The Utility of Social and Topical Factors in Anticipating Repliers in Twitter Conversations

Johannes Schantl
JOANNEUM RESEARCH
Graz, Austria
johannes.schantl@student.tugraz.at

Claudia Wagner
JOANNEUM RESEARCH
Graz, Austria
claudia.wagner@joanneum.at

Rene Kaiser
JOANNEUM RESEARCH
Graz, Austria
rene.kaiser@joanneum.at

Markus Strohmaier
University of Technology
Graz, Austria
markus.strohmaier@tugraz.at

ABSTRACT

Anticipating repliers in online conversations is a fundamental challenge for computer mediated communication systems which aim to make textual, audio and/or video communication as natural as face to face communication. The massive amounts of data that social media generates has facilitated the study of online conversations on a scale unimaginable a few years ago. In this work we use data from Twitter to explore the predictability of repliers, and investigate the factors which influence who will reply to a message. Our results suggest that social factors, which describe the strength of relations between users, are more useful than topical factors. This indicates that Twitter users' reply behavior is more impacted by social relations than by topics. Finally, we show that a binary classification model, which differentiates between users who will and users who will not reply to a certain message, may achieve an F1-score of 0.74 when using social features.

Author Keywords

Twitter, social media communication, reply behavior, reply prediction

ACM Classification Keywords

J.4 Computer Applications: Social and Behavioral Sciences

INTRODUCTION

Social media platforms like Twitter or Facebook are used for interacting and communicating with other users. Many different kinds of conversations, ranging from informal chats to formal discussions, can emerge on these platforms. The massive amounts of data that social media generates has facilitated the study of online conversations on a scale unimaginable a few years ago.

Identifying patterns in online conversations is important for at least two reasons: First, such patterns can be incorporated into the design of online conversation tools (e.g. *orchestrated* video communication systems as described in [9]) and social media services. Second, such patterns can provide an empirical test of social theoretical models that have been proposed in the literature (see e.g. [12]). Therefore, this work sets out to explore patterns in online conversations and investigates the predictability of repliers in Twitter.

When it comes to the theoretical study of online conversations, a natural assumption would be that the closer the friendship between two users A and B, the more likely user A replies to a message of user B and vice versa. A competing hypothesis would be that conversations are driven by topical factors rather than social factors, and that therefore the probability of user A replying to user B increases with their topical similarity – i.e., with the extent to which they talk about the same topics.

In this work, we aim to explore these two competing hypothesis and investigate the following research questions:

- RQ1: To what extent is communication of Twitter users influenced by social and topical factors?
- RQ2: To what extent are repliers on Twitter predictable?

To this end, we measure the predictability of users' reply behavior in Twitter conversations. We propose a comprehensive set of features to quantify the major social and topical factors which may impact users' communication behavior. In addition to topical and social factors we also add activity features (e.g. number of tweets, number of replies or number of followers) as covariates which describe how active, how communicative and how popular a user is on Twitter. We decided to add activity features since we are interested in exploring to what extent social and topical features help predicting repliers above and beyond the effects of activity features.

To address our research questions, we constructed a dataset consisting of user pairs $\langle a, c \rangle$ where either a user c saw a message m authored by user a and replied to it (positive samples), or where a user c saw a message m authored by user a and

WebSci'13, May 2–4, 2013, Paris, France.

did not reply to it (negative samples). In this work we use the variable a to refer to the user who authored the start message of a conversation and the variable c to refer to a potential reply candidate.

Gathering the aforementioned negative samples is obviously difficult since no factual data is available on which tweet has been read by which users. Finding out who has seen a certain message would require approximating unobservable variables such as the time a user spends reading messages which are shown on his/her Twitter timeline, the number of messages which are published on his/her timeline every minute and the extent to which users consume tweets which are not shown on their timeline (e.g. via using Twitter search). In this work we use a simplification and assume that the followers of a user are those users who are likely to see a message authored by this user.

Given this dataset, we first examine which features may have the potential to differentiate between users who see a certain message and reply to it, and users who see the same message but do not reply to it, by conducting statistical hypothesis tests. The null hypothesis states that the users who see the message and reply, and the users who see the message and do not reply, do not differ significantly, i.e., the feature distributions of both user groups are similar. Further, to assess the predictive power of individual features, we conduct a logistic regression analysis using positive and negative user-message pairs as samples. In addition to analyzing the statistically significant coefficients which reveal information about the impact of individual features, we also test the predictive power of the logistic regression model using a 10-fold cross validation.

Our results are in line with results from previous research [18] and suggest that on Twitter social features, which describe the strength of the relation between users, are more useful than topical features for predicting if a user will reply to another user or not. This suggests that conversations on Twitter might be more driven by social relations than by topics. Further, our results show that a binary classification model which aims to differentiate between users who will and users who will not reply to a certain message of another user may achieve an F1-score of 0.76.

This paper is structured as follows: In the next section we introduce some basic terminology used within our work and provide some background information about Twitter. In the *Related Work* Section we discuss research about the nature and the predictability of online conversations in social media applications. In the *Experimental Setup* Section we present our dataset, features and methodology. Our results are described in the *Results* Section. We conclude this work by drawing final conclusions in *Conclusion and Further Work*.

BACKGROUND AND TERMINOLOGY

Twitter was launched in 2006 and is one of the most popular microblogging services in the world. Users may write short messages, called *tweets*, which are limited to 140 characters. Information consumption on Twitter is mainly driven by explicitly defined social networks. That means, a user sees the messages authored by the users he/she follows on their Twitter timeline in reverse chronological order. We call a user u_1 a *follower* of user u_2 if u_1 has established a follow relation with u_2. In the same example, user u_2 is a *followee* of user u_1. We call a user u_3 a *friend* of user u_1 if u_1 has established a follow relation with u_3 and vice versa.

In this work we define a conversation as an interaction between at least two users, consisting of at least two messages, the original start message and the reply message. The Twitter API provides information which allows reconstructing conversation threads since for each message which is a reply, the ID of the message to which it is replying can be retrieved. Therefore, one can recursively find for any reply message the original start message. However, it is not possible to find the end of a conversation without using a temporally restricted definition of a conversation. In our work we therefore decided to predict only the first user who replied to a message rather than all users who replied to it since it is impossible to know if more users will reply to the message in the future. Further, 89% of conversations in our dataset consist of only two users and therefore predicting the first user who replies is most times equal to predicting all users who will join a conversation.

RELATED WORK

Previous research has focused on exploring how users use Twitter in general, and to what extent this platform is used for communication purposes. For example, in one of the first papers about Twitter usage intention, Java et al. [8] found that Twitter is often used for discussing events of daily life, sharing information or URLs, reporting news and for conversations, which we focus on in this study. Java et al. show that 21% of Twitter users participate in conversations, and 1/8 of all Twitter messages are part of conversations. They use the @*mention* sign as indicator for a conversation.

Macskassy et al. [10] show that 92% of dialogues are between two people and that the average number of messages in dialogues is less than 5 tweets. Honeycutt and Herring [7] evaluate conversations in Twitter and give insight about the nature of the @*mention* usage. They found that @*mention* is used in 90.96% for addressivity reasons, and that the median/mean number of users participating in conversations is 2/2.5. Naaman et al. [13] developed a content based categorization system for Twitter messages and found that most users focus on themselves (so-called “meformers”) while less users are “informers”.

Understanding the nature and dynamics of conversations on social media applications like Twitter was also subject of previous studies. For example, in [1] the authors explore the problem of predicting directed communication intention between users who did not communicate with each other before. The authors use various network and content features and conduct a link prediction experiment to assess the predictive power of those features. Their work focuses on predicting only new communication links between users, while our work aims to predict who will reply to a certain message of a certain author no matter if the user has communicated with the author before or not.

Most similar to our work is the work of [18] which explores if the reply behavior of users is mainly driven by topical or social factors. Similar to our findings their findings suggest that social factors are on average more important. For users with larger and denser ego-centric networks, they observed a slight tendency for separating their connections depending on the topics discussed. Unlike our work, their work focuses on three broad topics (sport, religion and politics) and therefore they only analyze the replies of messages which belong to one of these topics. Further, their work focuses on Portuguese tweets while we focus on English tweets. Finally, their work uses a different approach for addressing the same research question as we do. For each pair of topics, they analyze and compare the ego-centric networks of users who have replied to messages from both topics, while we use topical and social features to fit a regression model using user-pairs as observations and the reply-status of a user as dependent binary variable.

Wang and Huberman [20] study the predictability of online interactions both at the group and individual level. They measure the predictability of online user behavior by using information-theoretic methods applied to real time data of online user activities from Epinions, a who-trust-whom consumer review site and Whrrl, a location based online social network game. Their work shows that the users' interaction sequences have strong deterministic components. In addition, they show that individual interactions are more predictable when users act on their own rather than when attending group activities. The work presented in [2] describes an approach for recommending interesting conversations to Twitter users. They are using topic and tie strength between users and preferred thread length as factors to recommend conversations. Their approach gives interesting insights about which conversations different types of users prefer but they don't take into account if the users are also willing to join a conversation.

Research about predicting social links in online social networks is also related to our research about predicting communication links. For example, Rowe et al. [15] study the follow behavior of Sina Weibo users and found that the users' follow behavior is more driven by topical than by social factors. In [16] the authors present an approach that allows inferring social links between users by considering patterns in friendship formation, the content of people's messages and user location. Unlike the aforementioned work, our work solely focuses on communication links rather than on social links (i.e. follower relations). In addition to predicting the existence of social links, researchers also started being interested in predicting the strength of a link. Gilbert et al. [4] try to classify social relations in Facebook into strong and weak ties, referring to user with strong social relation and users with weak social relation. In [3] the authors apply the same approach to Twitter, and found that their Facebook tie strength model largely generalizes to Twitter.

Related to users' reply behavior is also users' retweet behavior and users' question answering behavior. The work of [11] explores the retweet behavior of Twitter users. They present four retweeting models (general model, content model, homophily model, and recency model) and found that content based propagation models were better at explaining the majority of retweet behaviors in their data. That means in contrast to our work they found that content and topics drive the retweet behavior of Twitter users, while we found that the reply behavior is more driven by social factors. Paul et al. [14] conducted a study of question asking and answering behavior on Twitter. They examined what characteristics of the asker might improve his/her chances of receiving a response. They found that the askers' number of followers and their Twitter account age are good predictors of whether their questions will get answered. However, the number of tweets the asker had posted or his/her frequency of use of Twitter do not predict whether his/her question will get answered. Finally, they examined the relationship between asker and replier and found that 36% of relationships are reciprocal and 55% are one-way. Surprisingly, 9% of answerers are not following the askers. Paul et al. focus on one specific type of message, namely questions, while our work is not limited to any message type. Further, they explore characteristics of the questions and the askers in order to predict the number of answers a question will receive, while we are interested in exploring characteristics of user pairs in order to predict if they will communicate with each other or not.

EMPIRICAL STUDY

The aim of our empirical study is to explore how predictable repliers are on Twitter and to what extent users' reply behavior is driven by topical and social factors. In the following Section we describe our experimental setup – i.e., we describe our dataset, features and methodology.

Dataset and Sample Generation

To obtain a random sample of Twitter conversations we firstly crawled Twitter's public timeline[1] by using its publicly available API, and filtered English tweets[2] containing a *reply_to_status_id* – i.e., tweets which were published in reply to another message. Since those tweets are part of a conversation, we reconstructed the conversation thread by recursively crawling all past messages which belong to this conversation. The conversations were crawled on November 20th, 2012 and we obtained 3,850 random conversations in total.

For each conversation we have exactly one positive author-candidate pair which consists of the author of the start message of the conversation and the first user who replied to this message. Further, we randomly selected for each of the remaining conversations one negative sample by selecting one follower of the author of the start message who has not replied to it. We decided to only keep positive author-candidate pairs where the candidate is a follower of the author of the start message, because we wanted to make sure that positive and negative samples are constructed in a consistent way. Surprisingly we had to remove 19.22% sample conversations since users who were not following the author of the message replied to it. This finding confirms the finding of [14] who found that 9% of answerers are not following the askers.

[1] https://dev.twitter.com/docs/api/1/get/statuses/public_timeline

[2] For language detection the *guess_language* python library was used, see: http://pypi.python.org/pypi/guess-language

	median	mean	std
Conversation length	3.0	5.3	12.2
Tweets per user	1,991.9	1,702.2	1,047.7
List memberships per user	0.0	33.2	456.2
Created lists per user	0.0	0.1	0.7
Character length of bio information per user	73.4	68.7	52.4
Followers	266.0	1,524.1	13,819.7
Followees	295.7	1,205.2	8,237.7

Table 1. Characteristics of the dataset consisting of 3,850 conversations from 12,701 different users.

We ended up having 3,215 positive and 3,215 negative samples. For all users who are part of the positive or negative samples (containing 9122 users) we further crawled their most recently published messages (up to 3,200 tweets), their user list memberships, the user lists they created, their user profile information and their followers and followees. We checked that there are no duplicate author/candidate pairs in the positive and negative samples. We want to point out that this information was crawled one day after the conversations were crawled, on the 21th of November 2012. This implies that the information about user's social network, their users lists and their biography may have changed during that day. Therefore features which are based on this information may contain future information which was not available when the conversation happened.

Table 1 shows the basic characteristics of our dataset. The zero median value for the number of participating membership lists and the created membership lists per user indicates that many user do not use or create membership lists. Further one can see from the table that the number of followers per user have a high standard deviation coming from outliers having multiple millions of followers.

Feature Engineering

We introduce three different groups of features. *Topical features* capture the topical similarity between the author of a message and a reply candidate. *Social features* describe the social relationship between the author of a message and a reply candidate. Finally, *Activity features* describe how active and popular a user is on Twitter. We added activity features since we are interested in exploring to what extent social and topical features help predicting repliers beyond the effects of activity features which may function as confounding variables. If we would not take into consideration the users' activity level, we might observe that some social or topical features are highly correlated with a user's reply probability, although they are only correlated with the user's activity level.

Topical Features

Topical features capture the topical similarity between the author of a message and a reply candidate. To identify topics we evaluated three different topic-annotation methods: First we used the concept and keyword extraction service from Alchemy[3], a third party information extraction service, and secondly we used a Twitter-specific Part-of-Speech Tagger (POS)[4]. The tagger reaches an overall tagging accuracy of 90% on Tweets [5] and performs better than the commonly used Stanford POS Tagger for text including abbreviations, interjections, and text which is not grammatically correct written. We decided to keep only proper nouns and hashtags since they often reveal information about the topic of a tweet. In [17] Saif et al. evaluate several open APIs for extraction semantic concepts and entities from tweets. They found that the AlchemyAPI, which we use in our work, extracted the highest number of concepts, and has also the highest entity-concept mapping accuracy. The concept extraction method takes a raw text as input and returns DBpedia[5] concepts and relevance scores as output, while the keyword extraction method extracts relevant unigrams and bigrams from a given input text. We experimented with using Dbpedia concepts, Alchemy generated keywords and POS tagger generated keywords. In this paper we only report the results which we obtained when using topical features produced by the Twitter POS tagger because we obtained the best model fit using this type of topical feature. That means we picked the best performing topical feature. Further in this work we will use the term concept to refer to our topical features.

	median	mean	std
Tweet concepts per user	10.3	8.5	5.6
List concepts per user	0.0	5.4	11.7
Bio concepts per user	1.3	1.9	2.0

Table 2. Number of concepts per user extracted from three types of information provided by a user. First, the aggregation of all tweets written by the user. Second, the aggregation of all membership list names and descriptions the user participates and finally the user's profile description.

We use the following three methods for representing users as documents:

- First, we represent each user as an aggregation of messages which he/she recently published (up to 3,200).
- Second, we represent each user as an aggregation of the names and descriptions of the user lists he/she is a member of.
- Third, we represent each user by his/her personal description obtained from his/her user profile page.

Each topic annotation method combined with each document representation method provides us with a different concept-vector for a user and allows computing the topical similarity between the author of a message and the potential reply candidate based on their concept-vectors. Table 2 shows the mean number of concepts which can be obtained for a user using the different types of user information. Not surprisingly, tweets allow to obtain the highest number of concepts per user, followed by lists and bio information.

We calculate the similarity of the concept-vector of user a and the concept vector of user c using the cosine similarity which

[3] http://www.alchemyapi.com

[4] http://www.ark.cs.cmu.edu/TweetNLP/

[5] http://dbpedia.org

is defined as follows:

$$sim(a,c) = \frac{\langle concepts(a), concepts(c)\rangle}{||concepts(a)|| \cdot ||concepts(c)||} \qquad (1)$$

Using the three aforementioned methods for representing users via text and using cosine similarity as similarity measure, for each pair of users $\langle a, c\rangle$ we compute the following features: The *TweetConceptSimilarity* describes how similar two users are, given the concepts they are tweeting about. The *ListConceptSimilarity* describes how similar users' list memberships are, given the concepts the lists are about. Finally, the *BioConceptSimilarity* reveals how topically similar two users are, given the concepts extracted from their personal descriptions on Twitter.

Social Features

Social features capture the strength of the social relation between the author a and a reply candidate c. We introduce the following six social features: The *NumReplyRelation* feature describes how often the reply candidate has communicated with the author in the past. The *ReplyPartnerOverlap* feature reveals if the author and the reply candidate tend to have similar communication partners. The *FriendsOverlap* feature describes how many similar *friends* the author and the reply candidate have in their follower/followee network. The *isFriend* feature is a boolean value describing if the author and candidate have a bidirectional *follower/followee* relation or not. The *CommonListMembership* feature measures the overlap between the list memberships of the author and the candidate – i.e. in how many common lists they are both members. Finally, the *CandInAuthorsList* feature measures the overlap between the lists the author has created and the lists the candidate is member of.

For computing the overlap between the set of users or lists related with the author a ($users(a)$ or $lists(a)$) and the set of users or lists related with the potential reply candidate c ($users(c)$ or $lists(c)$) we use Jaccard similarity coefficient which is defined as follows:

$$Jaccard(a,c) = \frac{|users(a) \cap users(c)|}{|users(a) \cup users(c)|} \qquad (2)$$

Activity Features

The third category of features are the activity features. These features capture how active or communicative, and also how popular a reply candidate is. Activity features do not measure any association between the reply candidate and the author but rely solely on characteristics of the candidate. Activity features represent common confounding variables since they might be correlated with some topical and social features. Activity features represent of course not the only confounding factor. For example, external events or happenings or users' current locations might be other confounding variables. However, those factors can unfortunately not be obtained from our observational dataset. However, since we constructed our positive and negative samples randomly (with a slight bias towards active users in the case of positive samples) we can assume that other confounding factors are equally distributed across positive and negative samples.

We compute the following six activity features as follows: The *TweetActivity* feature measures the general activity level of a user on Twitter based on the number of tweets he/she has written in the past. The *AvgTweetActivityLastWeek* feature measures the user's average tweet activity per day within the last week. The *ReplyActivity* feature shows how communicative a user is given the number of reply messages the user has written in the past. The *Openness* feature reveals how open a user is giving the number of users he/she is communicating with. The *Followers* feature captures the popularity of a user given his/her *number of followers*. The *Followees* feature indicates the number of users a user is interested in given his/her *number of followees*.

All feature values are normalized by firstly subtracting the mean in each feature and secondly dividing the values of each feature by its standard deviation. Consequently, values of each feature have zero-mean and unit-variance

Methodology

In this section we describe the methodology which we use to answer our research questions.

Feature Analysis

To answer the first research question (*To what extent is communication of Twitter users driven by social and topical factors?*) we assess the association between each feature and the users' probability of replying. Therefore, we use statistical hypothesis tests and measure the potential of each feature to differentiate between the positive and negative class (i.e., user replies or does not reply). The null hypothesis states that the users who see the message and reply and the users who see the message and do not reply do not differ significantly – i.e., the feature distributions of both user groups are similar. We use the Wilcoxon rank sum test for ordinal features and the Chi-Squared test for categorical features. Unlike the t-test which works best for normally distributed ordinal data, the Wilcoxon rank sum test does not have any requirements in the distribution of the data.

Since the statistical tests compute the significance for each individual feature without taking the combination of features into account, we further use a logistic regression model. The dependent variable in our model is binary and indicates for each author-candidate pair $\langle a, c\rangle$ if the candidate has replied to the author or not. We add the previously described social, topical and activity features as independent variables. A logistic regression model reveals if the discriminative power of a feature persists, given all other variables are held constant.

When multicollinearity appears in a regression model, the standard error of the coefficients tend to be very large, and the coefficients are unreliable. Two commonly used ways for dissolving collinearity are combining the correlated features or neglecting one of them. As Figure 1 shows, the *ReplyPartnerOverlap* and *FriendsOverlap* (Pearson correlation coefficient 0.78) and the *ReplyActivity* and *TweetActivity* (Pearson correlation coefficient 0.76) were highly correlated (i.e. correlation coefficient > 0.75).

For the *ReplyPartnerOverlap* and *FriendsOverlap* we decided to neglect the *FriendsOverlap* because it is based on the *Fol-*

Feature	Description	Mathematical Description
Topical Features		
TweetConceptSimilarity	Cosine similarity between *tweet_concepts* of the candidate c and author a.	$\frac{\langle tweet\_concepts(a), tweet\_concepts(c)\rangle}{\|\|tweet\_concepts(a)\|\| \cdot \|\|tweet\_concepts(c)\|\|}$
BioConceptSimilarity	Cosine similarity between *profile_concepts* candidate c and author a.	$\frac{\langle profile\_concepts(a), profile\_concepts(c)\rangle}{\|\|profile\_concepts(a)\|\| \cdot \|\|profile\_concepts(c)\|\|}$
ListConceptSimilarity	Cosine similarity between *list_concepts* of candidate c and author a.	$\frac{\langle list\_concepts(a), list\_concepts(c)\rangle}{\|\|list\_concepts(a)\|\| \cdot \|\|list\_concepts(c)\|\|}$
Social Features		
CommonListMembership	Jaccard similarity between list memberships of candidate c and author a.	$\frac{\|lists(a) \cap lists(c)\|}{\|lists(a) \cup lists(c)\|}$
CandInAuthorsList	In how many list candidate c appears of author a.	$\frac{\|created\_lists(a) \cap created\_lists(c)\|}{\|created\_lists(a) \cup created\_lists(c)\|}$
NumRepliesRelation	Number of replies of candidate c to Author a in the past.	$replies(a, c)$.
ReplyPartnerOverlap	Jaccard similarity between between *reply_partners* of candidate c and author a.	$\frac{\|reply\_partner(a) \cap reply\_partner(c)\|}{\|reply\_partner(a) \cup reply\_partner(c)\|}$
isFriend	Is the candidate c a follower of author a and vice versa.	$isFollowing(a, c) \cap isFollowedby(a, c)$
FriendsOverlap	Jaccard similarity between candidate c and author a given their *friends*.	$\frac{\|friends(a) \cap friends(c)\|}{\|friends(a) \cup friends(c)\|}$
Activity Features		
TweetActivity	Number of tweets posted by the candidate c.	$num\_tweets(c)$
ReplyActivity	Number of replies the candidate c was participating.	$num\_replies(c)$
AvgTweetActivityLastWeek	Average tweets per day the candidate c writing within the last week.	$avg\_tweets\_week(c)$
Openness	Number of users the candidate c was replying to.	$num\_replyingto(c)$
Followers	Number of *followers* of the candidate c.	$num\_followers(c)$
Followees	Number of *followees* of the candidate c.	$num\_friends(c)$

Table 3. Overview of all features used in our empirical study.

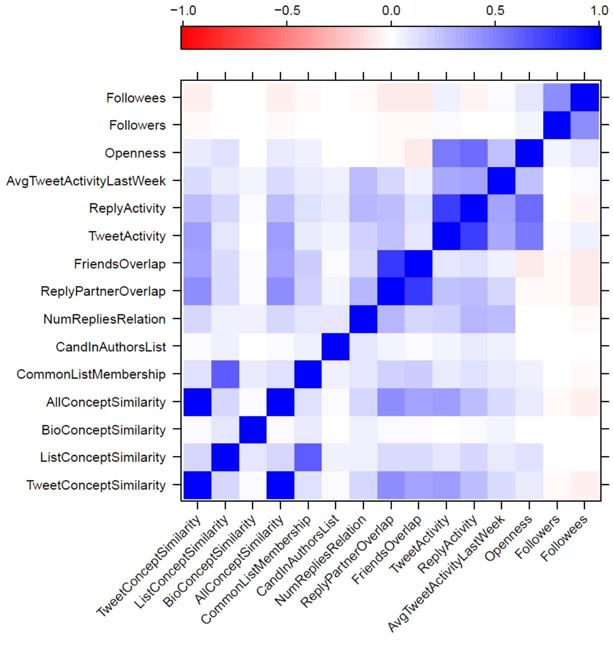

Figure 1. Pearson Correlation matrix of all features. One can see from this figure that the *ReplyPartnerOverlap* and *FriendsOverlap* and the *ReplyActivity* and *TweetActivity* are strongly correlated. When multicollinearity appears in a regression model, the standard error of the coefficients tend to be very large, and the coefficients are unreliable. We solved this issue by neglecting one of the highly correlated features.

lowers and *Followees* information which we crawled one day after the conversation took place. In theory, the social network as well as the list memberships may have changed within this day and therefore the features which rely on this information may contain future information. Finally, for the *ReplyActivitiy* and *TweetActivity* we decided to keep the *ReplyActivity* because we assume that this feature has more power to predict repliers than the more general *TweetActivity*.

After the removal of collinear features we fit the logistic regression model to our dataset. We use *Nagelkerkes pseudo* R^2 measure to assess how well the model fits our data. This value ranges from 0 to 1, where 1 denotes a perfect fit to the observed data and 0 the model doesn't fit at all.

Nagelkerkes pseudo R^2 is defined as follows:

$$Nagelkerkes\_pseudo\_R^2 = \frac{1 - \frac{L(M_{intercept})}{L(M_{full})}^{2/N}}{1 - L(M_{intercept})^{2/N}} \quad (3)$$

where N denotes the number of samples, $L(M_{full})$ refers to the likelihood to obtain the training data when using all features and $L(M_{intercept})$ without using any feature in the logistic regression model.

To gain further insights into the usefulness of individual features, we interpret the statistical significant coefficients of the model. The coefficients returned from a logistic regression model are log-odds ratios and can tell us how the log-odds of a "success“ (in our case a reply) changes with a one-unit change in the independent variable.

Prediction Experiment

In addition to looking into the utility of individual features, we are also interested in assessing the predictive power of the whole model in order to answer the second research questions

(*To what extent are repliers on Twitter predictable?*). Therefore we conduct a 10-fold cross validation and train and test the logistic regression model on our dataset. Since our dataset is balanced, i.e. it contains an equal number of positive and negative samples, a random guesser baseline would lead to a performance of 50%. We use *Precision*, *Recall* and the *F1-score* which is the harmonic mean of Precision and Recall as evaluation measures.

RESULTS

In this Section, we present the results from our empirical data analysis which aims to gain insights into patterns of conversations on Twitter and the factors which may potentially drive them.

Feature Analysis

Answering our first research question *RQ1* requires gaining insights into the utility of individual features. Towards that end, we conducted statistical significance tests and fitted a logistic regression model using all features as independent variables and the binary variable (replies or not) as dependent variable.

Statistical Hypothesis Tests

The results from the Wilcoxon rank sum test and the Chi-Squared test show that all features except *Followers* and *BioConceptSimilarity* are statistically significant (see Table 4). This indicates that almost all features are significantly associated with our binary variable (replies or not).

One potential explanation why the *BioConceptSimilarity* seems to be irrelevant is that the bio information of users tends to be short with a mean length of 75 characters per user and that around 14% of the users do not provide any bio information. In our previous work [19] we found that the users' bio information is almost as useful as tweets for predicting users' expertise. However, one needs to note that the dataset we used in [19] was biased towards active expert users who had a high Wefollow[6] rank, while our dataset in this work consist of average users who use Twitter for a conversational purpose. The number of followers seems to be unrelated with users' reply behavior which indicates that users' popularity does not impact their probability of replying.

Regression Analysis

Since the statistical tests compute the significance for each individual feature without taking the combination of features into account, we further fitted a logistic regression model. The dependent variable of our logistic regression model is binary and indicates for each author-candidate pair $\langle a, c \rangle$ if the candidate has replied to the author or not. The previously described social, topical and activity features are added as independent variables.

Table 5 shows the regression coefficients of each feature and their significance level. All features are normalized, so we can rank their influence using their coefficients. Figure 2 shows the distribution of the most significant features for each class. The more the class-specific feature distributions differ, the higher the ability of these features to discriminate the two classes.

Feature	**p-Values**	**Significance**
Wilcoxon Rank Sum Test (numerical features)		
TweetConceptSimilarity	4.873e-112	***
ListConceptSimilarity	3.882e-10	***
BioConceptSimilarity	0.4008	
CommonListMembership	1.904e-17	***
CandInAuthorsList	2.740e-08	***
NumRepliesRelation	0.00e+00	***
ReplyPartnerOverlap	3.644e-261	***
FriendOverlap	3.641e-120	***
TweetActivity	3.418e-98	***
ReplyActivity	2.948e-206	***
AvgTweetActivityLastWeek	8.255e-238	***
Openness	2.198731e-93	***
Followees	1.640e-36	***
Followers	0.151	
Chi-Squared Test (categorical features)		
isFriend	2.2e-16	***

Table 4. Results from the statistical hypothesis tests.

One can see from Table 5 that the activity features *AvgTweetActivityLastWeek* and *ReplyActivity* are significant and have a positive coefficient. This demonstrates that the activity level of a user is indeed a significant factor, which influences if a user will reply to a message or not. Not surprisingly, active users are more likely to reply than non active users. The features which are related with the popularity and social status of a user (*Openess* and *Followers*) are not significant which means that the users' reply behavior is not influenced by how open they are or by how many users they follow.

In addition to the activity features, the following social features have a significant positive coefficient – i.e., they help predicting repliers beyond the effects of activity features: *NumRepliesRelation*, *isFriend* and *ReplyPartnerOverlap*. This shows that previous communication relations as well as bidirectional friendship relations are very important for predicting who will reply to a message of a certain user. Friends of the author of the message who have communicated with each other before are more likely to reply than others. The only significantly negative feature is the *Followees* feature. This indicates that the more users a user is following the less likely he/she replies to their messages, as also shown in Figure 2. Intuitively this makes sense as we assume that every user has a maximum number of tweets to which he/she will reply e.g. per hour. The more people a user is following, the more new tweets will show up in his/her timeline. That means the users' reply probability is spread across more tweets and is therefore lower for each individual tweet.

Finally, the logistic regression model shows that topical features like the *TweetConceptSimilarity* and the *BioConceptSimilarity* are also significantly positively correlated with users' reply probability. This indicates that there is a slight tendency that users who are interested into similar topics are more likely to reply to each other. However, one needs to

[6] http://wefollow.com/

	Coefficient	Significance
(Intercept)	-0.0151	
TweetConceptSimilarity	0.1472	***
BioConceptSimilarity	0.0710	*
ListConceptSimilarity	-0.0575	
NumRepliesRelation	2.6073	***
ReplyPartnerOverlap	0.2638	***
CommonListMembership	0.0281	
CandInAuthorsList	0.0727	
isFriend	0.3962	***
ReplyActivity	0.3418	***
AvgTweetActivityLastWeek	0.3505	***
Openness	0.0726	
Followers	0.6063	
Followees	-1.9698	***

Table 5. Results from the logistic regression model using topical, social and activity features as independent variables and reply or not as binary dependent variable.

note that the coefficients of the significant topical features are much smaller than the coefficients of the significant social features. This indicates that users' reply behavior on Twitter is more influenced by social factors than by topical factors.

Prediction Experiment

To answer our second research question *RQ2* we conducted a prediction experiment using the same features as in the aforementioned logistic regression experiment. We trained our logistic regression model and tested the predictive power of the model using a 10 fold cross-validation.

Our results in Table 7 show that when using all three types of features we achieve an average F1-score of 0.76 while a naive baseline (random guesser) would achieve 0.5 since our dataset is balanced. The confusion matrix in Table 6 shows that the model classified more users who replied as non-repliers than users who did not reply as repliers. Interestingly, using social features alone was almost as good as using a combination of all features (F1=0.74). This indicates that social features contribute most to the performance of the classification model. Also, activity features alone performed very well (F1=0.70) as shown in Table 7. This confirms our hypothesis that the activity level of a user is a common confounding variable when analyzing the factors that influence users' reply behavior.

Finally, Table 7 shows that the performance is worst when using topical features alone (F1=0.63). Also Table 8 indicates that a logistic regression model using only topical features as independent variables is worst in explaining the variability in the training dataset, while a combination of all features is best, followed by using social features alone.

	predicted non replier	predicted replier
non replier	2582	633
replier	924	2291

Table 6. Confusion matrix of the logistic regression classification results using all features. The columns of the confusion matrix show the predicted values and the rows show the reference values.

	Precision	Recall	F-Score
All features			
non replier class	0.74	0.80	0.77
replier class	0.79	0.71	0.75
average	0.76	0.76	0.76
Topical features			
non replier class	0.61	0.73	0.67
replier class	0.67	0.54	0.60
Average	0.64	0.64	0.63
Social features			
non replier class	0.70	0.84	0.76
replier class	0.80	0.64	0.71
Average	0.75	0.74	0.74
Activity features			
non replier class	0.67	0.77	0.72
replier class	0.73	0.62	0.67
Average	0.77	0.70	0.70

Table 7. Classification accuracy of our logistic regression model using all features, topical features, social features and activity features.

Our results clearly demonstrate that conversations on Twitter are not driven by topics but by social relations. Further our work shows that in addition to social relations users' activity level plays an important role since more active users are also more likely to reply (i.e., have a higher prior probability of replying). Researchers need to consider activity information since they may function as confounding variables when neglected. Including activity features into our models allows us to conclude that social features help predicting repliers above and beyond the effects of activity features.

CONCLUSIONS, LIMITATIONS AND FUTURE WORK

In this work we conducted an empirical study about the nature and predictability of conversations on Twitter.

Concretely, our work answers the following research questions:

- *RQ1:To what extent is communication of Twitter users influenced by social and topical factors?* Our results show that social features, which describe the strength of the relation between users, help predicting repliers above and beyond the effects of activity features and are more useful than topical features for predicting if a user will reply to another user or not. This suggests that conversations on Twitter are more driven by friendships and social relations rather than topics. The best social features were the *NumRepliesRelation*, the *isFriend* and the *FriendsOverlap* features. This suggests that users are far more likely to reply to a message authored by a user who is a friend of them, to whom they have talked in the recent past frequently and with whom they share common friends.

- *RQ2: To what extent are repliers on Twitter predictable?* Our work shows that a binary classification model that dif-

	all	topical	social	activity
R^2	0.402	0.105	0.337	0.246

Table 8. Goodness of fit of the logistic regression model measured using the Nagelkerke pseudo R^2.

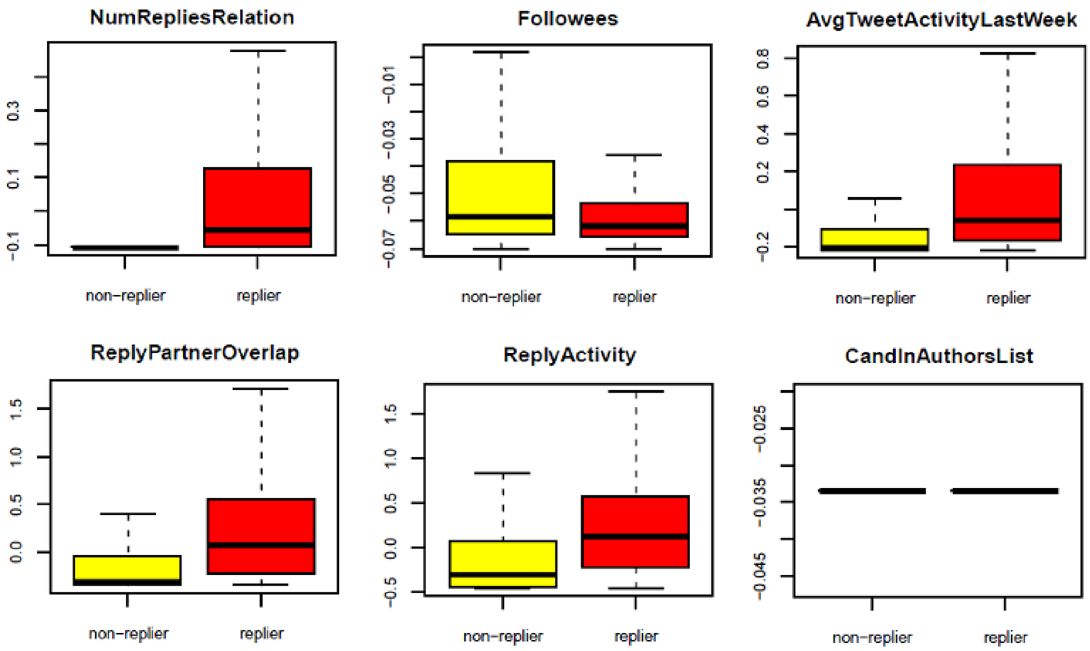

Figure 2. The six most discriminative numerical features from the logistic regression analysis. One can see that users are more likely to reply to a message if they have a high conversation partner overlap with the author of the message (*ReplyPartnerOverlap*) or if they communicated with the author of the message before (*NumRepliesRelation*). Further, users who reply tend to be more active – i.e. they have a higher *AvgTweetActivityLastWeek* and a higher *ReplyActivity*. One can also see that the users who have many *Followees* are less likely to reply.

ferentiates between users who will and will not reply to each other may achieve an F1-score of 0.75 using social, topical and activity features. Using topical features as independent variables leads to the worst statistical model, while using a combination of all features works best, followed by using social features alone. We were able to increase the average F1 score of a random baseline classifier by 24% when using social features alone.

Our work has certain limitations since our assumption that all users who follow a user are similar likely to see messages authored by this user is a simplification which may not reflect the reality. By adding activity features as covariates we addressed this limitation to some extent. Further, this work focuses on the first replier on a single branch of the conversation, and does not take the long-term dynamics of social media conversations into account. We also want to point out that any crawling strategy might introduce a certain bias, as comprehensively studied and described in [6].

In this work we focused on features which can be computed between pairs of users rather than triples (consisting of the two users and the current message) since we are interested in integrating this work into a real-time video communication tool [9] which exploits users' social media stream as background knowledge for orchestrating the video communication. Therefore, it is necessary to be able to compute the features at the beginning of each communication session rather than re-computing them after each message or sentence. For future work we plan to analyze the influence of the current message on users' reply behavior and update the initial communication prediction model during the course of a conversation.

ACKNOWLEDGMENTS

This work was performed within Vconect - Video Communications for Networked Communities (grant agreement no 287760) and received funding from the European Commission under the Seventh Framework Programme (http://www.vconect-project.eu). This work was supported in part by a DOC-fForte fellowship of the Austrian Academy of Science to Claudia Wagner.

REFERENCES

1. Chelmis, C., and Prasanna, V. K. Predicting communication intention in social networks. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, SOCIALCOM-PASSAT '12, IEEE Computer Society (Washington, DC, USA, 2012), 184–194.

2. Chen, J., Nairn, R., and Chi, E. Speak little and well: recommending conversations in online social streams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, ACM (New York, NY, USA, 2011), 217–226.

3. Gilbert, E. Predicting tie strength in a new medium. In *CSCW*, S. E. Poltrock, C. Simone, J. Grudin, G. Mark, and J. Riedl, Eds., ACM (2012), 1047–1056.

4. Gilbert, E., and Karahalios, K. Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, ACM (New York, NY, USA, 2009), 211–220.

5. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan,

J., and Smith, N. A. Part-of-speech tagging for twitter: annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, Association for Computational Linguistics (Stroudsburg, PA, USA, 2011), 42–47.

6. Gonzalez-Bailon, S., Wang, N., Rivero, A., Borge-Holthoefer, J., and Moreno, Y. Assessing the Bias in Communication Networks Sampled from Twitter. *Social Science Research Network Working Paper Series* (2012).
7. Honeycutt, C., and Herring, S. C. Beyond microblogging: Conversation and collaboration via twitter. In *Proceedings of the Forty-Second Hawai'i International Conference on System Sciences (HICSS-42). Los Alamitos, CA.*, IEEE Computer Society (Los Alamitos, CA, USA, 2009), 1–10.
8. Java, A., Song, X., Finin, T., and Tseng, B. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, WebKDD/SNA-KDD '07, ACM (New York, NY, USA, 2007), 56–65.
9. Kaiser, R., Weiss, W., Falelakis, M., Michalakopoulos, S., and Ursu, M. F. A rule-based Virtual Director enhancing group communication. In *ICME Workshops*, IEEE (2012), 187–192.
10. Macskassy, S. A. On the study of social interactions in twitter. In *The International AAAI Conference on Weblogs and Social Media (ICWSM)*, The AAAI Press (2012).
11. Macskassy, S. A., and Michelson, M. Why do people retweet? Anti-homophily wins the day! In *ICWSM*, L. A. Adamic, R. A. Baeza-Yates, and S. Counts, Eds., The AAAI Press (2011).
12. Monge, P., and Contractor, N. *Theories of Communication Networks*. Oxford university Press, 2003.
13. Naaman, M., Boase, J., and Lai, C.-H. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW '10, ACM (New York, NY, USA, 2010), 189–192.
14. Paul, S. A., Hong, L., and Chi, E. H. Is twitter a good place for asking questions? A characterization study. In *ICWSM*, L. A. Adamic, R. A. Baeza-Yates, and S. Counts, Eds., The AAAI Press (2011).
15. Rowe, M., Stankovic, M., and Alani, H. Who will follow whom? exploiting semantics for link prediction in attention-information networks. In *International Semantic Web Conference (ISWC)*, vol. 7649 of *Lecture Notes in Computer Science*, Springer (2012), 476–491.
16. Sadilek, A., Kautz, H., and Bigham, J. P. Finding your friends and following them to where you are. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, ACM (New York, NY, USA, 2012), 723–732.
17. Saif, H., He, Y., and Alani, H. Semantic sentiment analysis of twitter. In *The 11th International Semantic Web Conference (ISWC)* (2012), 508–524.
18. Sousa, D., Sarmento, L., and Mendes Rodrigues, E. Characterization of the twitter @replies network: are user ties social or topical? In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, SMUC '10, ACM (New York, NY, USA, 2010), 63–70.
19. Wagner, C., Liao, V., Pirolli, P., Nelson, L., and Strohmaier, M. It's not in their tweets: Modeling topical expertise of twitter users. In *Proceedings ASE/IEEE International Conference on Social Computing (SocialCom2012)* (2012).
20. Wang, C., and Huberman, B. How random are online social interactions? *Scientific Reports 2* (2012).