Meaning as Collective Use:
Predicting Semantic Hashtag Categories on Twitter

Lisa Posch, Claudia Wagner, **Philipp Singer**, Markus Strohmaier

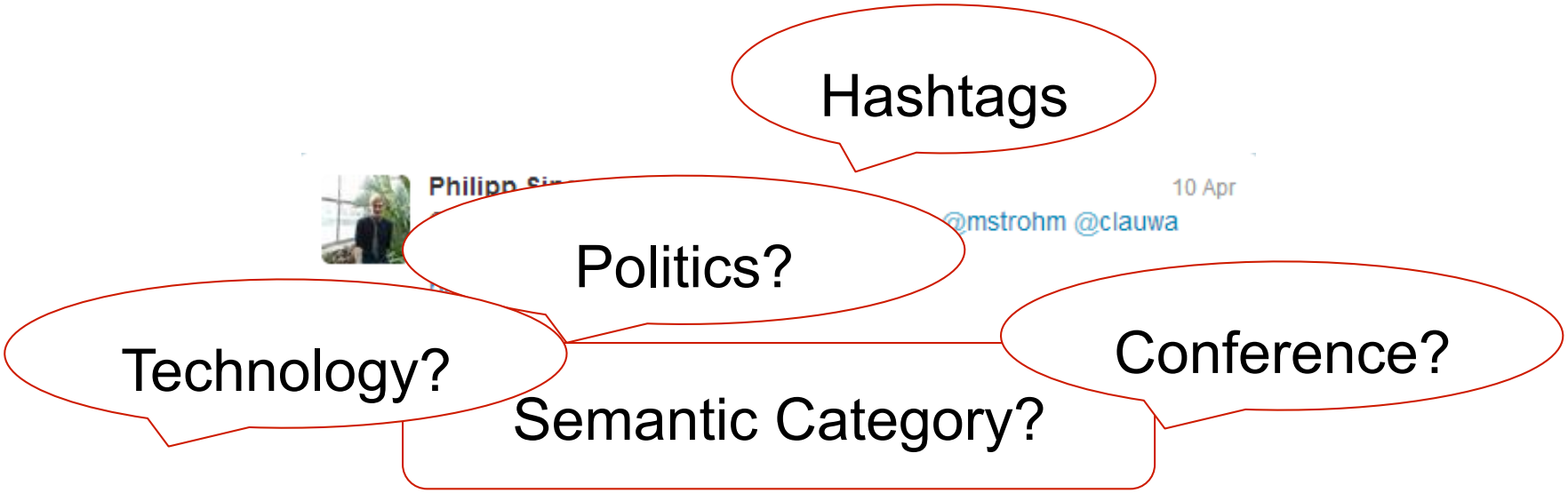
Knowledge Management Institute and Know Center
Graz University of Technology, Austria

Motivation

Twitter

Pragmatics?

Semantic Hashtag Category



Meaning is use [Wittgenstein]

Pragmatics of word: defined by the set of hashtags used by a group of users

Content: narrow lexical context of a word

Pragmatics

- structural patterns of social connections
  - Is the stream consumed by the same users that contribute to it?
  - Are social connections distributed evenly?
  - How much do the patterns change over time?
  - ...
- the structural context in which a hashtag occurs
  - How democratically is a hashtag used?
  - How conversational are tweets of a hashtag stream?
  - ...

Research Questions

1. Do different semantic categories of hashtags reveal substantially *different usage patterns*?

2. To what extent do pragmatic and lexical properties of hashtags help to *predict the semantic category* of a hashtag?

Dataset

- Twitter
- Semantic categories [Romero et al.]

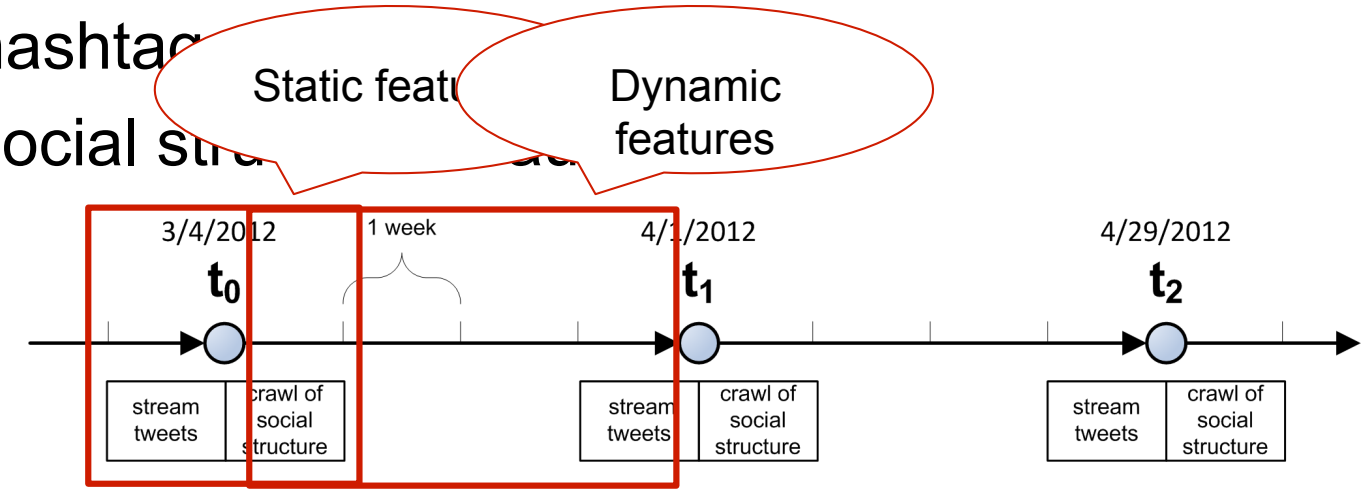
#factaboutme
#followfriday
#dontyouhate #iloveitwhen
#nevertrust
#iwish #omgfacts
#oneofmyfollowers
#rememberwhen
#wheniwaslittle

technology
idioms
sports
political
music
celebrity
movies

D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter. In Proceedings of the 20th international conference on World wide web, WWW '11, pages 695{704, New York, NY, USA, 2011. ACM.

Dataset

- three parts
- time frames of four weeks
- hashtag
- social str



D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter. In Proceedings of the 20th international conference on World wide web, WWW '11, pages 695{704, New York, NY, USA, 2011. ACM.

Features

- Static Pragmatic Measures
  - author, follower, followee, friend **entropies**
    - measure democracy of distributions
  - author-follower, -followee, -friend **overlaps**
    - measures if stream is consumed and produced by same users
  - informational, hashtag, retweet, conversational **coverages**
    - measures the nature of messages
- Dynamic Pragmatic Measures
  - **symmetric KL divergence** for authors, followers, followees, friends
    - measure how stable the social structure of a stream is
- Lexical Measure
  - **term frequency**

Do different semantic categories of hashtags reveal substantially *different usage patterns*?

Usage Patterns

- Pragmatic fingerprints
- Differences between categories
- Statistically significant?
- Pairwise comparison of categories
  - *Mann-Whitney-Wilcoxon-Test*

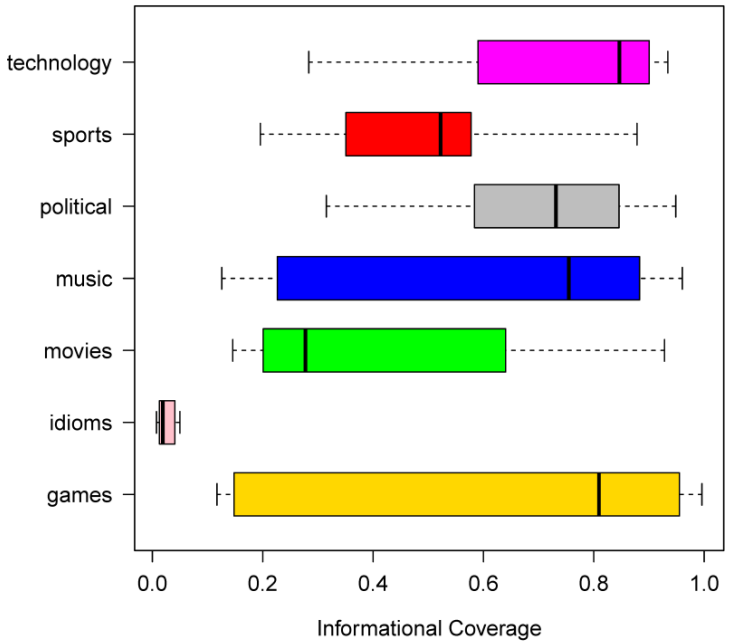
Results: Usage Patterns

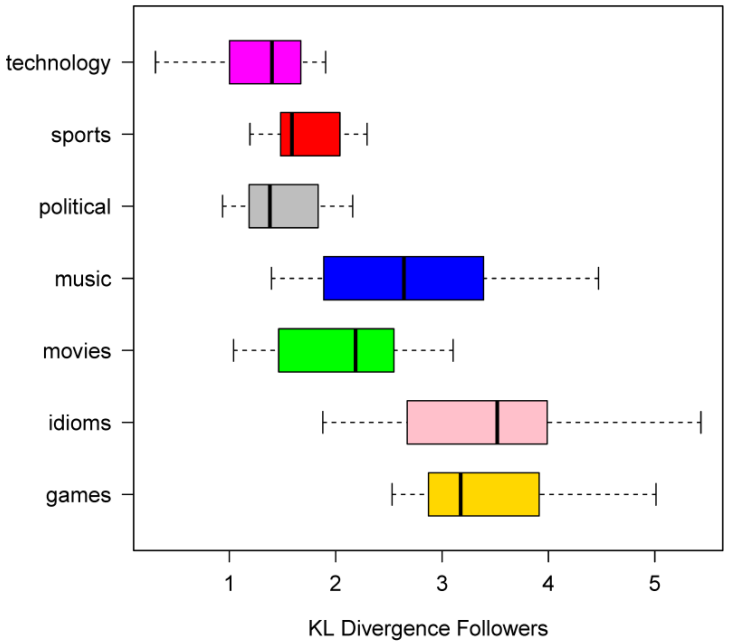
- With $p < 0.05$:
  - 26 statistical significances

	games	idioms	movies	music	political	sports
idioms	informational retweet					
movies		informational				
music		informational				
political	kl_followers	kl_authors kl_followers kl_followees informational hashtag				
sports	kl_followers	kl_authors kl_followers informational				
technology	kl_followers	kl_authors kl_followers kl_followees kl_friends informational retweet hashtag	kl_friends	kl_friends	overlap_authorfollower overlap_authorfriend	

- Best distinguishable categories: *idioms, technology*
- Most discriminative features: *informational coverage, KL divergences for followers, authors, and friends*

Results: Usage Patterns





Preliminary observations

- Pragmatic features can help to distinguish semantic categories
- Idioms and technology exhibit more distinct usage patterns than other semantic categories
- Informational coverage and KL divergence are the most discriminative features

To what extent do pragmatic and lexical properties of hashtags help to *predict the semantic category* of a hashtag?

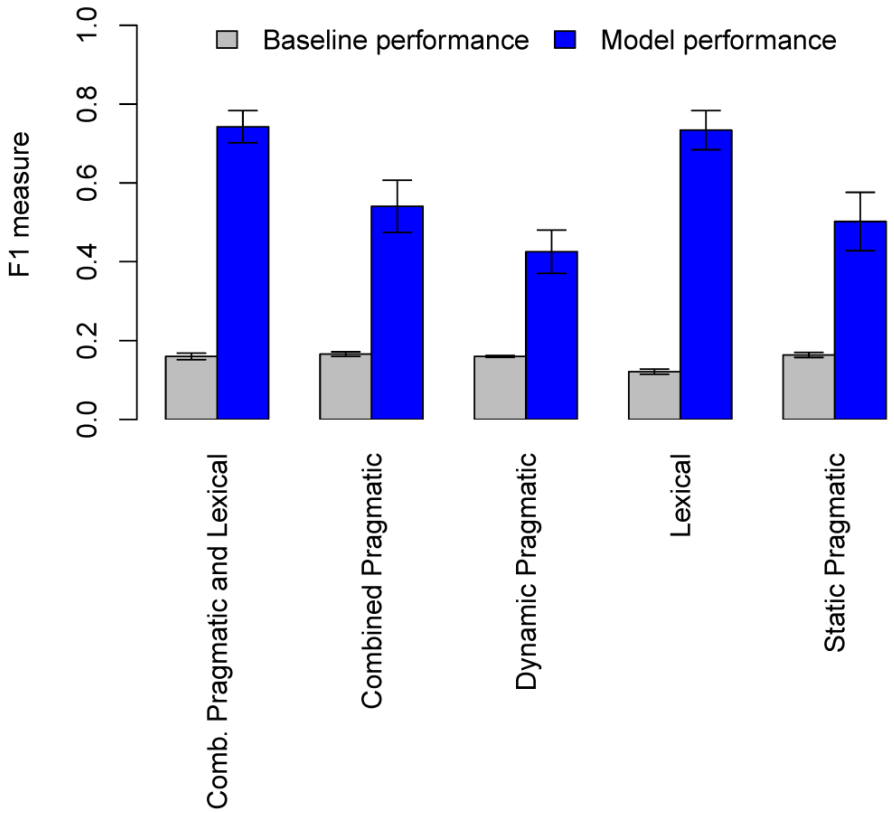
Hashtag Prediction

- Classify temporal snapshots of hashtag streams into their correct semantic categories
- By analyzing how they are used over time
- Extremely Randomized Trees
- Stratified 6-fold Cross Validation
- Baseline (randomly permuted categories 100 times)

Hashtag Prediction Models

- Static Pragmatic
- Dynamic Pragmatic
- Combined Pragmatic
- Lexical
- Combined Pragmatic and Lexical

Results: Hashtag Prediction



Results: Hashtag Prediction

- Feature Ranking

- Information Gain
  1. Informational coverage
  2. KL divergence followers
  3. KL divergence friends
  4. Hashtag coverage
  5. Friend entropy

Discussion

- Lexical features perform better
- But lexical features exhibit limitations
  - text and language dependent
  - only for settings with textual content
- Pragmatic features have advantages
  - rely on usage information
  - independent of the type of content
  - may also be computed for social video or image streams
  - multi-language corpora

Conclusions & Implications

- Collective usage of hashtags reveals information about their semantics
- Further insights necessary; especially for domains where no textual content is available
- Pragmatic features can supplement lexical features

Thanks for your attention!

Pragmatic features can play a vital role in supplementing or replacing lexical features!