

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

Claudia Wagner¹, Philipp Singer², Lisa Posch², and Markus Strohmaier²

¹ JOANNEUM RESEARCH, IIS, 8010 Graz, Austria
claudia.wagner@joanneum.at

² Graz University of Technology, KTI, 8010 Graz, Austria
philipp.singer@tugraz.at, lposch@sbox.tugraz.at,
markus.strohmaier@tugraz.at

Abstract. Interpreting the meaning of a document represents a fundamental challenge for current semantic analysis methods. One interesting aspect mostly neglected by existing methods is that authors of a document usually assume certain background knowledge of their intended audience. Based on this knowledge, authors usually decide what to communicate and how to communicate it. Traditionally, this kind of knowledge has been elusive to semantic analysis methods. However, with the rise of social media such as Twitter, background knowledge of intended audiences (i.e., the community of potential readers) has become explicit to some extents, i.e., it can be modeled and estimated. In this paper, we (i) systematically compare different methods for estimating background knowledge of different audiences on Twitter and (ii) investigate to what extent the background knowledge of audiences is useful for interpreting the meaning of social media messages. We find that estimating the background knowledge of social media audiences may indeed be useful for interpreting the meaning of social media messages, but that its utility depends on manifested structural characteristics of message streams.

1 Introduction

To understand the meaning of social media documents and annotate them with ontological concepts or lightweight semantic annotations is a crucial problem since social media documents tend to be short, the language used tends to be informal and new topics may arise on social media which have not been covered anywhere else before. While existing semantic analysis methods can be used to understand and model the semantics of individual social media messages to some extent, one drawback of these methods is that they are limited to analyzing the content of the document without taking the social context into account. However, social media documents are created and published in a social environment where users communicate with imagined audience [6] [7]. As we know from communication theory, e.g., the Maxim of Quantity by Grice [2] or from Speech Act Theory [11], authors or speakers usually make their messages as informative as required but do not provide more information than necessary. This suggests

that the background knowledge of an imagined audience for a given message may contribute to reveal the topics or concepts the message is about.

This paper sets out to study this hypothesis. We use three datasets obtained from Twitter, a popular microblogging service. Since information consumption on Twitter is mainly driven by explicitly defined social networks, we approximate the imagined audience of a message using the social network of its author. In addition, we estimate the collective background knowledge of an audience by using the content published by the members of the audience. While the aim of this work is not to predict who will read a message, we want to approximate the collective background knowledge of a set of key audience users of a hashtag stream who are likely to be exposed to a message and might have the background knowledge to interpret it. We do that to assess the value of background knowledge for interpreting the semantics of microblog messages. More specifically, this work addresses following research questions:

RQ1: To what extent is the background knowledge of the audience useful for guessing the meaning of social media messages? To investigate this question, we conduct a classification experiment in which we aim to classify messages into hashtag categories. As shown in [5], hashtags can in part be considered as a manually constructed semantic grounding of microblog messages. We assume that an audience which can guess the hashtag of a given message more accurately can also interpret the meaning of the message more accurately. We will use messages authored by the audience of a stream for training the classifier and we will test the performance on actual messages of a stream.

RQ2: 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? To answer this question, we introduce several measures describing structural characteristics of an audience and its corresponding social stream. Then, we measure the correlation between these characteristics and the corresponding classification performance analyzed in RQ1. This shows the extent to which useful audiences can be identified based on structural characteristics.

The results of our experiments demonstrate that the background knowledge of a stream’s audience is useful for the task of interpreting the meaning of microblog messages, but that the performance depends on structural characteristics of the audience and the underlying social stream. To our best knowledge, this is the first work which explores *to what extent* and *how* the background knowledge of an audience can be used to understand and model the semantics of individual microblog messages. Our work is relevant for researchers interested in learning semantic models from text and researchers interested in annotating social streams with lightweight semantics.

This paper is structured as follows: In Section 3 we give an overview about related research. Section 4 describes our experimental setup, including our methodology and a description of our datasets. Section 5 presents our experiments and empirical results. In Section 6 we discuss our results and conclude our work in Section 7.

2 Terminology

We define a *social stream* as a stream of data or content which is produced through users' activities conducted in an online social environment like Twitter where others see the manifestation of these activities. We assume that no explicitly defined rules for coordination in such environments exist. In this work we explore one special type of social streams, i.e., *hashtag streams*. A hashtag stream is a special type of a resource stream [13] and can be defined as a tuple consisting of users (U), messages (M), resources (R), a ternary relation (Y') and a function (ft). Specifically, it is defined as $S(R') = (U, M, R, Y', ft)$, where $Y' = \{(u, m, r) \mid r \in R' \vee \exists r' \in R', \tilde{m} \in M, u \in U : (u, \tilde{m}, r') \in Y\}$ and $R' \subseteq R$ and $Y' \subseteq Y$. In words, a hashtag stream consists of all messages containing one or several specific hashtags $r' \in R'$ and all resources (e.g., other hashtags, URLs or keywords) and users related to these messages.

In social online environments, information consumption is driven by explicitly defined social networks and therefore we can estimate the *audience* of a social stream by analyzing the incoming and outgoing links of the authors who created the stream. We call a user U_1 a *follower* of user U_2 if U_1 has established a unidirectional link with U_2 (in contrast user U_2 is a *followee* of user U_1), while we call a user U_3 a *friend* of user U_1 if U_1 has established a link with U_3 and vice versa. In this work, we assume that the union of the friends of all authors of a given hashtag constitute a hashtag stream's *audience*.

3 Related Work

Understanding and modeling the semantics of individual messages is important in order to support users in consuming social streams efficiently – e.g., via filtering social streams by users' interests or recommending tweets to users. However, one drawback of many state-of-the-art text mining approaches (such as *Bag of Words*) is that they suffer from the sparsity of microblog messages (i.e., the limited length of messages). Hence, researchers got interested in exploring those limitations and develop methods for overcoming them. Two commonly used strategies for improving short text classification are: (a) improving the classifier or feature representation and (b) using background knowledge for enriching sparse textual data.

Improving the classifier or feature representation: Sriram et al. [12] present a comparison of different text mining methods applied on individual Twitter messages. Similar to our work, they use a message classification task to evaluate the quality of the outcome of each text mining approach. Limitations of their work are that they only use five broad categories (news, opinions, deals, events and private message) in which they classify tweets. Further, they perform their experiments on a very small set of tweets (only 5,407 tweets) which were manually assigned to the aforementioned categories. Their results show that the authorship plays a crucial role since authors generally adhere to a specific tweeting pattern i.e., a majority of tweets from the same author tend to be within

a limited set of categories. However, their authorship feature requires that tweets of the same authors occur in the trainings and test dataset.

Latent semantic models such as topic models provide a method to overcome data sparsity by introducing a latent semantic layer on top of individual documents. Hong et al. [3] compare the quality and effectiveness of different standard topic models in the context of social streams and examine different training strategies. To assess the quality and effectiveness of different topic models and training strategies the authors use them in two classification tasks: a user and message classification task. Their results show that the overall accuracy for classifying messages into 16 general Twitter suggest categories (e.g., Health, Food&Drinks, Books) when using topics as features is almost twice as accurate as raw TF-IDF features. Further their results suggest that the best performance can be achieved by training a topic model on aggregated messages per user. One drawback of their work is that they only use 274 users from 16 selected Twitter suggest directories³, who may be very popular users that may be likely to mainly post messages about the assigned topic.

Enriching sparse textual data with background knowledge: In [9] the authors present a general framework to build classifiers for short and sparse text data by using hidden topics discovered from huge text and Web collections. Their empirical results show that exploiting those hidden topics improves the accuracy significantly within two tasks: “Web search domain disambiguation” and “disease categorization for medical text”. Hotho et al. [4] present an extensive study on the usage of background knowledge from WordNet for enriching documents and show that most enrichment strategies can indeed improve the document clustering accuracy. However, it is unclear if their results generalize to the social media domain since the vocabulary mismatch between WordNet and Twitter might be bigger than between WordNet and news articles.

Summary: Recent research has shown promising steps towards improving short text classification by enhancing classifiers and feature representation or by using general background knowledge from external sources to expand sparse textual data. However - to the best of our knowledge - using the background knowledge of imagined audiences to interpret the meaning of social media messages represents a novel approach that has not been studied before. The general usefulness of such an approach is thus unknown.

4 Experimental Setup

The aim of our experiments is to explore different approaches for modeling and understanding the semantics or the main theme of microblog messages using different kinds of background knowledge. Since the audience of a microblog message are the users who are most likely to interpret (or to be able to interpret) the message, we hypothesize that incorporating the background knowledge of the audience of such messages helps to better understand what a single message is about. In the following we describe our datasets and methodology.

³ <http://twitter.com/invitations/suggestions>

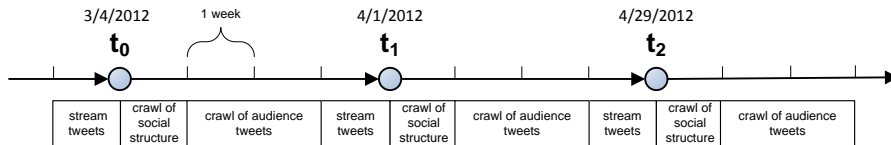


Fig. 1. Timeline of the crawling process

4.1 Datasets

In this work we use three Twitter datasets each consisting of a temporal snapshot of the selected hashtag streams, the social network of a stream’s authors, their follower and followees and the tweets authored by the selected followers and followees (see Figure 1). We generate a diverse sample of hashtag streams as follows: In [10] the authors created a classification of frequently used Twitter hashtags by category, identifying eight broad categories (see Table 1). We decided to reuse these categories and sample from each category 10 hashtags. We bias our random sample towards active hashtag streams by re-sampling hashtags for which we found less than 1,000 messages when crawling (4. March 2012). For those categories for which we could not find 10 hashtags which had more than 1,000 messages (games and celebrity) we select the most active hashtags per category. Since two hashtags #bsb and #mj appeared in the sample of two different categories, we ended up having a sample of 78 different hashtags (see Table 1).

Table 1. Randomly selected hashtags per category (ordered alphabetically)

technology	idioms	sports	political	games	music	celebrity	movies
blackberry	factaboutme	f1	climate	e3	bsb	ashleytisdale	avatar
ebay	followfriday	football	gaza	games	eurovision	brazilmisessedemi	bbcqt
facebook	dontyouhate	golf	healthcare	gaming	lastfm	bsb	bones
flickr	iloveitwhen	nascar	iran	mafiaawars	listeningto	michaeljackson	chuck
google	iwish	nba	mmot	mobsterworld	mj	mj	glee
iphone	nevertrust	nhl	noh8	mw2	music	niley	glennbeck
microsoft	omgfacts	redsox	obama	ps3	musicmonday	regis	movies
photoshop	oneofmyfollowers	soccer	politics	spymaster	nowplaying	teamtaylor	supernatural
socialmedia	rememberwhen	sports	teaparty	uncharted2	paramore	tilatequila	tv
twitter	wheniwaslittle	yankees	tehran	wow	snsd	weloveyoumiley	xfactor

Each dataset corresponds to one timeframe. The starting dates of the timeframes are March 4th (t_0), April 1st (t_1) and April 29th, 2012 (t_2). We crawled the most recent English tweets for each hashtag of our selection using Twitter’s public search API on the first day of each timeframe and retrieved tweets that were authored within the last week. During the first week of each timeframe the user IDs of the followers and followees of streams’s authors were crawled. Finally, we also crawled the most recent 3,200 tweets (or less if less were available) of all users who belong either to the top hundred authors or audience users of each hashtag stream. We ranked authors by the number of tweets they contributed to the stream and ranked audience users by the number of a stream’s authors with whom they have established a bidirectional follow relation. Figure 1 illustrates

this process. Table 2 depicts the number of tweets and relations between users that we crawled during each timeframe.

Table 2. Description of the datasets

	t_0	t_1	t_2
Stream Tweets	94,634	94,984	95,105
Audience Tweets	29,144,641	29,126,487	28,513,876
Stream Authors	53,593	54,099	53,750
Followers	56,685,755	58,822,119	66,450,378
Followees	34,025,961	34,263,129	37,674,363
Friends	21,696,134	21,914,947	24,449,705
Mean Followers per Author	1,057.71	1,087.31	1,236.29
Mean Followees per Author	634.90	633.34	700.92
Mean Friends per Author	404.83	405.09	454.88

4.2 Modeling Twitter Audiences and Background Knowledge

Audience Selection. Since the audience of a stream is potentially very large, we rank members of the audience according to the number of authors per stream an audience user is friend with. This allows us to determine key audience members per hashtag stream (see Figure 2). We experimented with different thresholds (i.e., we used the top 10, 50 and top 100 friends) and got similar results. In the remainder of the paper, we only report the results for the best thresholds (c.f., Table 3).

Background Knowledge Estimation. Beside selecting an audience of a stream, we also needed to estimate their knowledge. Hence, we compared four different methods for estimating the knowledge of a stream’s audience:

- The first method (*recent*) assumes that the background knowledge of an audience can be estimated from the most recent messages authored by the audience users of a stream.
- The second method (*top links*) assumes that the background knowledge of the audience can be estimated from the messages authored by the audience which contain one of the top links of that audience – i.e., the links which were recently published by most audience-users of that stream. Since messages including links tend to contain only few words due to the character limitations of Twitter messages (140 characters), we test two variants of this method: (1) we represented the knowledge of the audience via the plain messages which contain one of the top links (*top links plain*) and (2) (*top links enriched*) we resolved the links and enriched the messages with keywords and title information derived from meta-tags of html pages links are pointing to.
- Finally, the last method (*top tags*) assumes that the knowledge of the audience can be estimated via the messages authored by the audience which contain one of the top hashtags of that audience – i.e., the hashtags which were recently used by most audience users of that stream.

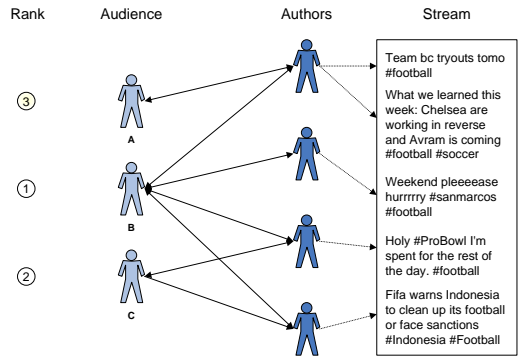


Fig. 2. To estimate the audience of a hashtag stream, we rank the friends of the stream’s authors by the number of authors they are related with. In this example, the hashtag stream #football has four authors. User B is a friend of all four authors of the stream and is therefore most likely to be exposed to the messages of the stream and to be able to interpret them. Consequently, user B receives the highest rank. User C is a friend of two authors and receives the second highest rank. The user with the lowest rank (user A) is only the friend of one author of the stream.

4.3 Methods

In this section we present the text mining methods we used to extract content features from raw text messages. In a preprocessing step we removed all English stopwords, URLs and Twitter usernames from the content of our microblog messages. We also removed Twitter syntax such as *RT* or *via*. For stemming we used Porter Stemming. In the following part of this section we describe the text mining methods we used for producing semantic annotations of microblog messages.

Bag-of-Words Model. Vector-based methods allow us to represent each microblog message as a vector of terms. Different methods exist to weight these terms – e.g., term frequency (*TF*), inverse document frequency (*IDF*) and term frequency-inverse document frequency (*TF-IDF*). We have used different weighting approaches and have achieved the best results by using *TF-IDF*. Therefore, we only report results obtained from the *TF-IDF* weighting schema in this paper.

Topic Models. Topic models are a powerful suite of algorithms which allow discovering the hidden semantic structure in large collection of documents. The idea behind topic models is to model documents as arising from multiple topics, where each document has to favor few topics. Therefore, each document exhibits different topic proportions and each topic is defined as a distribution over a fixed vocabulary of terms, where few words are favored.

The most basic topic modeling algorithm is Latent Dirichlet Allocation (LDA) [1]. In our experiments we used MALLETT’s [8] LDA implementation and fitted an

LDA model to our tweet corpus using individual tweets as trainings document. We chose the default hyperparameters ($\alpha = 50/T$, $\beta = 0.01$) and optimized them during training by using Wallach’s fixed point iteration method [14]. We chose the number of topics $T= 500$ empirically by estimating the log likelihood of a model with $T= 300, 500$ and 700 on held out data. Given enough iterations (we used 2000) the Markov chain (which consists of topic assignments z for each token in the training corpus) has potentially converged and we can get estimates of the word distribution of topics ($\hat{\phi}$) and the topic distribution of documents ($\hat{\theta}$) by drawing samples from the chain. The estimated distributions $\hat{\phi}$ and $\hat{\theta}$ are predictive distributions and are later used to infer the topics of social stream messages.

4.4 Message Classification Task

To evaluate the quality and utility of audience’s background knowledge for interpreting the meaning of microblog message, we conducted a message classification task using hashtags as classes (i.e., we had a multi-class classification problem with 78 classes). We assume that an audience which is better in guessing the hashtag of a Twitter message is better in interpreting the meaning of the message. For each hashtag stream, we created a baseline by picking the audience of another stream at random and compared the performance of the random audience with the real stream’s audience. Our baseline tests how well a randomly selected audience can interpret the meaning of a stream’s messages. One needs to note that a simple random guesser baseline would be a weaker baseline than the one described above and would lead to a performance of $1/78$.

We extracted content features (via the aforementioned methods) from messages authored by the audience of a stream before t_1 and used them to train a classifier. That means messages of the audience of a stream were used as training samples to learn a semantic representation of messages in each hashtag class. We tested the performance of the classifier on actual messages of a stream which were published after t_1 . In following such an approach, we ensured that our classifier does not benefit from any future information (e.g., messages published in the future or social relations which were created in the future). Out of several classification algorithms applicable for text classification such as Logistic Regression, Stochastic Gradient Descent, Multinomial Naive Bayes or Linear SVC, we could achieve the best results using a Linear SVC⁴. As evaluation metric we chose the weighted average *F1-score*.

4.5 Structural Stream Measures

To assess the association between structural characteristics of a social stream and the usefulness of its audience (see RQ2), we introduce the following measures which describe structural aspects of those streams. We differ between static measures which only use information from one time point and dynamic measures which combine information from several time points.

⁴ <http://www.csie.ntu.edu.tw/~cjlin/liblinear/>

Static Measures.

- **Coverage Measures:** The coverage measures characterize a hashtag stream via the nature of its messages. For example the *informational coverage* measure indicates how many messages of a stream have an informational purpose - i.e., contain a link. The *conversational coverage* measures the mean number of messages of a stream that have a conversational purpose - i.e., those messages that are directed to one or several specific users. The *retweet coverage* measures the percentage of messages which are retweets. The *hashtag coverage* measures the mean number of hashtags per message in a stream.
- **Entropy Measures:** We use normalized entropy measures to capture the randomness of a stream’s authors and their followers, followees and friends. We rank for each hashtag stream the authors by the number of tweets they authored and the followers, followees and friends by the number of authors they are related with. A high *author entropy* indicates that the stream is created in a democratic way since all authors contribute equally much. A high *follower entropy* and *friend entropy* indicate that the followers and friends do not focus their attention towards few authors but distribute it equally across all authors. A high *followee entropy* and *friend entropy* indicate that the authors do not focus their attention on a selected part of their audience.
- **Overlap Measures:** The overlap measures describe the overlap between the authors and the followers (*Author-Follower Overlap*), followees (*Author-Followee Overlap*) or friends (*Author-Friend Overlap*) of a hashtag stream. If these overlaps are one, the stream is consumed and produced by the same users who are interconnected. A high overlap suggests that the community around the hashtag is rather closed, while a low overlap indicates that the community is more open and that the active and passive part of the community do not extensively overlap.

Dynamic Measures. To explore how the social structure of a hashtag stream changes over time we measure the distance between the tweet-frequency distributions of a stream’s authors at different time points and the author-frequency distributions of a stream’s followers, followees or friends at different time points. We use a symmetric version of the *Kullback-Leibler (D_{KL}) divergence* which represents a natural distance measure between two probability distributions (A and B) and is defined as follows: $\frac{1}{2}D_{KL}(A||B) + \frac{1}{2}D_{KL}(B||A)$. The KL divergence is *zero* if the two distributions A and B are identical and approaches infinity as they differ more and more. We measure the KL divergence for the distributions of authors, followers, followees and friends.

5 Experiments

The aim of our experiments is to explore different methods for modeling and understanding the semantics of Twitter messages using background knowledge of different kinds of audiences. Due to space restrictions we only report results

obtained when training our model on the dataset t_0 and testing it on the dataset t_1 . We got comparable results when training on the dataset t_1 and testing on dataset t_2 .

5.1 RQ1: To What Extent Is the Background Knowledge of the Audience Useful for Guessing the Meaning of Social Media Messages?

To answer this question we compared the performance of a classification model using messages authored by the audience of a stream (i.e., the top friends of a hashtag stream’s authors) as training samples with the performance of a classification model using messages of a randomly selected audience (a baseline, i.e. the top friends of the authors of a randomly selected hashtag stream) as training samples. 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.

Our results show that all classifiers trained on messages authored by the audience of a hashtag stream clearly outperform a classifier trained on messages authored by a randomly selected audience. This indicates that the messages authored by the audience of a hashtag stream indeed contain important information. Our results also show that a TF-IDF based feature representation slightly outperforms a topical feature representation.

The comparison of the four different background knowledge estimation methods (see Section 4.2) shows that the best results can be achieved when using the most recent messages authored by the top 10 audience users and when using messages authored by the top 100 audience users containing one of the top hashtags of the audience (see Table 3). Tweets containing one of the top links of the audience (no matter if enriched or not) are less useful than messages containing one of the top hashtags of the audience. Surprisingly, our message link enrichment strategies did not show a large boost in performance. A manual inspection of a small sample of links showed that the top links of an audience often point

Table 3. Average weighted F1-Scores of different classification models trained on data crawled at t_0 and tested on data crawled at t_1 . We either used words weighted via TF-IDF or topics inferred via LDA as features for a message. The table shows that all audience-based classification models outperformed a random baseline. For the random baseline, we randomly swapped audiences and hashtag streams. A classifier trained on the most recent messages of the top 10 friends of a hashtag stream yields the best performance.

Classification Model	F1 (TF-IDF)	F1 (LDA)
Baseline (Random audience: top 10 friends, Messages: recent)	0.01	0.01
Audience: top 10 friends, Messages: recent	0.25	0.23
Audience: top 100 users, Messages: top links enriched	0.13	0.10
Audience: top 100 users, Message selection: top links plain	0.12	0.10
Audience: top 100 users, Message selection: top tags	0.24	0.21

to multimedia sharing sites such as youtube⁵, instagr.am⁶ or twitpic⁷. Unfortunately, title and keywords which can be extracted from the meta information of those sites often contain information which is not descriptive.

To gain further insights into the usefulness of an audience’s background knowledge, we compared the average weighted F1-Score of the eight hashtag categories from which our hashtags were initially drawn (see Table 4). Our results show that for certain categories such as sports and politics the knowledge of the audience clearly helps to learn the semantics of hashtag streams’ messages, while for other streams – such as those belonging to the categories celebrities and idioms – background knowledge of the audience seems to be less useful. This suggests that only certain types of social streams are amenable to the idea of exploiting the background knowledge of stream audiences. Our intuition is that audiences of streams that are about fast-changing topics are *less useful*. We think that these audiences are only loosely associated to the topics of the stream, and therefore their background knowledge does not add much to a semantic analysis task. Analogously, we hypothesize audiences of streams that are narrow and stable are *more useful*. It seems that a community of tightly knit users is built around a topic and a common knowledge is developed over time. This seems to provide useful background knowledge to a semantic analysis task. Next, we want to understand the characteristics that distinguish audiences that are useful from audiences that are less useful.

Table 4. Average weighted F1-Score per category of the best audience-based classifier using recent messages (represented via TF-IDF weighted words or topic proportions) authored by the top ten audience users of a hashtag stream. We got the most accurate classification results for the category *sports* and the least accurate classification results for the category *idioms*.

category	support	TFIDF		LDA	
		F1	variance	F1	variance
celebrity	4384	0.17	0.08	0.15	0.16
games	6858	0.25	0.33	0.22	0.31
idioms	14562	0.09	0.14	0.05	0.05
movies	14482	0.22	0.19	0.18	0.18
music	13734	0.23	0.25	0.18	0.26
political	13200	0.36	0.22	0.33	0.21
sports	13960	0.45	0.19	0.42	0.21
technology	13878	0.22	0.20	0.22	0.2

⁵ <http://www.youtube.com>

⁶ <http://instagram.com/>

⁷ <http://twitpic.com/>

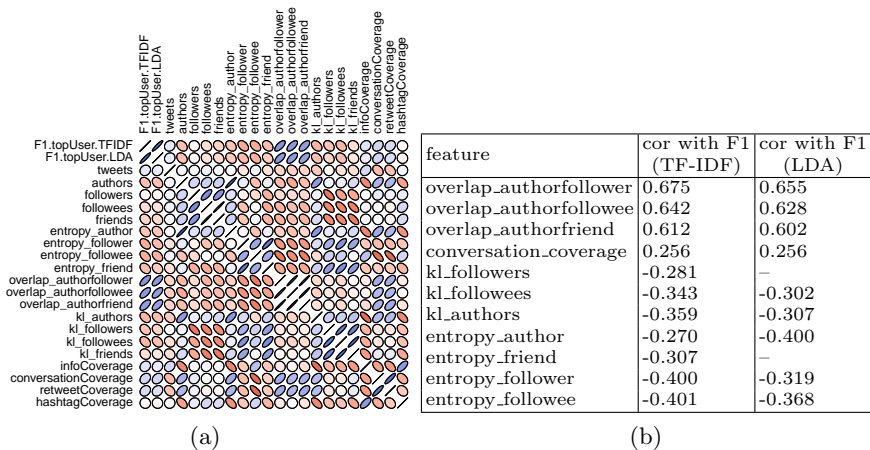


Fig. 3. This matrix shows the Spearman rank correlation strength between structural stream properties and F1-Scores of two audience-based classification models averaged across all categories. The color and form of the ellipse indicate the correlation strength. Red means negative and blue means positive correlation. The rounder the ellipse the lower the correlation. The inspection of the first two columns of the correlation matrix reveals that several structural measures are correlated with the F1-Scores and Figure 3b shows which of those are indeed statistical significant.

5.2 RQ2: What Are the Characteristics of an Audience Which Possesses Useful Knowledge for Interpreting the Meaning of a Stream’s Messages and Which Types of Streams Tend to Have Useful Audiences?

To understand whether the structure of a stream has an effect on the usefulness of its audience for interpreting the meaning of its messages, we perform a correlation analysis and investigate to what extent the ability of an audience to interpret the meaning of messages correlates with structural stream properties. We use the F1-scores of the best audience based classifiers (using TFIDF and LDA) as a proxy measure for the audience’s ability to interpret the meaning of a stream’s messages.

Figure 3a shows the strength of correlation between the F1-scores and the structural properties of streams across all categories. An inspection of the first two columns of the correlation matrix reveals interesting correlations between structural stream properties and the F1-scores of the audience-based classifiers. We further report all significant *Spearman rank correlation coefficients* ($p < 0.05$) across all categories in Figure 3b.

Figure 3a and Figure 3b show that across all categories, the measures which capture the overlap between the authors and the followers, friends and followees show the highest positive correlation with the F1-scores. That means, the higher the overlap between authors of a stream and the followers, friends and followees of the stream, the better an audience-based classifier performs. This is not sur-

prising since it indicates that the audience which is best in interpreting stream messages is an active audience, which also contributes to the creation of the stream itself (high author friend overlap). Further, our results suggest that the audience of a stream possesses useful knowledge for interpreting a stream’s messages if the authors of a stream follow each other (high author follower and author followee overlap). This means that the stream is produced and consumed by a community of users who are tightly interconnected. The only significant coverage measure is the conversational coverage measure. It indicates that the audiences of conversational streams are better in interpreting the meaning of a stream’s messages. This suggests that it is not only important that a community exists around a stream, but also that the community is communicative.

All entropy measures show significant negative correlations with the F1-Scores. This shows that the more focused the author-, follower-, followee- and/or friend-distribution of a stream is (i.e., lower entropy), the higher the F1-Scores of an audience-based classification model are. The entropy measures the randomness of a random variable. For example, the author-entropy describes how random the tweeting process in a hashtag stream is – i.e., how well one can predict who will author the next message. The friend-entropy describes how random the friends of hashtag stream’s authors are – i.e., how well one can predict who will be a friend of most hashtag stream’s authors. Our results suggest that streams tend to have a better audience if their authors and author’s followers, followees and friends are less random.

Finally, the KL divergences of the author-, follower-, and followee-distributions show a significant negative correlation with the F1-Scores. This indicates that the more stable the author, follower and followee distribution is over time, the better the audience of a stream is. If for example the followee distribution of a stream changes heavily over time, authors are shifting their social focus. If the author distribution of a stream has a high KL divergence, this indicates that the set of authors of a stream are changing over time.

In summary, our results suggest that *streams which have a useful audience tend to be created and consumed by a stable and communicative community* – i.e., a group of users who are interconnected and have few core users to whom almost everyone is connected.

6 Discussion of Results

The results of this work show that messages authored by the audience of a hashtag stream indeed represent background knowledge that can help interpreting the meaning of streams’ messages. We showed that the usefulness of an audience’s background knowledge depends on the applied content selection strategies (i.e., how the potential background knowledge of an audience is estimated). However, since the audience of a hashtag stream is potentially very large, picking the right threshold for selecting the best subset of the audience is an issue. In our experiments we empirically picked the best threshold but did not conduct extensive experiments on this issue. Surprisingly, more sophisticated content selection

strategies such as top links or top hashtags were only as good or even worse than the simplest strategy which used the most recent messages (up to 3,200) of each top audience user.

Our work shows that not all streams exhibit audiences which possess knowledge useful for interpreting the meaning of a stream’s messages (e.g., streams in certain categories like celebrities or especially idioms). Our work suggests that the utility of a stream’s audience is significantly associated with structural characteristics of the stream.

Finally, our work has certain limitations. Recent research on users’ hashtagging behavior [15] suggests that hashtags are not only used as topical or context marker of messages but can also be used as a symbol of community membership. In this work, we have mostly neglected the social function of hashtags. Although the content of a message may not be the only factor which influences which hashtag a user chooses, we assume a “better” semantic model might be able to predict hashtags more accurately.

7 Conclusions and Future Work

This work explored whether the background knowledge of Twitter audiences can help in identifying the meaning of social media messages. We introduced different approaches for estimating the background knowledge of a stream’s audience and presented empirical results on the usefulness of this background knowledge for interpreting the meaning of social media documents.

The main findings of our work are:

- The audience of a social stream possesses knowledge which may indeed help to interpret the meaning of a stream’s messages.
- The audience of a social stream is most useful for interpreting the meaning of a stream’s messages if the stream is created and consumed by a stable and communicative community – i.e., a group of users who are interconnected and have few core users to whom almost everyone is connected.

In our future work we want to explore further methods for estimating the potential background knowledge of an audience (e.g., using user lists or bio information rather than tweets). Furthermore, we want to compare our method directly to the proposed research in [4] and [9]. Combining latent and explicit semantic methods for estimating audience’s background knowledge and exploiting it for interpreting the main theme of social media messages are promising avenues for future research.

Acknowledgments. This work was supported in part by a DOC-fForte fellowship of the Austrian Academy of Science to Claudia Wagner and by the FWF Austrian Science Fund Grant I677 and the Know-Center Graz.

References

1. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.
2. H. Paul Grice. Logic and conversation. In Peter Cole, editor, *Speech acts*, volume 3 of *Syntax and semantics*, pages 41–58. Academic Press, New York, 1975.
3. Liangjie Hong and Brian D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the IGKDD Workshop on Social Media Analytics (SOMA)*, 2010.
4. Andreas Hotho, Steffen Staab, and Gerd Stumme. Wordnet improves text document clustering. In *In Proc. of the SIGIR 2003 Semantic Web Workshop*, pages 541–544, 2003.
5. David Laniado and Peter Mika. Making sense of twitter. In Peter F. Patel-Schneider, Yue Pan, Pascal Hitzler, Peter Mika, Lei Zhang, Jeff Z. Pan, Ian Horrocks, and Birte Glimm, editors, *International Semantic Web Conference (1)*, volume 6496 of *Lecture Notes in Computer Science*, pages 470–485. Springer, 2010.
6. Eden Litt. Knock, knock. who’s there? the imagined audience. *Journal of Broadcasting and Electronic Media*, 56, 2012.
7. Alice Marwick and danah boyd. I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media and Society*, 2010.
8. Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. <http://mallet.cs.umass.edu>, 2002.
9. Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on World Wide Web, WWW ’08*, pages 91–100, New York, NY, USA, 2008. ACM.
10. Daniel M. Romero, Brendan Meeder, and Jon 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.
11. John Searle. *A taxonomy of illocutionary acts*, pages 334–369. University of Minnesota Press, Minneapolis, 1975.
12. Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, SIGIR ’10*, pages 841–842, New York, NY, USA, 2010. ACM.
13. Claudia Wagner and Markus Strohmaier. The wisdom in tweetonomies: Acquiring latent conceptual structures from social awareness streams. In *Semantic Search Workshop at WWW2010*, 2010.
14. Hanna M. Wallach. *Structured Topic Models for Language*. PhD thesis, University of Cambridge, 2008.
15. Lei Yang, Tao Sun, Ming Zhang, and Qiaozhu Mei. We know what @you #tag: does the dual role affect hashtag adoption? In *Proceedings of the 21st international conference on World Wide Web, WWW ’12*, pages 261–270, New York, NY, USA, 2012. ACM.