Evaluation of the Fragment-level Classification Subtask

Shared Task on Fine-grained Propaganda Detection at
2019 Workshop on NLP4IF: censorship, disinformation, and propaganda.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Preslav Nakov
email: gmartino@hbku.edu.qa

Let document $\boldsymbol{d}$ be represented as a sequence of characters. The i-th propagandistic text fragment is then represented as a sequence of contiguous characters $t \subseteq \boldsymbol{d}$. A document includes a set of (possibly overlapping) fragments T. Similarly, a learning algorithm produces a set S with fragments $s \subseteq \boldsymbol{d}$, predicted on $\boldsymbol{d}$. A labeling function $l(x) \in \{1, \ldots, 18\}$ associates $t \in T$, $s \in S$ with one of the eighteen techniques. An example of (gold) annotation is in Figure 1: an annotation t_1 flags the words "stupid and petty" with the technique "Loaded_language".

We define the following function to handle partial overlaps between fragments with same labels:

$$C(s,t,h) = \frac{|(s \cap t)|}{h} \delta\left(l(s), l(t)\right), \qquad (1)$$

where h is a normalizing factor and $\delta(a,b) = 1$ if $a = b$, and 0 otherwise. For example, still with reference to Figure 1, $C(t_1, s_1, |t_1|) = \frac{6}{16}$ and $C(t_1, s_2, |t_1|) = 0$.

Given (1), we now define variants of precision and recall able to account for the imbalance in the corpus:

$$P(S,T) = \frac{1}{|S|} \sum_{\substack{s \in S, \\ t \in T}} C(s,t,|s|), \qquad (2)$$

$$R(S,T) = \frac{1}{|T|} \sum_{\substack{s \in S, \\ t \in T}} C(s,t,|t|), \qquad (3)$$

We define (2) to be zero if $|S| = 0$ and Eq. (3) to be zero if $|T| = 0$. Following Potthast et al. (2010), in (2) and (3) we penalize systems predicting too many or too few instances by dividing by $|S|$ and $|T|$, respectively; e.g., in Figure 1 $R(\{s_3, s_4, s_5\}, \{t_1\}) = \frac{7}{24} < R(\{s_1\}, \{t_1\}) = \frac{9}{24} < R(\{t_1\}, \{t_1\}) = 1$.

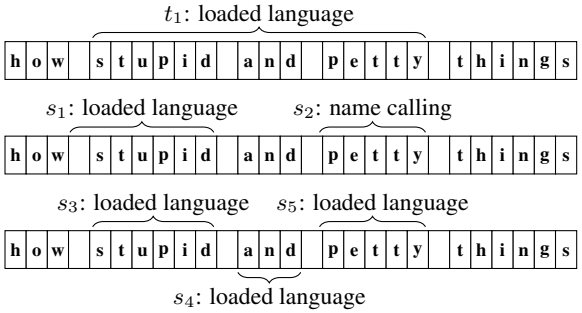

Figure 1: Example of gold annotation (top) and the predictions of a supervised model (bottom) in a document represented as a sequence of characters.

Finally, we combine Eqs. (2) and (3) into an F_1-measure, the harmonic mean of precision and recall:

$$F_1(S,T) = 2\frac{P(S,T)R(S,T)}{P(S,T) + R(S,T)} \qquad (4)$$

Notice that (4) can be computed with respect to one technique only simply by replacing the δ function in (1) with $\delta_L(a,b) = 1$ if $a = b = L$, where L is a predetermined propaganda technique.

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.