAI spring symposium: emotion, personality, and social behavior*, Vol. 8. AAAI. http://www.aaai.org/Papers/Symposia/Spring/2008/SS-08-04/SS08-04-010.pdf

[22] Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 562–570.

[23] Young Bin Kim, Jun Gi Kim, Wook Kim, Jae Ho Im, Tae Hyeong Kim, Shin Jin Kang, and Chang Hun Kim. 2016. Predicting fluctuations in cryptocurrency transactions based on user comments and replies. *PloS one* 11, 8 (2016), e0161197. https://doi.org/10.1371/journal.pone.0161197

[24] Ting-Peng Liang, Xin Li, Chin-Tsung Yang, and Mengyue Wang. 2015. What in consumer reviews affects the sales of mobile apps: A multifacet sentiment analysis approach. *International Journal of Electronic Commerce* 20, 2 (2015), 236–260. https://doi.org/10.1080/10864415.2016.1087823

[25] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal* 5, 4 (2014), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

[26] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 79–86. https://doi.org/10.3115/1118693.1118704

[27] James W Pennebaker, Roger J Booth, Ryan L Boyd, and Martha E Francis. 2015. Linguistic Inquiry and Word Count: LIWC 2015 [Computer software]. Pennebaker Conglomerates. http://liwc.wpengine.com

[28] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report. The University of Texas at Austin. http://hdl.handle.net/2152/31333

[29] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* (2018).

[30] Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444* (2017).

[31] Filipe N Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. SentiBench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science* 5, 1 (2016), 23. https://doi.org/10.1140/epjds/s13688-016-0085-1

[32] Sebastian Ruder and Barbara Plank. 2018. Strong Baselines for Neural Semi-supervised Learning under Domain Shift. *arXiv preprint arXiv:1804.09530* (2018).

[33] J.A. Russel. 1979. Affective space is bipolar. *Journal of personality and social psychology* 37, 3 (1979), 345–356. https://doi.org/10.1037/0022-3514.37.3.345

[34] María Ruz and Pío Tudela. 2011. Emotional conflict in interpersonal interactions. *Neuroimage* 54, 2 (2011), 1685–1691.

[35] BjÃűrn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication* 53, 9 (2011), 1062–1087. https://doi.org/10.1016/j.specom.2011.01.011

[36] Michele Settanni and Davide Marengo. 2015. Sharing feelings online: studying emotional well-being via automated text analysis of Facebook posts. *Frontiers in psychology* 6 (2015), 1045. https://doi.org/10.3389/fpsyg.2015.01045

[37] Aliaksei Severyn and Alessandro Moschitti. 2015. 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, 959–962. https://doi.org/10.1145/2766462.2767830

[38] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics* 37, 2 (2011), 267–307. https://doi.org/10.1162/COLI_a_00049

[39] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 1555–1565.

[40] Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann. 2014. *Twitter and society*. Vol. 89. Peter Lang Inc.

[41] Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu. 2011. *Combining lexicon-based and learning-based methods for twitter sentiment analysis*. Technical Report 89. HP Laboratories. https://uic.pure.elsevier.com/en/publications/combining-lexicon-based-and-learning-based-methods-for-twitter-se