Predicting and Characterising User Impact on Twitter

Vasileios Lampos[1], Nikolaos Aletras[2], Daniel Preoţiuc-Pietro[2] and **Trevor Cohn[3]**

[1] Department of Computer Science, University College London
[2] Department of Computer Science, University of Sheffield
[3] Computing and Information Systems, The University of Melbourne
v.lampos@ucl.ac.uk, {n.aletras,d.preotiuc}@dcs.shef.ac.uk, trevor.cohn@gmail.com

Abstract

The open structure of online social networks and their uncurated nature give rise to problems of user credibility and influence. In this paper, we address the task of predicting the impact of Twitter users based only on features under their direct control, such as usage statistics and the text posted in their tweets. We approach the problem as regression and apply linear as well as non-linear learning methods to predict a user impact score, estimated by combining the numbers of the user's followers, followees and listings. The experimental results point out that a strong prediction performance is achieved, especially for models based on the Gaussian Processes framework. Hence, we can interpret various modelling components, transforming them into indirect 'suggestions' for impact boosting.

1 Introduction

Online social networks have become a wide spread medium for information dissemination and interaction between millions of users (Huberman et al., 2009; Kwak et al., 2010), turning, at the same time, into a popular subject for interdisciplinary research, involving domains such as Computer Science (Sakaki et al., 2010), Health (Lampos and Cristianini, 2012) and Psychology (Boyd et al., 2010). Open access along with the property of structured content retrieval for publicly posted data have brought the microblogging platform of Twitter into the spotlight.

Vast quantities of human-generated text from a range of themes, including opinions, news and everyday activities, spread over a social network. Naturally, issues arise, like user credibility (Castillo et al., 2011) and content attractiveness (Suh et al., 2010), and quite often trustful or appealing information transmitters are identified by an impact assessment.[1] Intuitively, it is expected that user impact cannot be defined by a single attribute, but depends on multiple user actions, such as posting frequency and quality, interaction strategies, and the text or topics of the written communications.

In this paper, we start by predicting user impact as a statistical learning task (regression). For that purpose, we firstly define an impact score function for Twitter users driven by basic account properties. Afterwards, from a set of accounts, we measure several publicly available attributes, such as the quantity of posts or interaction figures. Textual attributes are also modelled either by word frequencies or, more generally, by clusters of related words which quantify a topic-oriented participation. The main hypothesis being tested is whether textual and non textual attributes encapsulate patterns that affect the impact of an account.

To model this data, we present a method based on nonlinear regression using Gaussian Processes, a Bayesian non-parametric class of methods (Rasmussen and Williams, 2006), proven more effective in capturing the multimodal user features. The modelling choice of excluding components that are not under an account's direct control (e.g. received retweets) combined with a significant user impact prediction performance ($r = .78$) enabled us to investigate further how specific aspects of a user's behaviour relate to impact, by examining the parameters of the inferred model.

Among our findings, we identify relevant features for this task and confirm that consistent activity and broad interaction are deciding impact factors. Informativeness, estimated by computing a joint user-topic entropy, contributes well to the separation between low and high impact accounts. Use case scenarios based on combinations of features are also explored, leading to findings such as that engaging about 'serious' or more 'light' topics may not register a differentiation in impact.

[1] For example, the influence assessment metric of Klout — http://www.klout.com.

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 405–413, Gothenburg, Sweden, April 26-30 2014.

2 Data

For the experimental process of this paper, we formed a Twitter data set ($\mathcal{D}1$) of more than 48 million tweets produced by $|U| = 38,020$ users geolocated in the UK in the period between 14/04/2011 and 12/04/2012 (both dates included, $\Delta t = 365$ days). $\mathcal{D}1$ is a temporal subset of the data set used for modelling UK voting intentions in (Lampos et al., 2013). Geolocation of users was carried out by matching the location field in their profile with UK city names on DBpedia as well as by checking that the user's timezone is set to G.M.T. (Rout et al., 2013). The use of a common greater geographical area (UK) was essential in order to derive a data set with language and topic homogeneity. A distinct Twitter data set ($\mathcal{D}2$) consisting of approx. 400 million tweets was formed for learning term clusters (Section 4.2). $\mathcal{D}2$ was retrieved from Twitter's Gardenhose stream (a 10% sample of the entire stream) from $02/01$ to $28/02/2011$. $\mathcal{D}1$ and $\mathcal{D}2$ were processed using TrendMiner's pipeline (Preoţiuc-Pietro et al., 2012).

3 User Impact Definition

On the microblogging platform of Twitter, user – or, in general, account – popularity is usually quantified by the raw number of followers ($\phi_{\text{in}} \geq 0$), i.e. other users interested in this account. Likewise, a user can follow others, which we denote as his set of followees ($\phi_{\text{out}} \geq 0$). It is expected that users with high numbers of followers are also popular in the real world, being well-known artists, politicians, brands and so on. However, non popular entities, the majority in the social network, can also gain a great number of followers, by exploiting, for example, a follow-back strategy.[2] Therefore, using solely the number of followers to quantify impact may lead to inaccurate outcomes (Cha et al., 2010). A natural alternative, the ratio of $\phi_{\text{in}}/\phi_{\text{out}}$ is not a reliable metric, as it is invariant to scaling, i.e. it cannot differentiate accounts of the type $\{\phi_{\text{in}}, \phi_{\text{out}}\} = \{m, n\}$ and $\{\gamma \times m, \gamma \times n\}$. We resolve this problem by squaring the number of followers $\left(\phi_{\text{in}}^2/\phi_{\text{out}}\right)$; note that the previous expression is equal to $(\phi_{\text{in}} - \phi_{\text{out}}) \times (\phi_{\text{in}}/\phi_{\text{out}}) + \phi_{\text{in}}$ and thus, it incorporates the ratio as well as the difference between followers and followees.

An additional impact indicator is the number of times an account has been listed by others ($\phi_\lambda \geq 0$). Lists provide a way to curate content on Twitter; thus, users included in many lists are attractors of interest. Indeed, Pearson's correlation between ϕ_{in} and ϕ_λ for all the accounts in our data set is equal to .765 ($p < .001$); the two metrics are correlated, but not entirely and on those grounds, it would be reasonable to use both for quantifying impact.

[2] An account follows other accounts randomly expecting that they will follow back.

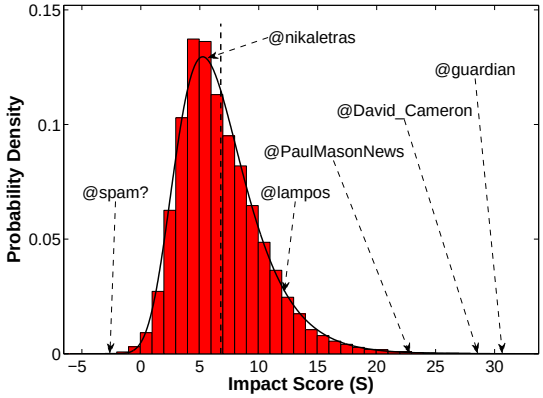

Figure 1: Histogram of the user impact scores in our data set. The solid black line represents a generalised extreme value probability distribution fitted in our data, and the dashed line denotes the mean impact score (= 6.776). User @spam? is a sample account with $\phi_{\text{in}} = 10$, $\phi_{\text{out}} = 1000$ and $\phi_\lambda = 0$; @lampos is a very active account, whereas @nikaletras is a regular user.

Consequently, we have chosen to represent user impact (S) as a log function of the number of followers, followees and listings, given by

$$\text{S}(\phi_{\text{in}}, \phi_{\text{out}}, \phi_\lambda) = \ln\left(\frac{(\phi_\lambda + \theta)\,(\phi_{\text{in}} + \theta)^2}{\phi_{\text{out}} + \theta}\right), \quad (1)$$

where θ is a smoothing constant set equal to 1 so that the natural logarithm is always applied on a real positive number. Figure 1 shows the impact score distribution for all the users in our sample, including some pointers to less or more popular Twitter accounts. The depicted user impact scores form the response variable in the regression models presented in the following sections.

4 User Account Features

This section presents the features used in the user impact prediction task. They are divided into two categories: non-textual and text-based. All features have the joint characteristic of being under the user's direct control, something essential for characterising impact based on the actions of a user. Attributes such as the number of received retweets or @-mentions (of a user in the tweets of others) were not considered as they are not controlled by the account itself.

a_1	# of tweets
a_2	proportion of retweets
a_3	proportion of non-duplicate tweets
a_4	proportion of tweets with hashtags
a_5	hashtag-tokens ratio in tweets
a_6	proportion of tweets with @-mentions
a_7	# of unique @-mentions in tweets
a_8	proportion of tweets with @-replies
a_9	links ratio in tweets
a_{10}	# of favourites the account made
a_{11}	total # of tweets (entire history)
a_{12}	using default profile background (binary)
a_{13}	using default profile image (binary)
a_{14}	enabled geolocation (binary)
a_{15}	population of account's location
a_{16}	account's location latitude
a_{17}	account's location longitude
a_{18}	proportion of days with nonzero tweets

Table 1: Non textual attributes for a Twitter account used in the modelling process. All attributes refer to a set of 365 days (Δt) with the exception of a_{11}, the total number of tweets in the entire history of an account. Attributes a_i, $i \in \{2-6, 8, 9\}$ are ratios of a_1, whereas attribute a_{18} is a proportion of Δt.

4.1 Non textual attributes

The non-textual attributes (a) are derived either from general user behaviour statistics or directly from the account's profile. Table 1 presents the 18 attributes we extracted and used in our models.

4.2 Text features

We process the text in the tweets of $\mathcal{D}1$ and compute daily unigram frequencies. By discarding terms that appear less than 100 times, we form a vocabulary of size $|V| = 71,555$. We then form a user term-frequency matrix of size $|U| \times |V|$ with the mean term frequencies per user during the time interval Δt. All term frequencies are normalised with the total number of tweets posted by the user.

Apart from single word frequencies, we are also interested in deriving a more abstract representation for each user. To achieve this, we learn word clusters from a distinct reference corpus ($\mathcal{D}2$) that could potentially represent specific domains of discussion (or topics). From a multitude of proposed techniques, we have chosen to apply spectral clustering (Shi and Malik, 2000; Ng et al., 2002), a hard-clustering method appropriate for high-dimensional data and non-convex clusters (von Luxburg, 2007). Spectral clustering performs graph partitioning on the word-by-word similarity matrix. In our case, tweet-term similarity is reflected by the Normalised Pointwise Mutual Information (NPMI), an information theoretic measure indicating which words co-occur in the same context (Bouma, 2009). We use the random walk graph Laplacian and only keep the largest component of the resulting graph, eliminating most stop words in the process. The number of clusters needs to be specified in advance and each cluster's most representative words are identified by the following metric of centrality:

$$\mathrm{C}_w(c) = \frac{\sum_{v \in c} \mathrm{NPMI}(w, v)}{|c| - 1}, \qquad (2)$$

where w is the target word and c the cluster it belongs ($|c|$ denotes the cluster's size). Examples of extracted word clusters are illustrated in Table 4. Other techniques were also applied, such as online LDA (Hoffman et al., 2010), but we found that the results were not satisfactory, perhaps due to the short message length and the foreign terms co-occuring within a tweet. After forming the clusters using $\mathcal{D}2$, we compute a topic score (τ) for each user-topic pair in $\mathcal{D}1$, representing a normalised user-word frequency sum per topic.

5 Methods

This section presents the various modelling approaches for the underlying inference task, the impact score (S) prediction of Twitter users based on a set of their actions.

5.1 Learning functions for regression

We formulate this problem as a regression task, i.e. we infer a real numbered value based on a set of observed features. As a simple baseline, we apply Ridge Regression (RR) (Hoerl and Kennard, 1970), a reguralised version of the ordinary least squares. Most importantly, we focus on nonlinear methods for the impact score prediction task given the multimodality of the feature space. Recently, it was shown by Cohn and Specia (2013) that Support Vector Machines for Regression (SVR) (Vapnik, 1998; Smola and Schölkopf, 2004), commonly considered the state-of-the-art for NLP regression tasks, can be outperformed by Gaussian Processes (GPs), a kernelised, probabilistic approach to learning (Rasmussen and Williams, 2006). Their setting is close to ours, in that they had few (17) features and were also aiming to predict a complex continuous phenomenon (human post-editing time). The initial stages of our experimental process confirmed that GPs performed better than SVR; thus,

we based our modelling around them, including RR for comparison.

In GP regression, for the inputs $\boldsymbol{x} \in \mathbb{R}^d$ we want to learn a function $f\colon \mathbb{R}^d \rightarrow \mathbb{R}$ that is drawn from a GP prior

$$f(\boldsymbol{x}) \sim \mathcal{GP}\left(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x'})\right), \quad (3)$$

where $m(\boldsymbol{x})$ and $k(\boldsymbol{x}, \boldsymbol{x'})$ denote the mean (set to 0 in our experiments) and covariance (or kernel) functions respectively. The GP kernel function represents the covariance between pairs of input. We wish to limit f to smooth functions over the inputs, with different smoothness in each input dimension, assuming that some features are more useful than others. This can be accommodated by a squared exponential covariance function with Automatic Relevance Determination (ARD) (Neal, 1996; Williams and Rasmussen, 1996):

$$k_{\text{ard}}(\boldsymbol{x}, \boldsymbol{x'}) = \sigma^2 \exp\left[\sum_i^d -\frac{(x_i - x_i')^2}{2\ell_i^2}\right], \quad (4)$$

where σ^2 denotes the overall variance and ℓ_i is the length-scale parameter for feature x_i; all hyperparameters are learned from data during model inference. Parameter ℓ_i is inversely proportional to the feature's relevancy in the model, i.e. high values of ℓ_i indicate a low degree of relevance for the corresponding x_i. By setting $\ell_i = \ell$ in Eq. 4, we learn a common length-scale for all the dimensions – this is known as the isotropic squared exponential function (k_{iso}) since it is based purely on the difference $|\boldsymbol{x} - \boldsymbol{x'}|$. k_{iso} is a preferred choice when the dimensionality of the input space is high. Having set our covariance functions, predictions are conducted using Bayesian integration

$$\text{P}(y_*|\boldsymbol{x}_*, \mathcal{O}) = \int_f \text{P}(y_*|\boldsymbol{x}_*, f)\text{P}(f|\mathcal{O}), \quad (5)$$

where y_* is the response variable, $\mathcal{O}$ a labelled training set and $\boldsymbol{x}_*$ the current observation. We learn the hyperparameters of the model by maximising the log marginal likelihood $\text{P}(y|\mathcal{O})$ using gradient ascent. However, inference becomes intractable when many training instances (n) are present as the number of computations needed is $\mathcal{O}(n^3)$ (Quiñonero-Candela and Rasmussen, 2005). Since our training samples are tens of thousands, we apply a sparse approximation method (FITC), which bases parameter learning on a few inducing points in the training set (Quiñonero-Candela and Rasmussen, 2005; Snelson and Ghahramani, 2006).

5.2 Models

For predicting user impact on Twitter, we develop three regression models that build on each other. The first and simplest one (A) uses only the non-textual attributes as features; the performance of A is tested using RR,[3] SVR as well as a GP model. For SVR we used an RBF kernel (equivalent to k_{iso}), whereas for the GP we applied the following covariance function

$$k(\boldsymbol{a}, \boldsymbol{a'}) = k_{\text{ard}}(\boldsymbol{a}, \boldsymbol{a'}) + k_{\text{noise}}(\boldsymbol{a}, \boldsymbol{a'}) + \beta, \quad (6)$$

where $k_{\text{noise}}(\boldsymbol{a}, \boldsymbol{a'}) = \sigma^2 \times \delta(\boldsymbol{a}, \boldsymbol{a'})$, δ is a Kronecker delta function and β is the regression bias; this function consists of $(|a| + 3)$ hyperparameters. Note that the sum of covariance functions is also a valid covariance function (Rasmussen and Williams, 2006).

The second model (AW) extends model A by adding word-frequencies as features. The 500 most frequent terms in $\mathcal{D}1$ are discarded as stop words and we use the following $2,000$ ones (denoted by $\boldsymbol{w}$). Setting $\boldsymbol{x} = \{\boldsymbol{a}, \boldsymbol{w}\}$, the covariance function becomes

$$\begin{aligned} k(\boldsymbol{x}, \boldsymbol{x'}) = k_{\text{ard}}(\boldsymbol{a}, \boldsymbol{a'}) + k_{\text{iso}}(\boldsymbol{w}, \boldsymbol{w'}) \\ + k_{\text{noise}}(\boldsymbol{x}, \boldsymbol{x'}) + \beta, \end{aligned} \quad (7)$$

where we apply k_{iso} on the term-frequencies due to their high dimensionality; the number of hyperparameters is $(|a| + 5)$. This is an intermediate model aiming to evaluate whether the incorporation of text improves prediction performance.

Finally, in the third model (AC) instead of relying on the high dimensional space of single words, we use topic-oriented collections of terms extracted by applying spectral clustering (see Section 4.2). By denoting the set of different clusters or topics as $\boldsymbol{\tau}$ and the entire feature space as $\boldsymbol{x} = \{\boldsymbol{a}, \boldsymbol{\tau}\}$, the covariance function now becomes

$$k(\boldsymbol{x}, \boldsymbol{x'}) = k_{\text{ard}}(\boldsymbol{x}, \boldsymbol{x'}) + k_{\text{noise}}(\boldsymbol{x}, \boldsymbol{x'}) + \beta. \quad (8)$$

The number of hyperparameters is equal to $(|a| + |\tau| + 3)$ and this model is applied for $|\tau| = 50$ and 100.

6 Experiments

Here we present the experimental results for the user impact prediction task and then investigate the factors that can affect it.

6.1 Predictive Accuracy

We evaluated the performance of the proposed models via 10-fold cross-validation. Results are presented in Table 2; Root Mean Squared Error

[3] Given that the representation of attributes a_{16} and a_{17} (latitude, longitude) is ambiguous in a linear model, they were not included in the RR-based models.

Model	Linear (RR)		Nonlinear (GP)	
	r	RMSE	r	RMSE
A	.667	2.642	.759	2.298
AW	.712	2.529	.768	2.263
AC, $\|\tau\| = 50$	.703	2.518	.774	2.234
AC, $\|\tau\| = 100$	.714	2.480	**.780**	**2.210**

Table 2: Average performance (RMSE and Pearson's r) derived from 10-fold cross-validation for the task of user impact score prediction.

Model	Top relevant features
A	$a^{\diamond}_{13}, a_{11}, a_7, a_1, a_9, a_8, a_{18}, a_4, a_6, a_3$
AW	$a_7, a_1, a_{11}, a^{\diamond}_{13}, a_9, a_8, a_{18}, a_4, a_6, a_{15}$
AC, $\tau = 50$	$a^{\diamond}_{13}, a_{11}, a_7, \tau'_1, a_1, a_9, a_8, \tau'_2, a_6, \tau'_3$
AC, $\tau = 100$	$a^{\diamond}_{13}, a_{11}, a_7, a_1, a_9, \tau_1, \tau_2, \tau_3, a_{18}, a_8$

Table 3: The 10 most relevant features in descending relevance order for all GP models. τ'_i and τ_i denote word clusters (may vary in each model).[6]

(RMSE) and Pearson's correlation (r) between predictions and responses were used as the performance metrics. Overall, the best performance in terms of both RMSE (2.21 impact points) and linear correlation ($r = .78$, $p < .001$) is achieved by the GP model (AC) that combines non-textual attributes with a 100 topic clusters; the difference in performance with all other models is statistically significant.[4] The linear baseline (RR) follows the same pattern of improvement through the different models, but never manages to reach the performance of the nonlinear alternative. As mentioned previously, we have also tried SVR with an RBF kernel for model A (parameters were optimised on a held-out development set) and the performance (RMSE: 2.33, $r = .75$, $p < .001$) was significantly worse than the one achieved by the GP model.[4]

Notice that when word-based features are introduced in model AW, performance improves. This was one of the motivations for including text in the modelling, apart from the notion that the posted content should also affect general impact. Lastly, turning this problem from regression to classification by creating 3 impact score pseudo-classes based on the .25 and the .9 quantiles of the response variable (4.3 and 11.4 impact score points respectively) and by using the outputs of model AC ($\tau = 100$) in each phase of the 10-fold cross-validation, we achieve a 75.86% classification accuracy.[5]

[4] Indicated by performing a t-test (5% significance level).

[5] Similar performance scores can be estimated for different class threshold settings.

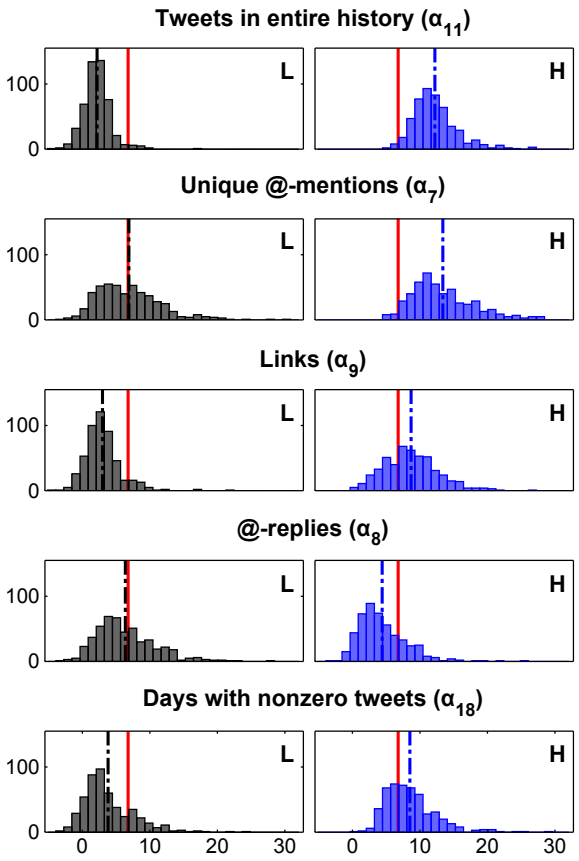

Figure 2: User impact distribution (x-axis: impact points, y-axis: # of user accounts) for users with a low (L) or a high (H) participation in a selection of relevant non-textual attributes. Dot-dashed lines denote the respective mean impact score; the red line is the mean of the entire sample (= 6.776).

6.2 Qualitative Analysis

Given the model's strong performance, we now conduct a more thorough analysis to identify and characterise the properties that affect aspects of the user impact. GP's length-scale parameters (ℓ_i) – which are inversely proportional to feature relevancy – are used for ranking feature importance. Note that since our data set consists of UK users, some results may be biased towards specific cultural properties.

Non-textual attributes. Table 3 lists the 10 most relevant attributes (or topics, where applicable) as extracted in each GP model. Ranking is determined by the mean value of the length-scale parameter for each feature in the 10-fold cross-validation process. We do not show feature ranking derived from the RR models as we focus on the models with the best performance. Despite this, it is worth mentioning

[6] Length-scales are comparable for features of the same variance (z-scored). Binary features (denoted by $\diamond$) are not z-scored, but for comparison purposes we have rescaled their length-scale using the feature's variance.

Label	$\mu(\ell) \pm \sigma(\ell)$	Cluster's words ranked by centrality	$\lvert c \rvert$
τ_1: Weather	3.73 ± 1.80	mph, humidity, barometer, gust, winds, hpa, temperature, kt, #weather [...]	309
τ_2: Healthcare Finance Housing	5.44 ± 1.55	nursing, nurse, rn, registered, bedroom, clinical, #news, estate, #hospital, rent, healthcare, therapist, condo, investment, furnished, medical, #nyc, occupational, investors, #ny, litigation, tutors, spacious, foreclosure [...]	1281
τ_3: Politics	6.07 ± 2.86	senate, republican, gop, police, arrested, voters, robbery, democrats, presidential, elections, charged, election, charges, #religion, arrest, repeal, dems, #christian, reform, democratic, pleads, #jesus, #atheism [...]	950
τ_4: Showbiz Movies TV	7.36 ± 2.25	damon, potter, #tvd, harry, elena, kate, portman, pattinson, hermione, jennifer, kristen, stefan, robert, catholic, stewart, katherine, lois, jackson, vampire, natalie, #vampirediaries, tempah, tinie, weasley, turner, rowland [...]	1943
τ_5: Commerce	7.83 ± 2.77	chevrolet, inventory, coupon, toyota, mileage, sedan, nissan, adde, jeep, 4x4, 2002, #coupon, enhanced, #deal, dodge, gmc, 20%, suv, 15%, 2005, 2003, 2006, coupons, discount, hatchback, purchase, #ebay, 10% [...]	608
τ_6: Twitter Hashtags	8.22 ± 2.98	#teamfollowback, #500aday, #tfb, #instantfollowback, #ifollowback, #instantfollow, #followback, #teamautofollow, #autofollow, #mustfollow [...]	194
τ_7: Social Unrest	8.37 ± 5.52	#egypt, #tunisia, #iran, #israel, #palestine, tunisia, arab, #jan25, iran, israel, protests, egypt, #yemen, #iranelection, israeli, #jordan, regime, yemen, #gaza, protesters, #lebanon, #syria, egyptian, #protest, #iraq [...]	321
τ_8: Non English	8.45 ± 3.80	yg, nak, gw, gue, kalo, itu, aku, aja, ini, gak, klo, sih, tak, mau, buat [...]	469
τ_9: Horoscope Gambling	9.11 ± 3.07	horoscope, astrology, zodiac, aries, libra, aquarius, pisces, taurus, virgo, capricorn, horoscopes, sagitarius, comprehensive, lottery, jackpot [...]	1354
τ_{10}: Religion Sports	10.29 ± 6.27	#jesustweeters, psalm, christ, #nhl, proverbs, unto, salvation, psalms, lord, kjv, righteousness, niv, bible, pastor, #mlb, romans, awards, nhl [...]	1610

Table 4: The 10 most relevant topics (for model AC, $|\tau| = 100$) in the prediction of a user's impact score together with their most central words. The topics are ranked by their mean length-scale, $\mu(\ell)$, in the 10-fold cross-validation process ($\sigma(\ell)$ is the respective standard deviation).

that RR's outputs also followed similar ranking patterns, e.g. the top 5 features in model A were a_{18}, a_7, a_3, a_{11} and a_9. Notice that across all models, among the strongest features are the total number of posts either in the entire account's history (a_{11}) or within the 365-day interval of our experiment (a_1) and the number of unique @-mentions (a_7), good indicators of user activity and user interaction respectively. Feature a_{13} is also a very good predictor, but is of limited utility for modelling our data set because very few accounts maintain the default profile photo (0.4%). Less relevant attributes (not shown) are the ones related to the location of a user (a_{16}, a_{17}) signalling that the whereabouts of a user may not necessarily relate to impact. Another low relevance attribute is the number of favourites that an account did (a_{10}), something reasonable, as those weak endorsements are not affecting the main stream of content updates in the social network.

In Figure 2, we present the distribution of user impact for accounts with low (left-side) and high (right-side) participation in a selection of non-textual attributes. Low (L) and high (H) participations are defined by selecting the 500 accounts with lowest and highest scores for this specific attribute. The means of (L) and (H) are compared with the mean impact score in our sample. As anticipated, accounts with low activity (a_{11}) are likely to be assigned impact scores far below the mean, while very active accounts may follow a quite opposite pattern. Avoiding mentioning (a_7) or replying (a_8) to others may not affect (on average) an impact score positively or negatively; however, accounts that do many unique @-mentions are distributed around a clearly higher impact score. On the other hand, users that overdo @-replies are distributed below the mean impact score. Furthermore, accounts that post irregularly with gaps longer than a day (a_{18}) or avoid using links in their tweets (a_9) will probably appear in the low impact score range.

Topics. Regarding prediction accuracy (Table 2), performance improves when topics are included. In turn, some of the topics replace non-textual attributes in the relevancy ranking (Table 3). Table 4 presents the 10 most relevant topic word-clusters based on their mean length-scale $\mu(\ell)$ in the 10-fold cross-validation process for the best performing GP model (AC, $|\tau| = 100$). We see that clusters with their most central words representing topics such as 'Weather', 'Healthcare/Finance', 'Politics' and 'Showbiz' come up on top.

Contrary to the non-textual attributes, accounts with low participation in a topic (for the vast majority of topics) were distributed along impact score values lower than the mean. Based on the fact that word clusters are not small in size, this is a rational outcome indicating that accounts with small word-frequency sums (i.e. the ones that do not tweet much) will more likely be users with small impact

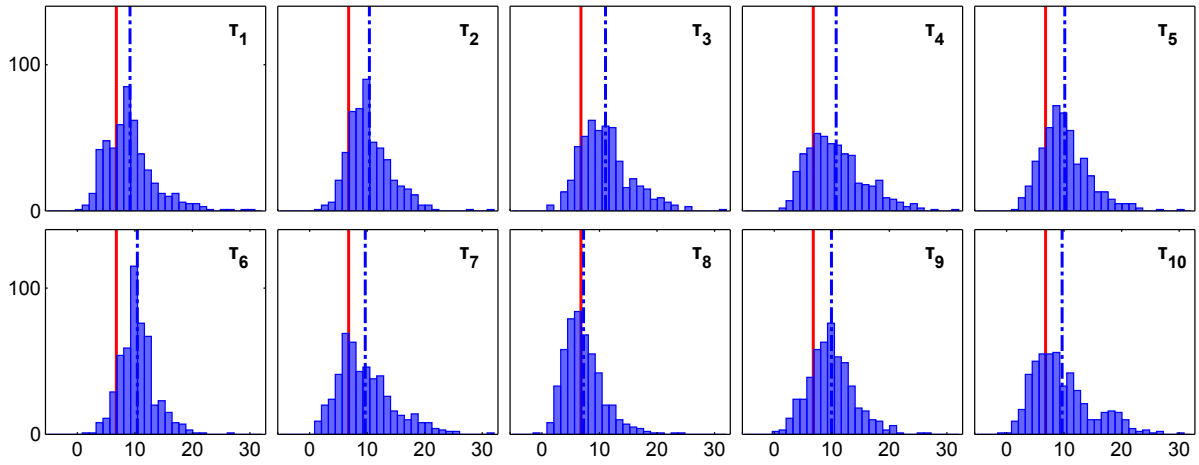

Figure 3: User impact distribution (x-axis: impact points, y-axis: # of user accounts) for accounts with a high participation in the 10 most relevant topics. Dot-dashed lines denote mean impact scores; the red line is the mean of the entire sample ($= 6.776$).

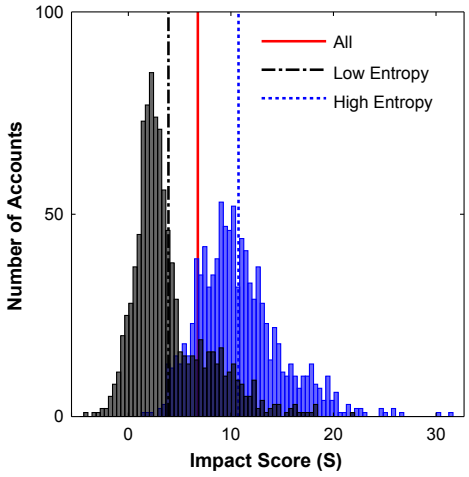

Figure 4: User impact distribution for accounts with high (blue) and low (dark grey) topic entropy. Lines denote the respective mean impact scores.

scores. Hence, in Figure 3 we only show the user impact distribution for the 500 accounts with the top participation in each topic. Informally, this is a way to quantify the contribution of each domain or topic of discussion in the impact score. Notice that the topics which 'push' users towards the highest impact scores fall into the domains of 'Politics' (τ_3) and 'Showbiz' (τ_4). An equally interesting observation is that engaging a lot about a specific topic will more likely result to a higher than average impact; the only exception is τ_8 which does not deviate from the mean, but τ_8 rather represents the use of a non-English language (Indonesian) and therefore, does not form an actual topic of discussion.

To further understand how participation in the 10 most relevant topics relates to impact, we also computed the joint user-topic entropy defined by

$$\mathrm{H}(u_i, \tau) = -\sum_{j=1}^{M} \mathrm{P}(u_i, \tau_j) \times \log_2 \mathrm{P}(u_i, \tau_j), \quad (9)$$

where u_i is a user and $M = 10$ (Shannon, 2001). This is a measure of user pseudo-informativeness, meaning that users with high entropy are considered as more informative (without assessing the quality of the information). Figure 4 shows the impact score distributions for the 500 accounts with the lowest and highest entropy. Low and high entropies are separated, with the former being placed clearly below the mean user impact score and the latter above. This pictorial assessment suggests that a connection between informativeness and impact may exist, at least in their extremes (their correlation in the entire sample is $r = .35$, $p < .001$).

Use case scenarios. Most of the previous analysis focused on the properties of single features. However, the user impact prediction models we learn depend on feature combinations. For that reason, it is of interest to investigate use case scenarios that bring various attributes together. To reduce notation in this paragraph, we use x_i^+ (x is either a non-textual attribute a or a topic τ) to express $x_i > \mu(x_i)$, the set of users for which the value of feature x_i is above the mean; equivalently $x_i^- : x_i < \mu(x_i)$. We also use τ_A^* to express the more complex set $\{\tau_A^+ \cap \tau_j^- \cap ... \cap \tau_z^-\}$, an intersection of users that are active in one topic (τ_A), but not very active in the rest. Figure 5 depicts the user impact distributions for five use case scenarios. Scenario A compares interactive to non interactive users, represented by $\mathrm{P}(a_1^+, a_6^+, a_7^+, a_8^+)$ and $\mathrm{P}(a_1^+, a_6^-, a_7^-, a_8^-)$ respectively; interactivity, defined by an intersection of accounts that tweet regularly, do many @-mentions and @-replies, but also

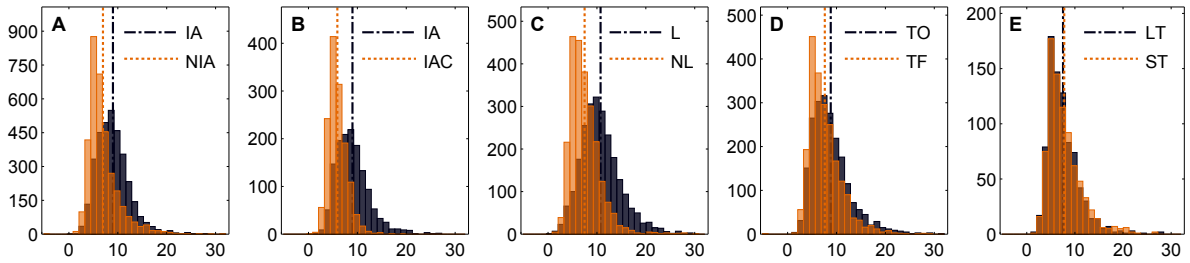

Figure 5: User impact distribution (x-axis: impact points, y-axis: # of user accounts) for five Twitter use scenarios based on subsets of the most relevant attributes and topics – IA: Interactive, IAC: Clique Interactive, L: Using many links, TO: Topic-Overall, TF: Topic-Focused, LT: 'Light' topics, ST: 'Serious' topics. (N) denotes negation and lines the respective mean impact scores.

mention many different users, seems to be rewarded on average with higher impact scores. Interactive users gain more impact than clique-interactive accounts represented by $\mathrm{P}(a_1^+, a_6^+, a_7^-, a_8^+)$, i.e. users who interact, but do not mention many different accounts, possibly because they are conducting discussions with a specific circle only (scenario B). The use of links when writing about the most prevalent topics ('Politics' and 'Showbiz') appears to be an important impact-wise factor (scenario C); the compared probability distributions in that case were $\mathrm{P}\big(a_1^+, (\tau_3^+ \cup \tau_4^+), a_9^+\big)$ against $\mathrm{P}\big(a_1^+, (\tau_3^+ \cup \tau_4^+), a_9^-\big)$. Surprisingly, when links were replaced by hashtags in the previous distributions, a clear class separation was not achieved. In scenario D, topic-focused accounts, i.e. users that write about one topic consistently, represented by $\mathrm{P}\big(a_1^+, (\tau_2^* \cup \tau_3^* \cup \tau_4^* \cup \tau_7^* \cup \tau_{10}^*)\big)$, have on average slightly worse impact scores when compared to accounts tweeting about many topics, $\mathrm{P}(a_1^+, \tau_2^+, \tau_3^+, \tau_4^+, \tau_7^+, \tau_{10}^+)$. Finally, scenario E shows thats users engaging about more 'serious' topics, $\mathrm{P}\big(a_1^+, \tau_4^-, \tau_5^-, \tau_9^-, (\tau_3^+ \cup \tau_7^+)\big)$, were not differentiated from the ones posting about more 'light' topics, $\mathrm{P}\big(a_1^+, (\tau_4^+ \cup \tau_5^+ \cup \tau_9^+), \tau_3^-, \tau_7^-\big)$.

7 Related Work

The task of user-impact prediction based on a machine learning approach that incorporates text features is novel, to the best of our knowledge. Despite this fact, our work is partly related to research approaches for quantifying and analysing user influence in online social networks. For example, Cha et al. (2010) compared followers, retweets and @-mentions received as measures of influence. Bakshy et al. (2011) aggregated all posts by each user, computed an individual-level influence and then tried to predict it by modelling user attributes (# of followers, followees, tweets and date of joining) together with past user influence. Their method, based on classification and regression trees (Breiman, 1984), achieved a modest performance ($r = .34$). Furthermore, Romero et al. (2011) proposed an algorithm for determining user influence and passivity based on information-forwarding activity, and Luo et al. (2013) exploited user attributes to predict retweet occurrences. The primary difference with all the works described above is that we aim to predict user impact by exploiting features under the user's direct control. Hence, our findings can be used as indirect insights for strategies that individual users may follow to increase their impact score. In addition, we incorporate the actual text posted by the users in the entire modelling process.

8 Conclusions and Future Work

We have introduced the task of user impact prediction on the microblogging platform of Twitter based on user-controlled textual and non-textual attributes. Nonlinear methods, in particular Gaussian Processes, were more suitable than linear approaches for this problem, providing a strong performance ($r = .78$). That result motivated the analysis of specific characteristics in the inferred model to further define and understand the elements that affect impact. In a nutshell, activity, non clique-oriented interactivity and engagement on a diverse set of topics are among the most decisive impact factors. In future work, we plan to improve various modelling components and gain a deeper understanding of the derived outcomes in collaboration with domain experts. For more general conclusions, the consideration of different cultures and media sources is essential.

Acknowledgments

This research was supported by EU-FP7-ICT project n.287863 ("TrendMiner"). Lampos also acknowledges the support from EPSRC IRC project EP/K031953/1.

References

Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. Everyone's an influencer: quantifying influence on Twitter. In *4th International Conference on Web Search and Data Mining*, WSDM'11, pages 65–74.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In *Biennial GSCL Conference*, pages 31–40.

Danah Boyd, Scott Golder, and Gilad Lotan. 2010. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In *System Sciences*, HICSS'10, pages 1–10.

Leo Breiman. 1984. *Classification and regression trees*. Chapman & Hall.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on Twitter. In *20th International Conference on World Wide Web*, WWW'11, pages 675–684.

Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. In *4th International Conference on Weblogs and Social Media*, ICWSM'10, pages 10–17.

Trevor Cohn and Lucia Specia. 2013. Modelling Annotator Bias with Multi-task Gaussian Processes: An Application to Machine Translation Quality Estimation. In *51st Annual Meeting of the Association for Computational Linguistics*, ACL'13, pages 32–42.

Arthur E. Hoerl and Robert W. Kennard. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67.

Matthew Hoffman, David Blei, and Francis Bach. 2010. Online Learning for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems*, NIPS'10, pages 856–864.

Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. 2009. Social Networks that Matter: Twitter Under the Microscope. *First Monday*, 14(1).

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media? In *19th International Conference on World Wide Web*, WWW'10, pages 591–600.

Vasileios Lampos and Nello Cristianini. 2012. Nowcasting Events from the Social Web with Statistical Learning. *ACM Transactions on Intelligent Systems and Technology*, 3(4):72:1–72:22.

Vasileios Lampos, Daniel Preoţiuc-Pietro, and Trevor Cohn. 2013. A user-centric model of voting intention from Social Media. In *51st Annual Meeting of the Association for Computational Linguistics*, ACL'13, pages 993–1003.

Zhunchen Luo, Miles Osborne, Jintao Tang, and Ting Wang. 2013. Who will retweet me?: finding retweeters in Twitter. In *36th International Conference on Research and Development in Information Retrieval*, SIGIR'13, pages 869–872.

Radford M. Neal. 1996. *Bayesian Learning for Neural Networks*. Springer.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. 2002. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, NIPS'02, pages 849–856.

Daniel Preoţiuc-Pietro, Sina Samangooei, Trevor Cohn, Nicholas Gibbins, and Mahesan Niranjan. 2012. Trendminer: An Architecture for Real Time Analysis of Social Media Text. In *6th International Conference on Weblogs and Social Media*, ICWSM'12, pages 38–42.

Joaquin Quiñonero-Candela and Carl E. Rasmussen. 2005. A unifying view of sparse approximate Gaussian Process regression. *Journal of Machine Learning Research*, 6:1939–1959.

Carl E. Rasmussen and Christopher K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. MIT Press.

Daniel M. Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A. Huberman. 2011. Influence and Passivity in Social Media. In *Machine Learning and Knowledge Discovery in Databases*, volume 6913, pages 18–33.

Dominic Rout, Daniel Preoţiuc-Pietro, Bontcheva Kalina, and Trevor Cohn. 2013. Where's @wally: A classification approach to geolocating users based on their social ties. In *24th Conference on Hypertext and Social Media*, HT'13, pages 11–20.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *19th International Conference on World Wide Web*, WWW'10, pages 851–860.

Claude E. Shannon. 2001. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55 (reprint with corrections).

Jianbo Shi and Jitendra Malik. 2000. Normalized cuts and image segmentation. *Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

Alex J. Smola and Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222.

Edward Snelson and Zoubin Ghahramani. 2006. Sparse Gaussian Processes using Pseudo-inputs. In *Advances in Neural Information Processing Systems*, NIPS'06, pages 1257–1264.

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'10, pages 177–184.

Vladimir N. Vapnik. 1998. *Statistical learning theory*. Wiley.

Ulrike von Luxburg. 2007. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.

Christopher K. I. Williams and Carl E. Rasmussen. 1996. Gaussian Processes for Regression. In *Advances in Neural Information Processing Systems*, NIPS'96, pages 514–520.