# DIGITAL - Institute for Information and Communication Technologies

## The Wisdom of the Audience: An Empirical Study of Social Semantics in Twitter Streams

Claudia Wagner, Philipp Singer, Lisa Posch and Markus Strohmaier

10th Extended Semantic Web Conference, Montpellier, 29.5.2013

**Authors make their messages as informative as required but do not provide more information than necessary (Maxim of Quantity by Grice (1975))**

#music

7m

Tickets from me website, Newcastle about to sell out so be snappy :)
instagram.com/p/Zlj069Mpod/

Expand Reply Retweet Favorite More

#fashion

21m

Rafael Cennamo Is Looking For A Production Intern In NYC!
@rcennamo bit.ly/1802XpJ

View summary



[src: http://www.techweekeurope.co.uk/wp-content/uploads/2012/07/Twitter.jpg]

# Research Questions

**RQ 1: To what extent is the background knowledge of audiences useful for analyzing the semantics of social media messages?**

**RQ 2: What are the characteristics of an audience which possesses useful background knowledge for interpreting the meaning of a stream's messages and which types of streams tend to have useful audiences?**



[scr: http://www.teachthought.com/twitter-hashtags-for-teacher/]

# Methodology

- Message Classification Task
  - Use hashtags as ground truth
    - Laniado and Mika (2010) showed that around half of all hashtags can be associated with Freebase concepts
  - Compare real audience with random audience - how well can an audience predict the hashtag of a tweet?
  - The audience which is better in guessing the hashtag of a Twitter message is better in interpreting the meaning of the message
  - Null hypothesis: If the audience of a stream does not possess more knowledge about the semantics of the stream's messages than a randomly selected baseline audience, the results from both classification models should not differ significantly

# Methodology

- Train different multiclass classifiers on the background knowledge of the audience
  - Logistic Regression, Stochastic Gradient Descent, Multinomial Naive Bayes and Linear SVM
- Compare different approaches for estimating the background knowledge
  - Different audience and content selection approaches
  - Different methods for estimating the background knowledge
- Test how well each model can predict the hashtag of future messages
- Weighted Macro F1

# Dataset

- Diverse sample of hashtags

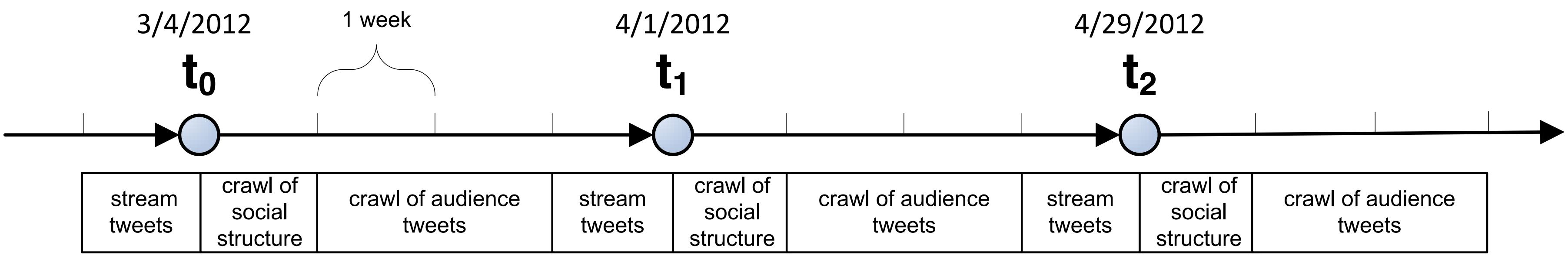
- Romero et al. (2011) identified eight categories of hashtags on a large data sample
  - *celebrity*, *games*, *idioms*, *movies/TV*, *music*, *political*, *sports*, and *technology*
- We randomly draw from each category ten hashtags which were still in use

# Dataset

| Technology | Idioms | Sports | Politics |
|---|---|---|---|
| #blackbery, #iphone, #google | #omgfacts, #factsaboutme, #iwish | #football, #nfl, #yankees | #climate, #iran, #teaparty |

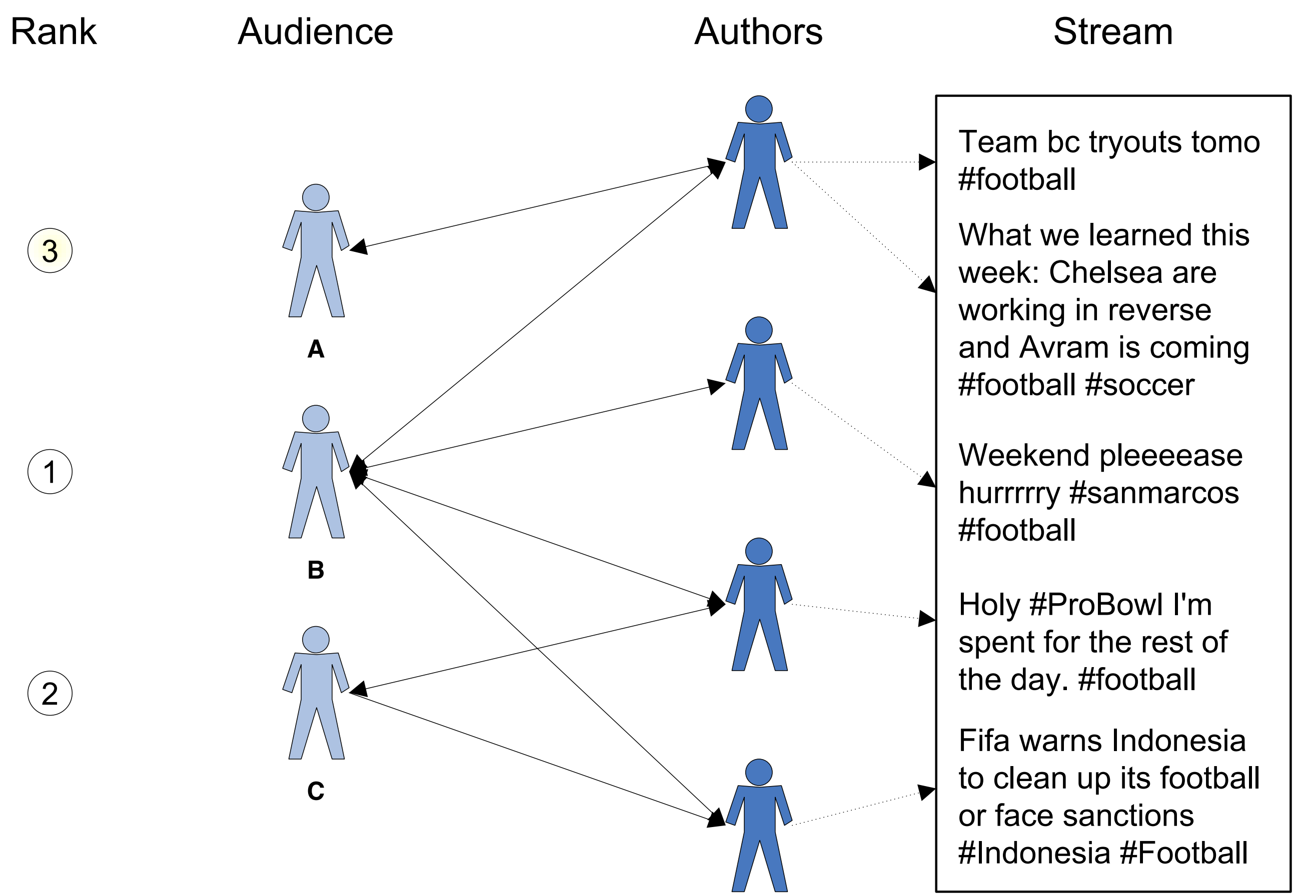
| Games | Music | Celebrity | Movies |
|---|---|---|---|
| #gaming, #mafiawars, #wow | #lastfm, #eurovision, #nowplaying | #bsb, #michaeljackson, #rogis | #avatar, #tv, #glennbeck |

# Dataset




| | t1 | t2 | t3 |
|---|---|---|---|
| Stream Tweets | 94,634 | 94,984 | 95,105 |
| Stream Authors | 53,593 | 54,099 | 53,750 |
| Friends | 7,312,792 | 7,896,758 | 8,390,143 |
| Audience Tweets | 29,144,641 | 29,126,487 | 28,513,876 |

# Audience Selection



# Background Knowledge Content Selection



- Recent
  - The most recent messages authored by the audience users
- Top Links (plain and enriched)
  - the messages authored by the audience which contain one of the top links of that audience
- Top Tags
  - the messages authored by the audience which contain one of the top hashtags of that audience

# Background Knowlegde Representation

- Preprocessing: remove stopwords, twitter syntax, stemming
- Represent background knowledge of the audience via the most likely topics or most important words of their messages
  - Bag of Words: TF and TFIDF
  - Topic Models: LDA

# Empirical Evaluation

- RQ 1: To what extent does the background knowledge of the audience support the semantic annotation of individual messages?
  - Combine audience selection and background knowledge estimation approaches to generate semantic features of the messages authored by an audience
  - Training data on audience's messages crawled at t0
  - Test model using messages of the hashtag streams crawled at t1

# Results

| | F1 (TF-IDF) | F1 (LDA) |
|---|---|---|
| Random Guessing | 1/78 | 1/78 |
| Baseline (random audience) | 0.01 | 0.01 |
| Audience – recent | 0.25 | 0.23 |
| Audience – top links enriched | 0.13 | 0.10 |
| Audience – top links plain | 0.12 | 0.10 |
| Audience – top tags | 0.24 | 0.21 |

**The audience of a hashtag stream contains knowledge which is useful for predicting the hashtags of future messages**

# Results

| | F1 (TF-IDF) | F1 (LDA) |
|---|---|---|
| celebrity | 0.17 | 0.15 |
| games | 0.25 | 0.22 |
| idioms | 0.09 | 0.05 |
| movies | 0.22 | 0.18 |
| music | 0.23 | 0.18 |
| political | 0.36 | 0.33 |
| sports | 0.45 | 0.42 |
| technology | 0.22 | 0.22 |

# Empirical Evaluation

- RQ 2: What are the characteristics of an audience which possesses useful background knowledge for interpreting the meaning of a stream's messages and which types of streams tend to have useful audiences?
  - Correlation analysis between the ability of an audience to interpret the meaning of messages and structural properties of the stream

# Structural Stream Properties



- Static Measures
  - Coverage: informational, hashtag, retweet and conversational extent of a stream
  - Entropy: randomness of a stream's authors and their followers, followees and friends
  - Overlap: overlap between authors and followers, authors  and followees and authors and friends
- Dynamic Measures
  - KL divergence between the author-, the follower-, and the friend-distributions of a stream at different time points

# Stat. Significant Spearman Rank Correlation ($p<0.05$)

| | F1 (TF-IDF) | F1 (LDA) |
|---|---|---|
| Overlap Author-Follower | 0.675 | 0.655 |
| Overlap Author-Followee | 0.642 | 0.628 |
| Overlap Author-Friend | 0.612 | 0.602 |

Streams which are produced and consumed by a community of users who are tightly interconnected tend to have a useful audience.

A useful audience possesses background knowledge which helps interpreting the meaning of messages.

# Stat. Significant Spearman Rank Correlation ($p<0.05$)

| | F1 (TF-IDF) | F1 (LDA) |
|---|---|---|
| Conversation Coverage | 0.256 | 0.256 |

Conversational streams tend to have a useful audience.

# Stat. Significant Spearman Rank Correlation ($p<0.05$)

| | F1 (TF-IDF) | F1 (LDA) |
|---|---|---|
| Entropy Author Distribution | -0.270 | -0.400 |
| Entropy Friend Distribution | -0.307 | - |
| Entropy Follower Distribution | -0.400 | -0.319 |
| Entropy Followee Distribution | -0.401 | -0.368 |

Streams which are produced and consumed by a focused set of authors, followers, followees and friends tend to have a useful audience.

# Stat. Significant Spearman Rank Correlation ($p<0.05$)

| | F1 (TF-IDF) | F1 (LDA) |
|---|---|---|
| KL Follower Distribution | -0.281 | - |
| KL Followee Distribution | -0.343 | -0.302 |
| KL Author Distribution | -0.359 | -0.307 |

Socially stable streams tend to have an audience which is good in interpreting the meaning of a stream's messages.

# Summary & Conclusions

- The audience of a social stream possesses knowledge which may indeed help to interpret the meaning of a stream's messages
- But not all streams have similar useful audiences
- The audience of a social stream seems to be most useful if the stream is created and consumed by a stable, focused and communicative community – i.e., a group of users who are interconnected and have few core users to whom almost everyone is connected
- We do not know if those relations are causal but we got similar results when repeating our experiments on t1 and t2

# Current and Future Work

- Compare the utility of ontological knowledge with audience background knowledge for the hashtag prediction task
- Algorithmic exploitation of our results
- Hybrid hashtag recommendation algorithm
  - Structural stream measures may inform weighting (how much can we count on the audience?)
  - Differentiate between social and topical hashtags
  - User-centric algorithms work only for active users who used hashtags before
    - An audience-integrated approach only requires an active audience

# References


- Grice, H. P. (1975). Logic and conversation. In Speech acts, 3, 41–58. New York: Academic Press.
- Laniado, D., & Mika, P. (2010). Making sense of twitter. In Proceedings of the 9th international semantic web conference (pp. 470-485). Shanghai, China.
- Romero, D. M., Meeder, B., & Kleinberg, J. (2011). 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 (pp. 695–704). Hyderabad, India.

THANK YOU

claudia.wagner@joanneum.at
http://claudiawagner.info

[src: http://www.crowdscience.com/2008/06/tips_and_more/]