Yifan Zhang<sup>1</sup>, Giovanni Da San Martino<sup>1</sup>, Alberto Barrón-Cedeño<sup>2</sup>, Salvatore Romeo<sup>1</sup>, Jisun An<sup>1</sup>, Haewoon Kwak<sup>1</sup>, Todor Staykovski<sup>3</sup>, Israa Jaradat<sup>4</sup>, Georgi Karadzhov<sup>5</sup>,

Ramy Baly<sup>6</sup>, Kareem Darwish<sup>1</sup>, James Glass<sup>6</sup>, Preslav Nakov<sup>1</sup>

<sup>1</sup>Qatar Computing Research Institute, HBKU, <sup>2</sup>Università di Bologna, Forlì, Italy <sup>3</sup>Sofia University <sup>4</sup>University of Texas at Arlington, <sup>5</sup>SiteGround Hosting EOOD

<sup>6</sup>MIT Computer Science and Artificial Intelligence Laboratory

{yzhang,gmartino,sromeo,jan,hkwak,kdarwish,pnakov}@hbku.edu.qa a.barron@unibo.it,todorstaykovski@gmail.com,{baly,glass}@mit.edu israa.jaradat@mavs.uta.edu,gogokaradjov@gmail.com

### Abstract

We introduce *Tanbih*, a news aggregator with intelligent analysis tools to help readers understanding what is behind a news story. Our system displays news grouped into events and generates media profiles that show the general factuality of reporting, the degree of propagandistic content, hyper-partisanship, leading political ideology, general frame of reporting, and stance with respect to various claims and topics of a news outlet. In addition, we automatically analyze each article in order to detect whether it is propagandistic and to determine its stance with respect to a number of controversial topics.

# 1 Introduction

Nowadays, more and more readers consume news online. The reduced costs and, generally speaking, the less strict regulations with respect to the standard press, have led to the proliferation of online news media. However, this does not necessarily entail that readers are exposed to a plurality of viewpoints as news consumed via social networks are known to reinforce the bias of the user (Flaxman et al., 2016) because of filtering bubbles and echo chambers. Moreover, visiting multiple websites to gather a more comprehensive analysis of an event might be too time-consuming for the average reader.

News aggregators —such as Flipboard<sup>1</sup>, News Lens<sup>2</sup> and Google News<sup>3</sup>—, gather news from different sources and, in the case of the latter two, cluster them into events. In addition, News Lens displays all articles about an event in a timeline and provides additional information, such as summary of the event and a description for each entity mentioned in an article.

While these news aggregators help readers to get a more comprehensive coverage of an event, some of the sources might be unknown to the user, and thus he/she could naturally question the validity and the trustworthiness of the information provided. Deep analysis of the content published by news outlets has been performed by expert journalists. For example, Media Bias/Fact Check<sup>4</sup> provides reports on the bias and factuality of reporting of entire news outlets, whereas Snopes<sup>5</sup>, Politifact,<sup>6</sup> and FactCheck<sup>7</sup> are popular fact-checking websites. All these manual efforts cannot cope with the rate at which news contents are produced.

Here, we propose *Tanbih*, a news platform that displays news grouped into events and provides additional information about the articles and their media source in order to promote media literacy. It automatically generates profiles for the news media with reports on their factuality, leading political ideology, hyper-partisanship, use of propaganda, and bias. Furthermore, *Tanbih* automatically categorizes articles in English and Arabic, flags potentially propagandistic ones, and examines their framing bias.

# 2 System Architecture

The architecture of *Tanbih* is sketched in Figure 1. The system consists of three main components: an online streaming processing pipeline for data collection and article level analysis, offline processing for event and media source level analysis, and a website for delivering news to the users. The online streaming processing pipeline continuously retrieves articles in English and Arabic, which are then translated, categorized, and analyzed for their general frame of reporting and use of propaganda.

<sup>&</sup>lt;sup>1</sup>http://flipboard.com

<sup>&</sup>lt;sup>2</sup>http://newslens.berkeley.edu

<sup>&</sup>lt;sup>3</sup>http://news.google.com.

<sup>&</sup>lt;sup>4</sup>http://mediabiasfactcheck.com
<sup>5</sup>http://www.snopes.com/
6

<sup>&</sup>lt;sup>6</sup>http://www.politifact.com/

<sup>&</sup>lt;sup>7</sup>http://www.factcheck.org/



Figure 1: The architecture of Tanbih. The arrows indicate the information flow.

We perform clustering on the articles that have been collected every 30 minutes. The offline processing includes factuality prediction, leading political ideology detection, audience reach and Twitter user based bias prediction at the media level, and stance detection and aggregation of statistics at the article level, e.g., propaganda index (see Section 2.3) for each news medium. The offline processing does not have strict time requirements, and thus the choice of the models we develop favors accuracy over speed.

In order to run everything in a streaming and scalable fashion, we use KAFKA<sup>8</sup> as a messaging queue and Kubernetes<sup>9</sup>, thus ensuring scalability and fault-tolerant deployment. In the following, we describe each component of the system. We have open-sourced the code for some of those, and we plan to do that for the remaining ones in the near future.

### 2.1 Crawlers and Translation

Our crawlers collect articles from a growing list of sources<sup>10</sup>, which currently includes 155 RSS feeds, 82 Twitter accounts and two websites. Once a link to an article has been obtained from any of these sources, we rely on the Newspaper3k Python library to extract its contents.<sup>11</sup> After deduplication based on both URL and text content, our crawlers currently download 7k-10k articles per day. As of present, we have more than 700k articles stored in our database. We use QCRI's Machine Translation (Dalvi et al., 2017) to translate English content into Arabic and vice versa. Since translation is performed offline, we select the most accurate system in Dalvi et al. (2017), i.e., the neural-based one.

#### 2.2 Section Categorization

We built a model to classify an article into one of six news sections: Entertainment, Sports, Business, Technology, Politics, and Health. We built a corpus using the New York Times articles from the FakeNews dataset<sup>12</sup> published between January 1, 2000 and December 31, 2017. We extracted the news section information embedded in the article URL and we used a total of 538k articles for training our models using TF.IDF representation. On a test set of 107k articles, our best-performing logistic regression model achieved F<sub>1</sub> scores of 0.82, 0.58, 0.80, and 0.90 for Sports, Business, Technology, and Politics, respectively. The overall F<sub>1</sub> for the baseline was 0.497.

#### 2.3 Propaganda Detection

We developed a propaganda detection component to flag articles that could be potentially propagandistic, i.e., purposefully biased to influence its readers and ultimately to pursue a specific agenda. Given a corpus of news that is labelled as propagandistic/non propagandistic (Barrón-Cedeño et al., 2019), we train a maximum entropy classifier on 51k articles, represented with various stylerelated features, such as character n-grams and a number of vocabulary richness and readability measures, and we obtain state-of-the-art  $F_1$ =82.89 on a separate test set of 10k articles. We refer to the score  $p \in [0, 1]$  of the classifier as propaganda index, and we define the following propaganda labels, which we use to flag articles (see Figure 2; right news): very unlikely (p < 0.2), unlikely  $(0.2 \le p < 0.4)$ , somehow  $(0.4 \le p < 0.6)$ , likely ( $0.6 \le p < 0.8$ ), and very likely ( $p \ge 0.8$ ).

<sup>&</sup>lt;sup>8</sup>http://kafka.apache.org

<sup>&</sup>lt;sup>9</sup>http://kubernetes.io

<sup>&</sup>lt;sup>10</sup>http://www.tanbih.org/about

<sup>&</sup>lt;sup>11</sup>http://newspaper.readthedocs.io

<sup>&</sup>lt;sup>12</sup>http://github.com/several27/ FakeNewsCorpus

### 2.4 Framing Bias Detection

Framing is a central concept in political communication, which intentionally emphasizes or ignores certain dimensions of an issue (Entman, 1993). In *Tanbih*, we infer the frames of news articles, thus making them explicit. In particular, we use the Media Frames Corpus (MFC) (Card et al., 2015) to train a fine-tuned BERT model to detect topic-agnostic media frames. For training, we use a small learning rate of 0.0002, a maximum sequence length of 128, and a batch size of 32. Our model, when trained on 11k articles from MFC, achieved an accuracy of 66.7% on a test set of 1,138 articles. This is better than the previously reported state-of-the-art (58.4%) on a subset of MFC (Ji and Smith, 2017).

# 2.5 Factuality of Reporting and Leading Political Ideology of a Source

The factuality of reporting and the bias of an information source are key indicators that investigative journalists use to judge the reliability of information. In Tanbih, we model the factuality and the bias at the media level, learning from the Media Bias/Fact Check (MBFC) website, which covers over 2,800 news outlets. The model improves over our recent research (Baly et al., 2018, 2019; Dinkov et al., 2019), and combines information from articles published by the target medium, from their Wikipedia page accounts, from their social media accounts (Twitter, Facebook, Youtube) as well as from the social media accounts of the users who interact with the medium. We model factuality on a 3-point scale (low, mixed and high), with 80.1% accuracy (baseline 46.0%), and bias on a 7-point *left*-to-*right* scale, with 69% accuracy (baseline 24.7%), and also on a 3-point scale, with 81.9% accuracy (baseline 37.1%).

#### 2.6 Stance Detection

Stance detection aims to identify the relative perspective of a piece of text with respect to a claim, typically modeled using labels such as *agree*, *disagree*, *discuss*, and *unrelated*. An interesting application of stance detection is medium profiling with respect to controversial topics. In this setting, given a particular medium, the stance for each article is computed with respect to a set of predefined claims. The stance of a medium is then obtained by aggregating the article-level stances. In *Tanbih*, the stance is used to profile media sources. We implemented our stance detection model as fine-tuning of BERT on the FNC-1 dataset from the *Fake News Challenge*<sup>13</sup>. Our model outperformed the best submitted system (Hanselowski et al., 2018), obtaining an  $F_{1_{macro}}$  of 75.30 and an  $F_1$  of 69.61, 49.76, 83.01, and 98.81 for *agree*, *disagree*, *disagree*, *discuss*, and *unrelated*, respectively.

# 2.7 Audience Reach

User interactions on Facebook enable the platform to generate comprehensive user profiles for gender, age, income bracket, and political preferences. After marketers have determined a set of criteria for their target audience, Facebook can provide them with an estimate of the size of this audience on its platform. As an illustration, there are about 160K Facebook users who are 20 years old, are very liberal, are female, and have an interest in The New York Times. In Tanbih, we use the political leaning of Facebook users who follow a news medium as a feature to potentially improve media bias and factuality prediction; we also show it in the media profiles. To get the audience of each news medium, we use Facebook's Marketing API to extract the demographic data of the medium's audience with a focus on audience members who reside in USA and their political leanings (ideology): (Very Conservative, Conservative, Moderate, Liberal, and Very Liberal).<sup>14</sup>

# 2.8 Twitter User-Based Bias Classification

Controversial social and political issues may spur social media users to express their opinion through sharing supporting newspaper articles. Our intuition is that the bias of news sources can be inferred based on the bias of social media users. For example, if articles from a news source are strictly shared by left- or right-leaning users, then the source is likely left- or right-leaning, respectively. Similarly, if it is being cited by both groups, then it is likely closer to the center. We used an unsupervised user-based stance detection method on different controversial topics in order to find core groups of right- and left-leaning users (Darwish et al., 2019). Given that the stance detection produces clusters with nearly perfect purity (> 97%purity), we used the identified core users to train a deep learning-based classifier, fastText, using the accounts that they retweeted as features to further tag more users.

<sup>&</sup>lt;sup>13</sup>http://www.fakenewschallenge.org/

<sup>&</sup>lt;sup>14</sup>These are only available for US-based Facebook users.

Next, we computed a valence score for each news outlet and for each topic. The valence scores range between -1 and 1, with higher absolute values indicating being cited with greater proportion by one group as opposed to the other. The score is calculated as follows (Conover et al., 2011):

$$V(u) = 2 \frac{\frac{tf(u,C_0)}{total(C_0)}}{\frac{tf(u,C_0)}{total(C_0)} + \frac{tf(u,C_1)}{total(C_1)}} - 1$$
(1)

where  $tf(u, C_0)$  is the number of times (term frequency) item u is cited by group  $C_0$ , and  $total(C_0)$  is the sum of the term frequencies of all items cited by  $C_0$ .  $tf(u, C_1)$  and  $total(C_1)$  are defined in a similar fashion. We subdivided the range between -1 and 1 into 5 equal size ranges and we assigned the labels *far-left*, *left*, *center*, *right*, and *far-right* to those ranges.

#### 2.9 Event Identification / Clustering

The clustering module aggregates news articles into stories. The pipeline is divided into two stages: (*i*) local topic identification and (*ii*) long-term topic matching for story generation.

For step (i), we represent each article as a TF.IDF vector, built from the title and the body concatenated. The pre-processing consists of casefolding, lemmatization, punctuation and stopword removal. In order to obtain the preliminary clusters, in stage (i), we compute the cosine similarity between all article pairs in a predefined time window. We set n = 6 as the number of days withing a window with an overlap of three days. The resulting matrix of similarities for each window is then used to build a graph G = (V, E), where V is the set of vertices, i.e., the news articles, and E is the set of edges. An edge between two articles  $\{d_i, d_i\} \in V$  is drawn only if  $sim(d_i, d_j) \geq T_1$ , with  $T_1 = 0.31$ . We selected all parameters empirically on the training part of the corpus from (Miranda et al., 2018). The sequence of overlapping local graphs is merged in the order of their creation, thus generating stories from the topics. After merging, a community detection algorithm is used in order to find the correct assignment of the nodes into clusters. We used one of the fastest modularity-based algorithms: the Louvain method (Blondel et al., 2008).

For step (*ii*), the topics created from the preceding stage are merged if the cosine similarity  $sim(t_i, t_j) \ge T_2$ , where  $t_i(t_j)$  is the mean of all vectors belonging to topic i(j), with  $T_2 = 0.8$ .



Figure 2: The Tanbih main page.

The model achieved state-of-the-art performance on the testing partition of the corpus from Miranda et al. (2018): an  $F_1$  of 98.11 and an  $F_{1_{BCubed}}$  of 94.41.<sup>15</sup> As a comparison, the best model described in (Miranda et al., 2018) achieved an  $F_1$  of 94.1. See Staykovski et al. (2019) for further details.

### **3** Interface

The home page of *Tanbih*<sup>16</sup> displays news articles grouped into stories (see the screenshot in Figure 2). Each story is displayed as a card. The users can go back and forth between the articles from the same event by clicking on the left/right arrows below the title of the article. A propaganda label is displayed if the article is predicted to be likely propagandistic. Such an example is shown on the right of Figure 2. The source of each article is displayed with the logo or the avatar of the respective news organization, and it links to a special profile page for this organization (see Figure 3). On the top-left of the home page, Tanbih provides language selection buttons, currently English and Arabic only, to switch the language the news are display in. Finally, a search box in the top-right corner allows the user to find the profile page of a particular news medium of interest.

On the media profile page (Figure 3a), a short extract from the Wikipedia page of the medium is displayed on the top, with recently published articles on the right-hand side. The profile page includes a number of statistics automatically derived from the models in Section 2. We use as an example Figure 3, which shows screenshots of the profile of CNN.<sup>17</sup>

 $<sup>{}^{15}\</sup>text{F}_{1_{BCubed}}$  is an evaluation measure specifically designed to evaluate clustering algorithms (Amigó et al., 2009).

<sup>&</sup>lt;sup>16</sup>http://www.tanbih.org

<sup>&</sup>lt;sup>17</sup>Here is a direct link to the profile: http://www.tanbih.org/media/1



Figure 3: A partial screenshot of the media profile page for CNN in Tanbih.

The first two charts in Figure 3a show the centrality and the hyper-partisanship (we can see that CNN is estimated to be fairly central and low in hyper-partisanship) and the distribution of propagandistic articles (CNN publishes mostly nonpropagandistic articles).

Figure 3b shows the overall framing bias distribution for the medium (CNN focuses mostly on cultural identity and politics), and the factuality of reporting (CNN is mostly factual). The profile also shows the leading political ideology of the medium on a 3-point and also on a 7-point scale.

Figure 3c shows the audience reach of the medium and the bias classification according to users' retweets (see Section 2.8). We can see that CNN is popular among readers with all political views, although it tends to have a left-leaning ideology on the topics listed. The profile also features reports on the stance of CNN with respect to a number of topics.

Finally, *Tanbih* features pages about specific topics. These are accessible via the search box on the top-right of *Tanbih*'s main page. An example is given in Figure 4, which shows the topic page for the Khashoggi's murder. Recent stories about this topic are listed on the top of the page, followed by statistics such as the number of countries, the number of articles, and the number of media reporting on it. A map shows how much reporting there is on the event per country, which allows users to get an idea of how important the topic is there.



Figure 4: A partial screenshot of the topic page for the Khashoggi Murder in *Tanbih*.

The topic page further features charts showing (i) the top countries in terms of coverage of the event, both in absolute and in relative numbers with respect to the total number of articles published, and (ii) the media sources that published most propagandistic content on the topic, again both in absolute and in relative terms with respect to the total number of articles published by the respective medium on the topic. Finally, the topic page displays plots showing the overall distribution of propagandistic articles and of the overall framing bias when reporting on the topic.

### 4 Conclusions and Future Work

We have introduced *Tanbih*, a news aggregator that performs media-level and article-level analysis of the news aiming to help users better understand what they are reading. *Tanbih* features factuality prediction, propaganda detection, stance detection, leading political ideology identification, media framing bias detection, event clustering, and machine translation.

In future work, we plan to include more media sources, especially from non-English speaking regions and to add interactive components, e.g., letting users ask a question about a specific topic. We also plan to add sentence-level and sub-sentence-level annotations for checkworthiness (Jaradat et al., 2018) and fine-grained propaganda (Da San Martino et al., 2019).

#### Acknowledgements

This research is part of the Tanbih project,<sup>18</sup> which aims to limit the effect of "fake news", propaganda and media bias by making users aware of what they are reading. It is developed in collaboration between the Qatar Computing Research Institute (QCRI), HBKU and the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL).

# References

- Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retr.*, 12(4):461–486.
- Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proc. of EMNLP'18*, pages 3528–3539.
- Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. 2019. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. In *Proc. of NAACL-HLT'19*, pages 2109– 2116.
- Alberto Barrón-Cedeño, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. 2019. Proppy: Organizing news coverage on the basis of their propagandistic content. *Information Processing and Management*, 56(5):1849–1864.
- Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal*

of Statistical Mechanics: Theory and Experiment, 2008(10):P10008.

- Dallas Card, Amber E Boydstun, Justin H Gross, Philip Resnik, and Noah A Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proc. of ACL '15*, pages 438–444.
- Michael Conover, Jacob Ratkiewicz, Matthew R Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on Twitter. In *Proc. of ICWSM'11*, pages 89–96.
- Giovanni Da San Martino, Seunghak Yu, Alberto Barron-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proc. of EMNLP'19*.
- Fahim Dalvi, Yifan Zhang, Sameer Khurana, Nadir Durrani, Hassan Sajjad, Ahmed Abdelali, Hamdy Mubarak, Ahmed Ali, and Stephan Vogel. 2017. QCRI live speech translation system. In Proc. of EACL'17, pages 61–64.
- Kareem Darwish, Peter Stefanov, Michaël J Aupetit, and Preslav Nakov. 2019. Unsupervised user stance detection on Twitter. In *Proc. of ICWSM*'20.
- Yoan Dinkov, Ahmed Ali, Ivan Koychev, and Preslav Nakov. 2019. Predicting the leading political ideology of Youtube channels using acoustic, textual and metadata information. In *Proc. of INTER-SPEECH'19*.
- Robert M Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of communication*, 43(4):51–58.
- Seth Flaxman, Sharad Goel, and Justin M Rao. 2016. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1):298– 320.
- Andreas Hanselowski, Avinesh P.V.S., Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance-detection task. In *Proc. of COLING'18*, pages 1859–1874.
- Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. Claim-Rank: Detecting check-worthy claims in Arabic and English. In *Proc. of NAACL-HLT '18*, pages 26–30.
- Yangfeng Ji and Noah A. Smith. 2017. Neural discourse structure for text categorization. In *Proc.* ACL'17, pages 996–1005.
- Sebastião Miranda, Arturs Znotins, Shay B. Cohen, and Guntis Barzdins. 2018. Multilingual clustering of streaming news. In *Proc. of EMNLP'18*, pages 4535–4544.
- Todor Staykovski, Alberto Barrón-Cedeño, Giovanni Da San Martino, and Preslav Nakov. 2019. Dense vs. sparse representations for news stream clustering. In *Proc. of Text2story*'19, pages 47–52.

<sup>&</sup>lt;sup>18</sup>http://tanbih.qcri.org/