Evaluation Measures for the 2019 *Hack The News!* Datathon

Giovanni da San Martino, Alberto Barrón-Cedeño and **Preslav Nakov**
Qatar Computing Research Institute, HBKU
tanbih@qcri.org

Abstract

We present the definitions and describe the evaluation measures for the three tasks of the 2019 *Hack The News!* Datathon: (*i*) predicting whether a document is propagandistic, (*ii*) determining which sentences in the document are propagandistic, and (*iii*) identifying the use of specific propaganda techniques in text.

1 Background

The *Hack The News!* Global Datathon[1] takes place on January 21–27, 2019. It features three tasks with different levels of complexity, aiming at identifying the use of propaganda in the news.[2] Below, we define the tasks and we present the evaluation measure for each of them.

2 Task 1

Definition: Given a document, determine whether it is propagandistic or not.

This is a binary classification task.

Meta-Data formatting: A single tab-separated file with two fields:

```
[doc id]    [label]
```

Evaluation: We adopt the F$_1$-measure:

$$F_1 = \frac{2 \cdot prec \cdot rec}{prec + rec} \tag{1}$$

where precision is defined as follows:

$$prec = \frac{|\{\text{propagandistic docs}\} \cap \{\text{retrieved docs}\}|}{|\{\text{retrieved docs}\}|} \tag{2}$$

and recall is defined as follows:

$$rec = \frac{|\{\text{propagandistic docs}\} \cap \{\text{retrieved docs}\}|}{|\{\text{propagandistic docs}\}|} \tag{3}$$

3 Task 2

Definition: Given a document, determine whether each of its sentences is propagandistic.

Once again, this is a binary classification task.

Meta-Data formatting: A tab-separated file with three fields:

```
[doc id] [sentence id] [label]
```

3.1 Evaluation:

Again, we adopt the F$_1$-measure, but this time defining precision and recall at the sentence level:

$$prec = \frac{|\{\text{propagandistic sents}\} \cap \{\text{retrieved sents}\}|}{|\{\text{retrieved sents}\}|} \tag{4}$$

$$rec = \frac{|\{\text{propagandistic sents}\} \cap \{\text{retrieved sents}\}|}{|\{\text{propagandistic sents}\}|} \tag{5}$$

4 Task 3

Definition: Given a document, (*i*) identify all text fragments that use a propagandistic technique (start and end character offsets) and also (*ii*) choose the propagandistic technique that was used in each such fragment, from an inventory of eighteen possible techniques.

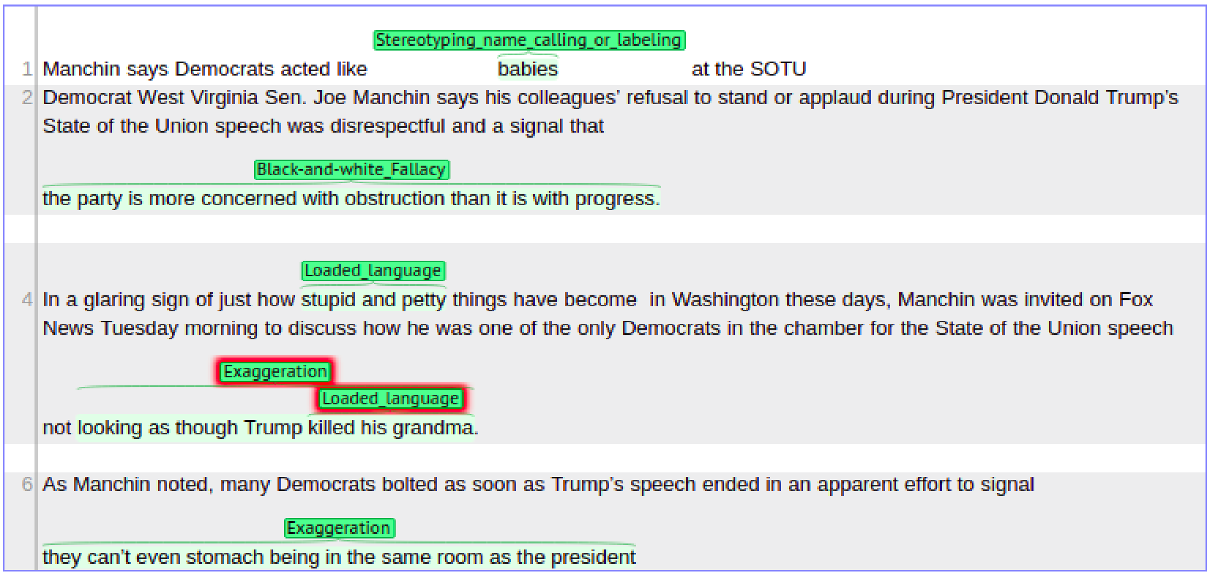
This is a multi-way sequence labeling task, akin to named entity recognition.

Meta-Data formatting: A tab-separated file with four fields, where the last two fields indicate the starting and ending position of the span:

```
[doc id] [label] [start] [end]
```

[1] http://www.datasciencesociety.net/datathon/

[2] http://www.datasciencesociety.net/hack-news-datathon-case-propaganda-detection/



Manchin says Democrats acted like babies at the SOTU

Democrat West Virginia Sen. Joe Manchin says his colleagues' refusal to stand or applaud during President Donald Trump's State of the Union speech was disrespectful and a signal that the party is more concerned with obstruction than it is with progress.

In a glaring sign of just how stupid and petty things have become in Washington these days, Manchin was invited on Fox News Tuesday morning to discuss how he was one of the only Democrats in the chamber for the State of the Union speech not looking as though Trump killed his grandma.

As Manchin noted, many Democrats bolted as soon as Trump's speech ended in an apparent effort to signal they can't even stomach being in the same room as the president

Figure 1: Example of annotation used when giving instructions to the annotators.

Evaluation: Here, we use a variant of F_1-measure like before, but the calculations for precision and recall are more complex, as we have a multi-way classification task,[3] and we also need to take the fragment spans into account. As these spans can be quite long for some of the techniques, we give partial credit in case the system has predicted a propaganda technique of the correct type, but the start and the end boundaries of the fragment do not match those of the gold annotation exactly, and instead there is only partial overlap.

Our definitions of precision and recall are inspired by (Potthast et al., 2010). Let document $\boldsymbol{d}$ be represented as a set of references to its elements. A propagandistic text fragment is then represented as $r = [r_i, \ldots, r_j] \subseteq \boldsymbol{d}, i < j$. A document includes a set of (possibly overlapping) fragments R. Similarly, the set of fragments predicted by a learning algorithm are denoted by $p = [p_k, \ldots, p_l] \subseteq \boldsymbol{d}, k < l$; the set of predictions P for a document $\boldsymbol{d}$. Let $c(p) = \{1, \ldots, 18\}$ be the propaganda technique associated with the fragment p by a learning algorithm $c()$. We denote the gold label of a fragment with $g(r)$. Let $\delta(a, b) = 1$ if $a = b$; 0 otherwise. Given two annotations, p and r, we score them as

$$S(p, r) = \frac{|(r \cap p)|}{\max(|r|, |p|)} \delta\left(g(r), c(p)\right) \quad . \tag{6}$$

The scoring depends on whether the labels $g(r)$ and $c(p)$ are identical and it is proportional to the intersection of the spans of the two annotations.

Precision and recall are defined as follows:

$$prec = \frac{1}{|P|} \sum_{\substack{p \,\in\, P, \\ r \,\in\, R}} S(p, r) \tag{7}$$

$$rec = \frac{1}{|R|} \sum_{\substack{p \,\in\, P, \\ r \,\in\, R}} S(p, r). \tag{8}$$

Figure 1 shows an example of the fragment-level annotation.

4.1 Extending the Measure to Multiple Documents

Eq. (7) and (8) are defined at the document level. In order to extend them to a set of n documents $\mathcal{P} = \{P_1, \ldots, P_n\}$ and $\mathcal{R} = \{R_1, \ldots, R_n\}$, we modify them as follows:

$$prec = \frac{\sum\limits_{i=0}^{i=n} \sum\limits_{\substack{p \,\in\, P_i, \\ r \,\in\, R_i}} S(p, r)}{\sum\limits_{i=0}^{i=n} |P_i|} \tag{9}$$

$$rec = \frac{\sum\limits_{i=0}^{i=n} \sum\limits_{\substack{p \,\in\, P_i, \\ r \,\in\, R_i}} S(p, r)}{\sum\limits_{i=0}^{i=n} |R_i|} \tag{10}$$

[3] See the following link for definition and examples for each of the eighteen propaganda techniques we consider: http://propaganda.qcri.org/annotations/definitions.html

Acknowledgments

This work is part of the Propaganda Analysis Project,[4] which is part of the Tanbih project,[5] developed by the Qatar Computing Research Institute, HBKU.

References

Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. An Evaluation Framework for Plagiarism Detection. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, volume 2, pages 997–1005, Beijing, China. Association for Computational Linguistics.

[4] http://propaganda.qcri.org/

[5] http://tanbih.qcri.org/