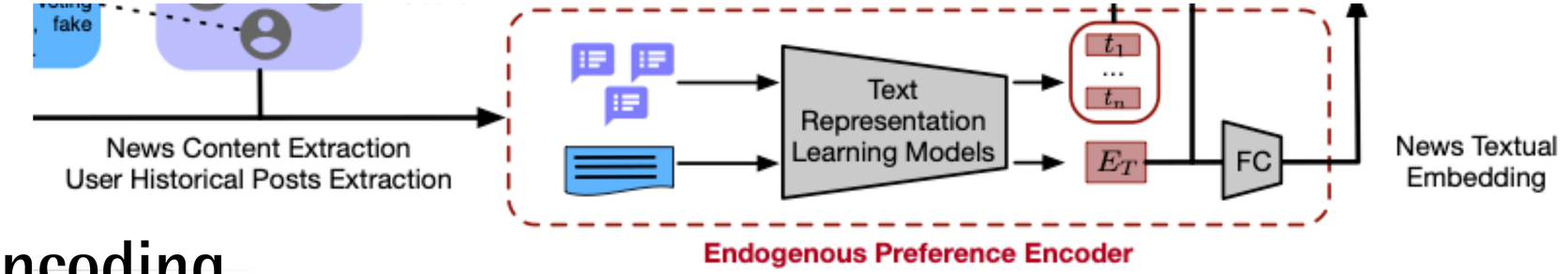
# Approach

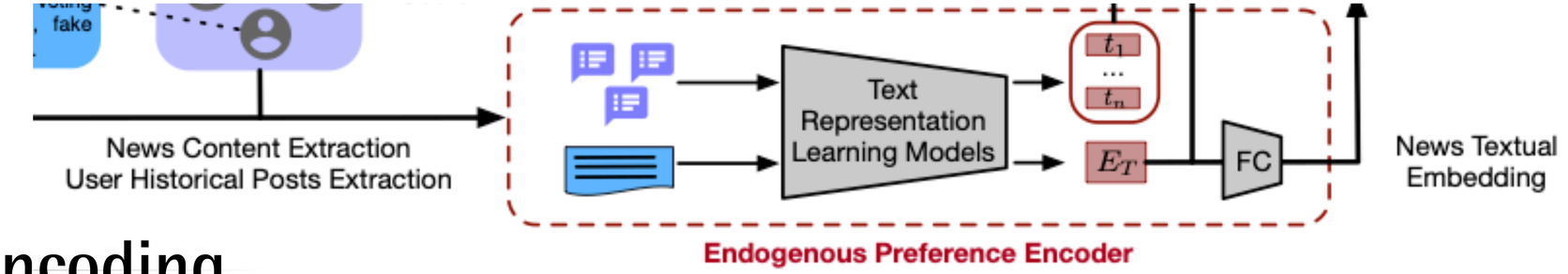
## Endogenous Preference Encoding




- To encode the news textual information and user preferences, employ two types of text representation learning approaches based on language pertaining.
  - word2vec: choose the 680k 300 dimensional vectors pretrained by spaCy
  - BERT: employ pretrained embeddings (BERT-large) using bert-as-a-service
- Instead of training on the local corpus, the word embedding pretrained on large corpus are supposed to encode more semantic similarities between different words and sentences.

# Approach

## Endogenous Preference Encoding




- word2vec (spaCy)
  - Average the vectors of existing words in combined recent 200 tweets to get user preference representation.
  - The news textual embedding is obtained similarly.
- BERT (BERT-large)
  - Due to BERT's input sequence length limitation (512 tokens), couldn't use BERT to encode 200 tweets as one sequence, so authors resort to encode each tweet separately and average them afterward to obtain a user's preference representation.
  - Generally, the tweet text is way shorter than the news text, authors empirically set the max input sequence length of BERT as 16 tokens to accelerate the tweets encoding time