Meta-learning for Fake News Detection Surrounding the Syrian War

AUTHORS

Fatima K. Abu Salem,*,1,2 Roaa Al Feel,1,2 Shady Elbassuoni,1,2 Hiyam Ghannam,1,2 Mohamad Jaber,3 and
May Farah4

* Lead contact and corresponding author fa21@aub.edu.lb
1 Equally contributing authors
2 Department of Computer Science, American University of Beirut, Beirut, Lebanon
3 Google, Zurich, Switzerland
4 Department of Sociology, Anthropology, and Media Studies, American University of Beirut, Beirut, Lebanon

THE BIGGER PICTURE

As world and regional powers get more embroiled in the Syrian war, serious questions arise surrounding the credibility of news documenting the facts of war in the decade long conflict. The spread of fake news around documentation of the war, going beyond plain news bias, compromises not only the integrity of the actual reporting, but can contribute to psychological warfare that drives the exodus and constant mobility of refugees, and hampers humanitarian planning for delivering aid to distraught communities.

Detecting fake news is a difficult task for humans and automated technological tools alike. In this manuscript, we propose meta-learning and machine learning approaches towards the automatic detection of fake news arising from the Syrian war. Our work reveals the importance of certain features specific to armed conflict in the Middle East, such as sectarian language, and consistency with respect to ground truth from a fact checking repository. It also highlights the efficacy and quality of meta-learning techniques when tackling datasets of a modest size.
DATA SCIENCE MATURITY SCALE

*Maturity level 3*

Development/Pre-production: Data science output has been rolled out/validated across multiple domains/problems.

SUMMARY

In this manuscript, we pursue the automatic detection of fake news reporting on the Syrian war using machine learning and meta-learning. Our model is based on a suite of features that include a given article’s linguistic style, its level of subjectivity, sensationalism and sectarianism, the strength of its attribution, as well as its consistency with other news articles from the same “media camp” and reporting on the same single large-scale event detected from a timeline of the Syrian war. To train our model, we use FA-KES, a fake news dataset about the Syrian war, consisting of 804 articles roughly balanced between *true* and *fake*. We explore a suite of basic machine learning models as well as the model-agnostic meta-learning algorithm (MAML) suitable for few-shot learning, using datasets of a modest size. Our feature importance analysis confirms that our collection of features specific to the Syrian war are indeed very important predictors for the output label. Our meta-learning model achieves the best performance improving upon the baseline approaches that are trained exclusively on text features in FA-KES in about 20% inaccuracy, 15% in F1-measure, and 30% in AUC. It also achieves more robust probabilistic outcomes and higher top $k$ precision and recall than the vanilla machine learning models.

INTRODUCTION

As world and regional powers get more embroiled in the Syrian war, serious questions arise surrounding the credibility of news documenting the facts of war in the decade long conflict. The
spread of fake news around documentation of the war, going beyond plain news bias, compromises not only the integrity of the actual reporting, but can contribute to psychological warfare that drives the exodus and constant mobility of refugees, and hampers humanitarian planning for delivering aid to distraught communities. Detecting fake news is a difficult task for humans and automated technological tools alike. Most of the current literature around automatic fake news detection focuses on US political news, entertainment news or satire articles, and to the best of our knowledge, no such coverage exists for news reporting around conflicts in the Middle East. Despite that ground truth and fact checking repositories exist around which one can investigate certain claims made on news forums, the process by which one can extract aggregated knowledge from those repositories is tedious and expensive for the average, non-tech savvy reader. In the absence of specialized content management systems that can sift through this information in order to reveal reliable war-related information, an alternative would be to pursue an automatic fake news detection mechanism with the help of labeled dataset. In this manuscript, we propose a meta-learning approach towards the automatic detection of fake news arising from the Syrian war, one that does not make any assumptions on the reputation of news sources, in an attempt to render the process as objective as possible. In building our machine learning model, we tap into FA-KES, the first publicly available fake news dataset around the Syrian war (Abu Salem et al., 2019). FA-KES was generated from a variety of media outlets with varying inclinations, and consists of 804 English news articles roughly balanced between fake versus true. FA-KES is labeled with the help of a fact checking mechanism against the Violations Data Center (VDC), one of the leading repositories documenting the human burden of the Syrian war. Using this dataset, we propose a set of carefully engineered (input) features that are specific to the content and linguistic style of news articles around the Syrian conflict -- for example, attributes such as inconsistency of a given article with respect to articles from the same media camp, the strength of its quoted sources/attribution, as well as lexica-based features for fake news detection inspired by
works related to fake news detection in the social and computer sciences. Additionally, we build our customized sectarian words lexicon which, to the best of our knowledge, constitutes the first such lexicon associated with armed conflict in the Middle East. Obtaining datasets in this domain and of such a caliber is an extremely challenging process. The relatively smaller data generation processes in armed conflict (e.g. only a handful of leading news media outlets, a few contentious albeit highly impactful events around which fakes news propagate), render the size of fake news datasets generated from news media to be not-so-big, which poses a big challenge for traditional machine learning models that are known to be data hungry. Although transfer (machine) learning has been used to overcome limitations associated with relatively small datasets (Pan et al., 2010), this paradigm employs pre-trained models on data that are restricted to some general knowledge, a process which may fail to generalise when porting the pre-trained model onto datasets with highly particular features. To overcome the limitations associated with both the size of FA-KES and the highly peculiar features associated with news articles around the Syrian conflict, we resort to elements from few-shot learning, specifically, the powerful optimisation based meta-learning approach via the MAML algorithm (Finn et al. 2018), which, in contrast to transfer learning, guides its base model by learning to use subtasks sampled from the same data distribution. We assess our overall approach quantitatively as well as qualitatively, and draw comparisons with baseline approaches to assess both the effectiveness of the engineered features, as well as the modeling paradigm chosen.

METHODS

Data Sources

The FA-KES Dataset

The rationale behind our previous work in (Abu Salem et al., 2019) is anchored in fact checking against any published content documenting casualties in armed conflict. As a prototype of such content, the
dataset derived in (Abu Salem et al, 2019) uses for example, the Syrian Violation Documentation Center (VDC), a non-profit, non-governmental organization instituted in 2011, and registered in Switzerland.

The VDC regularly documents war-related deaths as well as missing and detained people. As stipulated on its website, the VDC adheres to international standards for the documentation of its data. Each record in the VDC database consists of "violations" information relating to the demographics, date, location, cause of death (e.g., type of weapon used), and status of the victim (civilian or non-civilian).

The VDC has been vetted and adopted by a large number of researchers working within the framework of the Lancet commission on Syria. Each VDC database record consists of the following fields: Name of causality, cause of death (e.g., shooting, shelling, chemical weapons), gender and age group, status (civilian or non-civilian), actor (e.g., rebel groups, Russian forces, ISIS), location of death in Syria, and date of death.

The data in the VDC database is only available at an individual level. In order to quantify the burden of the armed conflict using those individual records, the current authors aggregate the data by various keys, such as by "actor", by "type of attack", by "number of civilian casualties" (Abu Salem et al., 2019).

Once the VDC records are aggregated, one is able to identify peaks in the timeline of the war from 2011-2018, corresponding to outliers in the burden which in turn can be associated with some of the most intense episodes in the Syrian war. From those identified peaks, the authors track the events taking place in the corresponding locations and dates, and conclude with a dataset representing major events in the Syrian war and that is verifiable by the VDC. Some sample events generated in the process include the Ghouta chemical attack in August 2013, major offensives against ISIS in July 2015, and the Aleppo offensive in July-August 2016. We refer the reader to Table 1 from (Abu Salem et al., 2019) for more of the major events driving the news coverage and associated scraping leading to FA-KES.

After extracting the corresponding (violations) information from the scraped news articles that can be mapped to the VDC content, the authors continue by mapping each news article to its closest VDC
event, and quantifying how accurately the article reports on the casualties compared to the truth from VDC. The resulting metric representing proximity to the VDC events dataset on multiple dimensions guides the process by which an article is labeled fake or true. Particularly, using unsupervised learning, the scraped dataset is split into two clusters based on how close they match the information in VDC with regards to the list of questions shown below. The cluster centroids yield sufficiently clear insight into the label of each cluster. This process resulted in a dataset consisting of 426 true articles and 378 fake articles, covering a wide spectrum of media outlets representing mobilisation press, loyalist press, and diverse print media. Questions used to map news articles to the VDC are shown below:

- What was the date (day, month and year) of the event reported in the article?
- What was the location of the event reported?
- What was the main cause of death associated with this event (cause_of_death)?
- How many civilians died in the event reported (nb_civilians)?
- How many children died (nb_children)?
- How many women died (nb_women)?
- How many non-civilians died (nb_civilians)?
- Who did the article blame for the casualties (actor)?

The responses to each of those questions in every article from FA-KES appear in this crowd-sourced annotations of FA-KES. Hereafter, we refer to this annotated version as A, where each article in A is tagged by a date-location feature as well as the six-feature tuple described above.

In the next sections, we describe the process to engineer features around articles in FA-KES, and the methods underlying our machine and meta-learning models to predict the labels in FA-KES, as a function of those features.

**Feature Engineering**

Our fake news detection model is driven by the intuitive hypothesis that the linguistic style of a news
article can shed light on whether it is likely to be fake or not. Also, if several articles coming from the same media camp exhibit poor consistency among them in reporting on a single, controversial event, this could signal greater diversion from the truth. Our hypothesis on what features contribute to a fake news article is inspired by the wealth of work on fake news detection in both the social science and computer science communities (see (Globeck et al., 2018; Hassan et al., 2017; Mukherjee and Weikum, 2015; Niculue et al., 2015; Popat, Strötgen, and Weikum, 2016; Prat and Strömberg, 2013; Rashking et al., 2017; Shu et al., 2018; Torabi and Taboada, 2018; Yang Wang 2017), to name a few, and we refer the reader to Sec. 1 of Supplementary Information for a wide survey of related work. Hereafter, we describe our suite of content-based and linguistic features.

(Content-based) inconsistency score: This feature measures the extent to which facts reported by a given article are consistent with respect to a suite of other articles belonging to the same media camp. Here, consistency reflects some sort of proximity among articles with respect to a body of information representing a large scale ``event” that those articles are reporting on. Such events can be identified through media outlets, social media, or NGOs tracking human rights violations, but can also be inferred from spikes in casualties observed in a certain timestamp and location, as reflected in fact checking repositories like the VDC. In the present manuscript, we demonstrate the calculation of the inconsistency score by referring to peaks in events as evident from the VDC, though again, this process can be easily adapted to any other source for around the armed conflict in question. We begin by grouping media sources into the following categories: pro-Syrian regime, against Syrian regime, and neutral. Table 1 in the Supplementary Information reveals how media outlets considered in FA-KES were split according to these three categories. After aggregating individual records from the VDC by joining on the `date’ and `location’ field, we identify peaks representing major events detected via all the relevant individual records in the VDC, leading to a collection of events that we label E. The date-location feature pair associated with each such event will be used for matching articles to events. The remainder
of the information that we manually extracted via crowdsourcing in response to the questions used to map articles to the VDC described in the previous section will be used for calculations that will yield a notion of proximity. From the above section, recall that the six pieces of information by which articles from FA-KES were mapped to the VDC were (cause_of_death, nb_civilans, nb_children, nb_women and nb_noncivilians, actor), yielding an annotated version of FA-KES which we previously denoted by $A$.

Given an arbitrary article $a$ in $A$, we now aim to retrieve all other articles $a'$ in $A$ that are from the same media camp and report on the same event, and to establish an inconsistency score between them. This requires a two-step process as follows:

1. Two articles $a$ and $a'$ are said to match if both can be traced to the same event from $E$. To match any article to an event, we rely on its date-location feature pair in a process borrowed from (Abu Salem et al., 2019). Given an article $a$ with date-location feature pair denoted by $(d_a, loc_a)$, and an integer $w$ representing an offset (window) of days, we retrieve all the events $e$ from the aggregated VDC data with date-location index $(d_e, loc_e)$ satisfying $loc_e = loc_a$ and $d_a - w \leq d_e \leq d_a + w$. The rationale behind accounting for a window of days around the time of the event is to be able to account for delays in reporting on any specific casualty (either a few-day delay in media reporting or a few days in documenting a casualty in the VDC). In (Abu Salem et al., 2019) we observe that the best window after which no improvements were noticed in the total number of matches was $w = 4$, and that even for a value as small as $w = 1$, only about 10% of the articles collected had no matching VDC events, which we ended up excluding from FA-KES altogether. Thereafter, we refer to the set of all matching articles $a'$ to $a$ as $A'$.

2. To determine the inconsistency score between $a$ and all matching articles $a'$ in $A'$, we rely on the notion of proximity between their coordinates in the six-feature tuple, thus leading to the idea of the distance $d(a, a')$ between the two vectors representing the two articles, as defined in (Abu Salem et al., 2019). As the six-feature tuple consists of a mix of numerical and categorical
types, we rely on the Gower's distance to compute \( d(a, a') \). The first two coordinates of \( d(a, a') \) consist of binary features set to 1, if \( a \) and \( a' \) agree on the cause of death and actor, and 0 otherwise. The rest of the coordinates of \( d(a, a') \) capturing consistency in the real-valued features representing the number of casualties reported in the articles is calculated as follows: If \( n_a \) and \( n_{a'} \) denote the number of casualties (in women, children, civilians, or non-civilians), then the corresponding coordinates in \( d(a, a') \) are set to \( (n_a - n_{a'})/n_a \).

Once \( d(a, a') \) is computed for all \( a' \) in \( A' \), the inconsistency score feature value of \( a \) is now finally computed as the average \( \text{AVE} \{d(a, a')\}_{a' \in A'} \). Note that the lower the Gower distance is, the more consistent an article is considered to be.

*(Content-based) quoted sources/attribution*: This feature measures the strength of source attribution in the articles. When reporting a certain event, articles should quote a source for their claims. For instance, a large number of articles were found to contain `a quote from local media sources', `a source claimed', `activists say', etc. These cannot be considered as actual attribution to a specific person or an organization, and weaken the associated source attribution signal. An article is deemed more credible when it features a clear quotation/attribution of sources of the information reported. Accordingly, we adopted the following spectrum of values on a scale from 0 to 1 for the `source attribution' feature:

- **0**: no source is mentioned at all; the article just describes the event and does not quote any sources for any of the information it reports.
- **0.5**: no real attribution is observed such as article quotes `sources', `activists', without naming the source/activists (e.g. local media sources, a source in the army).
- **1**: real attribution (named source) is observed; article specifies the name of the organization as a source or the name of the person acting as its source (e.g. WHO organization, name of the activist).

To extract this feature from each article, we proceeded as follows. We first used the report verbs lexicon
developed by Mukherjee and Weikum (2015) in order to extract the sentences that contained report verbs. Using the Stanford Dependency Parser, we determined the subject of the verb (e.g., if the wording was along the lines of `activists say', its subject is `activists'). Once we identified the subject of the report verb:

- We used Stanford's Named Entity Recognizer (NER) to classify the subject of the report verb as being 'Organization', 'Person', 'Location', or 'Other'. Based on the output of Stanford's NER, we were able to determine whether each article was revealing a real attribution (quoting an organization or a person).
- If the source was classified as 'other', we considered the `source attribution' feature to be 0.5 (no real attribution).

(Linguistic) sectarian words: We define the sectarian language feature of a particular article to be the frequency of sectarian words in a given article. An article is deemed less credible if it exhibits excessively sectarian discourse. To this end, we seek to build a sectarian language lexicon faithful to the intricacies of such discourse in the Middle East. Whilst building our lexicon, we follow closely the methodology adopted by the Linguistic Inquiry and Word Count project (LIWC) as appears in (Buchanan, Westbury, and Burgess, 2001; Ching and Pennebaker, 2012; Pennebaker et al., 2015). To illustrate, we begin by manually selecting twenty samples from the following two types of documents: (1) media articles reporting about the Syrian war, entrenched in a sectarian tone and/or analysing the Syrian conflict from a sectarian/religious lens, and (2) documentary reports delving into the origins of the various sects in Syria, their religious dogma, and important key figures that the community emulates or eulogizes, and geographical locations where followers and influencers of those sects are found. We refer the reader to Sec. 2.1 of Supplemental Methods for samples of articles we obtained for the purpose of building this lexicon. They incorporate elements from theological/political discourse arising in various and frequently antagonistic ideologies prevalent among people of the Middle East, reflecting a schism that got more
pronounced during the Syrian war. Thereafter, the lexicon is built by manually extracting the following types of words: adjectives (Sunni, Shiite), labels (Wahhabism, Sufism, Wilayat al-Faqih i.e. Guardianship of the Islamic Jurist), sectarian labels when reporting war events (terrorists [religious], takfiris i.e. excommunicational, militias [religious], pilgrims [religious]), in addition to terms reserved exclusively in reference to various ethnic minorities in the region, for example, terms associated with Kurdophobia. After manually extracting this suite of words from our selection of twenty articles, we iterate over all their possible forms by referring to nouns, their associated adjectives, and verbs, if applicable. Using this process, we were able to construct our sectarian language lexicon using 242 words. In Sec. 2.2 of Supplementary Information we present sample sentences extracted from our dataset that incorporate sectarian words as revealed by our lexicon.

Once obtained, we used the sectarian lexicon to extract the sectarian feature as follows. For each article in our dataset, we calculated the frequency of the sectarian words by dividing the number of sectarian words in the article by the total number of words of the article.

Features from existing lexica: An article is deemed more credible if written in an objective and unbiased language, avoiding sensationalism. A number of exhaustive lexica arising in other works around fake news cover a wide array of words that reflect such attributes (Mukherjee and Weikum, 2015), from which we name the following:

- **Assertive verbs** capture the degree of certainty to which a proposition holds, e.g., `guess', `seem', `appear', `figure', `believe', `expect', `claim', `report', `suspect', etc.
- **Factive verbs** presuppose the truth of a proposition in a sentence, e.g., `note', `notice', `observe', `know', `reveal', `acknowledge', etc.
- **Hedges** soften the degree of commitment to a proposition, e.g., `almost', `apparently', `appear', `around', `likely', `suspected', etc.
● **Implicative verbs** trigger presupposition in an utterance. For example, usage of the word ‘complicit’ indicates participation in an activity in an unlawful way. Examples of these words include: ‘bother’, ‘dare’, ‘neglect’, ‘force’, ‘fail’, ‘cause’, ‘support’, ‘allow’, etc.

● **Report verbs** emphasize the attitude towards the source of the information, e.g., ‘accuse’, ‘acknowledge’, ‘add’, ‘admit’, ‘advise’, ‘agree’, ‘declare’, ‘deny’, etc.

● Subjectivity and bias captured via positively and negatively opinionated words, that reflect subjective clues and bias-inducing words. This attribute reflects to what extent a given article is objective, rather than rooted in its author’s own prejudices. Examples of these words include: ‘abusive’, ‘bad’, ‘abuse’, ‘achievement’, ‘absurd’, ‘devastate’, ‘assault’, etc.

For each article in FA-KES, we create a feature representing the frequency by which a given linguistic attribute from the above list appears in the entire article. In Sec. 2.2 till 2.7 of Supplementary Information, we present sample excerpts from articles in FA-KES that demonstrate each of those linguistic features.

**Feature Selection**

Feature selection refers to a suite of statistical and machine learning based techniques that seek to identify those input features in a dataset that contribute mostly to the target variable or class (Zhen and Casari, 2018, Kuhn and Johnson, 2019). We choose the `SelectKBest` function from the `sklearn.feature_selection` module in Python which implements the univariate feature selection method, a filter-based method that invokes a dependence score between each input variable and the target variable or class. We set the dependence score to be the normalized mutual information index, which measures how much information is communicated, on average, in one of our input variable about the target variable or class.
Meta-learning

Machine learning systems require far more data to achieve the same level of learning developed by humans. It is computationally intractable to develop machine learning algorithms that emulate Bayesian inference. As a result, ML models consistently face the challenge of generalisation when presented with a new task, and no amount of training data can suffice to reflect the true data generation process in real life. Few-shot learning (FSL) aims to build powerful predictive models with improved generalization capabilities using training datasets with limited information. In the process, it helps end users reduce their operational costs required for training and serving time, data collection, and data preparation. FSL is implemented using a variety of approaches. Meta-learning is a well-established branch of FSL that uses prior knowledge of data such as its structure and variability, to enable construction of powerful models from a few examples, by constraining the learning algorithm to choose parameters that improve generalisation from a few examples. These methods are optimization based: they begin learning the initialisation of a network, and thereafter, proceed to fine tune this initialisation of weights at test time on a new task. While tuning, one optimises by differentiating as the learner falls back on a sensible gradient-based learning algorithm, even when the algorithm receives out-of-sample data. This is in contrast to conventional approaches where tasks are often learned from scratch, and a model’s initial weights are not optimized in any manner. Meta-learning algorithms have a better head-start than traditional models, learning quickly thereafter, and as such, reflect an ability to “learn to learn”.

The Model-agnostic meta-learning (MAML) algorithm by Finn et al. (2018) is an optimization based algorithm that can be applied to any model incorporating a gradient descent optimization strategy, such that a small number of gradient updates will achieve fast learning on a new task. A number of tasks $T_i$ is initially sampled from the same distribution $p(T)$. The model is trained using $K$ samples from $T_i$ called the support set, then validated on a subset from $T_i$, called the query set. The optimization problem in this set-up is as follows: given a model $f_\theta$, find a set of model parameters $\theta$ (randomly initialised) such that for each $T_i$, the model will reduce its loss with stochastic optimization such as stochastic gradient
descent (SGD) or adaptive moment estimation (Adam). The process is captured by computing fine-tuned parameters $\theta_i$, of $\theta$ after adapting to $K$ sampled points from $T_i$, where $K$ tends to be a small value. A common loss function for classification is the cross-entropy loss or log loss. The final overall loss of the metalearner is obtained through two steps: first, one implements an initial pass of updates on $\theta$ by minimising the loss using one (or more) gradient descent operators during the training phase in each $T_i$, represented by $\theta'_i \leftarrow \theta - \alpha \Delta \theta L_{T_i}(f_\theta)$, where $\alpha$ is a step size parameter in the formula of gradient descent. The resulting model with the updated parameters is now applied on the test samples from each $T_i$, where the loss function is now also minimised using stochastic gradient descent to return the ultimate fine-tuning of $\theta$ using the formula: $\theta \leftarrow \theta - \beta \Delta \theta \sum_{T_i} L_{T_i}(f_{\theta'_i})$. As this optimization is performed over the initial set of model parameters $\theta$ whilst the objective is obtained through the updated model parameters $\theta'_i$, it is referred to as a meta-optimization (or meta-update) and is referred to as the meta step size. This process re-iterates $n$ times, with a new set of tasks sampled in each iteration, until one converges towards an acceptable performance threshold. We refer the reader to Alg. 2 in (Finn et al., 2018) for a complete summary of the MAML algorithm, and to our Code and Datasets section below for a link to our meta-learning code.

The original MAML framework of (Finn et al., 2018) assumes a random sampling of tasks, and as a result, assumes the transferable knowledge of parameters weights to be globally shared across all tasks, rendered possible via the assumption that randomly sampled tasks come from the same distributions. When this assumption is violated and tasks are heterogeneous, one can employ automated relational meta-learning (ARML) to balance knowledge generalisation among closely related tasks and knowledge customisation to different clusters, tailoring the transferable knowledge to different clusters of tasks (Yao et al., 2020)
Baseline Approaches

Baseline approaches relative to a given experimental setup encompass all other models (model centric) or feature settings (data centric) that reflect a less skilled setting than originally proposed. In our case, a model centric baseline performance can be generated using a suite of vanilla machine learning models, such as the decision tree, logistic regression, ridge regression, the SGD classifier, extra trees, random forest, AdaBoost, gradient boosting, XGBoost, linear support vector classifier (SVC), Nu-SVC, SVC, and naïve Bayes. For a data-centric baseline performance we fit a suite of end-to-end models that rely exclusively on the text features of the articles, rather than any of the features we engineered. In addition to training our own best performing meta-learner and vanilla machine learning models using only textual features from FA-KES, we also fit text-based deep learning baselines using feedforward (FNN), convolutional (CNN), and long-short term memory (LSTM) Neural Networks, employing Glove word embeddings (300 dimensions) as features (Pennington, Socher, and Manning, 2014). Finally, we also fit a text-based BERT model (Devlin et al., 2018).

Performance Metrics

To assess the performance of our models, we refer to standard metrics for binary classification such as overall accuracy, precision and recall, their harmonic mean given by the F1-score, as well as the area under the curve of the ROC (AUC). For a more thorough qualitative assessment of the best performing accurate models, we analyse the robustness and goodness of the probabilistic score that a given article is true, which can be achieved by analysing the behaviour of those best performing models in terms top \( k \) precision and recall. The rationale behind this assessment is as follows. For the average reader with a limited amount of time to dispense on reading the news, an algorithm is considered useful if it is able to adapt accurately to the reader’s availability to read only the top \( k \) articles deemed highly probably true. Those estimates need to be robust for a model to be deemed useful. Additionally, the higher the precision and recall are for a specific \( k \), and the larger the value of \( k \) for which precision and recall
remain also high, the more useful an algorithm is. In addition to producing crisp binary class labels, all of the machine and meta-learning models used in this manuscript are capable of producing such probability estimates, hereby dubbed *risk scores*, and we refer the reader to (Breiman 2001; Chawla and Cieslak, 2006; Lakkaraju et al., 2015; Niculescu-Mizil and Caruana, 2005) for details on how those estimates are produced for our suite of machine learning models. Additionally, the feed forward neural network used at the base of our meta-learning model produces output in its last layer which is normalised by the softmax function, turning this numerical output into raw probabilities for landing in any one of the two classes.

*Empirical risk curves* can be used to assess the accuracy of the probabilistic estimates, and thus convey information about the robustness of algorithms. We consider the approach of (Lakkaraju, 2015) in the construction of such curves as follows. We begin by ranking the articles in descending order of their probability estimates of landing in the positive class. We then discretise the ranges of those rankings as follows. By creating say, ten bins, we assign each article to the $i$’th bin if and only if the article ranks between the $10.\ i$’th and $10.\ (i + 1)$’st percentiles. For each bin $i$, we now track the mean empirical estimate, capturing the fraction of articles in bin $i$ whose ground truth label is indeed the positive class. Thus, for an algorithm to possess good quality, accurate probability estimates, it should possess the property that the higher the index of a bin, the higher the mean empirical risk. Equivalently, one can plot the *empirical risk curve*, defined as the two-dimensional plot mapping each bin index $i$ to its mean empirical estimate, and aim for models that exhibit a monotonically non-decreasing empirical risk curve.

The performance of the various models when predicting the top $k$ articles having highest probabilities of landing in the positive class can be captured using top $k$ precision and top $k$ recall curves. To produce those curves, we begin by ranking all articles using their respective probability estimates, and calculate the corresponding precision and recall for the top $k$ such articles. Two observations related to model
performance at this level can be made: (1) for a fixed \( k \), a model is more useful than another if it has higher precision and recall at this specific \( k \). (2) We would prefer models to maintain higher values for precision and recall as \( k \) increases.

**EXPERIMENTAL PROCEDURES**

We ran our experiments on the High Performance Computing Cluster at the American University of Beirut. The system comprises an architecture of Intel/AMD, with a SLURM scheduler, and contains twelve hosts with 16 vCPUs each with 64 GB RAM. Training of the machine learning models took on the order of 0.1-0.2 minutes for all of the models. The meta-learning experiments required about 25 minutes to conclude.

**Meta-learning experimental setup**

The original dataset is split into 80% for meta-learning and 20% for testing, in such a way that the proportion of positive to negative class labels remains the same in all splits of the data (stratified sampling). The sampling was stratified to ensure that The meta-learning experimental setup on the training dataset is inspired by the Finn et al.’s code in this link. The hyper-parameters we tune are the number of iterations \( n \), number of tasks \( N \), number \( K \) of samples per task, inner learning rate \( \alpha \), meta-learning rate \( \beta \), number of gradients steps to execute in both the inner learning step as well as the meta-learning step. Tasks are randomly selected without replacement, and the gradient descent steps are guided by the cross entropy loss suited for classification problems. We choose the feed forward neural network to the base model for our meta-learner. In the hyper-parameter search for the FFN, we search over the number of hidden layers and the number of nodes in each layer, and the activation function. The “winning” hyper-parameters are those that correspond to the meta-learner with the best accuracy reported on the testing dataset, or those with default values whose efficacy is already
validated by the github implementation of Finn et al., 2018. The chosen hyper-parameters for our experiments are shown in Sec. 3.1 of Supplementary Information.

Baseline models

To train our machine learning models, we adopted the same (80%-20%) stratified splits used in the meta-learning phase. To tune the hyper-parameters of the different models, we performed a grid search using ten times repeated 10-fold cross-validation on the training set. In the following models we applied regularization to reduce overfitting: decision tree, SVC, linear SVC, logistic regression, ridge regression, the SGD classifier, extra trees, and random forest.

For the baseline experiments where we train exclusively on the textual features of FA-KES, a suitable set of hyper-parameters for each of the deep learning models was determined after several tuning attempts that were driven by the shape of the learning curve. The batch size and decay chosen for all the models were 64 and 0 respectively. The dropout was 0.2 in the LSTM layer of the LSTM model. The learning rate chosen was 0.001 for the LSTM, 0.0001 for the FNN, and 0.001 for the CNN. The total number of layers chosen were 4 for the LSTM, 5 for the FNN, and 10 for the CNN. The BERT model consisted of an Input layer, input masks and segment IDs (needed for BERT layer), BERT layer, a ReLu layer, and a sigmoid output layer. It was trained for a hundred epochs with a batch size of 32.

In all of our experiments, we avoided hyper-parameters that yielded validation error metrics which were less than training errors, or erratic spikes in the learning curves, and incorporated L2 regularization and dropout when needed, in the following algorithms: decision tree, linear svc, logistic regression, ridge regression, the SGD classifier, extra trees, random forest, gradient boost, and XGBoost.

Sensitivity Analysis to Important Features

To corroborate the results of feature importance analysis, and validate our hypothesis that the newly introduced features specific to the discourse around the Syrian war (sectarian language, strength of
quoted sources or attribution, and inconsistency score) introduce impactful knowledge to the fake news detection problem, we consider training our best performing model using subsets of the proposed features. The combinations presented below aim to explore our best model using the absence of those three features. Specifically, Scenarios 2 and 3 explore how necessary those three features are.

- **Scenario 1: All features:**
  - Sectarian words frequency
  - Strength of quoted sources/attribution
  - Inconsistency score
  - Linguistic based features: assertive verbs, factive verbs, hedges, implicative verbs, report verbs, subjectivity and bias.

- **Scenario 2: All features minus**
  - Sectarian words frequency

- **Scenario 3: All features minus**
  - Sectarian words frequency
  - Strength of quoted sources/attribution
  - Inconsistency score.

**RESULTS**

**Feature Exploration and Ranking**

An exploration of the content-based and linguistic features we engineered around each article in FA-KES revealed the following distinctive patterns:

- **Sectarian words frequency:** for the articles labeled *fake*, 95.2% exhibited sectarian language frequencies between 0 and 0.08 and 4.8% exhibited sectarian language frequencies between 0.07 and 0.15. For articles labeled *true*, 97.9% exhibit sectarian language frequencies between 0
and 0.04 and 2.1% exhibit sectarian language frequencies between 0.05 and 0.13. Thus, the majority of *true* articles had a smaller upper bound frequency of sectarian words than the majority of *false* articles, indicating lesser reference to sectarian discourse.

- Implicative verbs frequencies: for the articles labeled *fake*, 97.7% exhibited implicative verbs frequencies between 0 and 0.01 and 2.3% exhibited implicative verbs frequencies between 0.01 and 0.02. For articles labeled *true*, 88.9% exhibited implicative verbs frequencies between 0 and 0.008 and 11.1% exhibited implicative verbs’ frequencies between 0.008 and 0.016. Thus, the majority of *true* articles had a smaller upper bound frequency of implicative verbs than the majority of *false* articles, indicating lesser presuppositions.

- Subjectivity and bias words frequencies: for the articles labeled *fake*, 32.5% exhibited bias frequencies between 0 and 0.25 and 67.5% exhibited bias frequencies between 0.25 and 0.45. For articles labeled *true*, 53.9% exhibited bias frequencies between 0 and 0.25 and 46.1% exhibited bias frequencies between 0.25 and 0.45. Thus, a higher percentage of articles labeled *true* featured lower subjectivity and bias words frequencies than those labeled *false*, indicating lesser subjectivity and bias.

- Quoted source attribution strength: for the articles labeled *fake*, 13.8% exhibited source attribution values between 0 and 0.4 and 86% exhibited source attribution values between 0.4 and 1. For articles labeled *true*, 31% exhibited source attribution values between 0.5 and 0.8 and 69% exhibited source attribution values between 0.8 and 1. Thus, a larger percentage of *true* articles have a significantly higher quote source attribution strength than *fake* articles.

Many of these observations are corroborated by our feature selection analysis. Particularly, our newly engineered features consisting of frequency of sectarian words, strength of quoted source attribution, and inconsistency score, achieved the highest normalized mutual information index with the target
Quantitative performance metrics

In Table 1, we present the performance metrics of all machine learning models as well as the MAML algorithm on the holdout testing dataset using all of our proposed features. We refer the reader to Sec. 3.2 of Supplementary Information for a list of winning hyper-parameters associated with the best performing models. In Table 2, we present the performance metrics of the text-based models on the holdout testing dataset. In Table 3, we present the performance metrics after retraining our best performing model, the MAML algorithm, using various combinations of input features, to corroborate the importance of the three newly introduced features. The various scenarios presented in Table 3 can be found in the Experimental Procedures section.

| Model                    | Accuracy | Precision | Recall | F1-measure | AUC ROC |
|--------------------------|----------|-----------|--------|------------|---------|
| MAML                     | 0.89205  | 0.82955   | 0.98333| 0.89026    | 0.90909 |
| XGBoost                  | 0.88199  | 0.84444   | 0.93827| 0.88889    | 0.88164 |
| Logistic Regression      | 0.8323   | 0.76471   | 0.96296| 0.85246    | 0.83148 |
| Naive Bayes              | 0.8323   | 0.75472   | 0.98765| 0.85561    | 0.83133 |
| Model         | Accuracy | Precision | Recall | F1-measure | AUC ROC |
|---------------|----------|-----------|--------|------------|---------|
| Ridge Regression | 0.81988  | 0.74528   | 0.97531| 0.84492    | 0.8189  |
| Nu SVC        | 0.8323   | 0.75472   | 0.98765| 0.85561    | 0.83133 |
| SVC           | 0.81988  | 0.79545   | 0.8642 | 0.8284     | 0.8196  |
| AdaBoost      | 0.80745  | 0.73148   | 0.97531| 0.83598    | 0.8064  |
| Linear SVC    | 0.78882  | 0.79012   | 0.79012| 0.79012    | 0.78881 |
| SGD           | 0.78261  | 0.69828   | 1      | 0.82234    | 0.78125 |
| Extra trees   | 0.50311  | 0.50311   | 1      | 0.66942    | 0.5     |
| Random Forest | 0.50311  | 0.50311   | 1      | 0.66942    | 0.5     |
| Decision Tree | 0.50311  | 0.50311   | 1      | 0.66942    | 0.5     |
| Gradient Boost| 0.50311  | 0.50311   | 1      | 0.66942    | 0.5     |

Table 2. Performance metrics of text-based models on the holdout testing dataset.
From the above metrics we identify a substantial loss in accuracy, precision, F1-measure, and AUC ROC, when compared to models trained using our suite of engineered features.

Table 3. Performance metrics of MAML on the holdout testing dataset after training on a select combination of input features

| Features Combination | Accuracy | Precision | Recall | F1-measure | AUC ROC |
|----------------------|----------|-----------|--------|------------|---------|
| Scenario 1           | 0.89205  | 0.82955   | 0.98333| 0.89026    | 0.90909 |
| Scenario 2           | 0.65340  | 0.57142   | 0.94545| 0.70440    | 0.68333 |
| Scenario 3           | 0.44886  | 0.42889   | 0.71060| 0.52268    | 0.48030 |

From the above metrics we observe that the newly introduced features are necessary (training without those features causes a significant deterioration in the performance of the model bringing it to sufficiently lower than baseline performance by text-based models as well as the MAML algorithm trained on all the proposed features).

Qualitative Performance Metrics

Below we demonstrate the quality of our best performing models chosen to be the XGBoost and the MAML algorithm. Fig. 1 reveals the empirical risk curves for ten bins, where both algorithms maintain relatively monotonic non-decreasing trends, with the empirical risk curve for MAML however looking to be generally more robust and consistent. Figs. 3 and 4 reveal precision and recall curves at top \( k \). In both figures, MAML maintains higher precision and recall as \( k \) increases. Also, MAML reveals higher precision and recall for smaller values of \( k \), corresponding to improved performance when resources are constrained (e.g. lay-reader has a limited time to browse the news and requires a model that is more
accurate in identifying the top ranked articles deemed to be *true*). Fig 4. shows the ROC curves of both algorithms, where both curves maintain near ideal behaviour approaching the upper left corner (featuring 100% sensitivity and 100% specificity). Generally speaking, whilst both XGBoost and MAML reveal robust performance and usefulness in their predictions, the MAML algorithm seems to outperform XGBoost in the goodness of its scores and its ability to rank articles correctly.

**Figure 1. Empirical Risk Curves for best performing models.**

Better algorithms exhibit monotonically non-decreasing and higher rate of increase. MAML curve is more robust especially at lower valued and higher valued bins, than that of the baseline machine learning model.
Figure 2. Precision at top $k$ for best performing models.

Better algorithms retain higher precision for increasing $k$. MAML curve is consistently higher than that of the baseline machine learning model.

Figure 3. Recall at top $k$ for best performing models.

Better algorithms retain higher recall for increasing $k$. MAML curve is consistently higher than that of the baseline machine learning model.

Figure 4. ROC curves of best performing models.

Both algorithms have ROC curves approaching the upper left corner.
DISCUSSION

Our work constitutes some of the first steps towards automatic fake news detection in the context of the Syrian war with the help of highly customised features constructed over the labeled FA-KES dataset of fake news from Abu Salem et al., 2019. In addition to employing techniques from few-shot learning that improve the generalisation of learning from small datasets, our work reveals the importance of features that are specific to the Syrian conflict in identifying true from fake articles labeled in FA-KES. Of independent use is our sectarian language lexicon which covers a significant part of the sectarian discourse used in reporting not only around the war in Syria but also the remainder of conflicts in the Middle East. Whilst some of the features ingested in our models are common to the existing literature on automatic fake news detection, our feature importance analysis confirms that the three most important features are those that are newly introduced in this manuscript, and which capture intricacies of political discourse around the Syrian conflict (sectarian language, inconsistency scores among articles from the same media camp, and strength of quoted sources and attribution).

Our initial modeling using all of the proposed features reveal that that the model agnostic meta-learning model (MAML) achieved the best quantitative performance improving upon the baseline approaches that are trained exclusively on text features in FA-KES with about 20% improvement in accuracy, 15% in F1-measure, and 30% in AUC. Despite that the XGBoost achieved comparable quantitative results, the quality of its predictions lagged behind that of MAML. For example, our MAML model trained on all of the proposed features showed good quality predictions, and observation revealed through a more robust empirical risk curve and higher top k precision and recall. This means that the probabilistic outcomes of MAML as to the likelihood of an article being true are consistently more accurate than those of XGBoost, and that MAML is able to achieve higher precision and recall when predicting the top k articles with the highest probability of being true. As vanilla machine learning models are not
customized to learn from small datasets, and are tuned using conventional cross-validation schemes where tasks are often learned from scratch, and initial weights of a model are not optimized in any manner, several of the machine learning models we trained were overfitting, consequently requiring substantial regularization. Our meta-learning approach provides evidence that sufficient learning was achieved despite the modest dataset size of FA-KES. With that said, many of the models we fitted using our proposed features had an extremely high recall, capturing most of the fake news in the testing dataset as opposed to missing out on them as in the case of the baseline textual based models.

Several observations validate our feature importance results. Some of the intriguing observations we recall from our exploration reveal the following: (1) the majority of true articles had a smaller upper bound frequency of sectarian words than the majority of false articles, indicating lesser reference to sectarian discourse, (2) a higher percentage of articles labeled true featured lower subjectivity and bias words frequencies than those labeled false, indicating lesser subjectivity and bias, and (3), a larger percentage of true articles have a significantly higher quote source attribution strength than fake articles.

By modeling various scenarios of model fitting, our feature importance results are further corroborated. When training MAML without sectarian words frequency showed a degradation in performance of about 23 in accuracy, 22% in F1-measure, and 25% in AUC when compared to MAML trained on all the features. Training without sectarian words frequency, strength of quoted sources/attribution, and inconsistency score, resulted in a degradation in performance of about 50% in accuracy, 42% in F1-measure, and 47% in AUC.

Our framework can easily be adapted to treat fake news in armed conflicts in the Middle East, known to incorporate a considerable amount of sectarian discourse, provided ground truth data exists around which one can build a labeled dataset of fake news. There still remain multiple limitations and open
questions before the proposed approach can be fully automated. It is imperative that our approach is
tested on a wider variety of datasets coming from armed conflict: although ideological strife from other
parts of the world would require different lexica from the sectarian lexicon we conceive to be highly
relevant for the Middle East, the notion of inconsistency score and fact checking against reliable
databases documenting the conflict are more likely to remain relevant. Whilst the aspect of our work
concerned with computing the inconsistency score still hinges on manually extracting information from
the news articles to match it against the VDC, our approach can significantly improve by integrating
event extraction techniques that can automate this phase of our work. In fact, a sizeable body in the
literature – see (Hamborg et al., 2018; Reichart and Barzilay, 2012; Wang, 2012), to name a few --
demonstrate various techniques to aid in the automation of extraction of events, the latest of which
incorporate elements from sequence tagging. A number of the current authors of this manuscript have
also recently built a manually sequence tagged version of FA-KES, and used this dataset to train and test
three state-of-the-art, deep learning based, sequence tagging models: BERT, BiLSTM, and a plain
Conditional Random Field (CRF) model, with BERT delivering the best performance (Sawaya et al., 2020).
Skeptics of automatic techniques for fake news detection would champion a more investigative, less AI-
centered approach to the problem. For example, an immediate question that comes to mind is whether
fake news detection can immediately benefit from direct fact checking through repositories like the
VDC. To that end, some of the current authors of this manuscript are exploring the use of knowledge
graphs representations and RDF management systems through which claims around the Syrian war can
be fact-checked. In ongoing work, the authors are modeling conflict casualties’ data extracted from the
VDC as an RDF graph, allowing users to query such data using SPARQL end-point. An interesting open
question remains on whether labels in FA-KES can be fully validated by the use of such an RDF graph,
giving further legitimacy to machine learning approaches to fake news detection, and rendering the
overall machine learning pipeline proposed herein more accessible to the public. Whilst the work
presented in this manuscript attempts to harness the power of highly powerful machine learning models, its limitations are set by the validity of labels present in FA-KES. Certain limitations which can be the subject of future refinement in the methodology behind labeling articles in FA-KES are as follows. There could be many articles that cannot be mapped to information from repositories like the VDC, there could be several events not captured from such repositories, and the mechanism by which articles are mapped to ground truth is deemed too discrete (e.g. failing to capture articles that could be true in some aspects but fake in others). Whilst we acknowledge the need to revisit such limitations, we believe our proposed machine learning pipeline and feature engineering remains invariant under such nuances.

RESOURCE AVAILABILITY

Lead contact

Fatima K. Abu Salem is the lead contact for this manuscript and can be reached at <fa21@aub.edu.lb>

Materials availability

There are no physical materials associated with this study.

Data and Code

1. **FA-KES with features**

2. **FA-KES with crowd-sourced annotations** for questions used to map articles to the VDC.

3. Feature Engineering code:
   a. The [runner script](#) that calls all the functions that construct the features.
   b. The [script](#) for generating the consistency score.
   c. The [script](#) for generating all the lexicon-based features.
   d. The [script](#) for generating the quoted sources feature.
   e. The [sectarian words lexicon](#).
4. Feature Selection Code
5. Machine Learning Code
6. Meta-learning (MAML) Code
7. Qualitative Assessment (empirical risk estimates curves, top k precision and recall curves)

ACKNOWLEDGMENTS

This work is supported by the Collaborative Research Stimulus Grant (CRS) from the American University of Beirut.

AUTHORS CONTRIBUTIONS

Fatima Abu Salem led the work described in this manuscript. She contributed to the design of the methods investigated, the experiments and the manuscript writing. Shady Elbassuoni contributed to the design of the methods, the experiments and the manuscript writing. Mohamad Jaber contributed to the design of the methods and experiments. May Farah contributed by providing a journalistic point of view to the work and helped in verifying the dataset used in this work. Roaa Al Feel and Hiya Ghanem both contributed to the methods, experiments, write-up and implementation of all the proposed methods and experiments.

DECLARATION OF INTERESTS

The authors declare no competing interests.

SUPPLEMENTAL INFORMATION

Document S1.pdf:

- Related Work
- Addendum to Feature Engineering
- Addendum to Experimental Procedures
Addendum to Feature Selection Results

REFERENCES

[Abu Salem et al., 2019] Abu Salem F.K., Al Feel R., Elbassuoni S., Jaber M., Farah M. (2019). FA-KES: A fake news dataset surrounding the Syrian war. In: Proceedings of the 13th International AAAI conference on web and social media.

[Breiman, 2001] Breiman, L. (2001). Random Forests. In: Machine Learning, 45(1): 5—32, 2001.

[Buchanan, Westbury, and Burgess, 2001] Buchanan L., Westbury C., Burgess C. Characterizing semantic space: Neighborhood effects in word recognition. In: Psychonomic Bulletin & Review: 8 (531—44), 2001.

[Chawla and Cieslak, 2006] Chawla N. V. and Cieslak, D. A. (2006). Evaluating probability estimates from decision trees. In: American Association for Artificial Intelligence, 2006.

[Ching and Pennebaker, 2012] Chung C., Pennebaker J. Linguistic Inquiry and Word Count (LIWC): pronounced “Luke”, ... and other useful facts. In: Applied Natural Language Processing: Identification, Investigation, and Resolution, pp. 206—229, 2012.

[Devlin et al., 2018] Devlin J., Chang M.W., Lee K., and Toutanova K.(2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: arXiv preprint arXiv:181004805.

[Finn et al., 2018] Finn, C., Abbeel P., and Levine, S. (2018). Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. ICML 2017, pp. 1126-1135.

[Golbeck et al. 2018] Golbeck J, Mauriello M, Auxier B, Bhanushali KH, Bonk C, Bouzaghrane MA, et al. Fake News vs Satire: A Dataset and Analysis. In: Proceedings of the 10th ACM Conference on Web Science, pp. 17—21, 2018.

[Hamborg et al., 2018] Hamborg F., Lachnit S., Schubotz M., Hepp T., and Gipp B. (2018). Giveme5W: Main Event Retrieval from News Articles by Extraction of the Five Journalistic W Questions. In: Proceedings of the International Conference on Information: 356—366.

[Hassan et al., 2017] Hassan N., Zhang G., Arslan F., Caraballo J., Jimenez D., Gawsane S., et al. ClaimBuster: The First-ever End-to-end Fact-checking System. In: Proceedings of VLDB Endow. 10(12):1945—1948, 2017.

[Kuhn and Johnson, 2019] Kuhn M. and Johnson K. Feature Engineering and Selection: A Practical Approach for Predictive Modeling, CRC Press, 2019.

[Lakkaraju et al., 2015] Lakkaraju, H., Aguiar, E., Shan C., et al. (2015). A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes. In: Proceedings of KDD 2015, 1909—1918.

[Mukherjee and Weikum, 2015] Mukherjee S., Weikum G. Leveraging Joint Interactions for Credibility Analysis in News Communities. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp. 353—362, 2015.

[Niculae et al., 2015] Niculae V., Suen C., Zhang J., Danescu-Niculescu-Mizil C., Leskovec J. QUOTUS: The Structure of Political Media Coverage as Revealed by Quoting Patterns. CoRR. 2015;abs/1504.01383, 2015.

[Niculescu-Mizil and Caruana, 2005] Niculescu-Mizil, A. and Caruana R. (2005). Obtaining calibrated probabilities from boosting. In: UAI, page 413, 2005.

[Pan et al., 2010] Pan S.J., Yang Q., et al. (2010). A Survey on Transfer Learning. In: IEEE Transactions on Knowledge and Data Engineering 22 (10): pp. 1345—1359.

[Pennebaker et al., 2015] Pennebaker J., Booth R., Boyd R., Francis M. Linguistic Inquiry and Word Count: LIWC2015.

[Pennington, Socher, and Manning, 2014] Pennington J., Socher R., and Manning C. (2014). Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 1532—1543.

[Popat, Strötgen, and Weikum, 2016] Popat K., Mukherjee S., Strötgen J, Weikum G. Credibility Assessment of Textual Claims on the Web. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. ACM, pp. 2173—2178, 2016.

[Prat and Strömberg, 2013] Prat A., Strömberg D.S. The political economy of mass media. In: Advances in Economics and Econometrics: Tenth World Congress, Cambridge University Press, 2013.
[Rashkin et al., 2017] Rashkin H., Choi E., Jang J.Y., Volkova S., Choi Y. Truth of varying shades: Analyzing language in fake news and political fact-checking. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2931--2937, 2017.

[Reichart and Barzilay, 2012] Reichart R., and Barzilay R. Multi event extraction guided by global constraints. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: 70--79.

[Sawaya et al., 2020] Sawaya N., Elbassuoni, S., Abu Salem, F. K., and Al Feel, R. (2020). TAGWAR: An annotated Corpus for Sequence Tagging of War Incidents. In: Proceedings of KDIR 2020: 243—250.

[Shu et al., 2018] Shu K., Mahudeswaran D., Wang S., Lee D., Liu H. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. arXiv preprint arXiv:180901286, 2018.

[Torabi and Taboada, 2018] Torabi F. and Taboada M. The Data Challenge in Misinformation Detection: Source Reputation vs. Content Veracity. In: Proceedings of the first workshop on fact extraction and VERification (FEVER), pp. 10--15, 2018.

[Wang, 2012] Wang W. Chinese news event 5W1H semantic elements extraction for event ontology population. In: Proceedings of the 21st International Conference on World Wide Web: 197--202.

[Yang~Wang, 2017] Yang~Wang W. Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 422--426, 2017.

[Yao et al., 2020] Yao, H., Wu, X., Tao, Z., et al. (2020). Automated Relational Meta-learning. In: Proceedings of International Conference on Learning Representations.

[Zheng and Casari, 2018] Zheng A. and Casari, A. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists, O’Reilly publishers, April 2018.
Fake News Detection surrounding the Syrian War: Supplemental Information

Fatima K. Abu Salem1*, Roaa Al Feel1, Shady Elbassuoni1, Hiyam Ghannam1
Mohamad Jaber2, May Farah3

1 Department of Computer Science, American University of Beirut, Beirut, Lebanon
2 Google, Zurich, Switzerland
3 Department of Sociology, Anthropology and Media Studies, American University of Beirut, Beirut, Lebanon

*fa21@aub.edu.lb

1 Related Work

Most of the current literature around automatic fake news detection focuses on US political news, entertainment news or satire articles, and to the best of our knowledge, no such coverage exists for news reporting around conflicts in the Middle East. In [8], the authors build a dataset of news labeled true and fake. To do so, they develop crawlers that crawl fact-checking websites such as Politifact (for political news) and GossipCop (for entertainment news) in order to obtain news content for fake news and true news. In the process, both these websites rely on analysis conducted by journalists and domain experts. The authors also crawl E! online for entertainment news pieces and label all these news as true in their belief that this source is a trusted source. The authors then fit several machine learning models (e.g. SVM, Naive Bayes, CNN) on their dataset using five types of features extracted from the news articles, representing content features, user profiles information, sentiment scores for user posts on the news articles, network features such as the number of followers and followees, obtained likes, retweets, and temporal features such as the time and day of the week the news articles are posted. The model in [8] is trained using a dataset about US politics and entertainment news. It relies on media experts analysis for labeling the dataset and categorizes all articles from E! as true whenever the source is deemed trustworthy. This model employs several types of features such as user posts, user profiles, and network information, which might not always be available for news articles.

In [9], the authors introduce two datasets scraped from the web by leveraging links to news articles referenced by two fact-checking websites (Buzzfeed and Snopes) along with their labels. Both datasets used by [9] are made up of full articles labeled by human experts. These labels are: true, mostly true, mixture of true and false, mostly false, and false. The authors then use text-based features such TF-IDF, n-grams, word vectors, to train text-based classifiers on their dataset for fakes news detection. The authors of [1] build a dataset of fake news and satirical stories, restricted to American politics. The authors then classify each article using a Naive Bayes Multinomial model that is based exclusively on the language adopted by the article. Each article is thus represented as a word vector with a class of fake or satire. The authors then explore the themes portrayed by the content of the articles and manually label each article in the dataset as one or more of the following: hyperbolic position against one person or group, discredit a normally credible source, racist messaging, conspiracy theory. The authors then use these themes as an additional feature to the word vector dataset and
investigate whether this affects the classification outcomes in any way, but no significant improvements in the accuracy or AUC were observed. Both [1] and [9] rely exclusively on the language being used with the help of text-based classification techniques.

Again, in [3] and [5], the authors aim to detect the credibility of claims on the Web. To do so, they rely on stylistic and linguistic features and build lexicons that capture the frequencies of assertive and factive verbs, hedges, implicatives, report verbs, subjective and biased words. The authors in [5] also consider another feature related to the reliability of the source on which the claim is made. To extract this feature, the authors use PageRank, which determines the importance of each Web page by counting the number and quality of links to and from the website.

The authors in [7] present an analytic study on the language of news media in the context of political fact-checking and fake news detection. They compare the language of real news with that of satire, hoaxes, and propaganda, and attempt to find linguistic characteristics of untrustworthy text. Their results reveal that words that can be used to exaggerate a certain claim - subjectives, superlatives, and modal adverbs - appear more prominently in fake news. In contrast, words used to offer concrete figures - comparatives, money, and numbers - appear more frequently in truthful news. The authors also find that trusted sources are more likely to use assertive words and less likely to use hedging words, indicating that they are less vague about describing events. Similarly, the authors observe that trusted sources tend to cite primary sources more often. To assess the effectiveness of such linguistic cues in credibility prediction, they train multiple machine learning models such as an LSTM model, a Maximum Entropy one and Naive Bayes one, on a politifact.com dataset, in order to predict the credibility of claims. Their experiments show that while media fact-checking remains to be an open research question, stylistic cues can help determine the truthfulness of text.

Although our work shares several commonalities with the surveyed related work, we are able to extend several linguistic and content related aspects using features that are specific to the context of the Syrian war and which proved to be very important features. Most importantly as well, our work does not make any assumptions about the trustworthiness of news sources, in an attempt to reduce human bias in this regard.

2 Addendum to Feature Engineering

2.1 Sectarian Words Lexicon

The following are samples of articles obtained for the purpose of building the sectarian words lexicon. They incorporate elements from theological/political discourse arising in various and frequently antagonistic ideologies prevalent among people of the Middle East, reflecting a schism that got more pronounced during the Syrian war.

- Wilayat al-Faqih project aimed at destroying Arabs - Experts
- Iran repopulates Syria with Shia Muslims to help tighten regime’s control
- ‘Takfiri’ Crimes against Islam in Syria Escalate
- You Can’t Understand ISIS If You Don’t Know the History of Wahhabism in Saudi Arabia
- Why so Much Hate? Closer Look at Erdogan’s ‘Kurdophobia’
- The role of Christian militias in Syria goes largely ignored
- Assad’s Shabiha: “Starvation or Submission” to Civilians
Table 1. Categorization of news sources

| Anti-Syrian Regime      | Pro-Syrian Regime    | Neutral   |
|-------------------------|----------------------|-----------|
| Arabiya                 | SANA                 | Reuters   |
| Jordan Times            | Al Alam              |           |
| Al Ahram                | Al Manar             |           |
| Asharq Al Awsat         | Sputnik              |           |
| Lebanese National News Agency | TASS              |           |
| Etilaf                  |                      |           |
| Al Arabyy               |                      |           |
| TRT                     |                      |           |
| Daily Sabah             |                      |           |

2.2 Samples of Sentences Incorporating Sectarian Words

Below are sample sentences extracted from our dataset that incorporate sectarian words as indicated by our sectarian words lexicon:

- Meanwhile in the East a Syrian Kurdish news agency says clashes have erupted again between YPG and Syrian pro-regime *militias* in the northern Syrian city of Hasakeh where two groups have shared control of the city since the early years of the Syrian civil war.

- Damascus: a bomb claimed by the terrorist group Nusra Front tore apart a bus carrying Lebanese *Shi’ite* Muslim *pilgrims*

- The army units meantime thwarted an attempt by the *Takfiri* terrorists to penetrate into a Syrian army base in Jabal al-Rahmaliya region in Northern Lattakia.

- It did not say what was in the barrels but appeared to suggest that some sort of chemical agent was inside and supplied by Saudi Arabia the region’s *Sunni Muslim* power and a staunch supporter of Syria’s Sunni-led revolt

- The strike targeted the mainly *Druze* region of al-Hadr in the Quneitra province on the Golan Heights the Britain-based watchdog added.

2.3 Samples of Sentences Incorporating Assertive Verbs

Following are sample sentences extracted from our dataset that incorporate assertive verbs:

- The attack is *believed* to be among the deadliest in the city for years the British-based Syrian Observatory for Human Rights said

- The death toll is *expected* to rise because of the number of people seriously injured added the Observatory

- Victims of a suspected chemical attack in Syria *appeared* to show symptoms consistent with reaction to a nerve agent the World Health Organization said on Wednesday.

- Syrian rebels *claim* to have repelled a surprise regime attack north of Aleppo and taken prisoners launched after the regime agreed in principle to a UN-brokered six-week ceasefire.
• The Reuters news agency reported that at least 70 pro-regime fighters and more than 80 rebels were killed as rebels countered the attack on Tuesday and Wednesday which was intended to cut supply lines to the city.

• US airstrikes target the IS The Syrian Observatory for Human Rights reported the death of two senior Islamic State group (IS) leaders in a suspected US airstrike on Monday

2.4 Samples of Sentences Incorporating Factive Verbs
The following are sample sentences extracted from our dataset that incorporate factive verbs.

• US officials on Tuesday said they remained skeptical of the IS claims Mueller died in an air strike noting there had been no evidence of civilians at that site before it was targeted.

• Everyone knows Kobane it’s where the Kurds stopped IS

• The BBC quoted a former commanding officer of the British army’s Joint Chemical Biological Radiological Nuclear Regiment Hamish de Bretton-Gordon as saying the footage would be very difficult to stage-manage but that samples taken from the scene would be reveal if chemical weapons had been deployed

• The Syrian government acknowledges for the first time that it has chemical weapons and threatens to use them in the event of military operations by Western countries but not against its own population.

2.5 Samples of Sentences Incorporating Hedges
The following are sample sentences extracted from our dataset that incorporate hedges.

• Observatory director Rami Abdel Rahman said the blasts appeared to be coordinated.

• The WHO said it was likely that some kind of chemical was used in the attack because sufferers had no apparent external injuries and died from a rapid onset of similar symptoms including acute respiratory distress.

• Officials said the military fired dozens of cruise missiles against the base in response to the suspected gas attack in a rebel-held area that Washington has blamed on Assad’s forces.

2.6 Samples of Sentences Incorporating Implicative Verbs
The following are sample sentences extracted from our dataset that incorporate implicative verbs.

• Meanwhile fighting between Islamic State militants and a Kurdish-Arab alliance troops has forced 13000 residents to flee the IS-bastion city of Manbij

• The United States has said the deaths were caused by sarin nerve gas dropped by Syrian aircraft.

• He said the attack was a form of “support for the armed terrorist groups and it is an attempt to weaken the capabilities of the Syrian Arab Army to combat terrorism”.
• 'Restoring security' The official Syrian news agency Sana quoted a military source on Wednesday as saying the army was able to “restore security” to Sanei Tal al-Hawa and Tal-Arous in the southwestern countryside of Damascus Tal Meri west of Damascus and al-Danaji and Tal Antar near Deir al-Adas.

• He had not managed to get a commitment by the rebels to a trial ceasefire before heavy fighting broke out Tuesday.

• I would never allow those IS savages to violate my land.

2.7 Samples of Sentences Incorporating Report Verbs
The following are sample sentences extracted from our dataset that report verbs.

• However subsequent reports have disputed the group’s claims saying that the downed aircraft had in fact belonged to the Syrian regime

• Assad’s government has always denied responsibility for that attack.

• Missile attack on one of its air bases had killed six people and caused extensive damage adding that it would respond by continuing its campaign to “crush terrorism” and restore peace and security to all of Syria.

• President Donald Trump said he ordered missile strikes against an airfield from which a deadly chemical weapons attack was launched this week declaring he acted in America’s “national security interest” against Syrian President Bashar al-Assad.

2.8 Samples of Sentences Incorporating Subjective and Biased Words

• Once a powerhouse of industry Aleppo has been devastated by years of fighting between regime forces and a succession of rebel groups.

• It morphed into a conflict after the regime unleashed a brutal crackdown on dissent.

• On Tuesday Syrian dictator Bashar al-Assad launched a horrible chemical weapons attack on innocent civilians.

• ISIS published a grisly video showing the savage execution of 17 young men from the town of Al Tibni on the first day of Eid.

3 Addendum to Experimental Procedures

3.1 Metalearning Winning Hyper-parameters
In this section we present the hyper-parameters tuned in the metalearning module and their corresponding “winning” values as guided by the hyper-parameters grid or default values appearing in Finn et al.’s github implementation here.

• MAML hyper-parameters with random sampling of tasks:
  – Number of iterations attempted \( n = 1000 \).
  – Number of tasks (meta batch size) \( N = 4 \).
- Number of classes: NC = 2
- Sampled batch size K for each task: K = N = 4.
- Size of each task = 2 * K * NC.
- Support set size = NC * K.
- Query set size = NC * K.
- Update learning rate α = 0.1.
- Meta-learning rate β = 0.1.
- Gradient steps on the support set: 4.
- Gradient steps on the query set: 10.

* Base model (FNN) hyper-parameters for MAML with random sampling of tasks:
  - Number of hidden layers: 3
  - Number of nodes per layer: 128, 64, and 64 nodes for each hidden layer respectively.
  - Activation function: relu.

### 3.2 Machine Learning Cross Validation (Training) Results and Winning Hyper-parameters

In this section we present the winning hyper-parameters for the machine learning models trained:

- Logistic regression:
  - penalty:l2, C:1000.
- Extra trees:
  - max-depth:2, max-features:sqrt, ccp-alpha:0.2.
- Random forest:
  - max-depth:2, max-features:sqrt, min-samples-split:0.2, ccp-alpha:0.4.
- Decision tree:
  - max-depth:2, max-features:sqrt, ccp-alpha:0.1.
- Gradient Boost:
  - n-estimators:2, max-features:sqrt, max-depth:2, subsample:1, ccp-alpha:0.1.
- XGBoost:
  - max-depth:3, min-child-weight:3, gamma:0.2, colsample-bytree:0.6, objective:reg:squarederror, reg-alpha :0.9.
- Adaboost:
  - base-estimator:RidgeClassifier(alpha=0.9), n-estimators:1, algorithm:SAMME.
- Nu SVC:
  - nu:0.5, kernel:linear, gamma:0.6, tol:1e^−05, probability:True.
• Ridge regression:
  – alpha:0.5.
• Naive Bayes:
  – var-smoothing:1e\(^{-09}\).
• SGD classifier:
  – alpha:0.9, penalty:l2.
• Linear SVC:
  – penalty:l2, C:1000.
• SVC:
  – C:1000, gamma:0.6, tol:1e\(^{-05}\).

4 Addendum to Feature Selection Results

In this section we present the feature selection scores linking each input feature with the
target class. Features are ranked in decreasing order of importance. Fig. [1] also presents
a visualisation of those scores.

• sectarian words: 0.296444476
• inconsistency score: 0.059969267
• quoted sources and attribution: 0.054075288
• hedges: 0.041442006
• implicative verbs: 0.03532212
• report verbs: 0.023943159
• factive verbs: 0.023873448
• subjectivity and bias: 0.00504511
• assertive verbs: 0.00380454

References

1. Golbeck J., Mauriello M., Auxier B., Bhanushali K.H., Bonk C., Bouzaghrane
   M.A., et al. Fake News vs Satire: A Dataset and Analysis. In: Proceedings of the
   10th ACM Conference on Web Science, pp. 17–21, 2018.
2. Hassan N., Zhang G., Arslan F., Caraballo J., Jimenez D., Gawsane S., et al.
   ClaimBuster: The First-ever End-to-end Fact-checking System. In: Proceedings
   of VLDB Endow, 10(12):1945–1948, 2017.
3. Mukherjee S., Weikum G. Leveraging Joint Interactions for Credibility Analysis
   in News Communities. In: Proceedings of the 24th ACM International on
   Conference on Information and Knowledge Management, pp. 353–362, 2015.
4. Niculae V., Suen C., Zhang J., Danescu-Niculescu-Mizil C., Leskovec J. QUOTUS: The Structure of Political Media Coverage as Revealed by Quoting Patterns. CoRR. 2015;abs/1504.01383, 2015

5. Popat K., Mukherjee S., Strötgen J., Weikum G. Credibility Assessment of Textual Claims on the Web. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pp. 2173–2178, 2016.

6. Prat A., Strömberg D.S. The political economy of mass media. In: Advances in Economics and Econometrics: Tenth World Congress, Cambridge University Press, 2013.

7. Rashkin H., Choi E., Jang J.Y., Volkova S., Choi Y. Truth of varying shades: Analyzing language in fake news and political fact-checking. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2931–2937, 2017.

8. Shu K., Mahudeswaran D., Wang S., Lee D., Liu H. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. arXiv preprint arXiv:180901286, 2018.

9. Torabi F., Taboada M. The Data Challenge in Misinformation Detection: Source Reputation vs. Content Veracity. In: Proceedings of the first workshop on fact extraction and VERification (FEVER), pp. 10–15, 2018.

10. Yang Wang W. Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 422–426, 2017.