Methodology

Cross-modal Similarity Extraction

- Compared to news articles delivering relevant textual and visual information, those with disparate statements and images are more likely to be fake.
- Define the relevance between news textual and visual information as follows by slightly modifying cosine similarity:
  - $M_s(t, v) = \frac{t \cdot v + \|t\|\|v\|}{2\|t\|\|v\|}$ v.s. $\cos(t, v) = \frac{t \cdot v}{\|t\|\|v\|}$
- In such a way, it's guaranteed that $M_s(t, v)$ is positive and $\in [0,1]$
  - $M_s(t, v) \to 0$: t, v are far from being similar, $\to 1$: t, v are exactly the same

Methodology

Cross-modal Similarity Extraction

- Then defined the loss function based on cross-entropy as below, which assumes that news articles formed with mismatched textual and visual information are more likely to be fake compared to those with matching textual statements and images, when analyzing from a pure similarity perspective:
  - $L_S(\theta_t, \theta_v) = -\mathbb{E}_{(a,y)\sim(A,Y)}(y \log(1 - M_s(t, v)) + (1 - y)\log M_s(t, v))$