Meaning as Collective Use: Predicting Semantic Hashtag Categories on Twitter

Lisa Posch
Graz University of Technology, Knowledge Management Institute
Inffeldgasse 13, 8010 Graz, Austria
lposch@sbox.tugraz.at

Claudia Wagner
JOANNEUM RESEARCH, Institute for Information and Communication Technologies
Steyrergasse 17, 8010 Graz, Austria
clauwa@sbox.tugraz.at

Philipp Singer
Graz University of Technology, Knowledge Management Institute
Inffeldgasse 13, 8010 Graz, Austria
philipp.singer@tugraz.at

Markus Strohmaier
Graz University of Technology, Knowledge Management Institute
Inffeldgasse 13, 8010 Graz, Austria
markus.strohmaier@tugraz.at

ABSTRACT

This paper sets out to explore whether data about the usage of hashtags on Twitter contains information about their semantics. Towards that end, we perform initial statistical hypothesis tests to quantify the association between usage patterns and semantics of hashtags. To assess the utility of pragmatic features – which describe how a hashtag is used over time – for semantic analysis of hashtags, we conduct various hashtag stream classification experiments and compare their utility with the utility of lexical features. Our results indicate that pragmatic features indeed contain valuable information for classifying hashtags into semantic categories. Although pragmatic features do not outperform lexical features in our experiments, we argue that pragmatic features are important and relevant for settings in which textual information might be sparse or absent (e.g., in social video streams).

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

Keywords

Twitter; hashtags; social structure; semantics

1. INTRODUCTION

A hashtag is a string of characters preceded by the hash (#) character and it is used on platforms like Twitter as descriptive label or to build communities around particular topics [15]. To outside observers, the meaning of hashtags is usually difficult to analyze, as they consist of short, often abbreviated or concatenated concepts (e.g., #MSM2013). Thus, new methods and techniques for analyzing the semantics of hashtags are definitely needed.

A simplistic view on Wittgenstein's work [17] suggests that *meaning is use*. This indicates that the meaning of a word is not defined by a reference to the object it denotes, but by the variety of uses to which the word is put. Therefore, one can use the narrow, lexical context of a word (i.e., its co-occurring words) to approximate its meaning. Our work builds on this observation, but focuses on the pragmatics of a word (i.e., how a word, or in our case a hashtag, is used by a large group of users) – rather than its narrow, lexical context.

The aim of this work is to investigate to what extent pragmatic characteristics of a hashtag (which capture how a large group of users uses a hashtag) may reveal information about its semantics. Specifically, our work addresses the following research questions:

- Do different semantic categories of hashtags reveal substantially different usage patterns?
- To what extent do pragmatic and lexical properties of hashtags help to predict the semantic category of a hashtag?

To address these research questions we conducted an empirical study on a broad range of diverse hashtag streams belonging to eight different semantic categories (such as *technology*, *sports* or *idioms*) which have been identified in previous research [12] and have shown to be useful for grouping hashtags. From each of the eight categories, we selected ten sample hashtags at random and collected temporal snapshots of messages containing at least one of these hashtags at three different points in time. To quantify how hashtags are used over time, we extended the set of pragmatic stream measures which we introduced in our previous work [16] and applied them to the hashtag streams in our dataset. These pragmatic measures capture not only the social structure of a hashtag at specific points in time, but also the changes in social structure over time.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

To answer the first research question, we used statistical standard tests which allow to quantify the association between pragmatic characteristics of hashtag streams and their semantic categories. To tackle the second research question, we firstly computed lexical features using a a standard bag-of-words model with term frequency (TF). Then, we trained several classification models with lexical features only, pragmatic features only and a combination of both. We compared the performance of different classification models by using standard evaluation measures such as the F1-score (which is defined as the harmonic mean of precision and recall). To get a fair baseline for our classification models, we constructed a control dataset by randomly shuffling the category labels of the hashtag streams. That means we destroyed the original relationship between the pragmatic properties and the semantic categories of hashtags.

Our results show that pragmatic features indeed reveal information about hashtags' semantics and perform significantly better than the baseline. They can therefore be useful for the task of semantically annotating social media content. Not surprisingly, our results also show that lexical features are more suitable than pragmatic features for the task of semantically categorizing hashtag streams. However, an advantage of pragmatic features is that they are language- and text-independent. Pragmatic features can be applied to tasks where the creation of lexical features is not possible – such as multimedia streams. Also for scenarios where textual content is available, pragmatic features allow for more flexibility due to their independence of the language used in the corpus. Our results are relevant for social media and semantic web researchers who are interested in analyzing the semantics of hashtags in textual or non-textual social streams (e.g., social video streams).

This paper is structured as follows: Section 2 gives an overview of related research on analyzing the semantics of tags in social bookmarking systems and research on hashtagging on Twitter in general. In Section 3 we describe our experimental setup, including our datasets, feature engineering and evaluation approach. Our results are reported in Section 4 and further discussed in Section 5. Finally, we conclude our work in Section 6.

2. RELATED WORK

In the past, a considerable effort has been spent on studying the semantics of tags (e.g., tags in social bookmarking systems), but also hashtags in Twitter have received attention from the research community.

Semantics of tags: On the one hand, researchers explored to what extent semantics emerge from folksonomies by investigating different algorithms for extracting tag networks and hierarchies from such systems (see e.g., [1], [3] or [13]). The work of [14] evaluated three state-of-the-art folksonomy induction algorithms in the context of five social tagging systems. Their results show that those algorithms specifically developed to capture intuitions of social tagging systems outperform traditional hierarchical clustering techniques. Körner et al. [5] investigated how tagging usage patterns influence the quality of the emergent semantics. They found that 'verbose' taggers (*describers*) are more useful for the emergence of tag semantics than users who use a small set of tags (*categorizers*).

On the other hand, researchers investigated to what extent tags (and the resources they annotate) can be semantically grounded and classified into predefined semantic categories. For example, Noll and Meinel [8] presented a study of the characteristics of tags and determined their usefulness for web page classification [9]. Similar to our work, Overell et al. [10] presented an approach which allows classifying tags into semantic categories. They trained a classifier to classify Wikipedia articles into semantic categories, mapped Flickr tags to Wikipedia articles using anchor texts in Wikipedia and finally classified Flickr tags into semantic categories by using the previously trained classifier. Their results show that their ClassTag system increases the coverage of the vocabulary by 115% compared to a simple WordNet approach which classifies Flickr tags by mapping them to WordNet via string matching techniques. Unlike our work, they did not take into account how tags are used, but learn relations between tags and semantic categories via mapping them to Wikipedia articles.

Pragmatics and semantics of hashtags: On Twitter, users have developed a tagging culture by adding a hash symbol (#) in front of a short keyword. The first introduction of the usage of hashtags was provided by Chris Messina in a blog post [7]. Huang et al. [4] state that this kind of new tagging culture has created a completely new phenomenon, called *micro-meme*. The difference between such micro-memes and other social tagging systems is that the participation in micro-memes is an *a-priori* approach, while other social tagging systems follow an *a-posteriori* approach. This is due to the fact that users are influenced by the observation of the usage of micro-meme hashtags adopted by other users. The work of [4] suggests that hashtagging in Twitter is more commonly used to join public discussions than to organize content for future retrieval. The role of hashtags has also been investigated in [18]. Their study confirms that a hashtag serves both as a tag of content and a symbol of community membership. Laniado and Mika [6] explored to what extent hashtags can be used as strong identifiers like URIs are used in the Semantic Web. They measured the quality of hashtags as identifiers for the Semantic Web, defining several metrics to characterize hashtag usage on the dimensions of frequency, specificity, consistency, and stability over time. Their results indicate that the lexical usage of hashtags can indeed be used to identify hashtags which have the desirable properties of strong identifiers. Unlike our work, their work focuses on lexical usage patterns and measures to what extent those patterns contribute to the differentiation between strong and weak semantic identifiers (binary classification) while we use usage patterns to classify hashtags into semantic categories.

Recently, researchers have also started to explore the diffusion dynamics of hashtags - i.e., how hashtags spread in online communities. For example the work of [15] aims to predict the exposure of a hashtag in a given time frame while [12] are interested in the temporal spreading patterns of hashtags.

3. EXPERIMENTAL SETUP

Our experiments are designed to explore to what extent pragmatic properties of hashtag streams can be used to gauge the semantic category of a hashtag. We are not only interested in the idiosyncrasies of hashtag usage within one semantic category but also in the deltas between different semantic categories. In this section, we first introduce our dataset as well as the pragmatic and lexical measures which

we used to describe hashtag streams. Then we present the methodology and evaluation approach which we used to answer our research questions.

3.1 Dataset

In this work we use data that we acquired from Twitter's API. Romero et al. [12] conducted a user study and a classification experiment and identified eight broad semantic categories of hashtags: *celebrity*, *games*, *idiom*, *movies/TV*, *music*, *political*, *sports* and *technology*. We used a list consisting of the 500 hashtags which were used by most users within their dataset and which were manually assigned to the eight categories as a starting point for creating our own dataset.

For each category, we chose ten hashtags at random (see Table 1). We biased our random sample towards active hashtag streams by re-sampling hashtags for which we found less than 1000 posts at the beginning of our data collection (March 4th, 2012). For those categories for which we could not find ten hashtags that had more than 1000 posts (i.e., *games* and *celebrity*), we selected the most active hashtags per category (i.e., the hashtags for which we found the most posts).

The dataset consists of three parts, each part representing a time frame of four weeks. The different time frames ensure that we can observe the usage of a hashtag over a given period of time. The time frames are independent of each other, i.e., the data collected at one time frame does not contain any information of the data collected at another time frame.

At the start of each time frame, we retrieved the most recent tweets in English for each hashtag using Twitter's public search API. Afterwards, we retrieved the followers and followees of each user who had authored at least one message in our hashtag stream dataset. Some pragmatic features capture information about who potentially consumes a hashtag stream (*followers*) or who potentially informs authors of a hashtag stream (*followees*) and therefore require the one-hop neighborhood of hashtag streams' authors. In this work, we call users who hold both of these roles (i.e., have established a bidirectional link with an author) *friends*. The starting dates of the time frames were March 4th (t_0), April 1st (t_1) and April 29th, 2012 (t_2). Table 2 depicts the number of tweets and relations between users that we collected during each time frame.

The stream tweets were retrieved on the first day of each time frame, fetching tweets that were authored a maximum of seven days previous to the date of retrieval. During the first week of each time frame, the user IDs of the followers and followees were collected. Figure 1 depicts this process.

Since we were interested in learning what types of characteristics are useful for describing a semantic hashtag category, we removed hashtag streams that belong to multiple

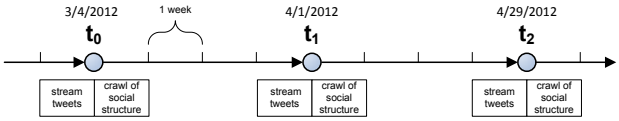


Figure 1: Timeline of the data collection process

categories (concretely, we removed the two hashtags #bsb and #mj). We also decided to remove inactive hashtag streams (those where less than 300 posts where retrieved) as estimating information theoretic measures is problematic if only few observations are available [11]. The most common solution is to restrict the measurements to situations where one has an adequate amount of data. We found four inactive hashtags in the category *games* and seven in the category *celebrity*. The removal of these hashtag streams resulted in the complete removal of the category *celebrity* as it was only left with one hashtag stream (#michaeljackson). A possible explanation for the low number of tweets in the hashtag streams for this category is that topics related to celebrities have a shorter life-span than topics related to other categories. Our final datasets consist of 64 hashtag streams and seven semantic categories which were sufficiently active during our observation period.

Table 2: Description of the complete dataset

	t_0	t_1	t_2
Tweets	94,634	94,984	95,105
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

3.2 Feature Engineering

In the following, we define the pragmatic and lexical features which we designed to capture the different social and message based structures of hashtag streams. For our pragmatic features we further differentiate between static pragmatic features (which capture the social structure of a hashtag at a specific point in time) and dynamic pragmatic features (which combine information from several time points).

3.2.1 Static Pragmatic Measures:

Entropy Measures are used to measure the randomness of streams' authors and their followers, followees and friends. For each hashtag stream, we rank the authors by the number of messages they published in that stream (norm_entropy_author) and we rank the followers (norm_entropy_follower), followees (norm_entropy_followee) and

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	brazilmissesdemi	bbcqt
facebook	dontyouhate	golf	healthcare	gaming	lastfm	bsb	bones
flickr	iloveitwhen	nascar	iran	mafiawars	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

friends (norm_entropy_friend) by the number of stream's 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 describe the overlap between the authors and the followers (overlap_authorfollower), followees (overlap_authorfollowee) or friends (overlap_authorfriend) of a hashtag stream. If overlap is *one*, all authors of a stream are also followers, followees or friends of stream authors. This indicates that the stream is consumed and produced by the same users. 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 active and passive part of the community do not extensively overlap.

Coverage Measures characterize a hashtag stream via the nature of its messages. We introduce four coverage measures. The *informational coverage* measure (informational) indicates how many messages of a stream have an informational purpose - i.e., contain a link. The *conversational coverage* (conversational) 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 (e.g., through @replies). The *retweet coverage* (retweet) measures the percentage of messages which are retweets. The *hashtag coverage* (hashtag) measures the mean number of hashtags per message in a stream.

3.2.2 Dynamic Pragmatic Measures:

To explore how the social structure of a hashtag stream changes over time, we measure the distance between the tweet-frequency distributions of authors at different time points, and the author-frequency distributions of followers, followees or friends at different time points. The intuition behind these features is that certain semantic categories of hashtags may have a fast changing social structure since new people start and stop using those types of hashtags frequently, while other semantic categories may have a more stable community around them which changes less over time.

We use a symmetric variation of the *Kullback-Leibler divergence (D_{KL})* 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 also known as *relative entropy* or *information divergence*. The KL divergence is *zero* if the two distributions are identical and approaches infinity as they differ more and more. We measure the KL divergence for the distributions of authors (kl_authors), followers (kl_followers), followees (kl_followees) and friends (kl_friends).

Figure 2 visualizes the different time frames and their notation. t_0 only contains the static features computed from data collected at t_0. Consequently, t_1 and t_2 only contain the static features computed from data collected at t_1 or t_2, respectively. $t_{0\to1}$ includes static features computed on data collected at t_0 and the dynamic measures computed on data collected at t_0 and t_1. $t_{1\to0}$ includes static features computed on data collected at t_1 and the dynamic measures computed on data collected at t_0 and t_1. $t_{1\to2}$ and $t_{2\to1}$ are defined in the same way.

3.2.3 Lexical Measures:

We use vector-based methods which allow representing each microblog message as a vector of terms and use term frequency (TF) as weighting schema. In this work lexical measures are always computed for individual time points and are therefore static measures.

3.3 Usage Patterns of Hashtag Categories

Our first aim is to investigate whether different semantic categories of hashtags reveal substantially different usage patterns (such as that they are used and/or consumed by different sets of users or that they are used for different purpose). To compare the pragmatic fingerprints of hashtags belonging to different semantic categories and to quantify the differences between categories, we conducted a pairwise *Mann-Whitney-Wilcoxon-Test* which is a non-parametric statistical hypothesis test for assessing whether one of two samples of independent observations tends to have larger values than the other. We used a non-parametric test since the *Shapiro-Wilk-Test* revealed that not all features are normally distributed, even after applying arcsine transformation to ratio measures. *Holm-Bonferroni* method was used for adjusting the p-values and counteract the problem of multiple comparisons. For this experiment, we used the timeframes $t_{0\to1}$ and $t_{1\to2}$.

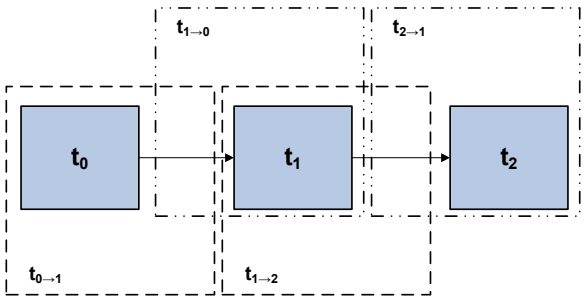


Figure 2: Illustration of our time frames

3.4 Hashtag Classification

Our second aim is to investigate to what extent pragmatic and lexical properties of hashtag streams contribute to classify them into their semantic categories. That means we aim to classify temporal snapshots of hashtag streams into their correct semantic categories (to which they were assigned in [12]) just by analyzing how they are used over time. We then compare the performance of the pragmatically informed classifier with the performance of a classifier informed by lexical features within a semantic multiclass classification task. We used the timeframes $t_{1\to0}$ and $t_{2\to1}$ for this experiment in order to avoid including information from the 'future' in our classification.

We performed grid search with varying hyperparameters using Support Vector Machine (linear and RBF kernels) and an ensemble method with extremely randomized trees. Since extremely randomized trees are a probabilistic method and perform slightly different in each run, we run them ten times and report the average scores. The features were standardized by subtracting the mean and scaling to unit variance. We used stratified 6-fold cross-validation (CV) to train and test each classification model.

Since we have two different types of pragmatic features, static and dynamic ones, we trained and tested three separate classification models which were only informed by pragmatic information:

Static Pragmatic Model: We trained and tested this classification model with static pragmatic features on data collected at t_1 using stratified 6-fold CV. The experiment was repeated on the data collected at t_2.

Dynamic Pragmatic Model: We trained and tested the classification model with dynamic pragmatic features on data collected at t_0 and t_1 using stratified 6-fold CV. The computation of our dynamic features requires at least two time points. We repeated this experiment on data collected at t_1 and t_2.

Combined Pragmatic Model: We combined the static and dynamic pragmatic features, and trained and tested the classification model on the data of $t_{1\to0}$ using stratified 6-fold CV. Again, we repeated the experiment on the data of $t_{2\to1}$.

We also performed our classification with a model using our lexical features (i.e., TF weighted words). Finally, we trained and tested a combined classification model using pragmatic and lexical features, which leads to the following classification models:

Lexical Model: We trained and tested the model on data from t_1 using stratified 6-fold CV and repeated the experiment on data collected at t_2.

Combined Pragmatic and Lexical Model: We trained and tested the mixed classifier on the data of $t_{1\to0}$ using stratified 6-fold CV, then repeated this experiment for $t_{2\to1}$. A simple concatenation of pragmatic and lexical features is not useful, since the vast amount of lexical features would overrule the pragmatic features. Therefore, we used a stacking method (see [2]) and performed firstly a classification using lexical features alone and Leave-One-Out cross-validation. We used a SVM with linear kernel for this classification since it worked best for these features. Secondly, we combined the pragmatic features with the resulting seven probability features which we got from the previous classification model and which describe how likely each semantic class is for a certain stream given its words.

To get a fair baseline for our experiment, we constructed a control dataset by randomly shuffling the category labels of the 64 hashtag streams. That means we destroyed the original relationship between the pragmatic properties and the semantic categories of hashtags and evaluated the performance of a classifier which tries to use pragmatic properties to classify hashtags into their shuffled categories within a 6-fold cross-validation. We repeated the random shuffling 100 times and used the resulting average F1-score as our baseline performance. For the baseline classifier we also used grid search to determine the optimal parameters prior to training. Our baseline classifier tests how well randomly assigned categories can be identified compared to our real semantic categories. 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 1/7.

To gain further insights into the impact of individual properties, we analyzed their information gain (IG) with respect to the categories. The information gain measures how accurately a specific stream property P is able to predict stream's category C and is defined as follows: $IG(C, P) = H(C) - H(C \mid P)$ where H denotes the entropy.

4. RESULTS

In the following section, we present the results from our empirical study on usage patterns of different semantic categories of hashtags.

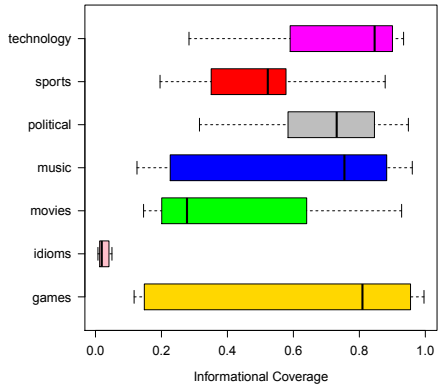
4.1 Usage Patterns of Hashtag Categories

To answer our first research question, we explored to what extent usage patterns of hashtag streams in different semantic categories are indeed significantly different.

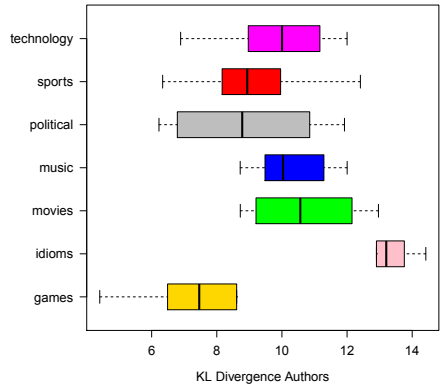
Our results indicate that some pragmatic measures are indeed significantly different for distinct semantic categories. This indicates that hashtags of certain categories are used in a very specific way which may allow us to relate these hashtags with their semantic categories just by observing how users use them. Table 3 depicts the measures that show statistically significant ($p < 0.05$) differences in both $t_{0\to1}$ and $t_{1\to2}$. In total, 35 pragmatic category differences were found to be statistically significant (with $p < 0.05$) for $t_{0\to1}$ and 33 for $t_{1\to2}$. 26 pragmatic category differences were found to be significant for both $t_{0\to1}$ and $t_{1\to2}$ which suggests that results are independent of our choice of time frame.

Table 3: Features which showed a statistically significant difference (with $p < 0.05$) for each pair of categories in both $t_{0\to1}$ and $t_{1\to2}$

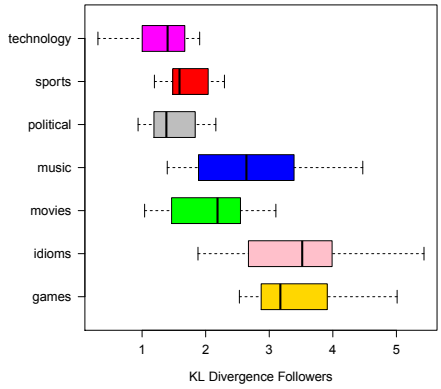
	games	**idioms**	**movies**	**music**	**political**	**sports**
idioms	informational retweet					
movies		informational				
music		informational				
political	kl_followers	kl_authors kl_followers kl_followees informational hashtag				
sports	kl_followers	kl_authors kl_followers informational				
technology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



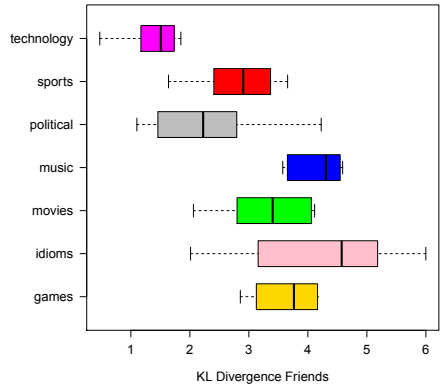
(a) This figure shows the percentage of messages of hashtag streams belonging to different categories that contain at least one link.



(b) This figure shows how much the authors' tweet-frequency distributions of hashtag streams of different categories change on average.



(c) This figure shows how much the followers' author-frequency distributions of hashtag streams of different categories change on average.

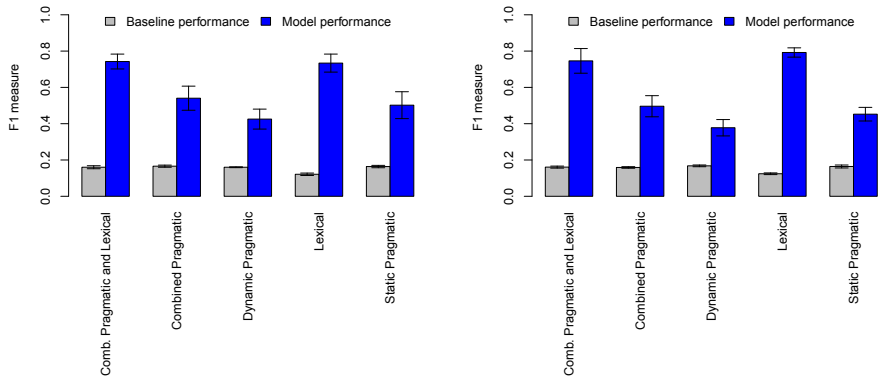


(d) This figure shows how much the friends' author-frequency distributions of hashtag streams of different categories change on average.

Figure 3: Each plot shows the feature distribution of different categories of one of the 4 best pragmatic features for $t_{0\to1}$. We obtained similar results for $t_{1\to2}$.

Not surprisingly, the category which shows the most specific usage patterns is *idioms* and therefore the hashtags of this category can be distinguished from all hashtags just by analyzing their pragmatic properties. Hashtag streams of the category *idioms* exhibit a significantly lower informational coverage than hashtag streams of all other categories (see Figure 3(a)) and a significantly higher symmetric KL divergence for author's tweet-frequency distributions (see Figure 3(b)). Also the followers' and friends' author-frequency distributions tend to have a higher symmetric KL divergence for idioms hashtags than for other hashtags (see Figures 3(c) and 3(d)). This indicates that the social structure of hashtag streams in the category *idioms* changes faster than hashtags of other categories. Furthermore, hashtag streams of this category are less informative - i.e., contain significantly less links per message on average.

The category *technology* can be distinguished from all other categories except *sports*, particularly because its followers' and friends' author-frequency distributions have significantly lower symmetric KL divergences than hashtags in the categories *games*, *idioms*, *movies* and *music* (see Figures 3(c) and 3(d)). This indicates that hashtag streams in the category *technology* have a stable social structure which changes less over time. This is not surprising since this semantic category denotes a topical area and users who are interested in such areas may consume and provide information on a regular base. It is especially interesting to note that the only pragmatic measures which allows distinguishing political and technological hashtag streams are the author-follower and author-friend overlaps since these overlaps are significantly lower for the category *technology* compared to the category *political*.



(a) $t_{1\to0}$ (b) $t_{2\to1}$

Figure 4: Weighted averaged F1-scores of different classification models trained and tested on $t_{1\to0}$ 4(a) and $t_{2\to1}$ 4(b) using 6-fold cross-validation

This indicates that the content of hashtag streams of the category *political* is far more likely to be produced and consumed by the same people than content of technological hashtag streams.

Comparing the individual measures reveals that the informational coverage (six category pairs) and the symmetric KL divergences of followers' author-frequency distributions (six category pairs), authors' tweet-frequency distributions (three pairs) and friends' follower-frequency distributions (three pairs) are the most discriminative measures. Figure 3 depicts the distributions of these four measures per category. Other measures that show significant differences in medians for both $t_{0\to1}$ and $t_{1\to2}$ are the symmetric KL divergence of followees' author-frequency distributions (two pairs), the author-follower and the author-friend overlap (one pair) as well as the retweet and hashtag coverage (two pairs).

Some measures like the conversational coverage measure did not show any significant differences for any of the category pairs, for any time frame. This indicates that in all hashtag streams an equal amount of conversational activities take place.

4.2 Hashtag Classification

In order to quantify the value of different pragmatic and lexical properties of hashtag streams for predicting their semantic category, we conducted a hashtag stream classification experiment and systematically compared the performance of various classification models trained with different sets of features.

Figure 4 shows the performance of the best classifier (extremely randomized trees) trained with different sets of features. One can see from this figure that in general lexical features perform better than pragmatic features, but also that pragmatic features (both static and dynamic) significantly outperform a random baseline. This indicates that pragmatic features indeed reveal information about a hashtag's meaning, even though they do not match the performance of lexical features in this case. In 4(a) we can see that for $t_{1\to0}$ the combination of lexical and pragmatic features performs slightly better than using lexical features alone.

Table 4: Top features for two different datasets ranked via Information Gain

Rank	$t_{1\to0}$	$t_{2\to1}$
1	informational	kl_followers
2	kl_followers	informational
3	kl_friends	hashtag
4	hashtag	kl_followees
5	norm_entropy_friend	kl_friends

4.2.1 *Feature Ranking:*

In addition to the overall classification performance which can be achieved solely based on analyzing the pragmatics of hashtags, we were also interested in gaining insights into the impact of individual pragmatic features. To evaluate the individual performance of the features we used information gain (with respect to the categories) as a ranking criterion. The ranking was performed on $t_{1\to0}$ and $t_{2\to1}$ with stratified 6-fold cross-validation. Table 4 shows that the top five features (i.e., the pragmatic features which reveal most about the semantic of hashtags) are features which capture the temporal dynamics of the social context of a hashtag (i.e., the temporal follower, followees and friends dynamics) as well as the informational and hashtag coverage. This indicates that the collective purpose for which a hashtag is used (i.e., if it used to share information rather than for other purposes) and the social dynamics around a hashtags – i.e., who uses a hashtag for whom – play a key role in understanding its semantics.

5. DISCUSSION OF RESULTS

Although our results show that lexical features work best within the semantic classification task, those features are text and language dependent. Therefore, their applicability is limited to settings where text is available. Pragmatic features on the other hand rely on usage information which is independent of the type of content which is shared in social streams and can therefore also be computed for social video or image streams.

We believe that pragmatic features can supplement lexical features if lexical features alone are not sufficient. In our experiments, we could see that the performance may slightly increase when combining pragmatic and lexical features. However, the effect was not significant. We think the reason for this is that in our setup lexical features alone already achieved good performance.

The classification results coincide with the results of the statistical significance tests. Ranking the properties by information gain showed that the most discriminative properties (the ones that showed a statistical significance in both $t_{0\to1}$ and $t_{1\to2}$ for the highest amount of category pairs) found in 4.1 were also the top ranked features (informational coverage and the KL divergences).

6. CONCLUSIONS AND IMPLICATIONS

Our work suggests that the collective usage of hashtags indeed reveals information about their semantics. However, further research is required to explore the relations between usage information and semantics, especially in domains where limited text is available. We hope that our research is a first step into this direction since it shows that hashtags of different semantic categories are indeed used in different ways.

Our work has implications for researchers and practitioners interested in investigating the semantics of social media content. Social media applications such as Twitter provide a huge amount of textual information. Beside the textual information, also usage information can be obtained from these platforms and our work shows how this information can be exploited for assigning semantic annotations to textual data streams.

7. ACKNOWLEDGEMENTS

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

8. REFERENCES

[1] D. Benz, C. Körner, A. Hotho, G. Stumme, and M. Strohmaier. One tag to bind them all: Measuring term abstractness in social metadata. In *Proc. of 8th Extended Semantic Web Conference ESWC 2011, Heraklion, Crete, (May 2011)*, 2011.

[2] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[3] P. Heymann and H. Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Computer Science Department, April 2006.

[4] J. Huang, K. M. Thornton, and E. N. Efthimiadis. Conversational tagging in twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, HT '10, pages 173–178, New York, NY, USA, 2010. ACM.

[5] C. Körner, D. Benz, M. Strohmaier, A. Hotho, and G. Stumme. Stop thinking, start tagging - tag semantics emerge from collaborative verbosity. In *Proceedings of the 19th International World Wide Web Conference (WWW 2010)*, Raleigh, NC, USA, Apr. 2010. ACM.

[6] D. Laniado and P. Mika. Making sense of twitter. In P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, and B. Glimm, editors, *International Semantic Web Conference (1)*, volume 6496 of *Lecture Notes in Computer Science*, pages 470–485. Springer, 2010.

[7] C. Messina. Groups for twitter; or a proposal for twitter tag channels. `http://factoryjoe.com/blog/2007/08/25/groups-for-twitter-or-a-proposal-for-twitter-tag-channels/`, 2007.

[8] M. G. Noll and C. Meinel. Exploring social annotations for web document classification. In *Proceedings of the 2008 ACM symposium on Applied computing*, SAC '08, pages 2315–2320, New York, NY, USA, 2008. ACM.

[9] M. G. Noll and C. Meinel. The metadata triumvirate: Social annotations, anchor texts and search queries. In *Web Intelligence*, pages 640–647. IEEE, 2008.

[10] S. Overell, B. Sigurbjörnsson, and R. van Zwol. Classifying tags using open content resources. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 64–73, New York, NY, USA, 2009. ACM.

[11] S. Panzeri and A. Treves. Analytical estimates of limited sampling biases in different information measures. *Network: Computation in Neural Systems*, 7(1):87–107, 1996.

[12] 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.

[13] P. Schmitz. Inducing ontology from flickr tags. In *Proceedings of the Workshop on Collaborative Tagging at WWW2006*, Edinburgh, Scotland, May 2006.

[14] M. Strohmaier, D. Helic, D. Benz, C. Körner, and R. Kern. Evaluation of folksonomy induction algorithms. *Transactions on Intelligent Systems and Technology*, 2012.

[15] O. Tsur and A. Rappoport. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 643–652, New York, NY, USA, 2012. ACM.

[16] C. Wagner and M. Strohmaier. The wisdom in tweetonomies: Acquiring latent conceptual structures from social awareness streams. In *Semantic Search Workshop at WWW2010*, 2010.

[17] L. Wittgenstein. *Philosophical Investigations*. Blackwell Publishers, London, United Kingdom, 1953. Republished 2001.

[18] L. Yang, T. Sun, M. Zhang, and Q. 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.