Supplementary Material Sentiment and Objectivity in Iranian State-Sponsored Propaganda on Twitter

I. Context

The number of people obtaining news updates through social media has increased [1] and various actors have seized upon this to push their own narratives and propaganda in order to influence the discourse around various topics [2]. This ability to alter the conversation around topics makes misinformation particularly problematic as ordinary users may not be able to differentiate between false and true information because of the way an account portrays itself. For instance, an account portraying itself as a legitimate news organisation that uses objective language may be perceived as more trustworthy than an account that posts about topics in a polarised manner and appears to be an ordinary user [3], [4]. This makes the analysis of Twitter posts' sentiment particularly valuable.

In April 2019, the US Special Counsel Robert Mueller declared that Russia had interfered in the 2016 US presidential election 'in a sweeping and systematic fashion' ([5], p. 1). His redacted report found that since 2014, Moscow had used social media to 'sow discord in the U.S. political system through what it termed "information warfare"...[to] undermine the U.S. electoral system...[and enact] a targeted operation that by early 2016 favoured candidate Trump and disparaged candidate Clinton' ([5], p. 4).

But this was not the first time that authoritarian regimes had used social media to influence real-life politics. In the year before the 2014 Russian annexation of Crimea, Moscow led a focused domestic social media campaign to discredit the Ukrainian government and generate domestic Russian support for subsequent military operations [6]. Meanwhile, Twitter accounts linked to the Iranian regime were leading an operation in languages including Persian, English, French, Spanish, Russian, Turkish and Arabic to foment pro-government narratives across the world, while South Korean intelligence officers had used social media to anonymously discredit the opposition and influence the country's 2012 election.

These developments came amid increasing recognition from policymakers and researchers that a direct connection had emerged between the online information environment and the integrity of democracy itself, and that authoritarian actors worldwide were now exploiting technological advances in online technology for their own illiberal ends.

II. Background and Related work

A. Misinformation on Social Media

Misinformation in the context of this paper is used as a broad term referring to misleading or inaccurate information [7]. We are using this umbrella term to cover other terms used in the literature in this area, such as: (a) disinformation, i.e. "false, inaccurate, or misleading information designed, presented and promoted to intentionally cause harm or for profit"([7], p. 10); (b) fake news, i.e. news articles that are intentionally and verifiably false, and could mislead readers [4]; (c) partial facts, i.e. unbalanced coverage and presentation of information [8]; and (d) propaganda, which is associated with the intention to influence others and persuade them to adopt a particular view [9], [10]. These all have the potential to be harmful to the people who come into contact with them, and to the society as a whole, as they can be used to alter the public perception on various issues.

Misinformation on social media has been identified in various contexts, with the phenomenon being particularly apparent in topics such as politics [11]–[13] and medicine [14]–[19].

The "pizzagate" scandal of the 2016 US Presidential election was spread through a misinterpretation of leaked emails that belonged to then Presidential candidate Hillary Clinton's campaign chairman John Podesta. Interpretations of these emails claimed that Hillary Clinton and other Democrats were running a paedophile ring from a pizza parlour in Washington D.C. [11]. These false allegations circulated widely on social media and led to an armed person storming the pizza parlour attempting to liberate children from the paedophile ring, which did not exist [11].

Misinformation related to the Coronavirus pandemic has been shared widely and alleges various things, including that cleaning the nostrils with salty water can kill the virus [14], that governments are being deceptive about the pandemic, that data are being exaggerated, that the virus was released from a Chinese laboratory [15]–[17] and that the vaccines are not safe [18]. Vaccines have been the subject of misinformation for some time, with anti-vaccination rhetoric stating that vaccines cause autism and vaccines have the potential to do more damage than the virus they are protecting against [15], [18], [19]. In the middle of a global pandemic, this misinformation has the potential to seed distrust in public health messaging, and potentially expose more people to the virus by understating its risk profile [16].

Researchers found that people spreading misinformation on social media tend to surround themselves with people holding the same views as them, creating an echo chamber [15], [19], [20]. Additionally, because actors who spread misinformation tend to surround themselves with people holding the same views, the echo chamber effect means that misinformation either goes unchallenged, or the challenge is ignored because the viewpoint appears to be held by a small number of

people. To address the spread of misinformation, stance[1] and credibility[2] labels have been used, however, these have been shown to be ineffective in the fight against misinformation in politics [21].

The aftermath of the 2020 United States Presidential election was the subject of political misinformation about election fraud, including claims that the election was stolen and that mail-in ballot fraud had caused Trump's loss [12]. Twitter was ultimately forced to place warnings on some of Trump's tweets in an effort to highlight that the claims being made were false, such as "this claim about election fraud is disputed". The misinformation spread by Trump and consumed by supporters had tangible outcomes, including leading to the violent mob attack on the Capitol building that resulted in five deaths [13]. This has led to Twitter and other social media platforms to permanently (or indefinitely) suspend Trump's accounts to prevent the incitement of violence based on misinformation [22].

Indeed these examples highlight the very real and severe consequences of misinformation on social media platforms, including the spread of preventable disease, people exposing themselves to additional risk, and even the loss of life.

Social media has proven particularly fertile ground for state-based misinformation campaigns. Although state-fuelled information operations are not a new phenomenon, they have been revolutionised in recent years by the spread of social media, which has represented a low-cost and comparatively low-effort tool for influencing the citizens of other states. The pay-off for such operations to date, however, has been significant, sowing doubt over the integrity of democratic processes such as the 2016 US presidential election and the Brexit Referendum, and perhaps even influencing outcomes. This has led to significant work examining the efforts of the West's most powerful adversaries, Russia and China [23], however, the activities and strategies of other authoritarian states with smaller online footprints, such as Iran, are less well understood.

B. The Nuclear Deals and Iran's Propaganda Efforts

The Islamic Republic of Iran has been pitted as an adversary to the West since the country's 1979 revolution. This conflict has taken many forms over the decades, ranging from the 1980-81 hostage crisis, the Iran-Iraq war, and – most prominently in recent years – the dispute over Iran's nuclear program.

Although Iran's nuclear program was conceived long before the 1979 revolution at a time when Iran was seen as a loyal ally of the West, the program quickly became subject of international concern after 1979, particularly in the 2000s after US President George W. Bush declared Iran to be a member of the "Axis of Evil" [24]. This led to various strategies to isolate Iran and to push it to dismantle its program, most notably the international sanctions regime, which had become economically crippling by the time President Hassan Rouhani was elected in 2013, on a platform of resolving the nuclear issue. The hallmark of this strategy was the Joint Comprehensive Plan of Action (JCPOA), which was eventually signed in 2015 after extensive negotiations with the international community, including the United States [25].

The nuclear deal placed limits on Iran's nuclear program while easing economic sanctions and allowing Iran to reintegrate into the international community [24]. Although the nuclear deal was signed by President Rouhani, and importantly had the approval of the country's most powerful person, Supreme Leader Ayatollah Ali Khamenei, it never enjoyed universal support in Iran. Most notably, it was opposed by political hardliners and those associated with the powerful paramilitary force, the Islamic Revolutionary Guards Corps, who viewed the deal as an unacceptable assault on Iran's sovereignty, and in particular its right to self-defence.

It is on this basis that the sentiment expressed in the tweets used for this research cannot be taken for granted, as the Iranian elite were very divided over the merits of the deal, and it is not clear which parties (or if many parties) influence its social media operations. This division would only be exacerbated following the election of the Trump administration in the US in 2016 [26], [27]. In May 2018, then President Trump withdrew the United States from the nuclear deal, claiming that is was "one-sided" and "should never have been made" [26]. Understanding the sentiment expressed in the tweets is therefore a valuable exercise, not only for evaluating sentiment analysis tools and potential for (semi-)automatic analysis, but also to understand Iranian online strategies from a policy perspective.

C. Sentiment Analysis

Sentiment analysis has been used in various scenarios, including crime prediction from Twitter data using the polarity of tweets and weather-related information [28], analysing people's experiences with 'coming out' [29] and for identifying the sentiment of tweets around the coronavirus pandemic [30].

III. Evaluation Metrics

The precision metric is used to identify the number of true positives (TP) (predicted positive, actually positive) among all positive predictions[3], i.e. of all positive predictions (including the ones wrongly predicted as positive, which are denoted as False Positives (FP)), how many were actually positive [31].

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Recall is a measure that identifies how many of the actually positive occurrences were correctly identified, i.e. of all actually positive instances (made of the ones that are correctly identified (True positives) and the ones that were

[1]stance labels are attached to separate news articles containing opposing ideologies, with the aim to expose people to diverse opinions [21]

[2]credibility labels were used to signal the trustworthiness of articles, with the aim of supporting the readers to decide the trustworthiness of the content [21]

[3]For a more detailed explanation and illustration of the precision and recall metrics see https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall

TABLE I
THE LIST OF KEYWORDS USED TO DETECT TWEETS ABOUT THE IRANIAN NUCLEAR DEAL.

Keywords	
Agreement	Deal
Deputy FM	Iran DepFM
Iran deputy	Iran FM
Kerry	Zarif
Araghchi	Deal bribing
Fm araghchi	Fm Zarif
Geneva	Geneva NuclearTalks
Iran nuclear	Iran talks
Irandeal	Irantalks
Irantalks Iran	Irantalks irandeal
Irantalks irantalksgeneva	Irantalks Irantalkslausanne
Irantalks irantalksvienna	Irantalks nucleartalks
Irantalks putin	IrantalksGeneva Putin
Irantalkslausanne nuclear talks	Irantalksvienna
Irantalksvienna irandeal	Irantalksvienna nucleartalks
JCPOA	JCPOA irandeal
Lausanne	Lausanne Irantalks
Lausanne nucleartalks	Moniz
Montreaux	Negotiation
Nuclear talks	Nuclear talks playboy
Nucleartalks	NuclearTalks btw
Nucleartalks iran	Nucleartalks irantalks
NuclearTalksSaudi	Salehi
Salehi moniz	Vienna
Vienna irantalks	Zarif Kerry

incorrectly missed (False Negatives (FN)), how many were correctly predicted as being positive [31].

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

As improving either precision or recall has an impact on the performance of the other metric, F-score uses both metrics to strike a balance between the two. Consequently, in the F-score both metrics are taken into consideration and a single metric can be used to judge the performance of the classification algorithm.

$$F = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall} \right)$$

Accuracy uses the number of actually positive predictions to determine how accurate the classification algorithm was in its predictions. This metric uses the number of true negatives (TN) (a negative instance correctly identified as such), which precision, recall, and F-score do not.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

The F-score is more reliable than accuracy when there is an imbalance in the data, for example, if one class has fewer instances than the other(s).

IV. NUCLEAR DEAL TWEET EXTRACTION KEY TERMS

Table I lists the keywords used to extract relevant tweets about the Iranian nuclear deal. As mentioned in the paper, the list of key terms was created by the third author, an expert in Iranian politics, after evaluating the fifteen most frequent unigrams and bigrams for each month across the Iranian Twitter Election Integrity datasets.

V. NEGATION KEYWORDS

Table II lists the negation keywords that were retained in order to ensure a more reliable detection of negative polarity. This list was adapted from [32].

TABLE II
THE LIST OF NEGATION KEYWORDS RETAINED DURING STOPWORD REMOVAL

Negation Keywords				
aint	cannot	cant	darent	didnt
doesnt	dont	hadnt	hardly	hasnt
havent	havnt	isnt	lack	lacking
lacks	neither	never	no	nobody
none	nor	not	nothing	nowhere
mighnt	mustnt	neednt	oughnt	shant
shouldnt	wasnt	without	wouldnt	

VI. WORST RESULTS

Table III shows the worst performance for all algorithms across all five feature sets. MPT stands for Match Percentage

TABLE III
WORST RESULTS FOR EACH OF THE MACHINE LEARNING ALGORITHMS USED AND DEATURE SETS (MACRO AVERAGE)

n-grams	Metric	k-NN	DT	SVM	NB	RF
UNI	MPT	90	90	90	90	90
	HP	9	-	0.5	1.5	[150, 300]
	P	**0.3311**	0.5214	0.5838	0.5583	0.5528
	R	**0.4205**	0.459	0.5988	0.5069	0.5072
	F	**0.3038**	0.4672	0.5691	0.4899	0.4965
	Acc	**0.4478**	0.5892	0.6481	0.64	0.6196
BI	MPT	80	90	90	90	90
	HP	11	-	0.5	1.5	default
	P	0.5476	0.6152	0.5545	**0.5238**	0.6256
	R	**0.4332**	0.554	0.5184	0.4645	0.5484
	F	**0.2934**	0.4786	0.4687	0.4464	0.4621
	Acc	**0.3562**	0.5173	0.5544	0.6166	0.5086
TRI	MPT	70	90	90	90	90
	HP	11	-	0.3	1.8	default
	P	0.6825	0.601	**0.5304**	0.5431	0.6107
	R	0.4206	0.5268	0.3979	**0.3885**	0.536
	F	**0.287**	0.3923	0.3359	0.3404	0.3937
	Acc	**0.3415**	0.4665	0.5284	0.5963	0.4659
UNI + BI	MPT	70	90	90	90	90
	HP	11	-	1.2	1.5	[200, 400]
	P	0.6059	**0.5393**	0.583	0.5585	0.5712
	R	**0.4264**	0.4647	0.5848	0.51	0.5286
	F	**0.3137**	0.4744	0.5629	0.493	0.512
	Acc	**0.3607**	0.601	0.6533	0.6466	0.6305
UNI + BI + TRI	MPT	90	90	90	90	90
	HP	9	-	0.5	1.5	[300, 600]
	P	**0.3493**	0.5558	0.5851	0.5557	0.5636
	R	**0.4516**	0.5016	0.5812	0.5079	0.5147
	F	**0.3015**	0.5078	0.5609	0.4911	0.5055
	Acc	**0.436**	0.6301	0.6516	0.646	0.6302

Threshold, HP for hyperparameter, P for precision, R for recall, F for F-score, Acc for accuracy, UNI for unigrams, BI for bigrams and TRI for trigrams.

REFERENCES

[1] N. Martin, "How Social Media Has Changed How We Consume News," https://www.forbes.com/sites/nicolemartin1/2018/11/30/how-social-media-has-changed-how-we-consume-news, 11 2018.

[2] S. A. Khan, M. H. Alkawaz, and H. M. Zangana, "The Use and Abuse of Social Media for Spreading Fake News," in *Proceedings of the 2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*. Selangor, Malaysia: IEEE, 6 2019, pp. 145–148.

[3] K. M. Caramancion, "Understanding the impact of contextual clues in misinformation detection," in *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*. Vancouver, Canada: IEEE, 9 2020, pp. 1–6.

[4] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–36, 2017.

[5] R. S. Muller, *Report on the investigation into the Russian interference in the 2016 presidential election*. US. Department of Justice, 2019, vol. I.

[6] M. Kofman, K. Migacheva, B. Nichiporuk, A. Radin, O. Tkacheva, and J. Oberholtzer, *Lessons from Russia's Operations in Crimea and Eastern Ukraine*. Santa Monica, CA: RAND Corporation, 2017.

[7] M. de Cock Buning, "A multi-dimensional approach to disinformation: Report of the independent high level group on fake news and online disinformation," 2018. [Online]. Available: https://data.europa.eu/doi/10.2759/739290

[8] V. F. Hendricks and M. Vestergaard, "Alternative facts, misinformation, and fake news," in *Reality Lost*. Springer, 2019, pp. 49–77.

[9] A. M. Guess and B. A. Lyons, "Misinformation, Disinformation, and Online Propaganda," in *Social Media and Democracy: The State of the Field, Prospects for Reform*. Cambridge, UK: Cambridge University Press, 2020, ch. 1, pp. 10–33.

[10] D. D. Chaudhari and A. V. Pawar, "Propaganda analysis in social media: a bibliometric review," *Information Discovery and Delivery*, 2021.

[11] E. Bergmann, "Populism and the politics of misinformation," *The Journal of South African and American Studies (Safundi)*, vol. 21, no. 3, pp. 251–265, 2020.

[12] BBC News, "US election results: Trump sues as path to victory over Biden narrows - BBC News," 11 2020. [Online]. Available: https://www.bbc.co.uk/news/election-us-2020-54818992

[13] ——, "US Congress in turmoil as violent Trump supporters breach building - BBC News," 1 2021. [Online]. Available: https://www.bbc.co.uk/news/world-us-canada-55555074

[14] N. Fleming, "Coronavirus misinformation, and how scientists can help to fight it," *Nature*, vol. 583, no. 7814, pp. 155–156, 7 2020.

[15] T. C. Smith and D. R. Reiss, "Digging the rabbit hole, COVID-19 edition: anti-vaccine themes and the discourse around COVID-19," *Microbes and Infection*, vol. 22, no. 10, pp. 608–610, 11 2020.

[22] Twitter Inc, "Permanent suspension of @realDonaldTrump," 1 2021. [Online]. Available: https://blog.twitter.com/en_us/topics/company/2020/suspension.html

[16] G. Pennycook, J. McPhetres, Y. Zhang, J. G. Lu, and D. G. Rand, "Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention," *Psychological Science*, vol. 31, no. 7, pp. 770–780, 7 2020.

[17] W. Cornwall, "Officials gird for a war on vaccine misinformation Fears of a rushed COVID-19 vaccine and rise of social media demand new messaging strategy," *Science*, vol. 369, no. 6499, pp. 14–19, 7 2020.

[18] A. L. Schmidt, F. Zollo, A. Scala, C. Betsch, and W. Quattrociocchi, "Polarization of the vaccination debate on Facebook," *Vaccine*, vol. 36, no. 25, pp. 3606–3612, 6 2018.

[19] T. Burki, "Vaccine misinformation and social media," *The Lancet Digital Health*, vol. 1, no. 6, pp. e258–e259, 10 2019.

[20] Z. Bastick, "Would you notice if fake news changed your behavior? An experiment on the unconscious effects of disinformation," *Computers in Human Behavior*, vol. 116, no. 106633, pp. 1–12, 3 2021.

[21] M. Gao, Z. Xiao, K. Karahalios, and W.-T. Fu, "To Label or Not to Label: The Effect of Stance and Credibility Labels on Readers' Selection and Perception of News Articles," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, p. 16, 2018.

[23] D. Gordon, *Targeted Systems and Democracy: Russia, Iran, and China's Cyber Threats and Disinformation Campaigns to Weaken and Undermine Western Democracies*. Utica College, 2020.

[24] A. Tabatabai, "Negotiating the "Iran talks" in Tehran: the Iranian drivers that shaped the Joint Comprehensive Plan of Action," *The Nonproliferation Review*, vol. 24, no. 3-4, pp. 225–242, 5 2017.

[25] S. Akbarzadeh and D. Conduit, "Rouhani's First Two Years in Office: Opportunities and Risks in Contemporary Iran," in *Iran in the World: President Rouhani's Foreign Policy*, 2016, pp. 1–15.

[26] S. Nikou, "Timeline of Iran's Nuclear Activities," United States Institute of Peace, Tech. Rep., 2018. [Online]. Available: https://iranprimer.usip.org/resource/timeline-irans-nuclear-activities

[27] S. H. Mousavian and M. M. Mousavian, "Building on the Iran Nuclear Deal for International Peace and Security," *Journal for Peace and Nuclear Disarmament*, vol. 1, no. 1, pp. 169–192, 2018.

[28] X. Chen, Y. Cho, and S. Y. Jang, "Crime prediction using Twitter sentiment and weather," in *Proceedings of the 2015 IEEE Systems and Information Engineering Design Symposium*. IEEE, 2015, pp. 63–68.

[29] T. Anand, K. Ramesh, and S. Singh, "Out of the Closet: Lexicon Based Sentiment Analysis on Tweets about Homosexuality," in *Proceedings of TENCON 2019)*. IEEE, 2019, pp. 733–738.

[30] K. Chakraborty, S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag, and A. E. Hassanien, "Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media," *Applied Soft Computing*, vol. 97, no. A, pp. 1–14, 2020.

[31] K. P. Shung, "Accuracy, Precision, Recall or F1? - Towards Data Science," 4 2020. [Online]. Available: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

[32] J. Reitan, J. Faret, B. Gambäck, and L. Bungum, "Negation Scope Detection for Twitter Sentiment Analysis," in *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015)*. Lisboa, Portugal: Association for Computational Linguistics, 9 2015, pp. 99–108.