Sentiment evolution on important topics from social media

Devanshu Agrawal

COSC 690

Motivation

- Understanding attitudes towards important topics of our time is key to communication.
  - Business, politics, art and entertainment, etc.
- Questions:
  - In what topics of conversation are people most engaged?
  - How have sentiments on these topics changed, and where are they headed?
- We assume social media reflects real attitudes.
  - Goal: We address the above questions using a Twitter data set.

Contributions

- We use latent Dirichlet allocation (LDA) to identify the most important topics of conversation on Twitter in the last five years.
- We record the monthly positive-negative sentiment ratio of tweets in each topic based on emojis.
- We show that sentiment for each topic exhibits a discernable trend over time and predict future sentiments.

Background and related work

- Topic modeling.
- Latent Dirichlet allocation (LDA).
- Sentiment analysis.
- Temporal topic-sentiment evolution.

Background and related work (cont.)

- **Note to self:** Briefly describe how LDA works.

Methodology: Data

- We used the Sentiment 140 data set of 1.6 million tweets.
- Each tweet is given a sentiment score based on its emoji content.

Note to self: Present first few rows of data set here.

Methodology: Preprocessing

- For computational efficiency, we selected a subset of $100,000$ tweets based on the distribution of tweets and sentiments over months.
- We preprocessed each tweet (e.g., removed stop words and punctuation).
- We trimmed the vocabulary to a manageable size.
- We formatted all tweets into a TFIDF matrix.

Methodology: Topic modeling and sentiment analysis

- We applied LDA to the set of tweets.
- We recorded the positive-negative sentiment ratio of tweets in each topic per month.
- We performed regression on the sentiments over time and made predictions on the past year.
- We compared the predictions to the true sentiments from the past year for evaluation.

Results: System environment

- **Note to self:** Get Jetstream system information.
- We implemented our methodology in the Spark framework.

Results: Choosing a subset of data

- We chose . . . as our subset of 100,000 tweets.

Note to self: Histograms of tweet count and sentiment.

Results: Key topics

- We ran LDA with [list parameter values].

Note to self: Figure of key topics (groups of words) along of percentage of tweets in each topic.

Results: Sentiment evolution

- (Some metric of regression for predicted sentiments).

Note to self: Figure of sentiment over time for each topic.

Lessons learned and future work

- We wanted to understand how public sentiment changes on key topics of conversation over time.
  - We applied LDA and regression.
- Key topics: [list key topics].
- Sentiments showed clear trends with [regression fit metric].
- Next steps:
  - Scale up to all 1.6 million tweets.
  - How do the key topics change over time?
  - Can we extract useful information from hashtags?