Sentiment and Objectivity in Iranian State-Sponsored Propaganda on Twitter

Michael Barrows, Ella Haig, *Member, IEEE,* and Dara Conduit

***Abstract*—In 2016, Russia attempted to use social media to influence the outcome of the US presidential election, highlighting the potential real-world impacts of state-led online misinformation campaigns. Misinformation on social media is a growing concern, especially in the areas of politics and medicine, given their impact not only at individual level, but also for society as a whole. In this paper, we investigate the potential to automatically label and detect the polarity (positive, neutral or negative) of Iranian state-sponsored propaganda tweets on the Iranian nuclear deal. The SentiWordNet lexicon is used to automatically assign a polarity label and an objectivity score to each tweet. Using the labels, five machine learning algorithms are used to create polarity detection models. The experimental results show that the best performing models correctly identify polarity in approximately 77% of the tweets.**

***Index Terms*—Sentiment, Objectivity, Machine Learning, Iran, Propaganda**

I. Introduction

As the number of users on social media platforms has increased, the quality and accuracy of the information presented to these users has become increasingly important [1]. Through social media, misinformation has been able to spread to a wider audience and it also tended to spread faster and further than true information [2], [3]. See information about the wider context of this research in the Supplementary Material.

Given the huge scale of social media data, the speed at which misinformation spreads [2] and the importance of addressing this [4], [5], as highlighted by policy makers across the world, e.g. the US government[1], the European Commission[2] and the UK government[3], there is a need for approaches that use, at least to some extent, automatic processes to support the detection and analysis of misinformation.

Since October 2018, as part of their election integrity efforts, Twitter has periodically released datasets consisting of tweets from accounts they believe to be spreading propaganda on behalf of various states and organisations[4]. Three of these dataset releases (as of March 2020) contain tweets from accounts Twitter believes to be linked to the Iranian state. These datasets contain all tweets from these accounts posted prior to their suspension from Twitter: 9 million tweets from 5,978 accounts in 67 languages (of which 13 had fewer than 100 tweets); almost 90% of the tweets were in 7 languages.

In this paper we explore the extraction of tweets about the Iranian nuclear deal from datasets of Iranian state-sponsored tweets. Although they included many languages, the English-language tweets are significant as they are likely aimed at an international audience, particularly citizens and elites in countries involved in the negotiation of the nuclear deal.

These tweets are labelled automatically for their polarity (positive/ negative/ neutral) and objectivity. The objectivity scores across the sentiment classes are analysed and several supervised machine learning algorithms are evaluated for their performance at sentiment analysis on these labelled datasets. We also analyse how various feature sets impact the performance of the machine learning algorithms. The aim is to facilitate the analysis of state-sponsored propaganda tweets in an (semi-)automatic manner and obtain insights that would not be possible otherwise due to the large volume of data.

Although Twitter is not the 'real world', we argue that it is, nonetheless, a significant phenomenon to examine because it has an out-sized influence on world events, including its politics. For example, in 2013, the Syrian Electronic Army hacked the Associated Press Twitter account, announcing that Barack Obama was injured in explosions at the White House, leading to 140 points being wiped from the Dow Industrial Average in the following two minutes [6]. International diplomatic incidents were also a regular outcome of tweets posted by President Trump during his time in office [7].

To the best of our knowledge, the Iranian releases of the Twitter Election Integrity datasets have not been analysed for their sentiment and objectivity previously. Our research shows that these tweets can be labelled in an automated manner, thus providing a basis for additional research into the contents of these datasets and methods to facilitate this.

The contributions of this research are multifaceted; firstly, we demonstrate that automated sentiment and objectivity labelling are feasible, and find that the neutral class is highly objective, reinforcing the approach used for automated labelling. Secondly, we find that the best performing machine learning algorithm and feature set combination can correctly classify 77% of the Iranian nuclear deal propaganda tweets. Third, we present the lessons we learned while conducting this research by identifying limitations to the approach we took. Finally, we hope that this paper will also be of interest to those outside

M. Barrows was with the School of Computing, University of Portsmouth, UK e-mail: michael.barrows@myport.ac.uk.

E. Haig is with the School of Computing, University of Portsmouth, UK e-mail: ella.haig@port.ac.uk.

D. Conduit is with School of Social and Political Sciences, University of Melbourne, Melbourne, Australia e-mail: dconduit@unimelb.edu.au.

Manuscript received April 19, 2021; revised August 16, 2021.

[1]See for example https://www.loc.gov/law/help/social-media-disinformation/compsum.php

[2]See for example https://digital-strategy.ec.europa.eu/en/library/final-report-high-level-expert-group-fake-news-and-online-disinformation

[3]See for example https://www.gov.uk/government/news/government-cracks-down-on-spread-of-false-coronavirus-information-online and a toolkit for countering disinformation: https://gcs.civilservice.gov.uk/publications/resist-counter-disinformation-toolkit/

[4]https://transparency.twitter.com/en/reports/information-operations.html

computer science, with its empirical findings of interest to those in the social sciences studying Iran and disinformation.

The rest of this paper is structured as follows: Section II contains background information and related work; Section III contains the proposed approach for data extraction, automated labelling of sentiment and objectivity, and the sentiment analysis; Section IV contains the results and discussion of the experiments and Section V concludes the paper.

II. Background and Related Work

There are four areas of background information and related work relevant to the research in this paper: misinformation on social media, Iran's propaganda efforts, the Iranian nuclear deal and sentiment analysis. Due to space, in this section we focus on the Iranian nuclear deal and sentiment analysis; this is not an extensive survey, but, rather, aims to present the wider context in which our research is conducted. More details and the other two areas are covered in the Supplementary Material.

A. Studies on Iran's propaganda

Various studies into Iran's propaganda efforts have been conducted, e.g., [8]–[10]. These studies concern propaganda on social media platforms and websites impersonating legitimate media organisations. For example, a network of personas on social networking platforms and websites that imitate legitimate news outlets were discovered and found to be used to spread misinformation and propaganda that aligned with Iranian narratives [8]. Analysis of the domains identified that some were registered using the same email address and used the same website hosting which indicates an orchestrated campaign by a single entity (assumed to be Iran) [8].

Research on Arabic tweets from the Iranian releases of the Twitter Election Integrity datasets to analyse Iran's interference in the Arab world found that the most popular hashtags used reflected Iranian interests and that Iran attempts to impersonate news outlets (supporting conclusions reached by [8]). It was also found that 69% of tweets containing links were to websites promoting Iranian narratives [9]. Additionally, analysis of domain registration information discovered links between the domains through email addresses, registrant organisations and false registration information; one of the websites shared in the tweets was found to be funded by Iran [9].

Another study using the same datasets explored Iranian interference in the Arab world, with tweets in English to investigate Iran's international efforts [10]. Findings show that Iran used Twitter accounts to spread propaganda and influence the conversation around Arab countries, with an increasing number of tweets coinciding with highly contentious events [10].

Indeed, Iran's social media operations have been conducted in many languages[5] and have focused on hundreds of topics. One of the most prominent of these in the English-language Tweets was the nuclear deal, although no study has to date examined Iran's online information operations related to the nuclear deal. This is despite that Iran's nuclear program has been one of the most significant security issues on the international stage for the past two decades.

[5]we found 67 language values in our datasets, including an "undefined" category representing 5.28% of all tweets; see more details in Section IV-A

B. Sentiment Analysis

Sentiment analysis (or opinion mining) is used to identify the polarity of text and gauge opinions on topics, products and events using computational methods. Text can be split into two classes, objective and subjective [11], and it is subjective text that contains polarity. For our research, sentiment analysis is a useful technique as it helps us to understand the kind of messaging that Iranian information operations were seeking to disseminate during the nuclear deal period.

Before performing sentiment analysis on textual data, preprocessing of the text is undertaken to transform it into a structured format suitable for further analysis. This involves the transformation of a piece of text (e.g., for a tweet, the entire text of the tweet is stored as a string) into numerical features that can then be used for developing classification models. Several representations can used for this purpose: (a) bag of words/terms [12]; (b) word embedding techniques such as Word2Vec [13] and Glove [14] and (c) representations based on BERT (Bidirectional Encoder Representation from Transformers) [15] and its variations (e.g., [16], [17]).

The bag of words representations transforms a piece of text into a vector of values that indicate the occurrence/frequency of each word in the training corpus [12]. Similarly, a bag of terms is represented as a vector of occurrence/frequency values for a variety of terms, such as n-grams (i.e, groups of n words) or combinations of n-grams for different values of n.

In word embedding techniques, each word is represented by a fixed vector size that captures its relation with other words. Massive data is needed for training to created suitable vectors for each word, making them less suitable for data such as tweets [18]. The most relevant criticism of word embeddings for sentiment analysis is that due to the nature of the method words with opposite polarity can be mapped into close vectors [19]. This could explain the relatively poor performance reported in some studies such as [20] and [21]. Moreover, word embeddings have been shown to contain and amplify biases present in data, such as stereotypes and prejudice [22].

The BERT-based approaches are the most recent and they involve the use of pre-trained models on large corpora. The BERTweet model [23] is particularly relevant, as it is a large-scale pre-trained language model for English tweets. Although this has shown good performance on sentiment analysis, this gain in performance comes with a significant increase in computational time and model size, as well as having the drawback of lack of transparency. Moreover, it is not clear why the BERT models bring improvements in performance, which limits hypothesis-driven improvements [24].

While model performance is an important factor, transparency is also important if we are to make inferences based on the outputs of the models. As the aim of our study is to investigate the potential for (semi-)automatic analysis of state-sponsored propaganda Tweets, we chose the more traditional and transparent approach using the bag of words/terms representation. Once the feasibility has been established, we will investigate the potential of other representations to improve the performance of the computational models.

There are some studies that have looked at the sentiment of tweets in the context of political events and misinformation.

For example, an analysis of the 2016 Austrian presidential election found that the winner of the election, Alexander Van der Belle, posted tweets that were predominately neutral, that the negative information (for both candidates) was spread for a longer time compared with positive and neutral information, and that the Van der Belle's Twitter followers participated in a substantial manner in the spread of misinformation about him [25]. In an analysis of fake news, the authors found that fake news contain more negative sentiment and adverse emotional words [26]. Sentiment analysis was also used to predict opinion inversion in tweets of political communication on the Israeli–Palestinian conflict [27].

These studies differ from the one presented in this paper with regard to labelling methods (using deep learning and other lexicons) and the algorithms used. Additionally, as far as we are aware, sentiment analysis has not been previously done on the Twitter Election Integrity datasets.

III. Proposed Approach

This section details the steps of our approach, including data extraction, preprocessing, sentiment and objectivity labelling, and machine learning algorithms evaluation.

A. Tweet Extraction

As Twitter states that they reinstate accounts that successfully appeal their suspension and subsequent removal from the Twitter platform[6], and the datasets are updated occasionally after their initial release (presumably to remove the reinstated accounts), the data was downloaded at the beginning of the research (March 5th, 2020) and was not updated. This ensured that the data remained consistent across all research stages.

All three of the Iranian dataset releases were modified since their initial release; the October 2018 release was last modified on February 11th, 2019; the January 2019 release was last modified on August 25th and 26th, 2019; the June 2019 release was last modified on February 25th, 2020.

To minimise the chance of retaining tweets that are not about the Iranian nuclear deal, three criteria were used to determine if a tweet was relevant. These criteria were: the tweet is labelled as being in English; the tweet is labelled as being published between August 2013 and December 2018; and the tweet has at least one match with a list of key terms. These are explained below.

1) Language Filtering: The language filtering was completed to ensure that the tweets retained were in the English language so that they are consistent and compatible with the selected lexicon for automatic labelling (see Section III-C). The Twitter Election Integrity datasets contain the field 'tweet_language' which contains a two character language code. To select tweets in the English language, this field was filtered for 'en', representing English. The selected tweets were stored in a new dataset.

[6]https://transparency.twitter.com/en/reports/information-operations.html

2) Date Filtering: The date filtering was completed to ensure that retained tweets were published between August 2013 and December 2018. These dates were chosen because Iranian President Rouhani was inaugurated in August 2013 and former US President Trump withdrew the USA from the JCPOA[7] in May 2018. This period of time covers major events including the interim agreement being reached (2013), significant negotiations (2014), the final agreement being reached and implemented (2015/2016), and former President Trump's decertification of Iranian compliance (2017) and subsequent withdrawal of the USA from the agreement (2018). The date filtering was completed using the 'tweet_time' field within the datasets. The year and month are extracted from this field and checked to ensure that they are within the time frame.

3) Key Terms Matching: In addition to time and language filtering, key term matching was used to identify tweets that match a list of key terms related to the Iranian nuclear deal and associated events within the time frame. A list of key terms, including terms such as *JCPOA*, *irandeal*[8], *nuclear talks* and *negotiation* (listed in Supplementary Material), were compared with the text of the tweets to determine matches between them. At least one match between the tweet text and the key terms list is required for a tweet to be retained. 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[9] for each month across the Iranian Twitter Election Integrity datasets. The new dataset of tweets was then stored for preprocessing.

B. Preprocessing

In order to remove noise, personal information, and to make the labelling simpler and less resource intensive, preprocessing was undertaken. The preprocessing included lowercase conversion of the text, normalisation of accented letters, username removal, URL removal, expansion of contracted words, special character removal (including '#' from hashtags), and stopword removal. The stopword removal excludes words that are indicative of negation (listed in the Supplementary Material), so that these were retained.

C. Sentiment Labelling

To label the tweets for their sentiment, the SentiWordNet lexicon [28] was employed. It was chosen as it provides a score between 0 and 1 for both positive and negative sentiment, indicating the degree of sentiment rather than an absolute label for either positive or negative. This was important as the final sentiment label for a given tweet is determined by the sum of its words, and as such, binary labels would not differentiate between words that are slightly or deeply positive or negative, which would impact the reliability of the overall labelling. SentiWordNet was used for sentiment labelling in

[7]Joint Comprehensive Plan of Action – see details in Section II-B of the Suplemmentary Material

[8]on social media key terms consisting of several words are often merged into one, as in this example; this is frequently seen in hashtags

[9]unigrams are single terms/words, such as 'Iran' and 'deal'; bigrams are sets of 2 terms/words, such as 'nuclear deal'

various other studies, including sarcasm detection [29], as a component of a fake news detection system [30] and for sentiment analysis of tweets about the coronavirus pandemic [31].

SentiWordNet extends WordNet [32], meaning that WordNet synsets have positive, negative and objectivity scores attached to them through SentiWordNet. This extension allows for word sense disambiguation (WSD) through the WordNet synsets definition (see details in the next subsection). This also allows for different sentiment scores for the same word, within different senses (synsets). By performing WSD, the context of the word is taken into consideration, and the sentiment scores for the word are potentially more accurate than using a single set of scores for a given word regardless of its sense [33].

1) Word Sense Disambiguation (WSD): To complete word sense disambiguation, the text of the tweet was tokenised and then tagged for its parts-of-speech (POS). WordNet was then used to identify synsets for each of the words in the tweet; this was completed with the POS tag assigned to the word in order to limit the number of synsets returned.

In order to identify the correct synset for an individual word in a tweet, a simplified version of Lesk's algorithm was used [34]. This implementation compares the words in the tweet against the words of the synsets definition. The synset with the most matches is selected as the synset for the word. Stemming was used to reduce words to the root (removing prefixes and suffixes). Stemming was implemented using the PorterStemmer which is available in Python through the Natural Language Toolkit (NLTK)[10]; this improves the chance of a match between the definition and the tweet as only the roots of the words are taken into consideration.

2) Scoring (see Algorithm 1): Once the synsets for the words of the tweets were identified, the label could be determined and the objectivity score calculated. For each of the synsets identified for an individual tweet, the positive and negative scores were obtained and added to the relevant set (PSS for the positive sentiment set and NSS for the negative sentiment set). The scores in each set were then summed, with the result being divided by the number of scores in the set (i.e. the cardinality of the sets: $|PSS|$, $|NSS|$) to generate the average of the scores which was then used as the positive ($Tweet\ Positive\ Score$) and negative ($Tweet\ Negative\ Score$) score. This is shown for positive and negative sentiment is Equations 1 and 2 respectively. The label with the higher score was assigned. The neutral label was assigned where the positive and negative scores were equal.

$$Tweet\ Positive\ Score = \frac{\sum_{i=1}^{n=|PSS|} PSS_i}{|PSS|} \quad (1)$$

$$Tweet\ Negative\ Score = \frac{\sum_{i=1}^{n=|NSS|} NSS_i}{|NSS|} \quad (2)$$

where i represents a synset.

[10]https://www.nltk.org/howto/stem.html

Algorithm 1: Sentiment and objectivity labelling of the extracted and preprocessed tweets

input : Dataset T of all extracted tweets
output: The dataset T with sentiment labels

for each $tweet$ in dataset T **do**
  $PSS = \emptyset$
  $NSS = \emptyset$
  $OSS = \emptyset$
  tokenise $tweet$
  assign part-of-speech tags to $tweet$
  for each $word$ in $tweet$ **do**
    retrieve WordNet $synsets$ for $word$
    for each $synset$ in $synsets$ **do**
      identify number of matches between the $synset$ definition and the $tweet$
    end for
    $synset$ = $synset$ with the most matches
    $SentiSynset$ = $synset$
    add $SentiSynset\ Positive\ Score$ to set PSS
    add $SentiSynset\ Negative\ Score$ to set NSS
    add $SentiSynset\ Objective\ Score$ to set OSS
  end for
  $tweet\ Positive\ Score = \sum_{i=1}^{n=|PSS|} PSS_i\ /\ |PSS|$
  $tweet\ Negative\ Score = \sum_{i=1}^{n=|NSS|} PSS_i\ /\ |NSS|$
  $tweet\ Objective\ Score = \sum_{i=1}^{n=|OSS|} OSS_i\ /\ |OSS|$
  if $tweet\ Positive\ Score$ = $tweet\ Negative\ Score$ **then**
    $tweet\ sentiment\ label$ = neutral
  else if $tweet\ Positive\ Score > tweet\ Negative\ Score$ **then**
    $tweet\ sentiment\ label$ = positive
  else
    $tweet\ sentiment\ label$ = negative
  end if
end for
Return dataset T

The objectivity score ($Tweet\ Objectivity\ Score$) was also determined in the same way; the objectivity scores (OSS) were summed and divided by the number of synset matches (cardinality: $|OSS|$) to provide a score between zero and one which is an average of all objectivity scores for the individual tweet. This is shown in Equation 3.

$$Tweet\ Objectivity\ Score = \frac{\sum_{i=1}^{n=|OSS|} OSS_i}{|OSS|} \quad (3)$$

where i represents a synset, similarly to Equations 1 and 2.

3) Match Percentage Threshold (MPT): To investigate the influence of the lexicon coverage on the labelling process and the subsequent sentiment analysis, we defined a metric, which we called Match Percentage Threshold (MPT), for capturing this coverage and employed it as a threshold in the experiments described in the next subsection.

The Match Percentage Threshold (MPT) is a metric that identifies the number of words in a given tweet that achieved matches with the lexicon ($Matched\ Tweet\ Words$), ex-

pressed as a percentage of the total words in the individual tweet ($Total\ Tweet\ Words$). This is shown in Equation 4.

$$MPT = \frac{|Matched\ Tweet\ Words|}{|Total\ Tweet\ Words|} \quad (4)$$

As a higher MPT value represents more word matches between the tweet and the sentiment lexicon, the overall label assigned to the tweet is likely to be more accurate because more sentiment scores are taken into consideration when determining the sentiment label.

Within the machine learning task, experiments were conducted with various minimum MPT values which were incremented by 10% with each set of experiments. The maximum MPT value was always 100%. Tweets below the minimum MPT value were excluded from the experiment, e.g. for an MPT of 40%, any tweets that had fewer than 40% words matched to the lexicon were removed.

D. Sentiment Analysis

To perform the sentiment analysis task using machine learning, five supervised machine learning algorithms were implemented and evaluated, namely k-Nearest Neighbours (k-NN), Decision Tree (DT), Naive Bayes (NB), Support Vector Machine (SVM) and Random Forest (RF). The experiments are conducted using k-fold cross-validation with ten folds. Experiments were run using Python version 3.6.9 on a computer running Ubuntu version 18.04.5 LTS with an AMD Ryzen 3 processor with a frequency of 3.5GHz and 8GB of RAM.

Experiments were conducted across the Match Percentage Thresholds (MPT), starting with the lowest threshold of 0% (i.e., all tweets regardless of coverage were included) that increased by 10% with each iteration until 100% was reached.

To identify the performance for different features, the bag-of-words (bag-of-ngrams) approach was used with the Count Vectoriser from the Scikit-Learn package in Python. The five sets of features were: unigrams (each individual word), bigrams (two consecutive words), trigrams (three consecutive words), unigrams+bigrams and unigrams+bigrams+trigrams.

The hyperparameters for four of the five algorithms were adjusted throughout the experiments to try and achieve the best results. This was applied to all classification algorithms with the exception of decision tree (as it does not take any hyperparameters). Three sets of experiments were conducted; baseline, initial experiments and further experiments. Each set of experiments used different hyperparameter values. The hyperparameters used for random forest were generated using a trial and error approach independent of the three experiment stages, with all hyperparameters for random forest used across all of the MPT values and features. This approach was used for random forest because it takes two hyperparameters, namely trees and features. This would therefore require more experiments to try all of the combinations to identify which combination improved the results and which did not.

The baseline experiments were conducted to obtain results for the algorithms using the default hyperparameters. The initial experiments were conducted to evaluate the performance of the algorithms with two different hyperparameters (one higher and one lower than the default). The results (i.e., F-scores) from these and the baseline experiments were used to optimise the hyperparameters used in the further experiments.

The further experiments consisted of three or four further experiments (dependent upon the previous results). The hyperparameters were set depending on the previous results. For example, if the best performance for a classifier was achieved using a hyperparameter higher than the default, two of the experiments were conducted with hyperparameters higher than the best performing hyperparameter; one experiment was conducted with a value lower than the lowest hyperparameter. The inverse was completed if the lower hyperparameter achieved the best result.

If the baseline experiment achieved the best result for an algorithm, two experiments would be completed with higher values than the highest hyperparameter and two experiments with lower values than the lowest hyperparameter.

To evaluate the performance of the algorithms, F-score and accuracy were used. Precision and recall, which are used to calculate the F-score, are also reported. These metrics are explained in the Supplementary Material.

IV. Results and Discussion

In this section, results from the nuclear deal tweet extraction, sentiment labelling, objectivity labelling and sentiment analysis are presented and discussed.

A. Tweet Extraction

Prior to any action being taken on the datasets, the volume of tweets across the three Iranian datasets totalled 9,314,829 from 5,978 accounts in 67 languages. The language distribution for the most prominent languages is displayed in Table I, showing that almost 90% of Tweets are in 7 languages and that for 5.28% of the tweets the language was not defined[11].

TABLE I
Dominant languages

Language	Percentage	Cumulative Percentage
Arabic	23.15	23.15
English	20.67	43.82
Persian	16.34	60.16
Urdu	11.14	71.30
Undefined	5.28	76.58
French	3.85	80.43
Indonesian	3.80	84.23
Hindi	3.07	87.30

The total number of tweets matching the English language, time and key terms criteria was 24,117 (0.26%) as shown in Table II. Of the 9.3 million tweets in the dataset, around 21% of these were in English (1.9 million) and around 94% (8.7 million) of all tweets were within the time frame.

Of all the tweets in the Iranian dataset releases, 0.27% (24,731) of the tweets had matches with the key terms list. Overall, 0.26% (24,117) of all tweets matched all three criteria and were extracted. This represents an extremely small portion

[11] see language codes at: http://web.archive.org/web/20210123205307/https://developer.twitter.com/en/docs/twitter-api/enterprise/powertrack-api/guides/operators

of the dataset, indicating that many of the tweets not extracted were potentially about other topics.

TABLE II
VOLUME OF ALL TWEETS AND VOLUME OF TWEETS EXTRACTED.

Description	Volume	Percentage
All tweets	9,314,829	100.00%
Tweets between Aug 2013 & Dec 2018	8,795,778	94.43%
Tweets in English	1,925,772	20.67%
Tweets matching key terms	24,731	0.27%
Tweets meeting all three criteria	**24,117**	**0.26%**

Given Iran's international isolation at the beginning of Rouhani's presidency [35], [36] and President Rouhani's policy of trying to improve international relations, it is not unexpected that Iran's state-sponsored propaganda is not limited to a single topic. It is however unexpected that only 0.27% of all tweets matched keywords related to the Iranian nuclear deal as Iran's nuclear program has been the subject of international concern and outrage for many years [37] and the sanctions related to Iran's nuclear programme have caused the Iranian economy significant harm [36]. This may indicate that (1) the nuclear deal was not a significant issue for those running Iranian information operations on Twitter, or (2) that those behind the information operations (which could be multiple actors) were running a 'scatter gun' approach across a range of topics. The corpus nonetheless included more than 24,000 tweets, representing a significant effort.

B. Extracted Data

Of the 24,117 tweets extracted, 12,972 (53.04%) were original tweets and 11,325 (46.96%) were retweets. All retweets were of other tweets contained in these dataset releases. These were removed from the extracted data to prevent the classification algorithms from developing a bias based on the features contained within a small number of tweets that have been retweeted many times.

Given that almost half of the extracted tweets were retweets of other tweets, it would appear that Iran is using retweets to amplify their propaganda; whether this is by automated means or not is something that requires further investigation.

Fig. 1 shows that major increases in the volume of tweets occur periodically. The five highest peaks in the graph coincide with political events including an interim agreement being reached (November 2013); talks on the sidelines of the U.N. General Assembly (September 2014); negotiators meeting in Lausanne, Switzerland (March 2015); the nuclear deal being agreed (July 2015) and former US President Trump withdrawing the USA from the nuclear agreement (May 2018) [38].

C. Sentiment Labelling

The results for the sentiment labelling are shown in Table III. After the labelling was completed, 5 of the tweets could not be automatically labelled (see further details below); because these tweets were not labelled, they were removed from the dataset and not further processed. The volume of positive and negative tweets are similar (38% and 37% respectively). The volume of the neutral class is significantly smaller than the positive and negative classes (25%).

TABLE III
VOLUME OF TWEETS FOR EACH SENTIMENT LABEL.

Class	Volume	%
Positive	4,863	38.02%
Neutral	3,162	24.72%
Negative	4,762	37.23%
Unclassified	5	0.04%

This could suggest several things, including: (a) that more than one entity operated the accounts, meaning that different parts of the dataset reflected different actors' objectives; (b) it could mean that Iran pursued a social media strategy akin to Russian operations in Eastern Europe, which aimed at creating confusion and eroding trust, rather than overtly pursuing a clear agenda pro- or against- a certain policy [39].

In other studies investigating Iranian propaganda, a network of websites were uncovered that imitated legitimate press organisations and were used to spread Iranian propaganda as legitimate news [8]. Given this, it is somewhat unexpected that the neutral class consists of a quarter of tweets, particularly in comparison to the size of the positive and negative classes. One possible explanation for this is that Iran's propaganda techniques have somewhat evolved from the legitimate news method and the Twitter accounts are portraying themselves as ordinary users contributing to the discourse around the nuclear deal in a polarising manner. The 25% of tweets that make up the neutral class indicate that spreading propaganda as legitimate news, potentially with objective language, still has a place in Iran's propaganda toolkit.

The 5 unclassified tweets could not be labelled after preprocessing because there were zero matches between the words in the tweet and the SentiWordNet lexicon. Reasons for this include tweets primarily consisting of usernames with other words being identified as stopwords, or URL(s) with no other words, or words not present in the SentiWordNet lexicon.

The neutral class is commonly identified as tweets containing neither positive or negative sentiment, and the method for calculating these scores differs between studies [31], [40]–[42]. For example, one method for calculating these scores is to subtract the negative sentiment score from the positive [31], and if the score is below zero, the text is deemed to be negative; the opposite is true for positive sentiment. The problem with this approach is that when a piece of text is equally positive and negative, the score is zero, and would therefore be classed as neutral. It may be that the text is both positive and negative, rather than neutral in the sense that it is objective. Consequently, currently there are no methods that distinguish between the two types of neutral text: (a) text that is objective and (b) text that contains positive and negative sentiment in equal measure. This problem, however, is outside of the scope of this study and therefore, if the positive and negative scores are equal, the tweet is labelled as being neutral. This is further discussed in Section V.

D. Objectivity Scoring

The results for the objectivity labelling are shown in Table IV and Fig. 2. From these results, it can be observed that

the vast majority of the objectivity scores within the neutral class are nearer to 1 (indicating maximum objectivity) than the positive and negative classes.

Of all of the labelled tweets, only tweets labelled as being neutral had objectivity scores of 1.0, i.e. maximum objectivity. Moreover, at least 50% of the tweets labelled as neutral had the maximum objectivity score, as indicated by the median objectivity value of 1.0. Tweets in the neutral class also had a higher minimum objectivity score than its positive and negative counterparts.

The positive and negative classes have high average objectivity scores (>90%), although, as seen in Fig. 2, the positive and negative objectivity scores are more spread out across the 90-100% range. As shown in Fig. 2, the lowest objective score in the neutral class (0.65) was much higher than that of the positive (0.125) and negative (0.25) classes.

On one hand, these scores seem to suggest that there is a contradiction between the sentiment scoring (positive or negative) and the objectivity scoring, which indicates that many positive and negative tweets have a high objectivity score. On the other hand, this may indicate the use of objective language as a technique for bringing legitimacy and persuasiveness to the propaganda messages. This is also aligned with the findings of [8], whose research indicated impersonating legitimate news as a strategy to spread Iranian propaganda.

TABLE IV
STATISTICS FOR THE OBJECTIVITY SCORING BROKEN DOWN BY CLASS

Sentiment Class	Mean	Standard Deviation	Min	Median	Max
Positive	0.9126	0.0673	0.125	0.925	0.9938
Neutral	0.9852	0.0409	0.65	1.0	1.0
Negative	0.8947	0.0747	0.25	0.9063	0.9983

Having discussed the points above, returning to the data, it does show that the neutral class has higher objectivity scores for mean, minimum and median scores than its positive and negative counterparts. Consequently, this relative alignment between the polarity scoring and the objectivity score gives some degree of confidence in the automatic labelling outcome.

E. Volume of Tweets per Match Percentage Threshold

Table V displays the volume of tweets and their distribution among the three polarity labels (positive, neutral and negative) for different values of the Match Percentage Threshold. For example, for a Match Percentage Threshold of 70%, the tweets that have at least 70% of their words matched to the lexicon, up to 100% matched words (i.e. all the words in a tweet are in the lexicon), are included. Fig. 3 displays this distribution. The results show that a minimum match percentage threshold up to and including 50% retains more than 90% of the data. As the threshold is increased above 50%, the volume of retained tweets decreases significantly.

The individual sentiment classes decrease in volume at a similar rate to each other, with the exception of the neutral class, which may be a result of its lower volume in comparison with the positive and negative classes.

This data shows that more than half of the tweets have an MPT value of at least 70%, and 75% of the tweets have an MPT value of at least 60%. If we consider that at least half of the words in a tweet (after preprocessing) must have matches with the lexicon to be considered reliable matches, the data

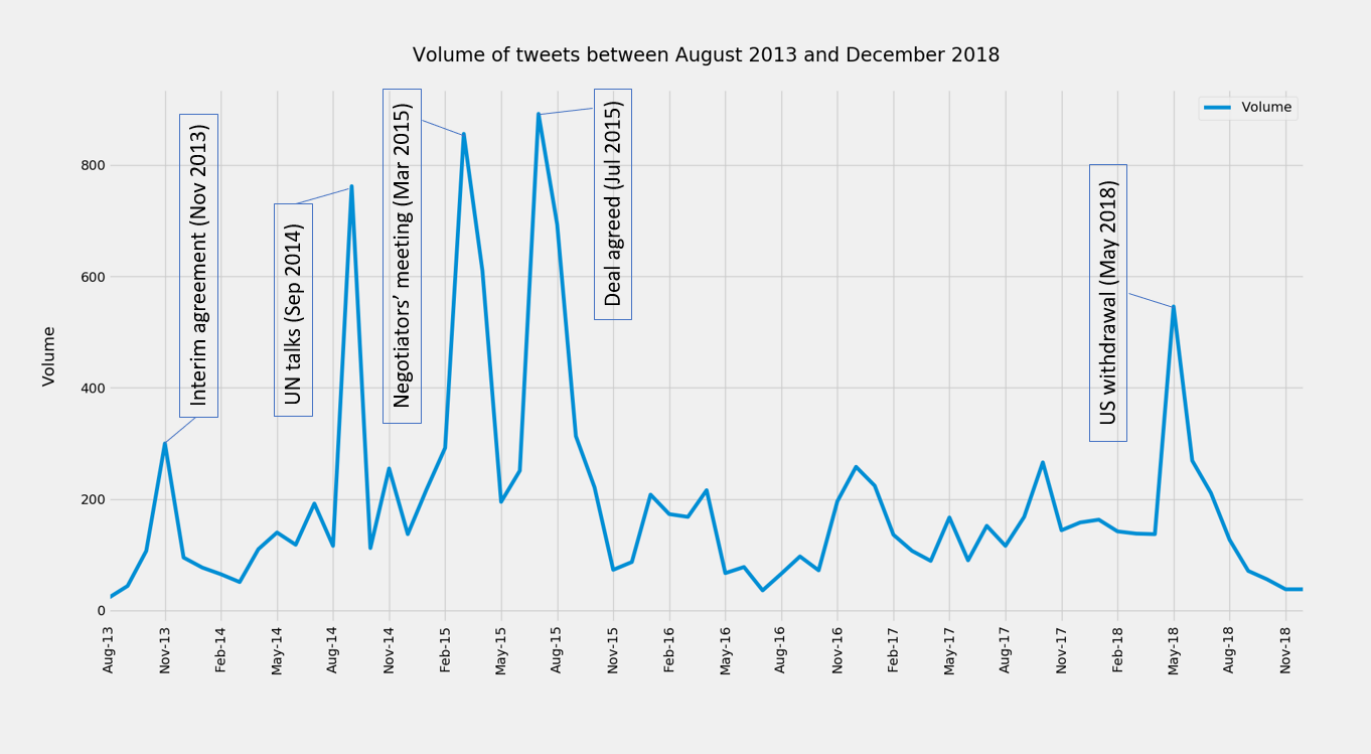

Fig. 1. Volume of tweets between August 2013 and December 2018 (retweets removed)

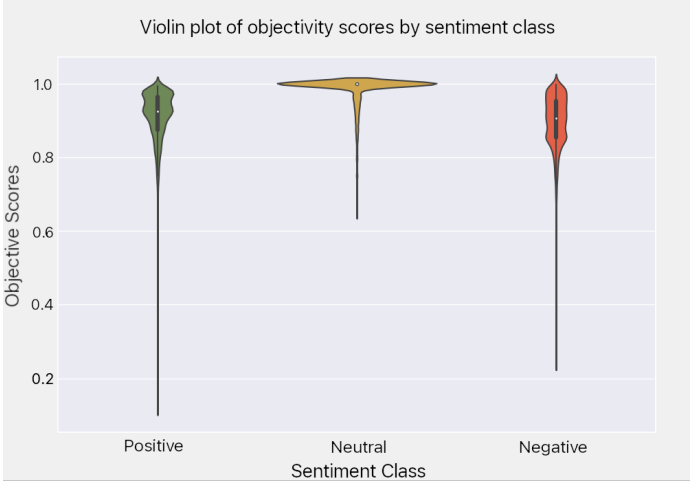

Fig. 2. Violin plot of objectivity scores across the three sentiment classes

TABLE V
NUMBER OF TWEETS FOR DIFFERENT MATCH PERCENTAGE THRESHOLDS (IN STEPS OF 10%) WITH CLASS BREAKDOWNS.

MPT Value	# tweets	%	Positive	Neutral	Negative
≥0%	12787	100.0%	4863	3162	4762
≥10%	12784	99.98%	4863	3159	4762
≥20%	12753	99.73%	4862	3132	4759
≥30%	12602	98.55%	4844	3007	4751
≥40%	12268	95.94%	4763	2807	4698
≥50%	11670	91.26%	4585	2498	4587
≥60%	9688	75.76%	3664	1934	4090
≥70%	6974	54.54%	2734	1367	2873
≥80%	4130	32.3%	1691	858	1581
≥90%	1522	11.9%	619	284	619

shows that more than 90% of tweets about the Iranian nuclear deal achieved this standard.

F. Sentiment Analysis

Table VI shows the hyperparameters used in the experiments. We report the best (Table VII) and worst (in the Supplementary Material) results for each algorithm and each feature set, as well as the performance of the best performing algorithms for each class (Table VIII). MPT stands for Match Percentage 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.

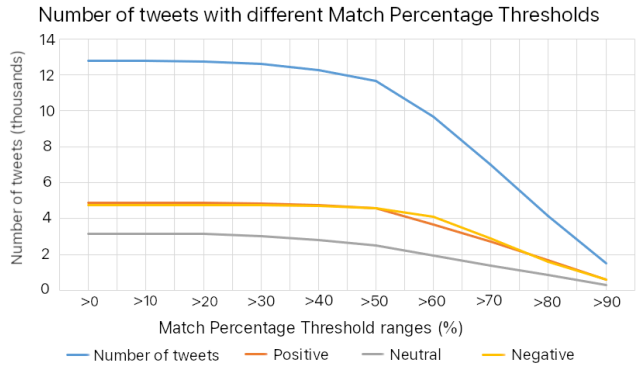

Fig. 3. Number of tweets according to different Match Percentage ranges (in steps of 10%).

TABLE VI
HYPERPARAMETERS USED IN THE EXPERIMENTS

KNN	Linear SVM	Naive Bayes	Random Forest
1	0.3	0.3	[300, 600]
3	0.5	0.5	[200, 400]
default (5)	0.8	0.8	[150, 300]
7	default (1.0)	default (1.0)	[100, 200]
9	1.5	1.2	[75, 150]
11	1.2	1.5	default ([100, n_features])
	1.8	1.8	[50, 100]

TABLE VII
BEST RESULTS FOR EACH OF THE MACHINE LEARNING ALGORITHMS USED AND N-GRAMS (MACRO AVERAGED)

n-grams	Metric	k-NN	DT	SVM	NB	RF
UNI	MPT	10	30	30	10	40
	HP	1	-	0.3	default	[300, 600]
	P	0.6076	0.7277	0.7587	0.6576	**0.7591**
	R	0.6004	0.7328	0.7617	0.6554	**0.7695**
	F	0.5917	0.7295	0.7594	0.6559	**0.7605**
	Acc	0.5964	0.7375	**0.7701**	0.664	0.7678
BI	MPT	10	10	10	10	10
	HP	1	-	0.5	0.5	[50, 100]
	P	**0.7282**	0.6237	0.661	0.6266	0.6507
	R	0.5394	0.6188	**0.6626**	0.6244	0.6368
	F	0.4818	0.611	**0.6582**	0.625	0.6284
	Acc	0.4832	0.6147	**0.665**	0.633	0.6311
TRI	MPT	40	20	0	0	40
	HP	1	-	0.3	0.5	[150, 300]
	P	**0.7507**	0.6264	0.6373	0.6031	0.6418
	R	0.535	0.5643	**0.5745**	0.5523	0.5711
	F	0.4636	0.5349	**0.577**	0.557	0.5371
	Acc	0.4639	0.5311	**0.5976**	0.5728	0.5317
UNI + BI	MPT	10	40	0	0	40
	HP	1	-	0.3	0.8	[300, 600]
	P	0.5894	0.727	**0.7628**	0.682	0.7523
	R	0.5487	0.7325	**0.7663**	0.6714	0.7626
	F	0.5339	0.7288	**0.7634**	0.6748	0.7511
	Acc	0.5428	0.7387	**0.7708**	0.6846	0.7578
UNI + BI + TRI	MPT	10	40	20	0	40
	HP	1	-	default	0.8	[300, 600]
	P	0.5937	0.729	**0.7617**	0.6838	0.7489
	R	0.5118	0.7347	**0.7662**	0.6676	0.7576
	F	0.4932	0.7307	**0.7622**	0.6723	0.7462
	Acc	0.5123	0.741	**0.7691**	0.683	0.7529

The best results were obtained using the support vector machine (SVM) algorithm, with a linear kernel, the whole dataset (minimum MPT: 0%), with a hyperparameter of 0.3 and using unigrams and bigrams as features. Across all of the classification algorithms and features, the best results all had minimum match percentage thresholds of 40% or lower which used more than 90% of the extracted tweets.

The SVM algorithm achieved the best results (in terms of accuracy) across every set of features evaluated, with either the default hyperparameter (1.0) or lower; SVM also achieved the best F-score across four of the five feature sets, with Random Forest performing best for the remaining feature set.

The k-NN classifier achieved the worst results among the algorithms across each of the feature sets in both the best (Table VII) and worst (Table III in the Supplementary Material) results while only using one neighbour to classify a given tweet (despite experiments with every other hyperparameter listed for k-NN in Table VI). It is also observed that the precision and recall values for k-NN using bigrams and trigrams had a difference of around 0.2, which is higher than the

TABLE VIII
RESULTS PER CLASS FOR BEST RESULT FOR EACH OF THE ALGORITHMS

	Metric	k-NN	DT	SVM	NB	RF
	MPT	10	40	0	0	40
	HP	1	default	0.3	0.8	[300, 600]
	n-grams	UNI	UNI + BI + TRI	UNI + BI	UNI + BI	UNI
Positive	P	0.7235	0.7606	0.7871	0.6634	**0.7952**
	R	0.5318	0.7518	**0.7842**	0.728	0.7643
	F	0.6128	0.756	**0.7855**	0.694	0.7792
Neutral	P	0.4487	0.6202	**0.6596**	0.6576	0.6282
	R	0.6354	0.6947	0.7333	0.5689	**0.7803**
	F	0.5244	0.6552	0.6943	0.6096	**0.6959**
Negative	P	0.6505	0.806	0.8416	0.7251	**0.854**
	R	0.634	0.7575	**0.7814**	0.7174	0.7638
	F	0.6378	0.7808	**0.8103**	0.7209	0.8063
Macro Average	P	0.6076	0.729	**0.7628**	0.682	0.7591
	R	0.6004	0.7347	0.7663	0.6714	**0.7695**
	F	0.5917	0.7307	**0.7634**	0.6748	0.7605
	Acc	0.5964	0.741	**0.7708**	0.6846	0.7678

difference between precision and recall for any other algorithm in the best results table (Table VII).

While SVM achieved the best results across most of the feature sets, RF achieved similar results across unigrams + bigrams and unigrams + bigrams + trigrams. RF achieved better results than SVM for precision, recall and F-score when using unigrams as a feature set. NB achieved similar results to the SVM algorithm with trigrams.

The class breakdown for the best results across algorithms is shown in Table VIII. These show that across all algorithms, the neutral class achieved a worse performance in comparison with the positive and negative classes, which could be due to the smaller volume of data in this class. Another possibility is the manner in which the neutral class was identified, thus containing tweets that were positive and negative in equal measure.

adsawsdawsdf sfews in which the neutral class was identified, thus containing tweets that were positive and negative in equal measure.

In the breakdown of performance by class, it can also be seen that the negative class achieves the best F-score across all of the algorithms. A potential reason for this is that negative sentiment may be more explicit in these tweets, and therefore easier to detect, however, the difference between the positive and negative F-scores is on average 0.25.

V. CONCLUSIONS AND RESEARCH DIRECTIONS

In this paper, tweets from Iranian state-sponsored actors about the Iranian nuclear deal have been automatically labelled for their sentiment and objectivity. Five machine learning classification algorithms were evaluated across five different feature sets for their performance at the task of sentiment analysis. Additionally, we analysed the percentage of matches between the tweet and the lexicon, and the objectivity scores across the sentiment classes.

This provided an important insight into Iranian social media strategy related to the nuclear deal. With our sentiment analysis finding that the dataset contained a similar number of tweets labelled as 'positive' or 'negative', it raised questions as to (1) Whether the dataset reflects a single cohesive strategy linked to a single entity, or whether the dataset captured multiple information operations run by actors with varying agendas, and (2) If the accounts are indeed all linked to a single information operation, then there may be parallels with Russia's strategy of sowing confusion by swamping Twitter with tweets, rather than overtly pursuing a messaging for- or against- the nuclear deal. Either way, these are important observations that highlight the utility of deploying computer science methodologies in the pursuit of answers to questions in the international policy arena.

With regard to the sentiment analysis experiments, the results of the algorithm evaluation show that the SVM algorithm with a linear kernel achieved an accuracy of 0.77 and an F-score of 0.76, showing that it can correctly classify more than three quarters of these tweets. The results also show that at least half of the words in a tweet matched words in the SentiWordNet lexicon in more than 90% of tweets.

Other studies have identified that sentiment can improve misinformation detection, however, our results do not show the same to be true for propaganda, particularly because the polarised classes were almost equal in terms of volume and Iran has previously used imitations of legitimate news sources to spread propaganda, with the objectivity of the neutral class implying that this could still be the case.

Limitations of this study include: Twitter does not provide details of how it determines which accounts are state-sponsored actors, and it potentially removes accounts from the datasets upon the successful appeal of their suspension and reinstatement on the platform. The automatic labelling of the tweets for both objectivity and subjectivity was not manually verified and could be subject to error. This risk was somewhat mitigated by using word sense disambiguation. Another potential limitation is that the data extraction was not verified, meaning that the data could contain tweets about other topics; the list of key terms used to extract the data was created after analysing frequent bigrams contained in the tweets, over monthly periods in an attempt to mitigate this risk.

The labelling of the neutral class presents a limitation as a tweet can contain both positive and negative sentiments where both scores are equal, and therefore should be classified as both sentiment classes. While this could have potentially been addressed by performing sentiment analysis at the sentence level, tweets are short in length, and may have only contained one sentence with both sentiments. Labelling the tweets in the way that we have is consistent with other studies implementing sentiment labelling with a lexicon [31].

Future research directions include emotion detection on these datasets, topic analysis for validating tweet extraction, evaluating various lexicons for labelling these tweets and using different text representations. Future research will also include subsequent releases of Iranian propaganda by Twitter.

REFERENCES

[1] 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.

[2] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[3] 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.

[4] S. Torpan, S. Hansson, M. Rhinard, A. Kazemekaityte, P. Jukarainen, S. F. Meyer, A. Schieffelers, G. Lovasz, and K. Orru, "Handling false information in emergency management: A cross-national comparative study of european practices," *International Journal of Disaster Risk Reduction*, vol. 57, p. 102151, 2021.

[5] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild *et al.*, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[6] H. A. Popkin, "AP latest victim in string of Twitter break-ins by Syrian Electronic Army," 4 2013. [Online]. Available: https://www.nbcnews.com/tech/tech-news/ap-latest-victim-string-twitter-break-ins-syrian-electronic-army-flna6c9567459

[7] K. Calamur, "The International Incidents Sparked by Trump's Twitter Feed in 2017," 12 2017. [Online]. Available: https://www.theatlantic.com/international/archive/2017/12/trump-tweets-foreign-policy/547892/

[8] G. Lim, E. Maynier, J. Scott-Railton, A. Fittarelli, N. Moran, and R. Deibert, "BURNED AFTER READING Endless Mayfly's Ephemeral Disinformation Campaign," University of Toronto, The Citizen Lab, Toronto, Canada, Tech. Rep., 5 2019. [Online]. Available: https://tspace.library.utoronto.ca/bitstream/1807/96661/1/Report%23118--endlessmayfly.pdf

[9] M. Elswah, P. N. Howard, and V. Narayanan, "Iranian Digital Interference in the Arab World," Project on Computational Propaganda, Oxford, UK, Tech. Rep., 2019. [Online]. Available: https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/04/Iran-Memo.pdf

[10] B. Kießling, J. Homburg, T. Drozdzynski, and S. Burkhardt, "State propaganda on twitter: How Iranian propaganda accounts have tried to influence the international discourse on Saudi Arabia," in *Lecture Notes in Computer Science*, vol. 12021. Springer, 2020, pp. 182–197.

[11] B. Liu, "Sentiment Analysis and Subjectivity," in *Handbook of Natural Language Processing*, 2nd ed., N. Indurkhya and F. Damerau, Eds. Boca Raton, Florida: CRC Press, 2010, ch. 26, pp. 627–666.

[12] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.

[13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[14] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing*, 2014, pp. 1532–1543.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[16] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[17] H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and T. Zhao, "Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization," in *The 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2177–2190.

[18] H. T. Phan, V. C. Tran, N. T. Nguyen, and D. Hwang, "Improving the performance of sentiment analysis of tweets containing fuzzy sentiment using the feature ensemble model," *IEEE Access*, vol. 8, pp. 14 630–14 641, 2020.

[19] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for twitter sentiment classification," in *The 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1)*, 2014, pp. 1555–1565.

[20] O. Araque, I. Corcuera-Platas, J. F. Sánchez-Rada, and C. A. Iglesias, "Enhancing deep learning sentiment analysis with ensemble techniques in social applications," *Expert Systems with Applications*, vol. 77, pp. 236–246, 2017.

[21] Y. Ren, R. Wang, and D. Ji, "A topic-enhanced word embedding for twitter sentiment classification," *Information Sciences*, vol. 369, pp. 188–198, 2016.

[22] O. Papakyriakopoulos, S. Hegelich, J. C. M. Serrano, and F. Marco, "Bias in word embeddings," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 446–457.

[23] D. Q. Nguyen, T. Vu, and A. T. Nguyen, "Bertweet: A pre-trained language model for english tweets," *Preprint arXiv:2005.10200*, 2020.

[24] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in bertology: What we know about how bert works," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 842–866, 2020.

[25] E. Kušen and M. Strembeck, "Politics, sentiments, and misinformation: An analysis of the Twitter discussion on the 2016 Austrian Presidential Elections," *Online Social Networks and Media*, vol. 5, no. 2018, pp. 37–50, 3 2018.

[26] O. Ajao, D. Bhowmik, and S. Zargari, "Sentiment Aware Fake News Detection on Online Social Networks," in *IEEE Intern. Conf. on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 2507–2511.

[27] Y. Matalon, O. Magdaci, A. Almozlino, and D. Yamin, "Using sentiment analysis to predict opinion inversion in tweets of political communication," *Scientific reports*, vol. 11, no. 1, pp. 1–9, 2021.

[28] A. Esuli and F. Sebastiani, "SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining," in *The 5th International Conference on Language Resources and Evaluation*, 2006, pp. 417–422.

[29] S. K. Bharti, B. Vachha, R. K. Pradhan, K. S. Babu, and S. K. Jena, "Sarcastic sentiment detection in tweets streamed in real time: a big data approach," *Digital Communications and Networks*, vol. 2, no. 3, pp. 108–121, 2016.

[30] V. Sabeeh, M. Zohdy, A. Mollah, and R. Al Bashaireh, "Fake News Detection on Social Media using Deep learning and Semantic Knowledge Sources," *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 18, no. 2, pp. 45–68, 2020.

[31] 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.

[32] Princeton University, "WordNet — A Lexical Database for English," 2010. [Online]. Available: https://wordnet.princeton.edu

[33] C. Sumanth and D. Inkpen, "How much does word sense disambiguation help in sentiment analysis of micropost data?" in *The 6th workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2015, pp. 115–121.

[34] A. Kilgarriff and J. Rosenzweig, "English SENSEVAL: Report and Results," in *Proceedings of the Second International Conference on Language Resources and Evaluation*, 2000, pp. 1–5.

[35] 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.

[36] R. Shanahan, "Iranian foreign policy under Rouhani," Lowy Institute for International Policy, Tech. Rep., 2015. [Online]. Available: http://www.jstor.org/stable/resrep10162

[37] 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.

[38] 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

[39] 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.

[40] G. Li and F. Liu, "Sentiment analysis based on clustering: a framework in improving accuracy and recognizing neutral opinions," *Applied intelligence*, vol. 40, no. 3, pp. 441–452, 2014.

[41] A. Balahur and J. M. Perea-Ortega, "Sentiment analysis system adaptation for multilingual processing: The case of tweets," *Information Processing & Management*, vol. 51, no. 4, pp. 547–556, 2015.

[42] A. Valdivia, M. V. Luzón, E. Cambria, and F. Herrera, "Consensus vote models for detecting and filtering neutrality in sentiment analysis," *Information Fusion*, vol. 44, pp. 126–135, 2018.