This lecture is about *sentiment analysis* and *social media text*. *Opinions* also come up frequently, since they provide a practical frame for examining sentiment.

For the purposes of this lecture you can think of a *sentiment* as an attitude, expressed in text, along a positive-negative dimension. Sentiment analysis is the measurement of the sentiments expressed by documents. The result of the measurement may be a binary variable (with values "positive" and "negative"), or it may be along a scale (e.g., "very positive", "somewhat positive", "neutral", "somewhat negative", "very negative"). Common documents in sentiment analysis include social media posts (particularly tweets since they are convenient and plentiful), product reviews, discussion forum posts, and blog posts.

An *opinion* (again, for present purposes) is an expression of a sentiment about an entity. Identifying the relevant entity in the expression of an opinion is a topic in itself, and we'll concentrate instead on identifying the sentiment.

A very simple approach to sentiment analysis is to use a word list identifying the emotional valences (a quantification of positivity or negativity) of key words, then sum the valences for each key word occurring in the document. However, we quickly run into trouble with phrases like "not bad" (which actually means good), sparsity in our word list, and the complexity of language. Sentiment analysis is often treated as a machine learning problem, with features such as n-grams, part of speech tags, and syntactic patterns in sentences.

Many obstacles exist to accurate sentiment analysis, especially in the noisy, informal domain of social media. These include:

- sentiments changing over time
- misleading sentiment associations of words ("sick" means "great" sometimes)
- language identification (a model trained on English will not work on Chinese)
- spam
- domain and topical differences

Just like in other NLP problems, we evaluate a sentiment analysis system using precision and recall.

Comparative opinions are common: they express an opinion about relative sentiment toward two different entities ("Item A is better than Item B", or "Item A is just as good as Item B"). However, not all comparisons contain opinions.