Modelability across time as a signature of identity construction on YouTube

Abstract

Linguistic self-representation and identity construction on social media have attracted much scholarly attention. However, relevant studies tend to overlook the temporal dimension of social media, potentially systematic patterning of linguistic behavior across time, and the attendant implications of such a temporal perspective on identity. Combining an automated lexical tool (LIWC) and the Box-Jenkins method of statistical time series analysis, this paper shows how the 'modelablity' of linguistic choices across time can be interpreted as signatures of identity construction and complement existing frameworks for identity analysis. Two levels of modelablity are discussed – the availability of a well-fitting time series model as evidence of temporal patterning, and specific parameters of that model interpreted in context. These are demonstrated with a case study of the construction of 'amateur expertise' over 109 consecutive makeup tutorial videos on the popular YouTube channel 'Nikkiestutorials'. Results show that the linguistic display of 'analytical thinking' reflects a strategy of 'short term momentum', the display of 'clout' and 'authenticity' a strategy of 'short term restoration', while the display of 'emotional tone' fluctuates randomly across time. The approach is further discussed in terms of its general principles and potential applications in other contexts of identity and related research.

Keywords: identity construction, social media, LIWC, modelablity, ARIMA, time series analysis

Introduction

New media technologies and their applications have attracted much multidisciplinary research in the digital age. This includes the linguistic and discursive forms, structures, and functions that constitute and sustain the digital universe. In particular, social media as multi-semiotic platforms for diverse purposes like self-expression, entertainment, education, business, and politics offer new grounds for age-old questions about self-representation and identity. Tolson's (2010) case study of make-up tutorials depicts YouTube as 'post-television' where communicative practices have become more authentic and 'entitled' than traditionally allowed. Adopting sociolinguistic notions of stance (Du Bois 2007), Valentinsson (2018) shows how artist Lady Gaga constructs an authentic and morally credible celebrity persona on Twitter by linguistically aligning with fans and disaligning with mass media establishment. Such phenomena are not limited to prominent individuals, and have been examined with respect to traditional demographics like gender (Strangelove 2010) and ethnicity (Dlaske 2017). The general consensus is that social media have created, on an unprecedented scale, a global "amateur digital culture" (Strangelove, 2010:17) where ordinary people can sculpt themselves in extraordinary ways. Bhatia's (2018) recent work succinctly ties these research threads together, highlighting a particular need to examine the emergence of 'amateur expertise' through interdiscursive practices. Similar to Tolson (2010), she uses a single case study of a YouTube beauty vlogger to demonstrate how discourse strategies like colloquial, jargon, and promotional talk in extracts of video transcripts and comments construct a strategic blend of an 'authentic' and 'expert self'.

The above studies have taken diverse methodological approaches including (critical) discourse, genre, thematic, multimodal, and cyber-ethnographic analysis. Each of these has its own strengths in showcasing particular aspects of identity construction. However, there are two points worth highlighting. The first is the general and controversial perception that qualitative studies lack reliability, replicability, and other such traits. While qualitative approaches do not usually lay claim to these traits, it is still worthwhile to explore how they might work together with quantitative computational techniques to capture identity-related phenomena like personality and sentiment on scales far greater than selective data fragments (Kern et al. 2016; Xu and Zhang 2018). The second subtler point is that they tend to overlook time, or temporal progress, as an implicit but key variable underlying identity construction and performance. Most social media platforms have a built-in timestamp to track online activity. YouTube, for example, clearly marks the upload time of every video and comments. However, the relationships between the linguistic/discursive elements of each video and their role in the construction of YouTubers' identities is seldom explicitly modeled in systematic and replicable ways. In statistical parlance, any series produced by a YouTuber constitutes naturally occurring 'time series data' (Box et al. 2015) analyzable in ways seldom considered by language/discourse researchers. Time series analysis is in fact a big part of contemporary data analytics with underexplored potential for diachronic discourse research (Tay 2019).

This paper introduces an approach that combines automated lexical analysis with an explicit time series analytic methodology to systematically depict identity construction across a YouTuber's oeuvre. The central idea is that strategic identity construction via linguistic performance can be depicted by the nature and extent of its quantitative 'modelability' (or conversely, randomness) across time. The lexical analysis is performed by Linguistic Inquiry and Word Count (LIWC) (Tausczik and Pennebaker 2010), a widely used computer program to classify words under socio-psychological variables that reflect identity-relevant aspects. The time series methodology

is the also widely used Box-Jenkins method with ARIMA models (Box et al. 2015). The present approach has been applied to other time-sensitive discourse contexts like psychotherapy and news (Tay 2019, 2020, 2021), and will be demonstrated as complementing existing ways to investigate online identity construction. Specifically, it is well-suited for the case study of particular individuals as often seen in the literature, and can offer key 'entry points' into the data for various qualitative analyses that are crucial for in-depth understanding of identity construction. In the following sections, I first discuss Bucholtz and Hall's (2005) synergistic identity framework and its relevance for the present case study of 'amateur experts' on social media. I then describe LIWC and the Box-Jenkins method in more detail and how they together determine the modelability of language as time series. This approach is then demonstrated on 109 consecutive videos on the YouTube channel *Nikkiestutorials* owned by beauty vlogger Nikkie de Jager.

The 'amateur expert' as a constructed identity

Bhatia (2018:108) depicts the social media 'amateur expert' as one who builds "a community of subscribers, or in effect learners, creating an environment of informal learning". They engage in a participatory culture with multiple roles like "creator, mentor, critic, follower, teacher, learner, and subscriber", promoting their own contents while interacting with viewers and learning from peers. It is in this sense that they position themselves, often delicately, between 'experts' and 'amateurs' en route to becoming full-fledged 'influencers' at the pinnacle of internet celebrity (Abidin 2018). Like Bhatia, Abidin (2018) emphasizes that celebrityhood is a social media-fuelled construct, further depicting it as a commodification of ordinary people otherwise. She likewise observes the inherent tension in the notion of 'amateur expertise'. In constructing and performing this identity, people typically balance "anchor" content which demonstrates their talents and skills, with "filler" content where a sense of everyday ordinariness is exhibited.

The literature on identity and related themes has a rich and multidisciplinary history, cutting across fields like social psychology (Tajfel 1982), linguistic anthropology (Ochs 1996; Silverstein 1976), and sociolinguistics (Omoniyi and White 2006). Bucholtz and Hall's synergizing of these strands into a framework that focuses on "the details of language and the workings of culture and society" (2005:586) aptly contextualizes the present approach and case study of a YouTube vlogger as 'amateur expert'. In their view, 'identity' is an act of social positioning and a discursive construct that emerges from (social media) interaction. Vloggers position themselves and others as "particular kinds of people … even in the most fleeting of interactional moves" (2005:595). Video-to-video transitions as units of such 'fleeting interactional moves' remain underexplored in this regard. The mechanisms of identity construction are articulated along five principles: i) **emergence**, which asserts that identity is an emergent product rather than a pre-existing source of linguistic practices. Time series analysis can model the *process* of emergence explicitly, relating the vlogger's linguistic choices/tendencies and their motivations to natural (i.e. video-by-video) time intervals; ii) **positionality**, where ethnographically specific positions and temporally/interactionally specific stances/participant roles are as important as macro-level identity categories. The nascent 'amateur expert' and its many facets is a case-in-point as described above. We will see how the vlogger displays fluctuating levels of analyticity, clout, and authenticity (as measured by LIWC) when delivering her content and addressing viewers, operationalizing the notion of

"temporally/interactionally specific stances"; iii) **indexicality**, where identity relations emerge through indexical processes like overt category labels, implicatures, presuppositions, evaluative and epistemic orientations, and so on. We will see that this emphasis on concrete linguistic forms is supported by LIWC's word classification approach, where the (dis)preference of certain words that index an (un)desirable identity (Bucholtz 1999) is operationalized via relative frequencies of various lexical categories; iv) **relationality**, where identities are intersubjectively constructed, acquiring meaning in relation to other positions and actors, and v) **partialness,** where this construction is dependent on others' responses, constantly shifts as interaction unfolds, and partly deliberate/intentional and unconscious/habitual. This again echoes the description of 'amateur experts' as an identity valid only with reference to the community of viewers and peers, with its 'constant shifts' uncoverable by time series analytic lenses. Less-than-deliberate aspects of identity construction further dovetail with LIWC's emphasis on conventional rather than creative language, as well as the ability of time series modeling to uncover randomness across time. These are elaborated below.

LIWC

Pennebaker et al. (2015) details the development and validation measures of LIWC. It has mostly been applied to analyze large volumes of text from a synchronic perspective, but recent studies have turned their attention to diachronic changes such as patterns of linguistic progression in different narrative genres (Boyd et. al 2020). For each input text, LIWC computes i) the normalized frequencies of over 90 lexical categories. These include function and content words in different semantic categories (emotion words, cognitive words, informal words etc.) in its built-in dictionary; ii) a score from 0-100 measuring the extent to which the text manifests four socio-psychological constructs known as 'summary variables': *analytical thinking*, *clout*, *authenticity,* and *emotional tone.* These summary variables are computed by combinations of specific lexical categories that have been shown to co-occur and reliably differentiate input texts along their implied traits. They are designed to provide an overall socio-psychological profile, and as argued below, identity-relvant aspects of texts. Table 1 lists the summary variables, their constituent lexical categories, and some relevant studies. Plus/minus signs indicate categories more/less frequent in texts that reflect a higher level of that summary variable.

Summary variable	**Defining lexical categories**
Analytical thinking	+articles, prepositions -pronouns, auxiliary verbs, conjunctions, adverbs, negations (Pennebaker et al. 2014)
Clout	+1st person plural pronouns, 2nd person pronouns -tentative words (e.g. *maybe, perhaps*) (Kacewicz et al. 2013)
Authenticity	+1st person singular pronouns, 3rd person pronouns, exclusive words (e.g. *but, except, without*) -negative emotion words (e.g. *hurt, ugly, nasty*), motion verbs (e.g. *walk, move, go*) (Newman et al. 2003)
Emotional tone	+positive emotion words (e.g. *love, nice, sweet*) -negative emotion words (e.g. *hurt, ugly, nasty*) (Cohn, Mehl, and Pennebaker 2004)

Table 1 Summary variables and defining lexical categories

A high *analytical thinking* score[1] suggests formal, logical, and hierarchical thinking, versus informal, personal, here-and-now, and narrative thinking. This is based on a study of American college admission essays. Those with more articles and prepositions were more formal and precise in describing objects, events, goals, and plans, while those with more pronouns, auxiliary verbs etc. were more likely to involve personal stories (Pennebaker et al. 2014). We will see how this relates to Nikkie de Jager's stepwise expert-like makeup instructions for viewers.

A high *clout* score suggests speaking/writing with high expertise and confidence, versus a more tentative and humble style. This is based on studies of decision-making tasks, chats, and personal correspondences. Higher status individuals used more *we/our, you/your* and fewer tentative words. This was explained by an association between relative status and attentional bias. Higher-ups are more other-focused and less unsure while lower individuals are more self-focused and tentative (Kacewicz et al. 2013). This will relate to Nikkie's giving of technical instructions, use of hedges/uncertain words, as well as pronominal strategies when addressing viewers.

A high *authenticity* score suggests more honest, personal, and disclosing discourse, versus more guarded and distanced discourse. This is based on studies of elicited true versus false stories where the latter has fewer first and third person pronouns, exclusive words, and more negative emotion and motion verbs. This was explained by the idea that liars tend to dissociate themselves with the lie, feel greater tension and guilt, and speak in less cognitively complex ways. These linguistic tendencies accurately distinguished truth-tellers versus liars in independent data more than 60% of the time (Newman et al. 2003). We will see this in Nikkie's use of intimate, spontaneous, and colloquial language to ostensibly replicate an actual face-to-face conversation

A high *emotional tone* score suggests a more positive and upbeat style, a low score anxiety/sadness/hostility, while a neutral score around 50 a lack of emotionality. This was based on a study of online journals prior to and after the September 11 attacks where negative emotion words increased sharply following the attack and gradually returned to pre-attack baselines after some time (Cohn, Mehl, and Pennebaker 2004). We will see this in Nikkie's general preference for emotionally positive words to project a correspondingly positive image.

The present use of LIWC is motivated by its quantification of socio-psychological variables, which as mentioned above gives a working operationalization of lexical (dis)preferences that index identities. These variables broadly resonate with recurrent themes in the literature like the representation of authenticity (Tolson 2010:277), emotion analysis as part of the 'affective turn' (Dlaske 2017:451), the construction of expertise, power relations, and so on (Bhatia 2018). The scoring of each video transcript provides a reliable way to profile their linguistic instantiations, with subsequent time series analysis uncovering systematic changes across time. Nevertheless, in the context of identity and related research, it is important to stress that these LIWC categories should not be seen as inherently 'objective' or neutral, able to capture invariant notions of 'analyticity', 'clout', 'authenticity', or 'emotional tone'. LIWC definitions of these variables are necessarily mediated by factors like culture and context – in this case, the more-or-less

[1] The exact formulae for computing scores (0-100) from their defining lexical categories have not been disclosed.

conventional meanings of English words categorized for expedience in computational analytic settings. For instance, the definition of authenticity as 'honest, personal, and disclosing' may deviate from the central concern with 'genuineness' in sociocultural linguistics, which may in turn deviate from how speakers themselves understand it (Bucholtz 2003). It will become apparent that in principle, one can use any other measure deemed appropriate for the research purpose at hand to derive variable scores for time series analysis, and subsequently, to interpret the time series models through different theoretical lenses.

The Box-Jenkins method of Time Series Analysis

A time series is a set of consecutive measurements of a random variable usually made at equal time intervals. Canonical examples include stock prices, rainfall, and birth/death rates. Analysts across these different contexts share two general objectives: to uncover mathematical patterns underlying a seemingly randomly fluctuating series, and then use these to forecast future values. There are many ways to do this but the Box-Jenkins method (Box et al. 2015) is widely used, and has recently been applied to analyze discourse in diachronic contexts like psychotherapy, university lectures, and newspapers (Tay 2017, 2019, 2021).

The method is implementable in many programs/languages like MATLAB, SAS, R and Python (currently used). The core idea is that any series comprises one or more of the following components: a long-term up/downward trend (e.g. sales growth from stable economic growth), predictable seasonal oscillations (e.g. sales increases during Christmas each year), and less predictable longer-term oscillations (e.g. business cycles). The task is to account for or 'filter out' these components until the residual series is pattern-less. Filtering involves analyzing the autocorrelation functions of the series; i.e. how values 1, 2 … *k* intervals apart are positively/negatively correlated with one another, and fitting a suitable model from a family of what are called ARIMA models, to express the value at any time interval in terms of its past values and/or other aspects. If the model fits the data well and has good predictive accuracy, it can then be used to forecast future values. The whole process is summarized in the following six steps. They constitute an atheoretical exploratory process of uncovering (but not yet interpreting) temporal patterns.

- 1: Inspect and transform the series to meet statistical conditions if necessary
- 2: Calculate and analyze autocorrelation functions of the series
- 3: Identify candidate ARIMA models
- 4: Estimate model parameters
- 5: Evaluate model fit and perform residual diagnostics. If satisfactory, proceed to next step. Otherwise, return to Step 3.
- 6: Use model for forecasting and other purposes

The Box-Jenkins method lends itself particularly well to a case study approach, and exemplifies how quantitative techniques can dovetail with the general philosophy of qualitative analysis. It neither tests how population-level factors like 'gender' or 'age' shape language/discourse, nor is it designed to model aggregated or averaged scores across large groups of people. Instead, it focuses on a specific realization of time series data that may well vary from one case to another. It could in fact be seen as an 'N=1' approach despite its analysis of observations over multiple

intervals because in most cases each interval can only be observed once (i.e. we can hardly 'turn back time' to sample another video series talking about the exact same topics)[2]. For more comprehensive tutorials beyond the present scope, see Bowerman and O'Connell (1987) or Tay (2019).

'Modelability' and identity construction across time

The central argument of this paper is that the extent to which a video series is 'modelable' using this approach, and the details of the model, can be interpreted as a signature of identity construction across time. Modelability is defined here at two levels. The first and more general level is simply whether any patterns exist in a series of LIWC scores such that a model can be adequately fitted. If the series is random and pattern-less from the beginning (i.e. a 'white noise' series), this implies that the linguistic performance of analytical thinking, clout, authenticity, and/or emotional tone is statistically random across time. If patterns in one or more of these variables exist, we have two broad possibilities. The first is that the language is reflecting predictable/patterned background events, an example being the phrase 'summer olympics' that spikes every four years in newspapers. The second possibility is more intriguing and relevant here – that *despite* the absence of a clear pattern in background events/topics (as is likely the case for a series of makeup tutorial videos), the LIWC scores across time are still patterned to varying degrees. Concrete measures of model fit and accuracy provide a means to quantify this degree of patterning. While randomness might reflect identity principles like 'partialness' (Bucholtz and Hall 2005) as discussed earlier, it could be conversely hypothesized that the stronger the patterning, the more deliberate or effortful the processes of identity construction are. This idea remains controversial and will not be fully pursued for now since *all* identity work could be argued as effortful, and language is only a part of this.

The next and more specific level of modelability would then be to flesh out the details of this connection by fitting a suitable ARIMA model to the data. ARIMA stands for 'Autoregressive Integrated Moving Average', which are technical names for different types of regressors modeling the series. The model coefficients and parameters can reveal nuances about the dynamics of period-to-period identity construction reflected in specific ways by examples in context. Appropriate qualitative analytic framework(s) could then be applied to interpret or account for the observed dynamics in detail. In fact, subsequent qualitative analysis is what sets language/discourse apart as a unique application of time series analysis. This is because unlike inherently quantitative cases like sales figures and trade cycles where time series data usually follow better understood patterns, discourse is often far 'messier' and less conforming towards time series models. Qualitative analysis is therefore needed for a full understanding of 'what is going on'. In this sense, we can describe the present approach as involving an initial theory-driven phase (i.e., deriving variable scores), a 'bridging' atheoretical phase (exploratory time series modeling), and a final theory-driven interpretation phase (to make sense of the models in context).

[2] This is why, to ensure reliable statistical analysis (e.g. parameter estimation) and minimize potential analytical pitfalls like outliers and regression-to-the mean, Step 1 requires us to ensure that the series meets statistical conditions like having a constant mean and variance throughout, or be transformed by differencing, log-transformation etc. if otherwise. See Bowerman and O'Connell (1987) or Tay (2019) for details.

We will now illustrate it with a case study of 109 consecutive makeup tutorial videos uploaded between 30 June 2008 and 16 Dec 2017 to the YouTube channel *NikkieTutorials.* Owned by makeup artist and beauty vlogger Nikkie de Jager, it is the 4th most subscribed beauty and style channel as of October 2020[3] with 13.7 million subscribers and over 1.2 billion views. The 109 videos span across 9 years from her debut to gradual development as amateur expert, and eventual global recognition as a Forbes magazine top-ten beauty influencer. This provides a comprehensive and statistically adequate time range to investigate the her developmental trajectory.

Transcripts of the videos were generated by YouTube where available, or Amazon Transcribe otherwise, and cleaned by removing punctuation marks and errors. The total word count was 135301 (*M*=1241.3, *SD*=428.9). The latest LIWC version (2015) was used to compute the summary variables and *Python 3.71* for time series analysis.

Results

Table 2 shows the mean summary variable scores of the 109 *Nikkietutorials* videos compared with large sample averages of other discourse contexts from Pennebaker et al. (2015).

Variable	*Nikkietutorials*	Blogs	Expressive writing	Novels	Natural speech	NY Times	Twitter
Analytic	49.63	49.89	44.88	70.33	18.43	92.57	61.94
Clout	47.40	47.87	37.02	75.37	56.27	68.17	63.02
Authentic	67.36	60.93	76.01	21.56	61.32	24.84	50.39
Tone	71.72	54.50	38.60	37.06	79.29	43.61	72.24

Table 2 Mean summary variable scores of *Nikkietutorials* vs. other contexts

Generally, the videos are closest to blogs in the linguistic display of most variables except for emotional tone which is closest to Twitter. This is expected as the makeup tutorial videos are considered vlogs or 'video blogs' and are a form of social media like Twitter. From this broad atemporal perspective, the language of *Nikkietutorials* does not appear to be remarkable in any way.

We now consider this language from a temporal perspective. Figure 1 shows the time series plots of the four summary variable scores (y-axis) across the 109 videos (x-axis).

[3] https://www.statista.com/statistics/627448/most-popular-youtube-beauty-channels-ranked-by-subscribers/ Accessed 19 November 2020.

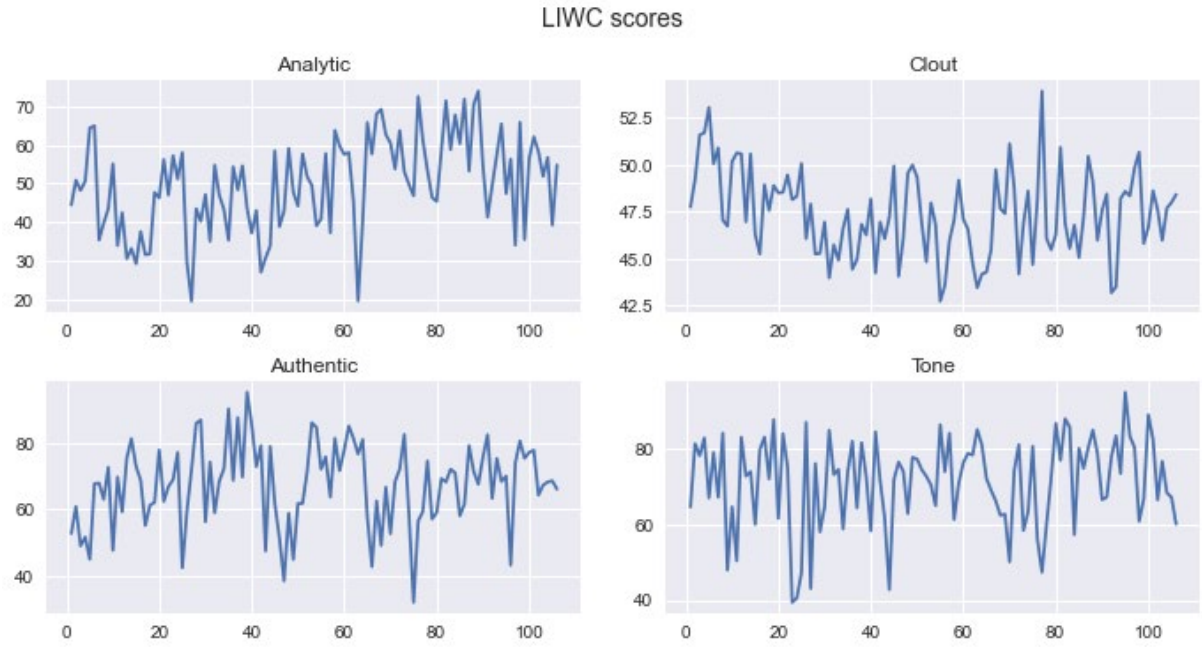

Figure 1 LIWC summary variable scores across 109 videos

Visual inspection confirms the characteristically fluctuating nature of the series. There is no obvious long-term trend anywhere, which rules out the analytically naïve notion of a linear pattern in the performance of the variables and, by extension, construction of identity. Any time series patterns would thus likely be localized and confined to shaping the characteristics of shorter session spans, implying a more dynamic character to the underpinning identity construction processes. As mentioned above, this is determined by the autocorrelation functions of the series, as depicted in the correlograms in Figure 2.

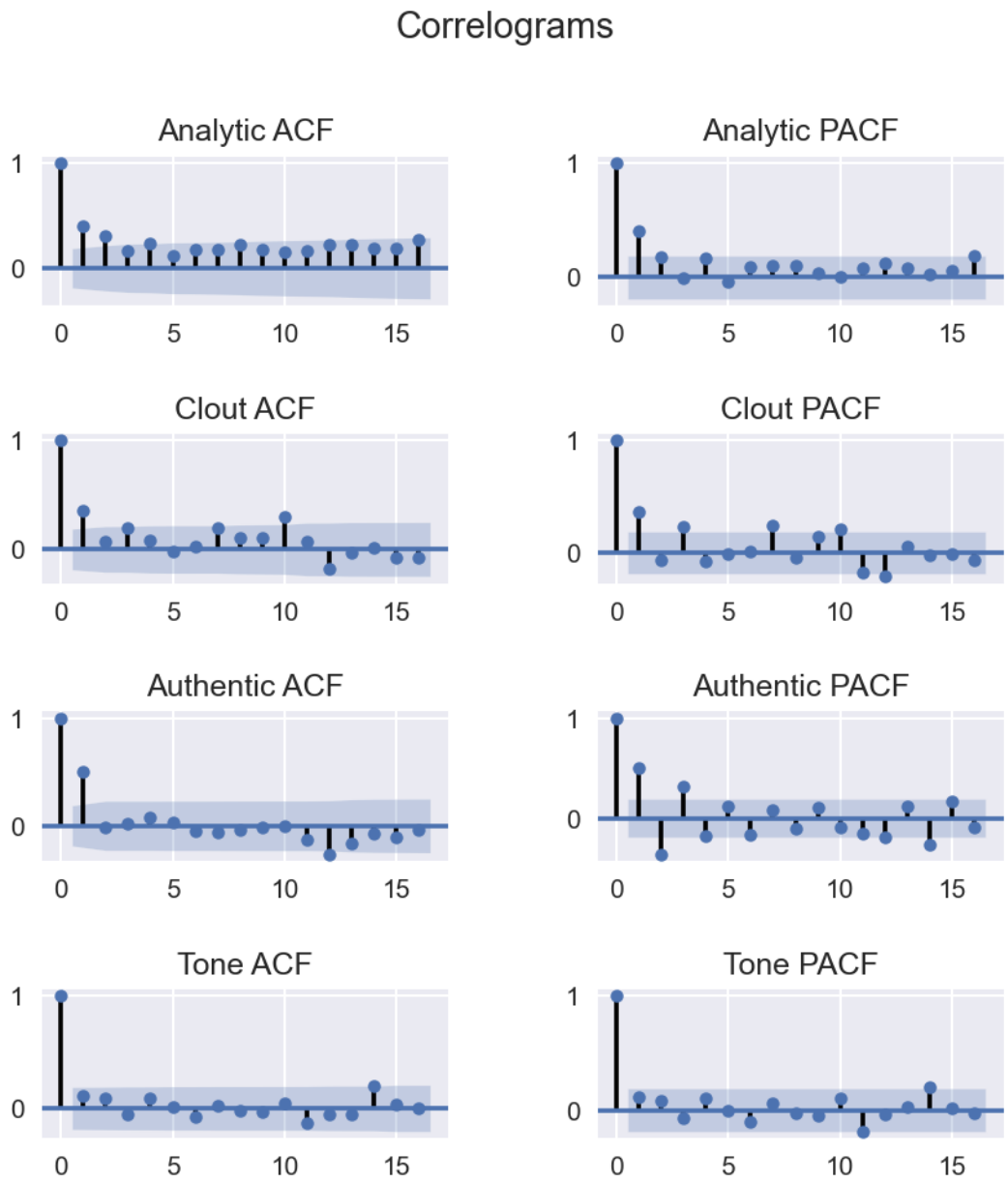

Figure 2 Autocorrelation functions of summary variables

The correlograms visualize the strength of autocorrelations (y-axis) at each lag from 0 to 15 (x-axis); i.e. how values 0,1, 2 … 15 intervals apart are positively/negatively correlated with one another. ACF stands for autocorrelation function and PACF, the partial autocorrelation function, denotes the autocorrelation at that lag controlled for previous lags. The blue bands demarcate the boundaries beyond which (P)ACF is statistically significantly different from zero at the 95% confidence level. The longer the series, the smaller the standard error of (P)ACF and the narrower these bands will be. This information reveals how consecutive series values shape one another and determines which ARIMA models should be chosen to fit each series.

The eventual fitted models are an AR(1) model for analytical thinking, MA(1) model for clout, and MA(2) model for authenticity. This is based on the observation that for these variables, the autocorrelations 'spike' beyond the region of statistical significance at lags 1 or 2. AR stands for 'autoregressive' and MA 'moving average'. These describe the nature of the relationship between values, while the number (1 or 2 in this case) indicates the number of time intervals across which this relationship holds. For emotional tone, however, no adequate model can be found. The absence of 'spikes' suggest no autocorrelations, meaning the series is random across time. Each variable will now be elaborated with supporting transcript extracts.

Analytical thinking: A display of 'short term momentum'

The AR(1) model for the linguistic display of analytical thinking in *Nikkietutorials* is formally expressed as $y_t = 30.121 + 0.394y_{t-1} + a_t$, where

- y_t is the value at time t (i.e. the t^{th} video)
- 30.121 is the intercept/constant term, and 0.394 is the AR(1) coefficient[4]
- y_{t-1} is the value at time *t-1* (i.e. one video prior)
- a_t is the error term inherent in any statistical model (i.e. the actual value minus the predicted value at time t)

Figure 3 details the estimated model parameters, coefficients, and other statistics less critical for the present analysis. It is meant for illustrating *Python* output and will not be discussed in detail here. The output will also be omitted for subsequent summary variables.

SARIMAX Results

Dep. Variable:	An	No. Observations:	106
Model:	SARIMAX(1, 0, 0)	Log Likelihood	-404.983
Date:	Wed, 18 Nov 2020	AIC	815.965
Time:	09:10:22	BIC	823.956
Sample:	0	HQIC	819.204
	- 106		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	30.1208	4.801	6.274	0.000	20.710	39.531
ar.L1	0.3944	0.096	4.114	0.000	0.207	0.582
sigma2	121.7308	18.950	6.424	0.000	84.588	158.873

Ljung-Box (L1) (Q):	0.47	Jarque-Bera (JB):	1.10
Prob(Q):	0.49	Prob(JB):	0.58
Heteroskedasticity (H):	1.22	Skew:	-0.14
Prob(H) (two-sided):	0.56	Kurtosis:	2.59

Figure 3 Estimated AR(1) model parameters

[4] The relationship between the constant term and AR parameters is c = (1-sum of parameters) x mean of the series

Figure 4 visualizes the predicted values (orange) using this model versus the actual 109 values (blue). The dotted line at the end shows forecasts using this model up to the hypothetical 112th video with the grey bands indicating 95% confidence intervals. By visual inspection, the model appears to fit the data well and successfully replicates its overall 'shape'. The mean absolute percentage accuracy (MAPE)[5], which is the averaged % error of the prediction for each video, is 14.734%. This is understandably higher than typical values in pure quantitative contexts (e.g. finance) but reasonably good for discourse data.

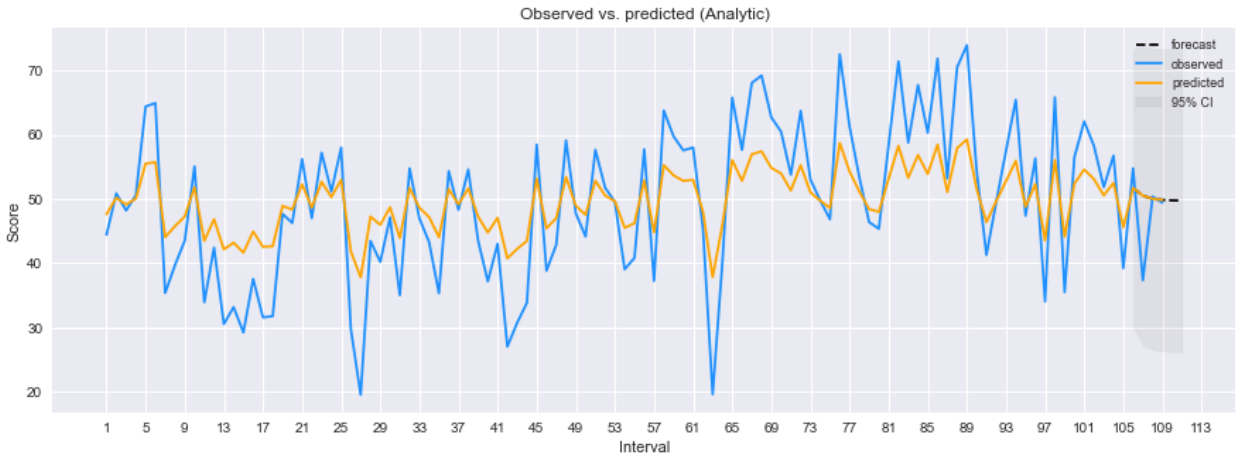

Figure 4 Observed vs. predicted values for analytical thinking

Setting the specific numbers aside, this model tells us in broad structural terms that

- Analytical thinking scores are positively correlated in successive videos, but only up to one video apart. In other words, the scores for videos *t* and *t*+1 tend to move in the same direction, but the score for video *t* does not provide useful information on *t*+2 and beyond

- The higher the value at this video, the higher the (predicted) value at the next video, and vice versa

As we saw in Table 2, the display of analytical thinking (i.e. formal, logical, hierarchical) in *Nikkiestutorials* is generally somewhere in the middle of common discourse contexts. Much of it is manifested when Nikkie gives stepwise instructions on how to perform make-up, and is 'offset' by informal, personal, and narrative-style language when she relates to viewers elsewhere in videos. Beyond this broad observation, the ARIMA model additionally tells us that

[5] There are many other measures of model fit and accuracy, but they will not be elaborated here. Interested readers may refer to Hyndman & Athanasopoulos (2018).

there are frequent video-to-video increases/decreases followed by a sudden movement towards the opposite direction. This dynamic, which can be described as a 'short term momentum' with potential directional shifts thereafter, is illustrated in the following three extracts from consecutive videos. Words constituting each summary variable (Table 1) are shown in bold throughout the rest of the paper.

Extract 1 (video #24, 15th May 2011): Disco lights (extreme clubbing) makeup tutorial

…**the** first **thing you** want **to** do is **to** apply base **to your** eyelids
to prevent **everything you** put **on from** creasing **and I'm** using **my** NYX Paoli pink pop **now** let **me** take **my** Mac pencil eyeliner impression **which** is **the** dark blue **and what I'm** going **to** do is **and I** learned **this** technique **at the** Academy **from my** teacher…

…thank **you** guys **so** much **for** watching **if you** want **to** find **this on** every product used **in this** look **it's on my** website nikkietutorials.com **I** love **you** thank **you for** watching **and hopefully I** will see **you** guys next time bye

The analytical thinking score in Extract 1 is 51.7, higher than the series average. In the first part she gives precise technical instructions on how to prevent creasing, emphasizing them as a technique learnt from her teacher at the academy and thus constructing her advice as highly credible. The second part is characteristic of the end of her videos where she thanks viewers and relates to them more personally, presenting a more 'authentic self' (Bhatia 2018). While these endings all appear similar, we will see subtle linguistic differences in the following examples as captured by the time series analysis. In other words, it is *because* we know that an AR(1) model fits the series well, that we direct our attention specifically to differences spanning one video apart. This is what is meant by the earlier claim that the present approach provides key 'entry points' to develop closer contextual analysis.

Extract 2 (video #25, 12th June 2011): Natural red carpet glamour makeup tutorial

…first **thing you** want **to** do is apply **a** base **to your**
eyelids **to** prevent **the** eyeshadow **that you're** going **to** put **on from** creasing **and** creasing is **I** have **it** right **now** creasing is **when that** line **of** concealer **that I** have **you** see **that** creasing is **when that** happens **to your** eyeshadow…

…I want **to** thank **you** guys **so** much **for** watching **and for a** full list **of** every product use go **to my** website nikkietutorials.com **I** love **you and** yeah **and you** guys **so** much **for** watching **and hopefully I'll** see **you** guys next time

In this immediately following video, the analytical thinking score rises to 58.01. She repeats her instructional style in the first part and continues to talk about preventing creasing, elaborating it in even more detail than before and explaining elsewhere that creasing seems to be a particular

issue in this case. We see a video-to-video momentum in constructing herself as a knowledgeable makeup artist where she picks up on a previous topic and elaborates it further. Her ending in the second part is almost identical to the previous video, and as mentioned above comes across as a template for thanking her viewers. However, we will see an obvious change in the next extract as her analytical thinking score changes direction as expected by the AR(1) model.

Extract 3 (video #26, 15th August 2011): Simple and clean back to school look

…hey **I'm** doing **a** tutorial **on this back-to**-school look um let **me** do **a** close-**up it's just a really** um natural look **with a** lot **of** shine **on the** eyes **and** some new lips **it's just a really** simple look um **the** eyes **and the** lips will take **like about** ten minutes **to** do…

… I hope **you** guys enjoyed **that if you** have any questions leave **them in the** comment section **below** umm **on my** website nikkietutorials.com **there's a** full list **of** all products used plugs **when I'm** wearing **them** is **a** lot **like** nails **and stuff** um yes **so I** hope **you** enjoyed **I** love **you and** uh stay **in** school **or** um **don't** fuck (beep) **when you're** 14 **I don't** know be good kids **it** will **it** will be good **for you in the** long haul yeah

In the next video, the analytical thinking score plunges to a low 29.64. This is predictable from the key feature of an AR(1) model; i.e. the value at *t-2* does not figure in the equation for the present value, implying that directional changes are unsurprising at two-interval spans. We can see that this is due to the switch from the complicated styles in the previous two videos to a 'simple and clean' style. In the first part, she no longer uses instructional language and emphasizes the simplicity of the style instead. The ending in the second part is also different as she concludes in a manner that coheres with the topic (back to school look) and target audience (young adolescents) – using more informal and narrative-like language ('*a lot like nails and stuff'*), and dispensing personal advice *('stay in school or um don't fuck when you're 14 I don't know be good kids…'*) as an older friend or sibling might.

In summary, her display of analytic thinking across the series can be described in terms of short-term, or 'localized' momentum shifts. The extracts show that for this particular three-video span, the 'up-up-down' pattern is strategically motivated by the different topics/target audiences at hand, but further analyses of other three-video spans could reveal other dynamics at work. The key point here is that it was the quantitative details of the time series model that focused the direction of the subsequent qualitative elaboration.

Clout: A display of 'short term restoration'

We now consider the linguistic display of clout; i.e. status, expertise, confidence. This is now an MA(1) model, formally expressed as $y_t = 47.419 - 0.455a_{t-1} + a_t$ where

- y_t is the value at time *t* (i.e. the t^{th} video)
- 47.419 is the intercept/constant term, and -0.455 is the MA(1) coefficient[6]
- a_t is the error term (i.e. the actual value minus the predicted value at time *t*)
- a_{t-1} is the error term at time *t-1* (i.e. one video prior)

The key difference between AR and MA models is that in AR models, the present value is predicted using past values, while in MA models it is predicted using past error terms. Before we elaborate this, Figure 5 shows the observed vs. predicted plot for clout.

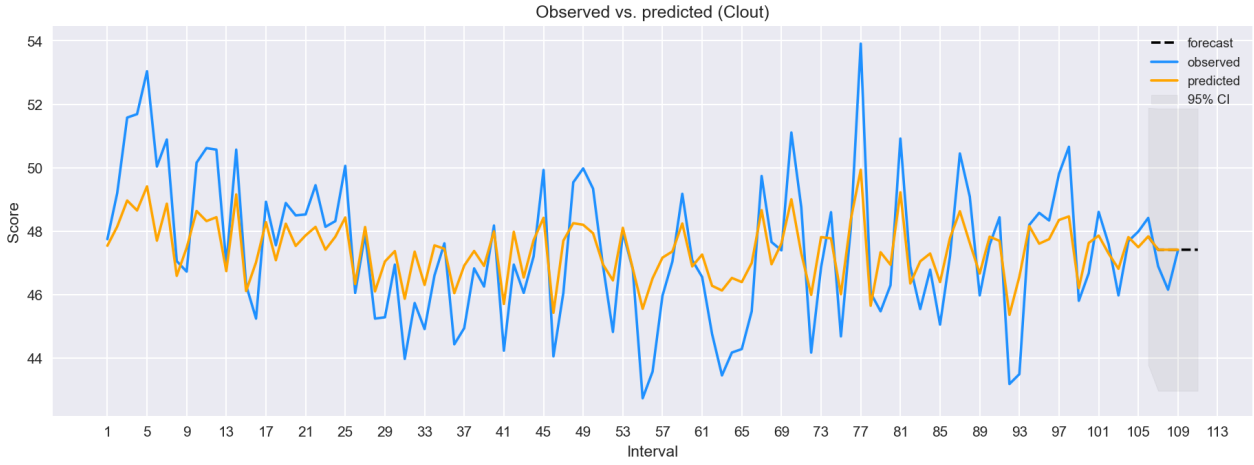

Figure 5 Observed vs. predicted values for clout

Once again, visual inspection reveals a good model-to-data fit. The MAPE is in fact better than the previous analytical thinking model at only 2.519%, even lower than many instances of inherently quantitative data (Fildes and Stekler 2002). This indicates that *Nikkietutorials'* clout is more 'modelable' than analytic thinking, and hence the dynamic described below is likely to be more robust than previously discussed for analytic thinking. The MA(1) model tells us in broad structural terms that

- Clout scores are not necessarily correlated in successive videos. Instead, the present clout value is negatively correlated with the size of previous *error terms*, up to one video apart. As before, because it is an MA(1) model, video *t* does not provide useful information on *t*+2 and beyond

[6] Different than AR models, the constant term for MA models is simply the mean of the series.

- A positive error term (i.e. actual value higher than predicted value) is likely to be followed by lower (predicted) clout in the next video, and vice versa

Table 2 shows that the display of clout in *Nikkiestutorials* is again somewhere in the middle of common discourse contexts. Much of it manifests in the giving of technical instructions, the absence of hedges/uncertain words, and the often subtle use of inclusive first-person plural pronouns when addressing her viewers. What the ARIMA model additionally tells us here is that i) there are occasions where clout is unexpectedly high/low due to various reasons (to be contextually investigated), resulting in a high positive/negative error term for that video; ii) the larger a positive error term, the more the display of clout will drop in the next video. The larger a negative error term, the more it will rise in the next video. Conversely, if clout is not unexpectedly high/low, the next video is not likely to witness a large increase. The following extracts from consecutive videos suggest a dynamic that can be described as a 'short term restoration' of equilibrium.

Extract 4 (video #65, 15th April 2015): Cat eyes makeup tutorial

…hey guys so today **I'm** doing a tutorial on the look **I** was wearing in **my** March hits and oh god knows **you** guys bombarded **me** with requests for that look so today I am here doing it for **you**… what people **seem** to love so much is that the lower lash line was very smoky and vampy and dark so go ahead and blend this as far down as **you like**…

…**I hope you** enjoyed for a full list of every single product mentioned and use go to **my** website nikkietutorials.com **you** can follow **me** on Twitter Instagram Facebook snapchat that all is nikkietutorials as always don't forget to thumbs up this video and subscribe to **my** channel again thank **you** guys so much for watching **I hope you** enjoyed and **hopefully I'll** see **you** guys on the next one

The clout score in Extract 4 is 44.2, lower than the series average. It was predicted to be 46.4 which gives us a relatively high negative error term of -2.2. The lower-than-predicted clout score could be due to her explanation that this video is a response to a 'bombardment' of requests, and she is therefore less sure than usual about being able to meet expectations. This is subtly evidenced in i) the first part where she downplays her usual 'expert' role by emphasizing that she is *'here doing it for you'*, that the look is '*what people seem to love'*, and that they can *'blend this as far down as you like';* ii) the second part where she repeats the tentative '*I hope you enjoyed'*. We now move on to the immediately following video.

Extract 5 (video #66, 27th May 2015): Berry/pink smokey eye makeup tutorial

…**I'm** taking this color right here called dusty rhodes and this color is going directly into the crease and because **we're** doing the smokey look **I'm** going to go all the way into **my** inner corners… **I** wanted to make this look super vampy and smoky

and dark so **let's** do that with the lips also hello can **we** focus right here right infront of **you** thank **you**…

…and that guys is how **you** can get berry smoky eyes with vampy lips and just a very vampy sexy look … **I** want to thank **you** guys so much for watching again and **hopefully I'll** see **you** guys on the next

In this video, in accordance with the aforementioned key feature of the MA(1) model, the clout score rises to 45.48 together with its predicted value of 47. The prior tentativeness is no longer apparent as this is not a request video, and there is thus less strategic need to project a sense of trying to reach popular expectations. In the first half of the extract, we see a subtle return of the inclusive and others-focused pronoun *we're doing.* The second *we* comes as the camera suddenly loses focus and she says '*hello can we focus right here'* in a bantering tone. While this is most likely a deliberate attempt at humor, it still contrasts clearly with the previous video in portraying her as being 'in charge'. The ending in the second half is likewise subtly different, with no more expressions of tentative *hope* that the audience liked the video.

In summary, these extracts qualitatively flesh out the essence of the MA(1) model, in that lower-than-expected clout in one video is quickly balanced by a higher clout level in the next. This short-term 'restoration' of a subtle sense of confidence and expertise stands in interesting contrast with the short term 'momentum' observed earlier for analytical thinking. Like before, however, the present extracts only illustrate one specific way in which the MA(1) model is strategically enacted.

Authenticity: 'Short term restoration' with a minor difference

We move on to consider the linguistic display of authenticity, defined in LIWC as more honest/ personal/disclosing versus guarded/distanced discourse. It is also the most complicated model thus far – an MA(2) model formally expressed as $y_t = 67.373 - 0.293a_{t-1} - 0.279a_{t-2} + a_t$ where

- y_t is the value at time t (i.e. the t^{th} video)
- 67.373 is the intercept/constant term, -0.293 is the MA(1) coefficient, and -0.29 is the MA(2) coefficient
- a_t is the error term (i.e. the actual value minus the predicted value at time t)
- a_{t-1} is the error term at time *t-1* (i.e. one video prior)
- a_{t-2} is the error term at time *t-2* (i.e. two videos prior)

The basic features of an MA(2) model are similar to the previous MA(1) model, except that the error term two videos prior (a_{t-2}) now also plays a significant role in predicting the present value. This means that the general dynamic of 'short term restoration' also applies to the display of authenticity, but with a slightly longer span; i.e. unexpectedly large error terms can now exert an

influence across two-video spans. As usual, Figure 6 shows the observed vs. predicted plot for authenticity.

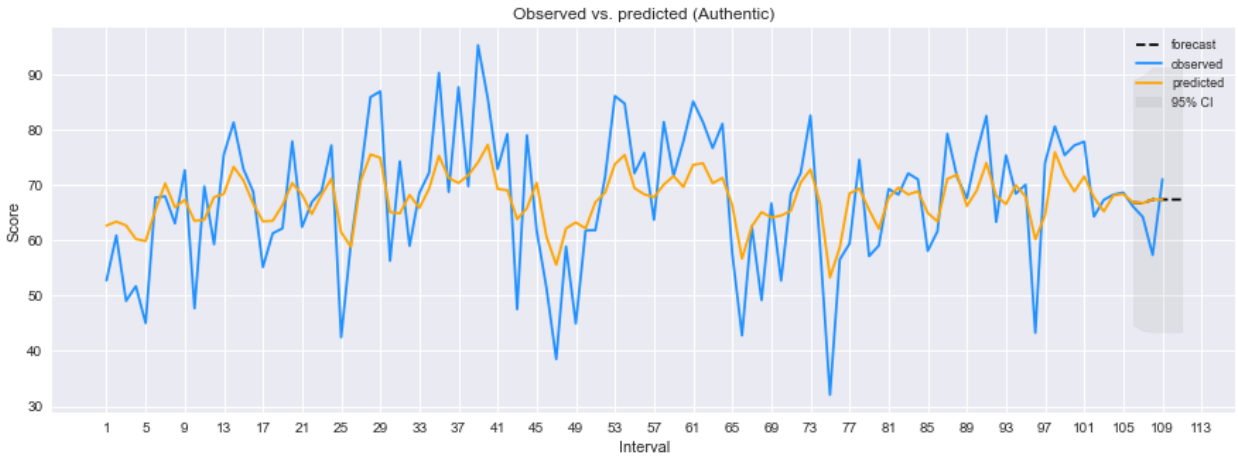

Figure 6 Observed vs. predicted values for authenticity

The model-to-data fit is again visually apparent. This time the MAPE is 11.397%, better than analytical thinking but not as good as clout. This more complex MA(2) model tells us in broad structural terms that

- Authenticity scores, like clout scores, are not necessarily correlated in successive videos. The present authenticity value is negatively correlated with the size of *two* previous error terms. As an MA(2) model, video *t* now provides useful information on *t*+2, but not *t*+3 and beyond

- Positive error terms in the previous two videos are likely to be followed by lower (predicted) clout in the next video, and vice versa. If one error term is positive and the other negative, the net effect will apply to the next video

Table 2 shows that authenticity in *Nikkiestutorials* is generally quite high and second only to expressive writing. As observed in previous studies (Bhatia 2018; Tolson 2010; Valentinsson 2018), much of YouTubers' authenticity is constructed with elements like intimate, spontaneous, and colloquial language to replicate an actual face-to-face conversation. Though LIWC defines authenticity somewhat differently (Table 1), we still see in the following extracts many of these elements. The MA(2) model captures a similar dynamic of 'short term restoration' as seen for clout, but this time an unexpectedly high/low display of authenticity can exert its influence up to two videos later. We illustrate this with extracts from early in her career. Videos #4 and #5 both

had actual values (51.73, 45.03) much lower than predicted values (60.25, 59.87) and also much lower than the mean of the series. Extract 6 is taken from video #5 to explain the low values, and Extract 7 is from the subsequent video #6 to illustrate the 'restoration' of authenticity.

Extract 6 (video #5, 25th January 2009): Makeup tutorial; dramatic black

…**I** use Beigeing Shadestick and **just** like in the Arabic tutorial **we're going** to **put** gesso on the lid to help it blend out the pencil and all the colors now you're **going** to take a black creamy eyeliner pencil **I** will take Gosh eyeliner in black ink make sure it's sharp so sharpen it up a bit and **just** like in **our** previous video you're **going** to look straight in your mirror and paint that crease line on your eye again…

…**I** hope you like it it's **really** dramatic **but** yeah **I** hope you like it okay thanks for watching and please comment rate and subscribe bye

In Extract 6, the lower-than-expected authenticity score can be explained by the more procedural ('*we're going to*') and distanced nature of the instruction, which is different than comparable sections later in her career where more first-person pronouns are used. In particular, she makes explicit reference to video #4 ('*just like in the Arabic tutorial/our previous video*') which implies that the two consecutive videos convey similar instructions. This somewhat downplays the spontaneity of the present video. In the second half, we see how differently she ends her videos compared to the previous extracts. It does not thank the viewers as much, does not express hope for their return, and generally comes across as less interpersonally oriented. Compare this with the following extract from the next video, where authenticity is 'restored' in line with the prediction of the MA(2) model.

Extract 7 (video #6, 21st February 2009): SCANDALOUS turqouise kiwi look

hello today **we're going** to do this **really** bright outgoing boom look it's gorgeous **I** love it and **I** hope you like it too and **if** you want to know how **I** did this keep on watching…now you're **going** back in Gesso with a stiff shader brush make it unique again other side **go** in that scandalous color and **really really** make that one **really** bright like bright in your face bright **we're going** to do **our lower** lid wipe the brush up **really** well…

…and **my** favorite lip gloss **or** plush glass by Mac which is ben-too-ful (beautiful). **I** hope you like it umm **I really** do like it **myself** it's **really** outgoing and artsy and **ugh I** love it so thanks for watching please comment rate and subscribe bye

In this video, the authenticity score rises to 51.79 together with its predicted value of 65.45. This mimics the earlier pattern for clout except that two previous videos influenced video #6, instead of just the previous one. The 'return to authenticity' is apparent right from the video title ('*SCANDALOUS*') which was capitalized to convey a sense of informal spontaneity. This was consistently enacted throughout the extract. In the first half, she emphatically expresses a much higher degree of love and excitement for the style at hand (*'really*'), and this continues in the second half with the embellished pronunciation of *beautiful* and the exclamative *ugh*. This restoration of authenticity invites viewers to regard the video content as new and refreshing, compared to the previous two videos that are similar to each other. As with the previous summary variables of analytical thinking and clout, we therefore see that her linguistic display of authenticity also takes on a 'localized' character – responding to quick shifts in the video-to-video situation at hand.

Emotional Tone: Random fluctuations across time

The final variable, emotional tone, is defined as the extent to which a text reflects positive, negative, or neutral emotionality. *Nikkietutorials* has a relatively positive tone closest to its social media counterpart Twitter (Table 2), which is unsurprising since most social media influencers would want to project an emotionally positive image. Unlike the previous three summary variables, however, the absence of significant autocorrelations in the series implies that her linguistic display of emotionality is time-independent. In other words, it fluctuates irregularly about the mean value across videos, and does not offer clear evidence of meaningful patterns across time. Figures 7 and 8 show what analysts typically do in such situations – to use either a i) mean model (Figure *7*), which simply predicts each value using the mean of the series, or ii) a naïve model, also known as a 'random walk' model (Figure 8), which predicts that the next value is simply the present value in the series.

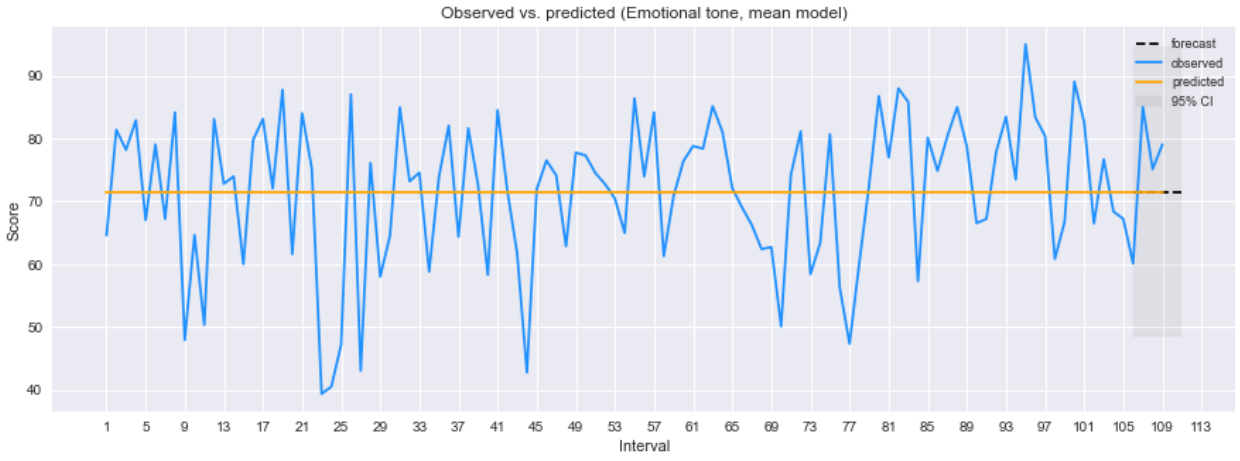

Figure 7 The mean model for emotional tone

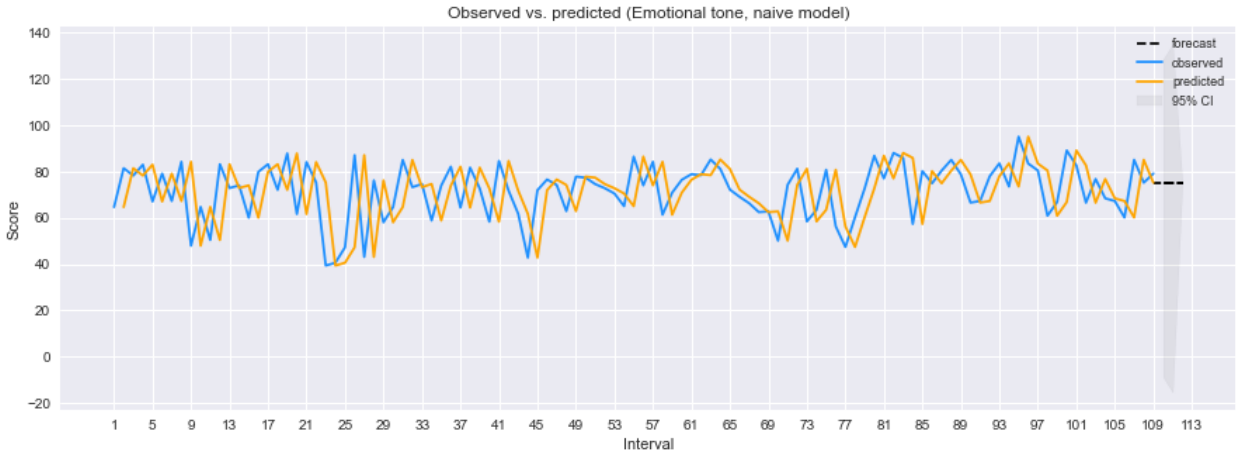

Figure 8 The naïve model for emotional tone

In both cases, the implications are that emotionality display has a spontaneous, *ad hoc* character that is more grounded upon the immediate context than shaped by prior presentations as was shown to be the case for the other summary variables.

Conclusion

This paper demonstrated an approach to analyzing the linguistic performance and construction of identity on YouTube that is based on the notion of 'modelability'. Combining LIWC for lexical analysis and the Box-Jenkins time series method, identity construction is depicted by the nature of ARIMA time series models for four LIWC summary variables. Well-fitted models were further investigated in context, with reference to actual transcripts as a signature of identity construction. One might object that LIWC or any other analysis of discrete structures and forms understates the pragmatic complexities of identity construction. It is important to emphasize that the present approach is complementary, and addresses the oft-overlooked question of how temporal progress shapes social media discourse in systematic ways.

The approach was applied to a case study of 109 consecutive *Nikkiestutorials* makeup tutorial videos. The primary finding ruled out any naïve conception of identity construction as a linear process. Instead, the autocorrelation functions painted a 'localized' picture where linguistic performance of analytical thinking, clout, and authenticity fluctuated across (only) short video spans. The autoregressive nature of analytical thinking scores suggested a dynamic of 'short term momentum' strategically motivated by changes in the topic/target audience, while the moving average nature of clout and authenticity suggested a subtly different dynamic of 'short term restoration' where unexpected dips were quickly balanced in the immediately following video. Further differences between clout and authenticity were captured in terms of their respective MA(1) and MA(2) models, with a longer span of influence of past videos in the latter. The linguistic display of emotional tone was a randomly fluctuating variable with no predictable dynamic across time.

There are several methodological advantages to the present approach. Firstly, the two levels of modelability – i) the general level where we consider if a series exhibits autocorrelation, and ii) the specific level where we interpret models in context – can depict identity construction in both broad and nuanced terms. Modelability indicates either passive linguistic reflection of patterned background events, or active patterned linguistic representation of seemingly random background events. The latter resonates particularly with major theoretical claims in cognitive and critical discourse analytic approaches (Hart 2011). At the specific level, model fit and evaluation measures like MAPE offer a fairly concrete evaluation of the extent to which identity construction across time is patterned.

Secondly, the approach can facilitate new ways of synergizing quantitative and qualitative analyses in related research. Time series models illuminate key entry points for subsequent scrutiny of phenomena in context. A model of order *k* alerts the analyst to pay attention to things *k* intervals apart, while the direction and extent of correlations motivate qualitative analysis of sequential increases/decreases. Time intervals may in fact be modified to suit the research objective, e.g., minute-to-minute versus video-to-video changes. Importantly, the time series modeling process does not itself impose a particular theoretical view, but invites the analyst to interpret temporal patterns with appropriate theoretical perspective(s) at the final phase. It should also be noted that this paper focused more on demonstrating how the temporal signatures are fleshed out in transcript extracts, rather than applying a comprehensive analytical framework to advance a specific interpretation. After all, while the four summary variables provide an adequate profile of identity construction, they are ultimately neither necessary nor sufficient to depict such a complex construct.

Thirdly, the approach is flexibly replicable on other linguistic/discursive variables in temporal discourses, with different analytic frameworks/foci, and using tools other than LIWC or the Box-Jenkins method (e.g. Gardner, 1985). Though suited for case studies, well-conceived aggregated or averaged could in principle be modeled and interpreted in similar ways. In other words, each major phase – the computation of variable scores, time series modeling, and subsequent interpretation – is in a sense modular, but together outline a schematic approach towards identity and related analysis. While replicability may not always be insisted upon as a feature of identity research, it should still contribute towards enhancing the comparability and transparency of findings.

Funding acknowledgement

This work was supported by the HKSAR Research Grants Council (Project number: 15601019).

References

Abidin, Crystal. 2018. *Internet Celebrity: Understanding Fame Online*. Bingley, UK: Emerald.

Bhatia, Aditi. 2018. "Interdiscursive Performance in Digital Professions: The Case of YouTube Tutorials." *Journal of Pragmatics* 124: 106–20.

Bowerman, B., & O'Connell, R. (1987). *Time Series Forecasting. Unified Concepts and Computer Implementation* (2nd ed.). Duxbury Press.

Du Bois, John. 2007. "The Stance Triangle." In *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction*, ed. Robert Englebretson. Amsterdam: John Benjamins, 139–82.

Boyd, Ryan L., Blackburn, Kate G., & Pennebaker, James W. (2020). The narrative arc: Revealing core narrative structures through text analysis. *Science Advances*, *6*(32), eaba2196.

Box, George E. P., Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. 2015. *Time Series Analysis: Forecasting and Control*. 5th ed. Hoboken, NJ: Wiley.

Bucholtz, Mary. 1999. "'Why Be Normal?': Language and Identity Practices in a Community of Nerd Girls." *Language in society* 28: 203–23.

———. 2003. "Sociolinguistic Nostalgia and the Authentication of Identity." *Journal of Sociolinguistics* 7(3): 398–416.

Bucholtz, Mary, and Kira Hall. 2005. "Identity and Interaction: A Sociocultural Linguistic Approach." *Discourse Studies* 7(4–5): 585–614.

Cohn, Michael A, Matthias R Mehl, and James W Pennebaker. 2004. "Linguistic Markers of Psychological Change Surrounding September 11, 2001." *Psychological Science* 15(10): 687–93.

Dlaske, Kati. 2017. "Music Video Covers, Minoritised Languages, and Affective Investments in the Space of YouTube." *Language in Society* 46(4): 451–75.

Fildes, Robert, and Herman Stekler. 2002. "The State of Macroeconomic Forecasting." *Journal of Macroeconomics* 24(4): 435–68.

Gardner, Everette S. 1985. "Exponential Smoothing: The State of the Art." *Journal of Forecasting* 4(1): 1–28.

Hart, Christopher. 2011. "Force-Interactive Patterns in Immigration Discourse: A Cognitive Linguistic Approach to CDA." *Discourse and Society* 22(3): 269–86.

Hyndman, Rob J., and George Athanasopoulos. 2018. *Forecasting: Principles and Practice*. OTexts.

Kacewicz, Ewa et al. 2013. "Pronoun Use Reflects Standings in Social Hierarchies." *Journal of Language and Social Psychology* 33(2): 125–43.

Kern, Margaret L et al. 2016. "Gaining Insights from Social Media Language: Methodologies and Challenges." *Psychological Methods* 21(4): 507–25.

Newman, Matthew L, James W Pennebaker, Diane S Berry, and Jane M Richards. 2003. "Lying Words: Predicting Deception From Linguistic Styles." *Personality and Social Psychology Bulletin* 29(1901): 665–75.

Ochs, Elinor. 1996. "Linguistic Resources for Socializing Humanity." In *Rethinking Linguistic Relativity*, eds. John J Gumperz and Stephen C Levinson. Cambridge: Cambridge University Press, 407–38.

Omoniyi, Tope, and Goodith White, eds. 2006. *The Sociolinguistics of Identity*. London: Continuum.

Pennebaker, James W. et al. 2014. "When Small Words Foretell Academic Success: The Case of College Admissions Essays." *PloS one* 9: 1–10.

Pennebaker, James W., Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The Development and Psychometric Properties of LIWC2015*. Austin, TX: University of Texas at Austin.

Silverstein, Michael. 1976. "Shifters, Linguistic Categories, and Cultural Description." *Meaning in Anthropology*: 11–55.

Strangelove, Michael. 2010. *Watching YouTube: Extraordinary Videos by Ordinary People*. Toronto, Canada: University of Toronto Press.

Tajfel, Henri, ed. 1982. *Social Identity and Intergroup Relations*. Cambridge and New York: Cambridge University Press.

Tausczik, Yla R, and James W Pennebaker. 2010. "The Psychological Meaning of Words : LIWC and Computerized Text Analysis Methods." *Journal of Language and Social Psychology* 29(1): 24–54.

Tay, Dennis. 2017. "Time Series Analysis of Discourse. A Case Study of Metaphor in Psychotherapy Sessions." *Discourse Studies* 19(6): 694–710.

———. 2019. *Time Series Analysis of Discourse. Method and Case Studies*. New York: Routledge.

———. 2020. "A Computerized Text and Cluster Analysis Approach to Psychotherapy Talk." *Language & Psychoanalysis* 9(1): 1–22.

———. 2021. "Automated lexical and time series modelling for critical discourse research: A case study of Hong Kong protest editorials" *Lingua* 255, 103056.

Tolson, Andrew. 2010. "A New Authenticity? Communicative Practices on YouTube." *Critical Discourse Studies* 7(4): 277–89.

Valentinsson, Mary Caitlyn. 2018. "Stance and the Construction of Authentic Celebrity Persona." *Language in Society* 47(5): 715–40.

Xu, Weiai, and Congcong Zhang. 2018. "Sentiment, Richness, Authority, and Relevance Model of Information Sharing during Social Crises—the Case of #MH370 Tweets." *Computers in Human Behavior* 89: 199–206.