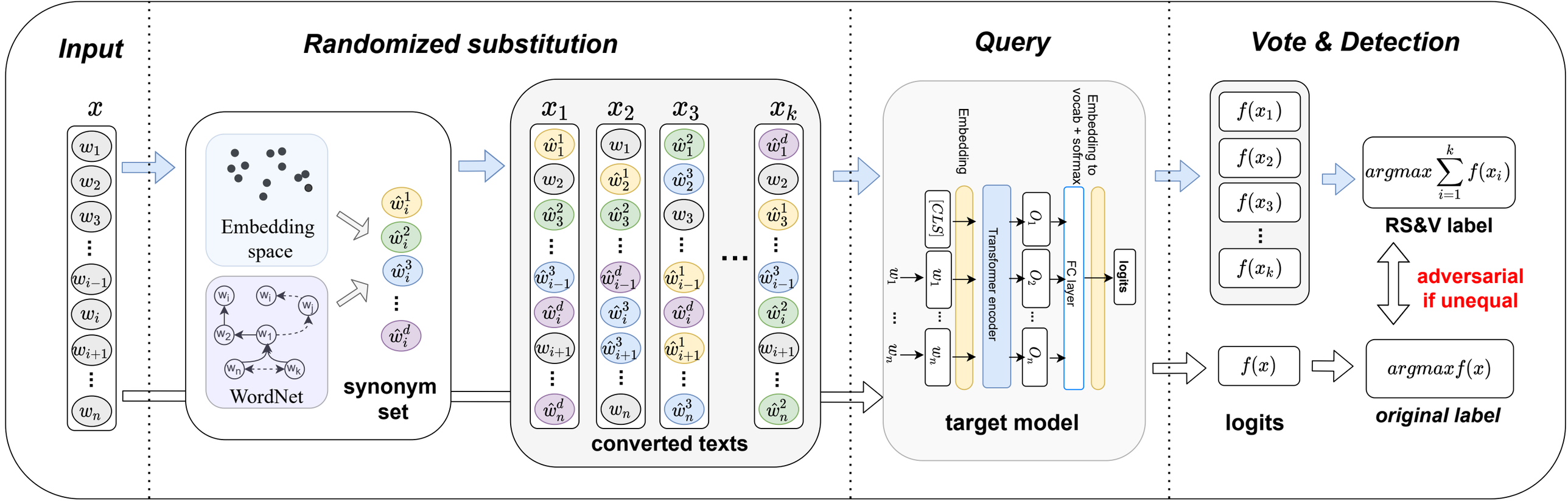
Input
Randomized substitution
Query
Vote & Detection
x
w_1 w_2 w_3 ⋮ w_{i-1} w_i w_{i+1} ⋮ w_n
Embedding space
WordNet
synonym set
$\hat{w}_i^1$ $\hat{w}_i^2$ $\hat{w}_i^3$ ⋮ $\hat{w}_i^d$
x_1 x_2 x_3 ... x_k
converted texts
Embedding
Embedding to vocab + sofrmax
[CLS]
Transformer encoder
FC layer
logits
target model
$f(x_1)$ $f(x_2)$ $f(x_3)$ ⋮ $f(x_k)$
$argmax \sum_{i=1}^{k} f(x_i)$
RS&V label
adversarial if unequal
$f(x)$
logits
$argmax f(x)$
original label