**Introduction**

- **Keyphrase prediction in social media**: distill salient information from massive posts.
- **Challenges**:
  - Social media language is noisy and informal (data sparsity).
  - Prior work only extract keyphrases from the source post.

**Source post with keyphrase “super bowl”**: 
[S]: Somewhere, a wife that is not paying attention to the game, says “I want the team in yellow pants to win.”

**Relevant tweets:**
- [T1]: I been a steelers fan way before black & yellow and this super bowl!
- [T2]: I will bet you the team with yellow pants wins.
- [T3]: Wiz Khalifa song “black and yellow” to spur the pittsburgh steelers and Lil Wayne is to sing “green and yellow” for the packers.

**Our solution: topic-aware keyphrase generation**
- **Topic-aware**: latent topics learned from the corpus can alleviate the data sparsity.
- **Sequence generation**: create new keyphrases.

**Our Approach**

**Overall framework**

**Neural topic model (NTM)**

**BoW Encoder**
- Prior latent variables
  - \( \mu = f_\theta(f_\sigma(x_{bow})) \)
  - \( \log \sigma = f_\sigma(f_\theta(x_{bow})) \)

**BoW Decoder**
- Draw latent variable \( z \sim N(\mu, \sigma^2) \)
- **Topic mixture** \( \theta = softmax(f_\theta(z)) \)
- For each word \( w \in x \):
  - Draw word \( w \sim softmax(f_\theta(\theta)) \)

**Keyphrase generation (KG) model**
- **Base model**: standard seq2seq with copy mechanism.
- **Advanced**: topic-aware sequence decoder.

**Joint learning topics and keyphrases**

\[
L_{NTM} = D_{KL}(p(z) || q(z|x)) - \sum_x \log(p(y|x |x, \theta_\theta))
\]

\[
L_{KG} = - \sum_x \log(p(y|x |x, \theta_\theta)) + \gamma \cdot L_{KG}
\]

**Data Description**

**Main results**

**Data Description**

- **Source posts**: # of posts /
- **Avg len per post**: # of KP Source vocab
- **Twitter**: 40,102 / 19.52 / 1.13 / 34,100
- **Weibo**: 46,296 / 33.07 / 1.06 / 98,310
- **StackExchange**: 49,447 / 87.94 / 2.43 / 99,775

**Target KP**: Avg len per KP / % of Target vocab
- **Twitter**: 4,347 / 92 / 71.35 / 4,171
- **Weibo**: 2,136 / 2.55 / 75.74 / 2,833
- **StackExchange**: 12,114 / 1.41 / 54.32 / 10,852

**Experiment Results**

- **Main results**
  - **BoW**: 0.80 training / 0.10 validation / 0.10 test
  - **Seq2seq**: 0.80 training / 0.10 validation / 0.10 test

**Conclusion & Future Work**

- **We propose the first topic-aware keyphrase generation model** that allows end-to-end training with latent topics.
- **We newly** construct three social media datasets for this task.
- **Extensive experiments** demonstrate the effectiveness of our proposed model for social media language.

**Explore how to explicitly leverage the topic-word information**
- **Extend to other text generation tasks**

**Find our code & data**