Are Experts Made or Born? Complexity And Consistency in Online Knowledge-Building Environments

Dave Gottlieb
Department of Philosophy
Stanford University
Stanford, CA
dmg1@stanford.edu

Bogdan State
Department of Sociology
Stanford University
Stanford, CA
bstate@stanford.edu

Abstract

We track the evolution over time of language use on the question-answering website, StackExchange. We find that user posts become more sophisticated by several measures as users increase in status and experience. Accordingly, we conclude that the experts of StackExchange are made (not born).

We draw some general lessons for the study of Knowledge-Building Environments (KBEs). Successful KBEs create a feedback loop where status and experience both reward and promote the creation of high-complexity contributions.

Knowledge-Building Environments (KBEs) such as Wikipedia, Reddit, Quora or StackExchange represent an important area of innovation among online platforms. These systems index vast, ever-growing knowledge bases. Their growth and vitality is based almost exclusively on voluntary user contributions. KBEs mobilize a large amount of resources to create vast corpora of knowledge. A key feature of most, and possibly all, KBEs is the emergence of domain experts, users whose contributions show a consistent pattern of quality and who have disproportionate effect on the viability of KBEs as timely and relevant venues for specialized knowledge exchange.

We aim to ask whether such experts are “born” or “made” in KBEs.

The continuing viability of online KBEs may depend on the answer to this question. If experts are “born,” created by processes exogenous to a platform, then the viability of the platform depends on attracting these born experts. If, on the other hand, KBEs “make” their own experts, then some KBEs may become self-sustaining.

Accordingly, the research question has important design implications for online communities. In the case of Wikipedia, for instance, [17] show that rules meant to protect incumbent, knowledgeable users ultimately led to a significant decline in the rate of contributions to the system. How expertise is produced may determine what sort of rules will lead to more and better contributions.

1 Prior Work

1.1 Expertise as a Social Process

We begin with the premise that expertise emerges through a fundamentally social process. The idea that people are socialized into certain behaviors and roles has a long history. In [1], Howard Becker described the process through which novice marijuana users learn to recognize the “high” produced by the drug and enjoy its effects. Social interaction shapes not only behavior but identity itself, as posited by Tajfel and Turner in their formulation of *social identity theory* [3]. Identity is constructed in piecemeal interactions with the group, through which the individual becomes increasingly aware of group norms [9]. Here “identity” need not be construed as a monolith: instead, a person’s identity is enacted through the multiple social roles they play [8]. The role of the expert in a StackExchange community is but one of many – and we expect to see users learn to play it as they become more vested in the community.

We should consider incentives as well as identity when discussing expertise. Experts possess a resource – knowledge – which, by definition, novices do not. The expert-novice interaction can be seen as an interpersonal exchange of cognitive resources [4]. Motivating experts to participate in this exchange is a challenge for online platforms because, with few exceptions, no direct rewards accrue to the experts. This creates a classic social dilemma, where individuals may free-ride on the contributions of others and contribute no expertise themselves [7]. Such social dilemmas have been well-documented on knowledge exchanges such as Wikipedia [11]. One established solution to this problem is status giving: rather than giving *resources* in return, beneficiaries may reward experts with *recognition* [2]. Testing whether this occurs on KBEs is beyond the goals of this paper – but some KBEs (including StackExchange, our case study) provide a formal mechanism for users to award recognition to each other with "upvotes."

The fact that learning to talk like an expert is the result of a social process does not undermine the value of expertise itself. Indeed, even in oenology, where it is controversial whether expertise even exists, [5] provides evidence that expert wine-tasters can discriminate more tastes than novices. There is certainly a great deal of norm, and even ritual, involved in the expert role – but as users learn this role, they more often than not also acquire knowledge that can be passed on. The great opportunity provided by online knowledge exchanges is the possibility of observing at least some individuals as they progress through this knowledge-acquisition cycle. This will allow us to ask whether experts are born or made.

1.2 Expertise in the Online World

Previous research on StackExchange suggests that experts are born. [13] fail to find any meaningful relationship between users' answer scores and their tenure on StackExchange websites. They likewise present evidence that the first answers users provide on the site are strong predictors of the upvotes they will garner for later answers. They thus conclude that users' expertise is fully formed by the time they join StackExchange, rather than shaped by structured social interactions on the site. While we find this to be an intriguing explanation, we note that upvotes are not an unambiguous measure of contribution quality, as they reflect status in the community as well as expertise.

Further, some indirect evidence seems to contradict this finding. [15] find that users of online communities exhibit two stages in their linguistic development: an initial, "learning" stage during which their language adapts to that of the community, and a second, conservative stage, during which users no longer modify their own language to match that of the community. Similarly, in an examination of the closely-related topic of connoisseurship on online product-rating platforms, [18] provide evidence of taste change throughout the user lifecycle. These results suggest that users learn from experience, and adapt their behaviors to the norms of online communities. Knowledge-sharing behaviors could be among those that are shaped by such experience.

1.3 Expertise as a Collective Phenomenon

Expertise is an intuitive concept that is nonetheless very hard to quantify. Above, we argued against reifying upvote scores as unambiguous measures of knowledge, since they are influenced by status-based processes. To better gauge the evolution of expertise, we develop measures of the linguistic complexity of contributions. Complexity is at least often a necessary condition for quality. Most questions asked on knowledge-exchange sites require complex answers, because simple answers are easily available via web search.

Investigating linguistic complexity, rather than quality directly, is also motivated by computational considerations. Extracting quality from computational semantics, even if we had a method for doing it, would be computationally intensive. By contrast, we are able to measure complexity with relatively straightforward metrics.

Studying complexity also has implications for a systemic understanding of knowledge exchange communities as thinking systems. We aim to provide evidence of group cognitive processes by studying the evolution of language in the context of user lifecycles. Previous evidence in the same vein has been provided by [16] with respect to the achievement of a "collective state" of cognition in the Wikipedia voting process.

Seen from a systems perspective, it is not only the linguistic complexity of StackExchange contributions that increases through the user lifecycle, but also the level of linguistic coordination between users. As individuals become more aware of group norms, we can expect them to converge on certain language patterns. And finally, we can expect users to develop topical niches in their contributions, and to gradually become more consistent in their choice of contributions.

2 Data and Methods

We begin our analysis with a diverse set of 10 medium-sized StackExchange communities: bicycling, cooking, cstheory (Computer Science theory), philosophy, diy, fitness, photography, skeptics, travel and workplace. We purposefully limit our initial analysis to this small number of cases to facilitate the intelligibility of our results. We also deliberately focus our analysis on mostly non-technical StackExchange communities, to achieve conclusions whose generalizability extends beyond the technical disciplines that form the main focus of the StackExchange platform.

Table 1 reports descriptive statistics for all of our measures across the 10 datasets.

2.1 Measuring Complexity, Coordination and Consistency

Post Length. A simple measure of contribution complexity is the number of words in each top-level post. The intuition here is that, as a question engages with more elaborate issues, the number of words required to describe those issues adequately increases.

Number of Distinct Tokens. One readily-apparent limitation of an approach based on word-counts alone is that it is not only cognitive complexity that may influence the number of words in a contribution. Disclaimers, for instance those posted by newcomers ("Sorry for the noob question, but ...") may artificially inflate word counts without increasing the information content of a communication.

Linguistic Entropy Not all distinct tokens are created alike. Neither extremely common stop-words nor singleton proper names communicate much generalizable knowledge. Instead, it is middling-frequency tokens that are maximally informative: these are the terms of art that form the focus of particular communities. Their increased presence and diversity arguably indicates a more refined contribution made by a user. To quantify this intuition we represent each communication as a set of n independent draws $X_1 \ldots X_n$ from a token-level distribution. Because of this simplifying independence assumption we can represent the joint Shannon entropy of all tokens, as a sum over individual entropies:

$$H(X_1, \ldots X_n) = \sum_{i=1}^{n} H(X_i)$$

To measure *coordination* we are interested in the normalized entropy, $\bar{H}(X_1, \ldots X_n) = H(X_1, \ldots X_n)/n$. The normalized entropy gives us a glimpse into the central tendency of the user's language. Because this central tendency will tend to be driven by the middling-terms mentioned above, we consider higher normalized entropies to be indicative of less coordination between users, and we expect entropy to decrease as expertise (measured through recognition and engagement) increases.

Embedding-based Measure We likewise compute a measure based on a Latent-Dirichlet Allocation (LDA) embedding of the token-to-posting matrix, separately for each of the 10 datasets, with $k = 10$. The measure was computed using the topicmodels R package [12]. For each user we mark the 'Consistent Topic' measure as 1 if the user's current posting has the same labeled LDA topic as the previous one and 0 otherwise. We only compute this measure for the second and greater posting made by each user.

Table 2 shows the most common term associated with each LDA topic in four different datasets. We checked the validity of the topic models by computing, for each post, the proportion of all posts by the same user assigned to the same topic. The mean across datasets was 0.13 ($N =$), 30% better than the chance distribution of 0.10 for 10 topics. (See Table 1.) Accordingly, we conclude that the topic models successfully capture some of the semantic structure in the data.

2.2 Pre-Processing

Before computing the previously-mentioned statistics we undertook a number of pre-processing tasks. All text was converted to lower case, HTML tags, punctuation and excess whitespace were removed. We likewise applied a standard English-language stemming algorithm, as provided by the R tm package [19].

2.3 Mixed-Effects Models

The problem of modelling the relationship between multiple time varying user characteristics and linguistic complexity is a challenging one. The dataset is structured as an imbalanced panel of observations, but a panel model would be inadequate, given that we are dealing with multiple time scales on which linguistic complexity varies. Specifically, because certain users join StackExchange at different points in time, they display heterogenous tenures as contributors, while the site is

Table 1: Summary statistics

Variable	Mean†	Median†	Bottom Quartile†	Top Quartile†
Days since first post^l	2.359	0.0	0.0	5.136
Votes received^l	1.495	0.0	0.0	2.944
Comments made^l	1.84	1.39	0.0	2.83
Unique Words^l	3.6341	3.6376	3.1781	4.0775
Total Words^l	3.8570	3.8712	3.3673	4.3567
Normalized Entropy‡	0.5769	0.4979	0.3828	0.6338
Same LDA Topic	0.13	0.00	0.00	0.00

Note: *p<0.1; **p<0.05; ***p<0.01

^{l} - log-transformed using $f(x) = \ln(x)$, or $\ln(x+1)$, where necessary.

† - averages are computed over posts, rather than e.g. users or topics.

‡ - multiplied by 1,000

Table 2: Most common term for each of $k = 10$ LDA topics by dataset.

philosophy	cooking	bicycles	cstheory
condit	chees	tyre	logic
simul	flour	pad	color
realiti	butter	shoe	distribut
model	cup	point	circuit
alreadi	chocol	seat	string
matter	bean	bar	&
method	flour	fork	\rightarrow
author	cream	spoke	weight
&	sauc	hub	formula
inform	powder	nut	cost

also likely to be influenced by a "period" effect (e.g., due to user interface changes going live on a certain day).

The heterogeneity of tenure with respect to period poses a very similar problem to the one modelled by Age-Period-Cohort models, a standard tool in demography. A long-standing observation is that this model is underidentified, given that age is but a linear combination of period and cohort. We adopt the very useful solution to the underidentification problem proposed by [10], who employ mixed models [6] to partition out the variance in the dependent variable due to age (or tenure in our case), period, and cohort effects.

The mixed effects framework has an additional advantage, in that it allows us to consider (via random effects) the influence of other time-varying covariates (such as the number of votes received by a user, or the number of posts they have made so far), with minimal assumptions regarding the functional form of conditional distributions. This is crucial to our analysis, as it provides us with a flexible framework for understanding the effects of multiple social processes.

2.4 Quantifying Changes in User Characteristics

We are interested in a number of different processes which are at work concurrently in influencing linguistic complexity.

2.4.1 Controls

Tenure. The time a user has been a member of the site is a first-pass measure of their experience. We use the number of days elapsed since the user's first posting as a measure of their tenure on the site.

Cohort. Users who join around the same time may resemble each other. For example, readers of some other website may be likely to sign up for StackExchange when it is linked on the site they read. We control for these cohort effects by including the month during which the user joined as a random effect.

Period. The complexity of posts on StackExchange may vary due to secular trends over time, apart from any changes in the behavior of individual users. We control for this source of variation by including the month of each post as a random effect in the model.

Posts made over lifetime. One thorny problem in virtually all online datasets is survivorship bias. In many cases, most "users" of a site have only visited the site once, signed up for an account, and used the site for its intended purpose, never to return. Slightly fewer maybe returned once and then left the site for good – and so on to the very frequent regular users of the site. One-time users can be expected to be different from regulars in many of the behaviors they display. Because we have the benefit of hindsight with respect to past user behavior, we can control for the actual number of posts users have made over their lifetime and thus account for differences according to unobserved characteristics that influence the likelihood to survive as a user.

Dataset We expect each dataset to have slightly different, subject-specific baseline levels of complexity, consistency and co-ordination. We include *effect-coded* factors for each of nine datasets. The effect for the reference dataset(`cooking`) can be obtained by summing all the other effects. Effect coding has the advantage of allowing the interpretation of effects as deviations from the grand mean (mean of dataset-specific means, in our case) of each of our datasets, rather than from the reference category, as is the case in the more frequently-used dummy coding.

2.4.2 Variables of interest

Recognition. We consider whether recognition of a user's posts by the community influence that user's posting behavior. We measure recognition by counting the votes the user has so far received on their posts, at the time of the given post. We expect recognition, an act of status-giving, to be correlated with prior displays of expertise, which we expect, in turn, to be correlated with greater complexity, coordination and consistency.

Engagement. We measure engagement by counting the number of comments users have left on other posts. In the StackExchange universe, comments serve to clarify existing posts (either questions or answers). We treat writing comments as a measure of users' critical engagement with existing knowedge, and expect a positive relationship between the number of comments and post complexity.

Recognition-Engagement Interaction. Our dataset shows a rank correlation coefficient of 0.8 between the votes received on one's posts and the number of comments the author has made so far. To appropriately address the danger of multicollinearity in including such closely-related variable we also include an interaction effect between the two predictors. We expect both recognition and engagement to increase post complexity, but we likewise suspect that engagement in the form of comments is more likely to be through the display of status-seeking behavior (e.g., thanking others for their answers), rather than instances of advancing cognition. As a result we expect to see a negative interaction effect associated with the product of comments made and votes received.

Table 3: Regression Results

	Dependent variable:			
	Unique Words^{l}	Total Words^{l}	Norm. Entropy‡	Same LDA topic
	linear mixed-effects	*linear mixed-effects*	*linear mixed-effects*	*generalized linear mixed-effects*
	(1)	(2)	(3)	(4)
(Intercept)	4.334^{***}	4.667^{***}	1.431^{***}	-1.765^{***}
	(0.038)	(0.042)	(0.008)	(0.161)
Days since first post^{l}	-0.036^{***}	-0.040^{***}	-0.00004	-0.030^{***}
	(0.002)	(0.002)	(0.0005)	(0.010)
Votes received^{l}	0.145^{***}	0.161^{***}	-0.004^{***}	0.022
	(0.003)	(0.004)	(0.001)	(0.020)
Comments made^{l}	0.090^{***}	0.100^{***}	-0.003^{***}	-0.024
	(0.004)	(0.004)	(0.001)	(0.019)
`cstheory`†	0.218^{***}	0.320^{***}	0.048^{***}	0.108^{***}
	(0.005)	(0.005)	(0.001)	(0.030)
`philosophy`†	0.157^{***}	0.192^{***}	-0.021^{***}	0.052
	(0.006)	(0.007)	(0.002)	(0.036)
`bicycles`†	0.103^{***}	0.102^{***}	0.089^{***}	-0.051
	(0.005)	(0.006)	(0.001)	(0.034)
`diy`†	0.120^{***}	0.138^{***}	0.003^{***}	0.022
	(0.004)	(0.004)	(0.001)	(0.026)
`fitness`†	0.148^{***}	0.153^{***}	0.815^{***}	-0.068^{*}
	(0.005)	(0.006)	(0.001)	(0.037)
`photo`†	0.064^{***}	0.068^{***}	0.107^{***}	-0.040
	(0.004)	(0.004)	(0.001)	(0.026)
`skeptics`†	0.179^{***}	0.186^{***}	0.047^{***}	0.016
	(0.006)	(0.006)	(0.002)	(0.033)
`travel`†	0.009^{**}	0.005	0.034^{***}	0.045
	(0.004)	(0.005)	(0.001)	(0.028)
`workplace`†	0.276^{***}	0.291^{***}	-0.150^{***}	-0.041
	(0.005)	(0.005)	(0.001)	(0.036)
Votes $\times$ Comments	-0.019^{***}	-0.021^{***}	0.001^{***}	-0.0004
	(0.001)	(0.001)	(0.0002)	(0.004)
Observations	93,672	93,672	93,672	46,399
Akaike Inf. Crit.	176,360.300	197,799.800	-60,952.760	34,989.370
Bayesian Inf. Crit.	176,530.400	197,969.900	-60,782.700	35,138.040

Note: *p<0.1; **p<0.05; ***p<0.01

^{l} - log-transformed using $f(x) = \ln(x)$, or $\ln(x+1)$, where necessary.

† - effect-coded variables

‡ - multiplied by 1,000

3 Results

We ran linear mixed-effects models predicting measures of linguistic complexity (number of total and unique tokens in post) as well as coordination (normalized entropy). Our measure of consistency – LDA topic persistence – is a binary variable, for which we fit a general mixed-effects model with a logistic link function. All models were fitted using the lme4 R package[14]. The mixed effects results are reported in table 3.

Our results confirm our expectations in the cases of complexity and coordination. The first model ('Unique words') show a positive, significant relationship between a user's number of votes received before making the current post and the expected number of unique words in their post. Specifically, for every increase in the number of votes by a factor of $e = 2.718$, the model predicts an increase by a factor of $e^{.145} = 1.16$ in the number of unique words. The same effect is $e^{.094}$ for comments. The effects for the second model ('Total words') are comparable.

Normalized entropy is our measure of coordination, as previously discussed. We see a significant-though-small downward drift in entropy as a result of increases in both votes received and comments made. The effects are fairly minute, to the order of 10^{-5} for each increase in votes or comments by a factor of e. This provides some confirmation for our expectations of increased coordination with more expertise.

Our expectations are not confirmed in the case of consistency, however. In that case we fail to find a significant effect for either votes received or comments made. As this is only an exploratory analysis, this result suggests the need for more nuanced measures of consistency.

We also note the fact that the interaction effect between votes and comments is large and opposite to the main effect. This provides some evidence towards our supposition that comments can also be evidence of status-seeking behavior (followed by votes), and not always of increases in expertise.

4 Discussion

We began with the question of whether experts in online KBEs are made or born. Is expertise something that develops over time, in response to structured social interactions? Or is it mostly static and exogenous, something that some users bring to their interactions with KBEs, but which KBEs play little role in developing?

[13] also posed this question about expertise on StackExchange, and concluded that StackExchange experts are born, not made. However, those earlier results made use of a narrow, language-external measure of expertise – the scores which users receive for their answers. As we have argued, this measure cannot be regarded as an unambiguous measure of expertise, because it reflects status as well as knowledge. Furthermore, since it is difficult to measure expertise directly, we have argued for, and used, a variety of indirect measures. Effects that are limited to only one measure of expertise should be regarded as provisional. Conversely, the failure to detect an effect using only one measure of expertise should not be regarded as proving that an effect does not exist. On the other hand, effects that are robust across a variety of measures of expertise should be regarded as more firmly established.

Our results suggest that expertise is made in KBEs, not born and then brought to KBEs ready-made. Experience on the site leads to more complex posts as well as more linguistic coordination with other users. Our findings are robust across different measures of expertise and across different StackExchange communities.

We failed to find any evidence that experience led users to focus more narrowly on a consistent set of topics, across their postings. At this stage in the analysis, it is premature to conclude that users do not have such a tendency. Perhaps our measures are not adequate to detecting it. On the other hand, if it is the case that more experienced users do not focus more narrowly on specific topics, this may tell us something about the KBE. It suggests that *subject matter* expertise is distinct from the behaviors that lead to the recognition of expertise on StackExchange. This would be broadly consonant with the other conclusions of our paper. Subject matter expertise, meaning actual domain knowledge about the community's topics of interest, may be "born" – brought in from offline, rather than mostly acquired in the community. But the behaviors that lead to the recognition of expertise in the community are "made" – they are learned in the community. On this view, a StackExchange expert is someone who meets two conditions. First, they have domain knowledge to share. Second, they are versed in the community norms that facilitate sharing that knowledge.

Alternatively, our failure to find that users consistently focus on fewer topics as they become more expert may indicate that, while users do build knowledge on the site, the knowledge they build is not focused on any particular subject matter, but rather is spread across the whole range of topics covered by the site. Disambiguating these effects will require further research.

If experts are made on StackExchange, we can tentatively conclude that the institutions of StackExchange are conducive to making experts. One question for further research, then, is, what about the StackExchange sites encourages the development of experts? More generally, our findings provide a framework for analyzing StackExchange and similar KBEs as collective cognitive processes. These processes certainly would bear further investigation.

Our results are also relevant to more general problems of expertise detection. To the extent that the linguistic measures we have developed agree with each other and with language-external measures of expertise, they can be regarded as robustly measuring expertise. It might be possible to deploy similar measures to detect and quantify expertise in other contexts.

5 Conclusion

Online knowledge-sharing communities allow for the in-depth, longitudinal study of expertise formation. Abundant data have opened to quantitative study research questions that were previously firmly within the remit of ethnography. The promise here goes well beyond the Internet. The study of online knowledge-exchange communities may better understand the role of knowledge and information in modern human societies. Ultimately, this is an essential ingredient for that very elusive goal that the "Big Data" revolution is rendering increasingly possible: a single science of society.

References

[1] Howard S Becker. "Becoming a marihuana user". In: *American journal of Sociology* (1953), pp. 235–242.

[2] Peter Michael Blau. *Exchange and power in social life*. Transaction Publishers, 1964.

[3] Henri Tajfel and John C Turner. "An integrative theory of intergroup conflict". In: *The social psychology of intergroup relations* 33.47 (1979), p. 74.

[4] Edna B Foa and Uriel G Foa. "Resource theory". In: *Social exchange*. Springer, 1980, pp. 77–94.

[5] Gregg Eric Arn Solomon. "Psychology of novice and expert wine talk". In: *The American Journal of Psychology* (1990), pp. 495–517.

[6] Robert A McLean, William L Sanders, and Walter W Stroup. "A unified approach to mixed linear models". In: *The American Statistician* 45.1 (1991), pp. 54–64.

[7] Peter Kollock. "Social dilemmas: The anatomy of cooperation". In: *Annual review of sociology* (1998), pp. 183–214.

[8] Sheldon Stryker and Peter J Burke. "The past, present, and future of an identity theory". In: *Social psychology quarterly* (2000), pp. 284–297.

[9] Michael A Hogg and Scott A Reid. "Social identity, self-categorization, and the communication of group norms". In: *Communication Theory* 16.1 (2006), pp. 7–30.

[10] Yang Yang and Kenneth C Land. "A mixed models approach to the age-period-cohort analysis of repeated cross-section surveys, with an application to data on trends in verbal test scores". In: *Sociological methodology* 36.1 (2006), pp. 75–97.

[11] Denise Anthony, Sean W Smith, and Timothy Williamson. "Reputation and reliability in collective goods the case of the online encyclopedia wikipedia". In: *Rationality and Society* 21.3 (2009), pp. 283–306.

[12] Kurt Hornik and Bettina Grün. "topicmodels: An R package for fitting topic models". In: *Journal of Statistical Software* 40.13 (2011), pp. 1–30.

[13] Daryl Posnett et al. "Mining stack exchange: Expertise is evident from initial contributions". In: *Social Informatics (SocialInformatics), 2012 International Conference on*. IEEE. 2012, pp. 199–204.

[14] Douglas Bates et al. "lme4: Linear mixed-effects models using Eigen and S4". In: *R package version* 1.4 (2013).

[15] Cristian Danescu-Niculescu-Mizil et al. “No country for old members: User lifecycle and linguistic change in online communities”. In: *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 2013, pp. 307–318.

[16] Simon DeDeo. “Collective phenomena and non-finite state computation in a human social system”. In: *PloS one* 8.10 (2013), p. 10.

[17] Aaron Halfaker et al. “The Rise and Decline of an Open Collaboration Community. How Wikipedia’s reaction to sudden popularity is causing its decline”. In: *American Behavioral Scientist* (2013).

[18] Julian John McAuley and Jure Leskovec. “From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews”. In: *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 2013, pp. 897–908.

[19] Ingo Feinerer. *Introduction to the tm Package Text Mining in R*. 2014.