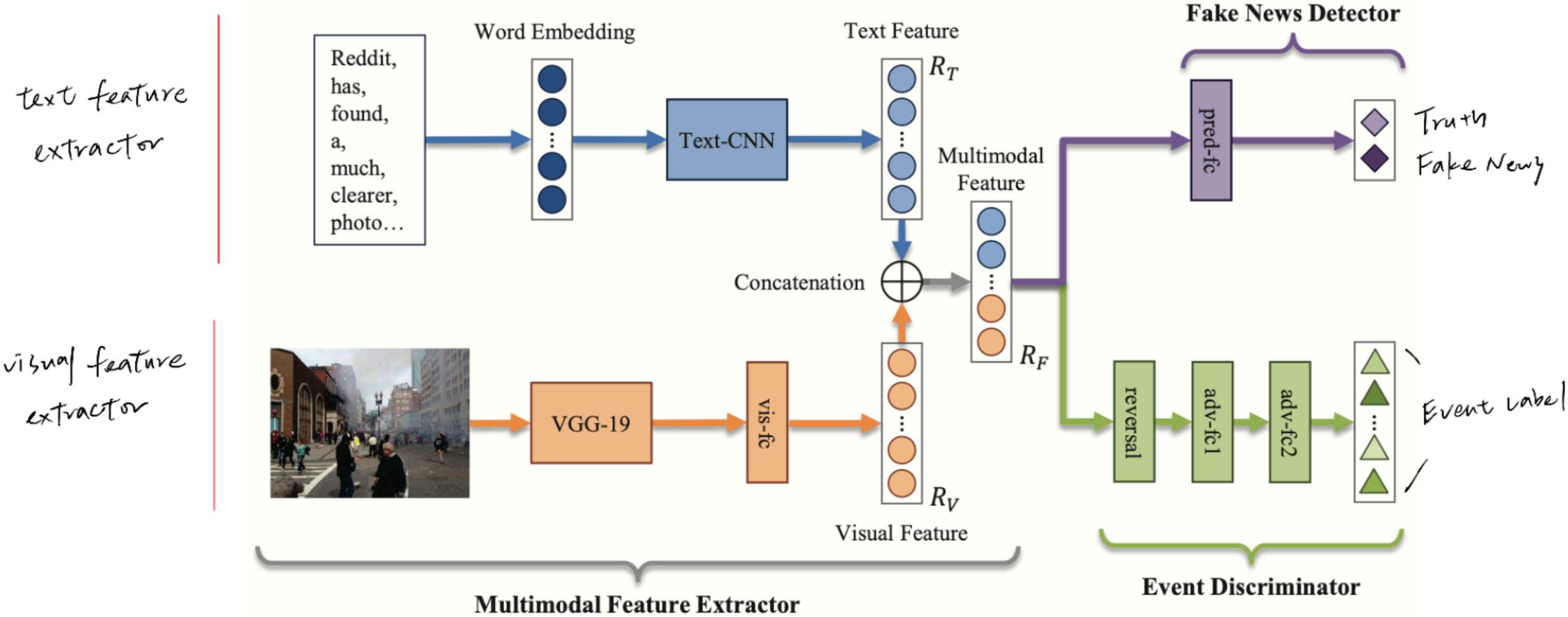
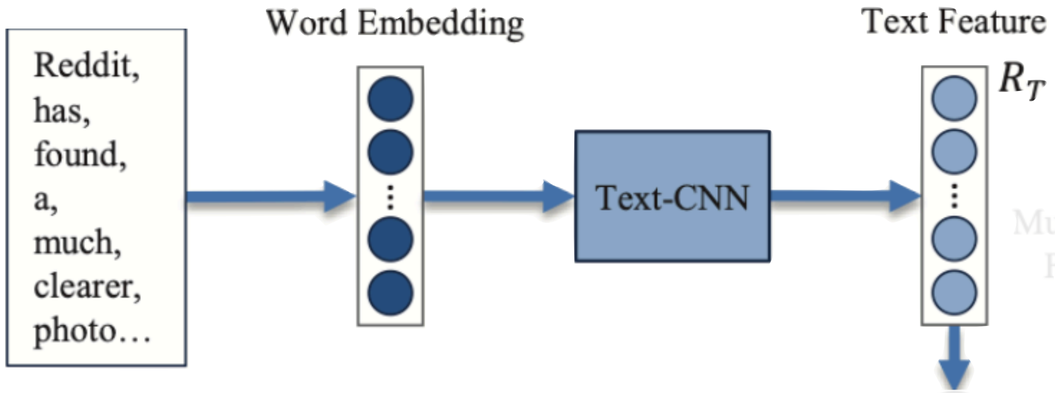
# Methodology.

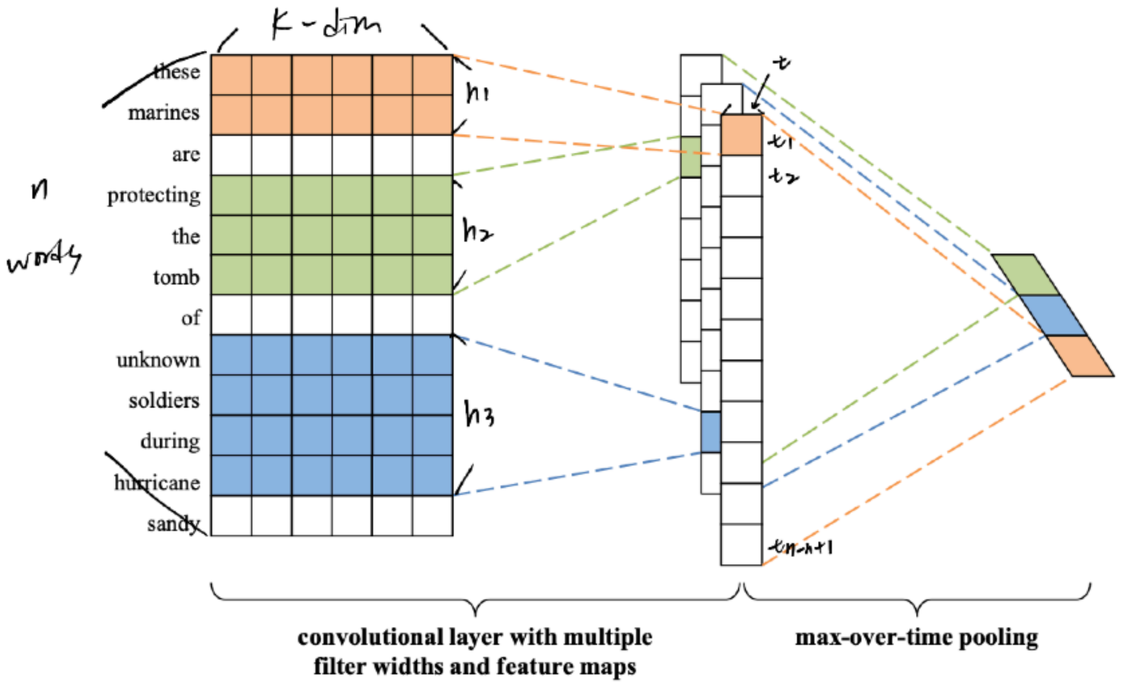
## Model Overview



# Methodology..

## Textual Feature Extractor




- Employ Text-CNN as text feature extractor
- $n$ words sentence: $T_{1:n} = T_1 \oplus T_2 \cdots \oplus T_n$
- Convolutional filter with window size $h$: $t_i = \sigma\left(W_c \cdot T_{i:i+h-1}\right)$
- Get feature vector of sentence: $t = [t_1, t_2, \cdots, t_{n-h+1}]$
- Following the max-pooling operations, a fully connected layer to ensure the final representation ($R_T \in \mathbb{R}^p$) has the same dimension $p$ with visual representation: $R_T = \sigma(W_{tf} \cdot R_{T_c})$