What Ignites a Reply? Characterizing Conversations in Microblogs

Johnny Torres, Carmen Vaca, Cristina L. Abad
Escuela Superior Politécnica del Litoral, ESPOL
Facultad de Ingeniería en Electricidad y Computación, FIEC
Guayaquil, Guayas 09-01-5863, Ecuador

ABSTRACT

Nowadays, microblog platforms provide a medium to share content and interact with other users. With the large-scale data generated on these platforms, the origin and reasons of users engagement in conversations has attracted the attention of the research community. In this paper, we analyze the factors that might spark conversations in Twitter, for the English and Spanish languages. Using a corpus of 2.7 million tweets, we reconstruct existing conversations, then extract several contextual and content features. Based on the features extracted, we train and evaluate several predictive models to identify tweets that will spark a conversation. Our findings show that conversations are more likely to be initiated by users with high activity level and popularity. For less popular users, the type of content generated is a more important factor. Experimental results shows that the best predictive model is able obtain an average score $F1 = 0.80$. We made available the dataset scripts and code used in this paper to the research community via Github[1].

CCS CONCEPTS

• **Applied computing** → **Sociology**; • **Computing methodologies** → *Machine learning*;

KEYWORDS

Big data; Machine Learning; Social Computing

1 INTRODUCTION

Amongst microblogging sites, Twitter has become one of the most popular worldwide. In this social network, its users share content publicly via short texts named *tweets*. Although, most tweets generate little or no interaction with other users, sometimes the published content can ignite a long chain of replies and interactions from other users. We will name *seed tweets* to those tweets that initiates conversations. In this work, we seek to understand the factors in *seed tweets* that contribute to ignite replies from other users.

The published content on social networks can have an impact on different aspects of society, such as: popular culture, brands communication, politics, activism, journalism, crisis communication, among others [19]. As the content generated increases on social networks, the factors that ignite conversations or discussions are of special interest.

In the last decade, microblogging—specifically Twitter—has attracted the attention of the research community due to the open data access through its public APIs[2]. Several aspects of Twitter have been studied by researchers, including but not limited to: its network structure, users' behaviors, content generated, and the infrastructure needed to handle its massive datasets. A particular aspect that researchers have been interested, is related to the nature of the interactions that occur in this social network. Among the type of interaction are the conversations spontaneously occurring among users.

Even though the idea of Twitter was originally for users to post what they were doing, soon its users began to use @ symbol to interact with other users [9]. This type of interactions often evolve into natural, complex, noisy, and long conversations. Thus, this type of interactions blurs the border between conversations in private chats and public blogs.

In understanding human conversations, several aspects are important. Some of them consist on identifying the conversations structure and intent [15]. Predicting whether content posted on social networks will become popular or generate interest from users constitute another aspect in the analysis of conversations.

The latter could be useful in many applications in the area of recommender systems (news feed, advertising placement). For instance, a user may be interested in reading several articles on different topics, for which a real-time news recommendation system should show relevant articles with the aim of fulfilling users' preferences or generate interest from the user. Similarly, ads published on social networks, aim to generate attention (reading) or interactions (in the form of a like, retweet, or replies) from its audience.

This paper examines the factors that contribute to spark a conversation on Twitter, i.e., identifying whether a tweet that will generate replies from other users. Our hypothesis is that contextual and content features extracted from the tweets can be used to predict the likelihood of a tweet evolving into a conversation. Our goal is not to predict the popularity level, but rather if a tweet will evolve into a conversation.

In this research, we propose a language independent model to identify *seed tweets* that have the potential to form conversations for different type of users. The main contribution of our work is to:

[1] https://github.com/johnnytorres/twconvcharact

[2] https://developer.twitter.com/en/docs

BDCAT'17, December 5–8, 2017, Austin, Texas, USA
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-5549-0/17/12…$15.00
https://doi.org/10.1145/3148055.3148071

- Characterize tweets that ignites conversations.
- Design and implement a classification model to identify *seed tweets*.
- Make the dataset script and code available to the research community via Github[1].

The rest of this paper is organized as follows. In Section 2, we begin discussing prior work about human conversations in general, and then more specifically about conversations on Twitter. In Section 3, we explain the data acquisition, storage, and processing of our dataset. In Section 4, we extract and characterize features to identify conversations. Then, using principal components analysis (PCA) on a subset of tweets, we build and train a predictive model for identifying conversations in Section 5. Finally, we discuss the results in Section 6 and draw the conclusions in Section 7.

2 RELATED WORK

2.1 General Conversations Modeling

Understanding human conversations has been extensively studied and continues attracting the attention of researchers in the quest to achieve human level reasoning and comprehension in machines.

Conversation modeling has been studied previously using cellphone SMS corpus [3, 11], IRC chat corpora [7], and blog datasets [21].

Several research directions has been studied in modeling human conversations. Amongst them, identifying conversation acts[3]. Several applications have rely on acts identification, such as: conversational agents [20], dialogue systems [2], automated customer support service [14], virtual assistants [13], among others.

Traditional approaches to identifying conversation acts are based on manual human annotation. This process includes collecting and labeling acts in the dataset following an annotation guide. Although successful, this process can be very time consuming and costly to carry out. Recent approaches focus on overcoming this limitation by using Neural Networks [16] and open data sources such as Twitter [14].

2.2 Twitter Conversations Modeling

Initially, Twitter was conceived as a medium to share personal status, but rapidly evolved as a platform to interact with others with the novel use of "@", as a way of targeting other users to reply to a prior status or establish conversations [9]. Since then, a large body of research has been developed to analyze this kind of interaction, and we will cover the most relevant to our work.

Boyd et al. studied the use of the *retweet* as a mean of engaging in conversations, and how dealt with different aspects such as authorship, attribution, and fidelity of the communication. They found that in general conversations are messy, even when the interactions take place in a bounded group by location, timespan, and participant characteristics. In bounded groups, it is more likely to find cohesive conversations with turns and references to previous messages, but that is not observed on unbounded groups where conversational structures are missing.

The aspect of information diffusion on Twitter has been studied by analyzing the retweet mechanism. Suh et al. conducted a large scale analysis about the factors impacting the retweet behavior for 74 million tweets. They identified a strong correlation of retweet behavior with content feature (e.g., URLs and hashtags), as well as, contextual features (e.g., number of followers and friends).

Ye and Wu studied the propagation patterns of general messages and breaking news on 58 million tweets. They found that messages propagate outside of the group of the originator, i.e., not restricted to the followers. Another aspect of their study, it is the analysis of the user influence calculated by several metrics such as number of followers, replies, retweets.

Also, previous works have studied the problem of predicting the popularity of messages. Based on the future number of retweets, and how those influence the content propagation, Hong et al. proposed a classification model including several content-based, contextual, and temporal features extracted from tweets. Additionally, they included network structure properties in its prediction model.

The process of content diffusion on Twitter can take the form of cascades when users reshare tweets. The characterization as well the predictability of these cascades has been studied, and shown that the predictability depends on temporal and structural features. Moreover, breath propagation rather than depth is a better indicator [5].

Most of these studies focus on measuring the popularity of content based on the propagation of the content in the network. But the form how a message is written, i.e., the effect of wording can have a impact on the popularity and propagation of a tweet. Tan et al. studied this factor by taking pairs of tweets posted with similar URLs and written by the same user but using different words. Their findings show that depending on the words' choice, some tweets can have more popularity than others.

In another aspect, the work done by [8] tackle the problem of predicting popularity of the conversations on Reddit Threads. Although using a different social network, the authors tackled the problem of identifying the popularity of a conversation thread based on the content analysis using deep reinforcement learning.

These prior works are closely related to our study, but differ in the task and the metric used. In this work, we count the replies received by tweets to predict whether a given tweet will generate interactions from other users. To the best of our knowledge, our work is the first in using this metric to identify *seed tweets* that spark conversations on Twitter.

3 SOURCE DATA

We use Twitter as our data source to collect a corpus of more than 150M tweets, from January to July 2017. We collect tweets using the Twitter Streaming API[4]. Through this API, Twitter provides researchers with the 1% of its public data collected at a given time. Although, we use several filters to focus on relevant tweets for our work, the scale of data received is massive.

To deal with the overwhelming amount of data obtained through Twitter Streaming API, we rely on Cassandra [12] as distributed storage. We choose Cassandra over other NoSQL databases because of its distributed architecture, scalability, and high availability without compromising performance as our database grows [1].

The figure 1 shows the data capture and storage architecture used in this research work. It is based on four nodes, on each node

[3] Known also as dialog acts

[4] https://developer.twitter.com/en/docs/tweets/filter-realtime/overview

running: a *Twitter Capture Service* that collects data from Twitter Streaming API, and a local Cassandra used to store the data.

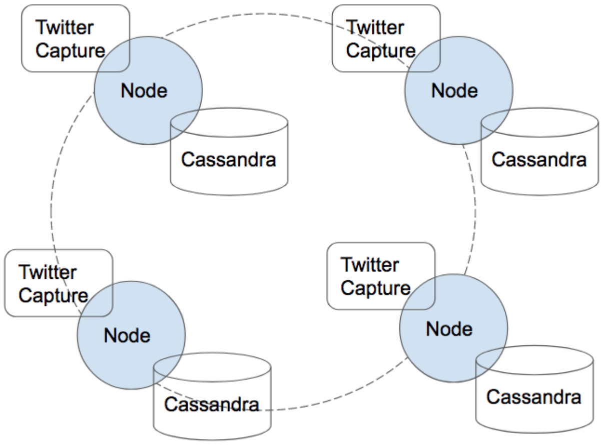

Figure 1: Cassandra cluster to capture and store Twitter data.

For the Twitter streaming data collection, we use two type of filters: a geolocated bounding box, and tweets containing specific words. The bounding box filter allows to capture geolocated tweets in South America, specifically three countries: Colombia, Ecuador, and Peru. We are interested only in English or Spanish tweets, or users who have specified those languages in their profiles. Therefore, tweets in other languages are excluded in our study.

Also, considering that retweets can be a significant part of streaming data, and the fact that retweets only propagate content generated by other users, we filter out retweets and only retain the *original* or *root* tweets. The reason is because we are interested mostly in the interactions in the form of conversations.

From our main dataset, we choose randomly 2M tweets to perform an exploratory analysis. In this subset, we group the tweets by conversations. To that end, we use tweet's field *in-reply-to-status-id* that specify if the tweet is a reply to another tweet. Based on this field, we establish the conversation that each tweet belongs to.

Then, we use Twitter REST API[5] to gather all the conversations' parent tweets not present in our dataset. With these additional tweets collected, our exploratory dataset increased to 2.7M tweets.

The figure 2 shows that the number of tweets in conversations follows a power law distribution [6]. As most human activities, short conversations are the bulk of sample, whereas there are few very long conversations.

As shown in table 1, we found that 64% of the tweets are non-conversational. And, the remaining tweets form conversational threads containing two or more tweets (length of the conversation). For conversational tweets (fourth column), we observe that 41% of them are short conversations (one tweet and one reply), similar to the results found in [15]. Lastly, the conversations have with more than 5 is marginal, as the distribution shown in figure 2.

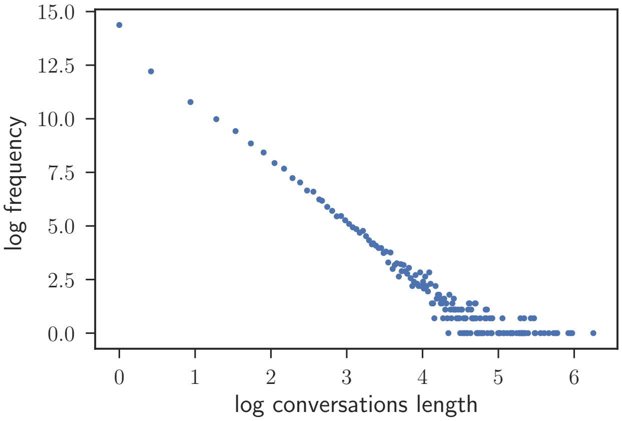

Figure 2: Distribution of tweets in conversations.

Table 1: Conversations statistics

# of conv.	Length	% tweets	% conv.
1,747,374	1	0.64	0.00
401,274	2	0.15	0.41
144,078	3	0.05	0.15
86,512	4	0.03	0.09
61,935	5	0.02	0.06

4 EXPLORATORY ANALYSIS

In this section, we determine the features that will be used to identify and filter tweets that belong to conversations. Also, this analysis helps to uncover features to filter out thread of tweets replies that does not represent conversations between two or more users. Thus, only valid conversations will be used in the predictive model in section 5.

4.1 Language

Although, the majority of users in Twitter post tweets in one language, some users can use multiple languages. To detect the language of the tweets, we use the information provided by the Twitter API in each tweet's metadata, specifically the field *lang*. There is a total of 44 different languages detected, from which english and spanish represent 85% of the tweets.

The language could not be identified for a small percentage of the tweets (7% approximately). Those tweets were marked as *undefined* language. We found that the content of those tweets is usually limited to: mentions, hashtags, URLs, emoticons, or multimedia (i.e. images, videos). In addition to English and Spanish languages, we include tweets marked as *undefined* for further analysis.

The table 2 shows the distribution of the number of languages used for conversations in our dataset. The first column refers to the number of languages detected. The second column indicates the number of conversations. The third indicates the percentage of conversations by language, and the fourth the cumulative percentage. We found the majority of conversations contain tweets

[5] https://dev.twitter.com/rest/public

in one language (75%). There is a 25% of conversations with two or more languages. Although, up to three languages represent the cumulative 99% of the total number of conversations.

Table 2: Languages in conversations

# of lang.	# of conv.	% of conv.	cum. %
1	216,138	0.75	0.75
2	62,992	0.22	0.97
3	6,165	0.02	0.99

4.2 Distance

We found that approximately 12% of tweets in our dataset have geolocated information associated. For conversations, we observe that only 3.2% are geolocated tweets. For geolocated tweets, we analyze the behavior of users that engage in conversations. Although, the geolocated tweets are a small percentage, figure 3 shows users posting tweets all over the world, mainly in English or Spanish speaking countries. Geographically, we focus on countries in American continent, but we found that interactions reach places all over the world.

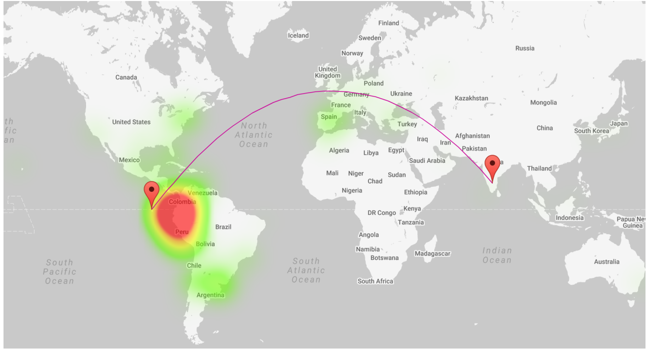

Figure 3: Heatmap of geolocated tweets (12% in exploratory dataset). The markers, connected by the geodetic line, represent a conversation between two users in very distant places.

Although, very distant conversations are not uncommon, conversations in the same exact point (zero meters from origin) may be an indicative that the same user is self-replying, or creating a message in multiple tweets. To avoid selecting tweets in the same spot as conversations, we consider only conversations involving more than one user in the conversation thread for further analysis.

4.3 Duration

We find that the temporary distribution of tweets in conversations is uniform throughout the week. Figure 4 shows the density slightly increasing at night on Tuesday and Wednesday, as well as Friday morning, and it goes down the Saturday.

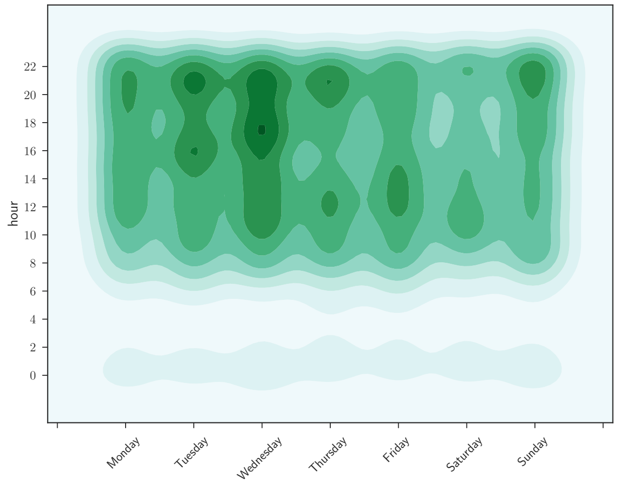

Figure 4: Temporal distribution for tweets initiating conversations.

Another aspect of interest is the duration of the conversations, as shown in table 3. Most of the conversations are short lived, i.e. have a duration of less than ten days. Also, we find that very short duration replies are usually self replies created by 3rd party apps. For instance, the following tweets belong to the same user, created almost at the same time[6]:

tweet: 2017-07-12 21:00:24: @trendinaliaEC: *'1. #ExperienciasElectoralesEC 2. #NoALaViolenciaDeGenero 3. Lula da Silva 4. #CPCCS-MarcandoElCamino 5.Alfaro Moreno...'*

reply: 2017-07-12 21:00:24: @trendinaliaEC: *'6. Roger Federer 7. #LeyEficienciaTramites 8. Defensor del Pueblo 9.#EmergenciaCBQ 10. James Rodriguez.'*

On the other hand, the increasing use of bots or spam accounts in Twitter, can create noisy conversations that span several years. The following tweets illustrate this case[7]:

tweet: 2009-03-07 @finkd: *'Yes; this is the real Mark Zuckerberg. Thanks for following me!'*

reply: 2016-08-13 @oropesa555: *'SOS SOS @$Pontifex_es$ @YourAnonGlobal @finkd ...'*

Despite these cases, we found valid long duration conversations. Usually, new friends or followers may visit tweets posted long ago and comment on them (using the reply option). For instance, the following conversation[8]:

tweet: 2009-10-25 @NARSissist: *'Eaten alive by a mosquito... Not fun'*

reply: 2017-01-08 @GigiFreireA: *'@NARSissist WTF NARS? xD'*

[6] https://twitter.com/trendinaliaEC/status/885242460644888577
[7] https://twitter.com/finkd/status/1293412597
[8] https://twitter.com/NARSissist/status/5157432533

Table 3: Duration of conversations

Days >	≤	Conversations	%	cperc
0	10	277,222	0.97	0.97
10	20	4,548	0.02	0.98
20	30	1,138	0.00	0.99
30	40	627	0.00	0.99
40	50	379	0.00	0.99

In our analysis, we filter out tweets that falls in the case of conversations containing sequential posts created by the same user in very short period of time. The other cases are more difficult to identify, so we will include all tweets in long duration conversations for further analysis.

4.4 Users

In spatial and temporal analysis, we have found that all tweets may actually belong to the same user in some conversations, and thus, cannot be considered as true conversations. We found that 80% of the conversations have two users involved, as shown in table 4.

For the conversations involving two users, the median of the conversation length is two tweets. But, there are some outliers, for example: we found some sport journalists and their followers narrating football matches on Twitter using replies. This kind of conversations can have more than 500 tweets forming a long conversation. We only consider conversations having two or more users for predictive analysis.

Table 4: Users in conversations

# Users	# Conversations	%	cum. %
1	20,476	0.07	0.07
2	245,417	0.80	0.87
3	29,128	0.09	0.96
4	7,029	0.02	0.98
5	2,420	0.01	0.99

5 CONVERSATIONAL MODEL

In this section, we build and train a predictive model for identifying *seed tweets* that evolve into conversations. First, we explain the subset of tweets used in this analysis. Then, we describe in detail the features extraction from tweets. Next, we perform PCA analysis to detect important features prior to step into the predictive modeling analysis.

5.1 Prediction dataset

For the purpose of the predictive analysis, we consider a subset of tweets that allows to understand the features that might spark a conversation. The idea is to use this subset of tweets to extract the features, i.e. independent variables. Then, we define whether a tweet is part or not of a conversation as the dependent variable tin our model.

From our exploratory dataset, we randomly select 1000 tweets that initiated a conversation (parent tweet of a conversation), stratified by the number of replies received. For these conversation tweets, we also extract all replies. Then, we randomly select 1000 tweets that does not evolve into conversations, i.e. with no replies. In total, our prediction dataset contains $10,805$ tweets.

Prior to further analysis, we apply a log transformation to features with large values (number of followers, friends, tweets posted). These features usually follow a power law distribution. Other features remain with the original range of values.

5.2 Features extraction

Using the prediction dataset, we extract a set of features that can be used in our conversational prediction model across different languages. These features are content, contextual, and language-invariant attributes present in the text and metadata of each tweet. Table 5 shows two type of features: user related and tweet related features. Those associated to the user level include metadata from the user profile. The tweet related features are the metadata and tweet's content itself.

Table 5: Conversations Features

User level features	
Statuses	# of statuses
Followers	# of followers
Friends	# of friends
Favorites	# of likes given to tweets by the user
Tweet level features	
Retweets	# of retweets received
Favorites	# of likes received
Urls	# of Urls in the tweet
Hashtags	# of hashtags in the tweet
Mentions	# of mentions in the tweet
Media	If there are images or video in the tweet
Replies	# of replies received

At user level, we consider contextual features, e.g. number of statuses posted, followers, friends, and favorites. At tweet level, we focus on content features: the number of urls, hashtags, mentions, and multimedia present in the text of the tweets. The number of replies, retweets, and favorites to indicate the popularity of a tweet. These attributes could be used as the dependent variables.

In this paper, we focus on conversations. Therefore, we chose *the number of replies* as the target feature that we want to predict. We do not use features that are populated after the creation of the tweet, such as number of retweets or favorites received.

We extract content features using regular expressions, by identifying words starting with @ (mentions), # (hashtags), or http (urls). For contextual features, we extract them directly from the tweets' metadata. While, the target feature requires counting all the replies for each tweet.

Prior to the PCA and predictive analysis, we perform data cleansing of our dataset. We remove tweets with missing user profile features, which correspond to few cases (0.3% of tweets). For user level features, we apply log transformations to avoid large values dominating in our predictive analysis.

5.3 Principal Components Analysis

To reducing the dimensionality, we perform PCA to find possibly correlated to features shown in table 5. These features are transformed into a small set of factors, identified as principal components. This technique aims to revel the underlying data structure and the weights each feature contribute to the data variance.

Table 6 shows the principal components or factors in our dataset. The factors are presented in descending order of importance that each factor represent (second column). This column contains the eigenvalues, i.e. the variance accounted by each factor. We also show the percentage calculated (third column) based on the eigenvalues of each feature, as well as, the cumulative percentage in the last column.

Table 6: Principal Components Analysis

Factor	Eigenvalue	% Variance	% Cum. Var.
1	1.71	0.27	0.27
2	1.47	0.24	0.51
3	1.02	0.16	0.67
4	0.54	0.09	0.76
5	0.49	0.08	0.84
6	0.42	0.07	0.90
7	0.27	0.04	0.95
8	0.21	0.03	0.98
9	0.12	0.02	1.00

The number of factors used may influence the error variance, if too many factors are retained for further analysis. While, retaining few factors risk leaving out valuable common variance. A criteria to determine the number of factors to retain is the *Kaiser's criterion*, which basically is a rule of thumb to retain that recommend to retain the factor with eigenvalues greater than 1. To avoid overestimating, a scree test can provide a more robust method, by choosing factors before the flattening of the slope of eigenvalues. To determine the number of factor to retain, we use both: the *Kaiser's criterion* as well the scree test [23].

By combining the aforementioned rules, we retain factors 1, 2, and 3 in table 6. Together these factors represent 67% of the total variance of the features. Table 7 shows factor loadings (correlations) between the original features in table 5 and each of the three factors retained in the previous step.

To visualize the importance of each feature based on the correlation with factors, we plot factors in pairs in figure 5. Factors represents the axis of the graphs. For instance, the factor vector for feature *mentions* is represented in the first graph with coordinates $(0.27, 0.91)$. Likewise, in the second graph the same feature is represented with coordinates $(0.91, -0.24)$, illustrating the correlation feature-factors.

Table 7: PCA Factors Loadings

Feature	Factor1	Factor2	Factor3
fav given	0.35	−0.09	−0.08
followers	0.50	−0.20	0.12
friends	0.52	−0.01	0.07
statuses	0.53	−0.20	0.04
tokens	0.04	−0.17	−0.06
urls	0.02	0.01	0.04
hashtags	−0.02	0.25	0.95
mentions	0.27	0.91	−0.24
media	0.03	0.01	0.03

We interpret the first graph as the *networking and activity level* of the user based on the high correlation with user profile features. Activity level is related to the number of favorites the user has given to other tweets or the number of tweets created. The networking level is related to user attributes that indicates the network relationships (followers, friends).

The second graph can be interpreted also as the *content patterns* of the user. There is a slight negative correlation with the feature *mentions*. While other content related features of the tweet (represented by the number of urls, media, or hashtags) are slightly positive. Content features, such as number of tokens have negative correlation in the third factor.

An important aspect is the fact that *mentions* to other users often generate a response (reply) from them. Factor vectors of users' *mentions* and *hashtags* are predominant in both graphs of figure 5.

$$c(n) = \begin{cases} 0 & \text{if } nr = 0 \\ 1 & \text{if } nr > 0 \end{cases} \tag{1}$$

5.4 Prediction Model

Based on the underlying structure extracted in PCA, we aim to predict the likelihood of conversation arising for a given tweet. We define our problem as binary classification, i.e., whether a tweet will initiate a conversation or not. To tackle this problem, we initially train and evaluate a supervised logistic classifier using our prediction dataset. The classification model renders a set of prediction coefficients for each feature. These coefficients can be used in a logistic equation to calculate the probability of a tweet initiating a conversation.

We define the dependent variable as binary: the tweet initiates or not a conversation. We transform the original feature *number of replies (nr)* into a binary feature, as follows:

Table 8 shows the coefficients in the predictive model. These coefficients corroborate certain findings of the importance of certain features revealed through PCA analysis. Mainly, those related to the activity level (favorites given, number of tweets posted), the interaction (mentions in tweets), the network (followers), and content (tokens in tweets). Friends has negative coefficient, i.e., little influence in initiating conversations. Content features such as: urls

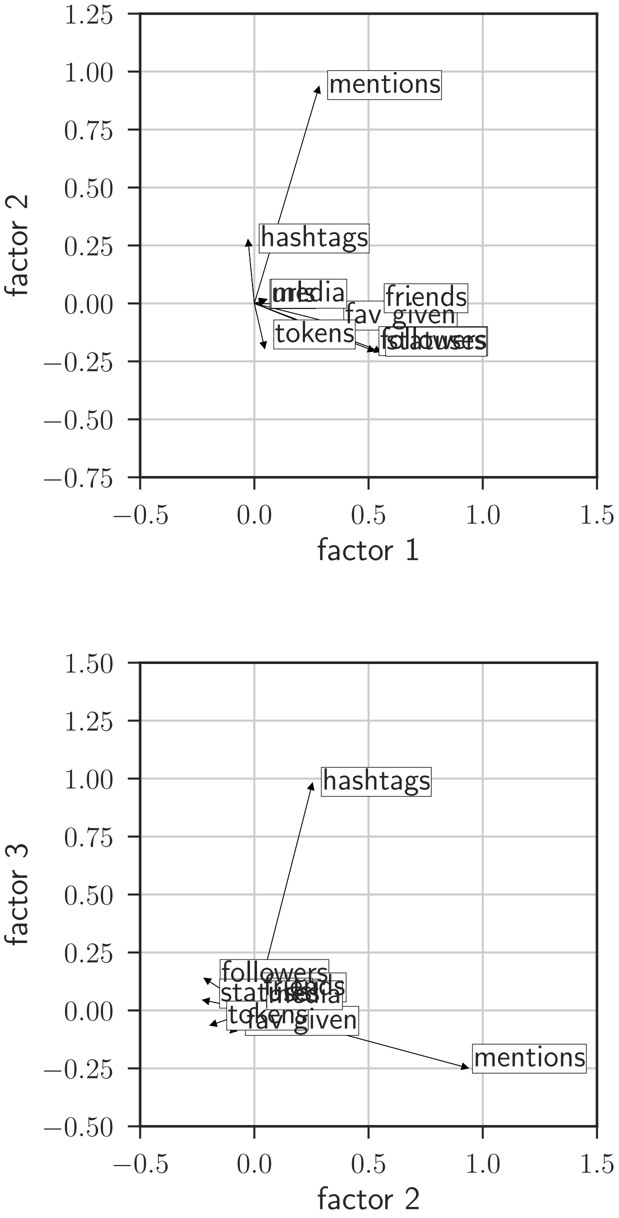

Figure 5: Factors mapping.

or hashtags present in tweets are less important in conversation forming, while mentions are slightly more important.

6 RESULTS

The classification pipeline consists in: extract and standardize the features, apply PCA, stratified splitting cross validation, and random grid search for hyper-parameters tuning. We evaluate several classifiers as shown in figure 6. We feed classifiers with two input datasets: content features only and full (user + content features). The first dataset considers only content features, i.e. tweet level features. The second consider both users and content level features.

Table 8: Predictive model

feat	coef	std err	z	P>|z|
intercept	0.185	0.082	2.254	0.024
fav given	0.771	0.087	8.832	0.000
followers	1.074	0.098	10.934	0.000
friends	−0.214	0.070	−3.047	0.002
statuses	−0.092	0.094	−0.973	0.331
tokens	0.449	0.083	5.409	0.000
urls	−0.550	0.112	−4.906	0.000
hashtags	−0.471	0.071	−6.650	0.000
mentions	0.088	0.047	1.893	0.058
media	−0.046	0.137	−0.337	0.736

The classification algorithms used in the evaluation are:

(1) Logistic Regression
(2) Support Vector Machine (SVM): RBF kernel
(3) Gaussian Naive Bayes (NB)
(4) Neural Net
(5) Naive Bayes

In figure 6, the visualizations use the first two factor obtained in PCA. The results show that Logistic Regression perform consistently using both datasets: content only features or all features. Using all features, SVM has the best performance (0.80), followed by Neural Net model (0.79). Neural Net model shows promising results, moreover if we want to include for more complex tasks that involves analyzing textual and visual content.

Additionally, we evaluate the performance of the best classifier by separating the dataset by percentiles (10, IQ, 90). We use the three features with higher coefficient in table 8. In the case of the feature *number of followers* the classifier performs better for the 90 percentile as those are users with high popularity, as their tweets are more likely to generate replies. The feature *favorites given* that denote the activity level of the user for the 90 percentile has similar behavior, but interestingly for users with few activity (10 percentile) performs better than for average users. The *number of tokens* created by users has similar behavior as *favorites given.*, but this could be due to the fact that IQ percentile contains more noisy tweets.

7 CONCLUSION

In this work, we analyze the factors that may influence conversation forming from a given tweet. We extracted both contextual and content features and establish their correlation using PCA, as well as, using predictive models.

In the exploratory analysis, we found the difficulties of working with noisy data found in Twitter. We also establish some considerations to avoid including noisy data in predictive analysis. Language, duration, distance, and number of users in conversations can help to filter irrelevant and non-conversational tweets. Regarding the predictive analysis, we found that the overall F1 score improve, if we consider both: users' profile features as well as, tweets' content features.

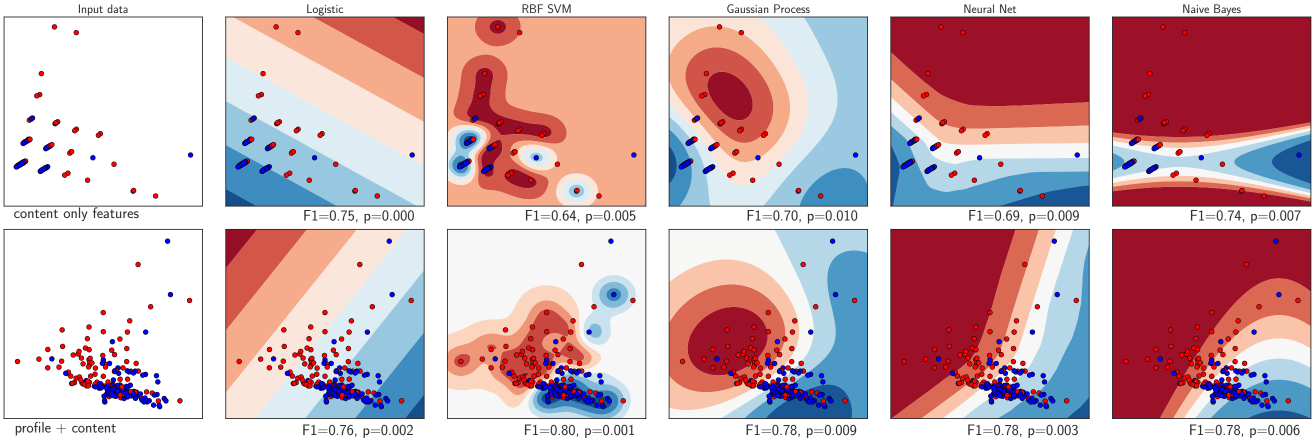

Figure 6: Classification models for identifying seed conversation tweets.

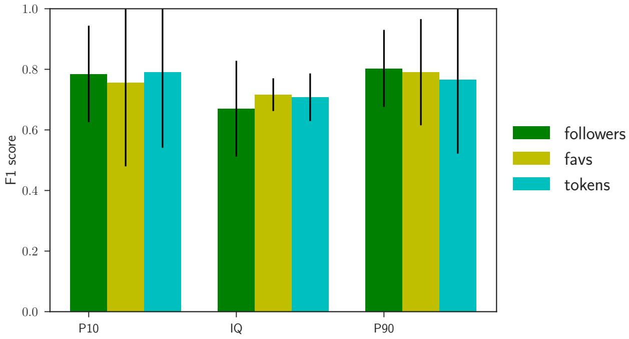

Figure 7: Classification comparison for different features percentiles.

In future work, we would like to explore large scale analysis for massive Twitter datasets using distributed machine learning. Also, we want to include additional features through analysis of textual and visual content of the tweets.

REFERENCES

[1] Veronika Abramova and Jorge Bernardino. 2013. NoSQL databases: MongoDB vs cassandra. In *Proceedings of the international C* conference on computer science and software engineering*. ACM, 14–22.

[2] James Allen, Nathanael Chambers, George Ferguson, Lucian Galescu, Hyuckchul Jung, Mary Swift, and William Taysom. 2007. Plow: A collaborative task learning agent. In *AAAI*. 1514–1519.

[3] Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Natural Language Processing for Mental Health: Large Scale Discourse Analysis of Counseling Conversations. *Transactions of the Association for Computational Linguistics* (2016).

[4] Danah Boyd, Scott Golder, and Gilad Lotan. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*. IEEE, 1–10.

[5] Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. 2014. Can cascades be predicted?. In *Proceedings of the 23rd international conference on World wide web*. ACM, 925–936.

[6] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-law distributions in empirical data. *SIAM review* 51, 4 (2009), 661–703.

[7] Micha Elsner and Eugene Charniak. 2008. You Talking to Me? A Corpus and Algorithm for Conversation Disentanglement.. In *ACL*. 834–842.

[8] Ji He, Mari Ostendorf, Xiaodong He, Jianshu Chen, Jianfeng Gao, Lihong Li, and Li Deng. 2016. Deep Reinforcement Learning with a Combinatorial Action Space for Predicting Popular Reddit Threads. *arXiv preprint arXiv:1606.03667* (2016).

[9] Courtenay Honey and Susan C Herring. 2009. Beyond microblogging: Conversation and collaboration via Twitter. In *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on*. IEEE, 1–10.

[10] Liangjie Hong, Ovidiu Dan, and Brian D Davison. 2011. Predicting popular messages in twitter. In *Proceedings of the 20th international conference companion on World wide web*. ACM, 57–58.

[11] Yijue How and Min-Yen Kan. 2005. Optimizing predictive text entry for short message service on mobile phones. In *Proceedings of HCII*, Vol. 5.

[12] Avinash Lakshman and Prashant Malik. 2010. Cassandra: a decentralized structured storage system. *ACM SIGOPS Operating Systems Review* 44, 2 (2010), 35–40.

[13] Gustavo López, Luis Quesada, and Luis A Guerrero. 2017. Alexa vs. Siri vs. Cortana vs. Google Assistant: A Comparison of Speech-Based Natural User Interfaces. In *International Conference on Applied Human Factors and Ergonomics*. Springer, 241–250.

[14] Shereen Oraby, Pritam Gundecha, Jalal Mahmud, Mansurul Bhuiyan, and Rama Akkiraju. 2017. How May I Help You?: Modeling Twitter Customer Service-Conversations Using Fine-Grained Dialogue Acts. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. ACM, 343–355.

[15] Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 172–180.

[16] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models.. In *AAAI*. 3776–3784.

[17] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H Chi. 2010. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Social computing (socialcom), 2010 ieee second international conference on*. IEEE, 177–184.

[18] Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic-and author-controlled natural experiments on Twitter. *arXiv preprint arXiv:1405.1438* (2014).

[19] Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann. 2014. *Twitter and society*. Vol. 89. P. Lang.

[20] Yorick Wilks. 2006. Artificial companions as a new kind of interface to the future internet. (2006).

[21] Tae Yano, William W Cohen, and Noah A Smith. 2009. Predicting response to political blog posts with topic models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 477–485.

[22] Shaozhi Ye and Shyhtsun Felix Wu. 2010. Measuring message propagation and social influence on twitter. com. *SocInfo* 10 (2010), 216–231.

[23] An Gie Yong and Sean Pearce. 2013. A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in quantitative methods for psychology* 9, 2 (2013), 79–94.