

Corpus linguistics and the study of social media: a case study using multi-dimensional analysis

Tony Berber Sardinha

1 What is social media?

Social media refers to the communication generated on platforms that enable users to send text messages and multimedia content to a group of other users. Many different social media outlets exist today, like *Facebook*, *Twitter*, *Instagram*, *YouTube*, *Tumblr*, *Snapchat* and *TikTok*, among others. Each was introduced to serve a particular purpose for a particular audience. When it was created in 2003, *Facebook*'s predecessor *Facemash* allowed Harvard University to rate the appearance of fellow students. YouTube, which was registered in 2005, was originally meant for people to share home videos on the Web. Two years later, Tumblr was introduced as a blogging site that provided a “tumblelog” – that is, a space where they could post “tumblin” (one-paragraph) messages, photos and videos.

Unlike *Facebook*, *YouTube* and *Tumblr*, which were devised as websites, *Twitter* was designed to run on cell phones as an SMS service. *Twitter* came about in 2006, conceived as a tool for users to update each other on their whereabouts and activities (Weller *et al.* 2014: x). Other social media outlets were developed first and foremost as mobile technology applications. *Instagram* was released in 2010 as a mobile phone application for photo sharing. Snapchat was introduced a year later as a photo app in which the post would quickly vanish, so users would not feel pressured to post “Kodak-perfect moments”. The mobile app *TikTok* came to the scene more recently, when it absorbed the short video-sharing app Music.ly in 2017. Like its predecessor, *TikTok* allows users to share short video content such as lip-sync and dance routines.

When these social media applications were first introduced, many catered to a particular audience and sought to meet or create a particular demand. However, over time many converged to offer some form of text and multimedia content sharing. The current leading social media platforms all provide the means for users to post texts, pictures and audio and video files. In addition, all offer mobile technology applications, which have become the predominant environment in which users engage with the social networks.

Mobile technology has helped social media platforms gain enormous popularity because smart phones and wireless broadband have become more affordable worldwide, thereby helping these social networks penetrate markets in the developing world, adding

billions of users to the networks in addition to enabling more users to shoot, edit and post video footage. Such expansion has caused a major shift in the social networks: They have become less a space for sharing content and more a tool to influence human behavior. Nowadays, engaging with social media is increasingly a way of life rather than a distraction, where millions of digital influencers dictate social, cultural and political norms. *Twitter*, for instance, has repositioned itself as a leading news source – so much so that its posts are widely quoted in mainstream news reports. Major public figures use *Twitter* as an official channel of communication. *Instagram* has been adopted as a prime outlet for major fashion brands to advertise their products and build their corporate image. *YouTube* and *TikTok* have become major players in the music business, with more artists using these platforms as their primary source of income.

The widespread use and influence of social media have led scholars to argue that the social networks are a real world rather than a made-up world:

When the study of the internet began people commonly talked about two worlds: the virtual and the real. By now it is increasingly evident that the online is just as real as the offline. In the same way no one today would regard a telephone conversation as taking place in a separate world from “real life”.

(Miller et al. 2016: 7)

This shift suggests that investigating social media enables us to understand both virtual and “real” life, as “real” life is increasingly influenced by social media. Corpus-based research on social media can seek answers to key questions such as what kind of language is used in these social media communities, how language use varies across groups and individuals, what text varieties exist and how they compare with non-digital varieties and how widespread non-standard forms such as contractions, hashtags, emoticons and emojis are.



Despite its growing importance, social media has received scant attention in our field due in part to the kinds of text used in social media posing challenges for corpus methods and techniques. Social media texts are generally “messy” – that is, they make extensive use of non-standard spelling, punctuation, abbreviations, emoticons, emojis and hashtags, which can cause serious problems for part-of-speech taggers and lemmatisers. In addition, social media texts are typically very short, which raises questions about the reliability of relative frequency counts of lexical and grammatical features. For such operations to be successful, the corpus generally needs to undergo several rounds of adaptation, as discussed next.

2 Collecting and handling a corpus of social media

A corpus of social media usually consists of many thousands of texts because social media posts are abundant (although some platforms impose restrictions on downloading). As manually collecting large numbers of individual texts is a tedious and error-prone process, many analysts resort to “scraping,” which is automated collection using computer software. A range of scraping tools exist in the natural language processing community, but since these generally require programming skills (e.g. *Python*) to install and run, many corpus linguists do not make use of them. Researchers can find the currently available scraping tools by searching the Web (e.g. “scraping *Twitter* data”). Some scraping tools have the added benefit of downloading metadata for the posts, such as the username of the poster, their geographical location (geodata) and a count of likes for the post.

Social media posts are notorious for their use of non-conventional and innovative word forms. Consequently, once a corpus of social media has been collected, researchers need to decide whether the orthography and punctuation in the texts will be adapted (“normalised”) and, if so, to what extent. This process of converting the orthographic features to a standard is generally referred to as “normalisation” in natural language processing. However, in order to avoid confusion with the term normalisation used in corpus studies to refer to the computation of relative frequencies (see Chapters 10 and 39, this volume), it will be referred to as “lexical normalisation”.

Lexical normalisation refers to a set of procedures whose goal it is to adapt the texts to orthographic convention. For example, the fully normalised version of “john gave ALL HIS LOVE 2u” would be “John gave all his love to you”. During lexical normalisation, many of the natural features of social media texts are stripped away, like emphasis by uppercase and character repetition, shortened forms and non-standard capitalisation. The downside of lexical normalisation is that by removing these features, layers of meaning are also removed. But a benefit is that corpus tools can retrieve all the instances of a token that were spelled differently, like “for,” previously spelled “4,” “just” (“jus”) and “great” (“GREAAATTT”). A lack of normalisation can significantly affect part-of-speech tagging and lemmatisation, as unconventional spelling and non-standard capitalisation can cause tagging errors with taggers not trained with social media data. Lexical normalisation involves several individual processes, covering a range of textual and graphic aspects, such as:

1. Automatic spell-checking. This refers to detecting non-canonical spellings and converting these to conventionally spelled forms. For instance, the spell-corrected version of “I jus cant see u” would be “I just can’t see you”;
2. Truecasing. This is a special case of spell-checking that refers to determining the standard capitalisation of words and changing the capitalisation if needed. For instance, “i think new york is AWESOME” could be truecased as “I think New York is awesome”;
3. Acronyms and abbreviations. Many posts normally include strings like “xoxo”, “imho” and “afaik”. If these strings were normalised, they would be replaced with “hugs and kisses”, “in my humble opinion” and “as far as I know”, respectively;
4. Emoticons. These are strings of keyboard characters that express such features as emotions, attitudes and feelings. For instance, normalising :-) and :-] would convert these strings to “smile” or “smiling face”; :-D to “smile”, “smiling face”, “grin” or “grinning face”; and :-O to “surprise”, “shock” or “yawn”. A definitive dictionary of emoticons does not exist; therefore, a single emoticon may be normalised in different ways;
5. Emojis. These are in-text graphic characters depicting all sorts of objects, body parts and animals as well as human emotions, attitudes and feelings (like their predecessors, the emoticons). There is no consensus on how to translate emojis to text, so a single emoji can be described in a variety of ways. For instance, the emoji  can be rendered as “laughing”, “laughing face”, “happy”, etc. Additional shades of meaning can be added to emojis, like skin color; for instance,  is referred to in emoji indexes as “Raised Back of Hand: Dark Skin Tone”. In addition to denotative meaning, emojis embody pragmatic meaning; a raised hand emoji can mean “me”, “I want to talk”, “I’m here”, etc., depending on the context, and skin colour can add further instances of meaning;

6. Hashtag segmentation. Hashtags are strings of characters that signal the contents of the post and make it possible to link different posts that refer to a similar topic. They are initiated by the hash or pound sign (#) and generally comprise a sequence of words with no spaces between them. Hashtag segmentation is the process whereby the boundaries between the constituent word forms in a hashtag are detected and blanks are inserted between the word forms. For instance, the hashtag #ILOVENYC would be segmented as “I LOVE NYC”.

Part-of-speech tagging can be carried out using general-purpose or specialised taggers. General-purpose taggers include such tools as the *Biber Tagger*, *TreeTagger* and *CLAWS*. General-purpose taggers will be affected by a lack of normalisation if they were not trained on social media texts, but if run on normalised posts, they can provide high-accuracy tagging for hundreds of linguistic features. Specialised taggers include CMU's *Twitter* tagger (aka “Gimpel tagger,” <http://www.cs.cmu.edu/~ark/TweetNLP>). These provide better performance on non-normalised corpora, but their feature sets are usually more restricted than general-purpose taggers.

3 Using multi-dimensional analysis to investigate social media

MD analysis is a framework for corpus analysis that uses multivariate statistical analysis to identify correlations among linguistic features across the texts in a corpus (Berber Sardinha and Veirano Pinto 2014, 2019). MD analysis was introduced by Douglas Biber in the 1980s (Biber 1988), and since then it has been widely applied in corpus linguistics, primarily in register variation research. Two major MD analysis types exist: a grammatical variant, which focuses mostly on the structural characteristics of the texts, and a lexical variant, which focuses on the lexical units in the texts (single words, n-grams, collocations) (Berber Sardinha 2021). In the study reported in this chapter, a grammatical MD analysis was carried out to detect the major dimensions underlying the variation across platforms, user groups and individual users in social media.

A dimension of variation is the parameter underlying the variation across the texts. In a grammatical MD analysis, the dimensions correspond to the major communicative functions in the texts. For example, Biber (1988) identified five major dimensions of register variation for spoken and written English:

1. Involved versus Informational Production;
2. Narrative versus Non-Narrative Concerns;
3. Explicit versus Situation-Dependent Reference;
4. Overt Expression of Persuasion;
5. Abstract versus Non-Abstract Information.

Each dimension comprises a set of linguistic features identified statistically through factor analysis. The factors are interpreted qualitatively based on the overall function performed by the features. For example, the involved end of dimension 1 is a result of the cooccurrence of such features as first- and second-person pronouns, contracted forms, *that* deletions, private verbs, hedges and amplifiers, among many others. These features enable the production of interactive, oral forms of communication in both speech and writing. In speech, this can take the form of various registers, like face-to-face and telephone conversations, spontaneous speeches and interviews, whereas in

writing it materialises as personal letters, stage plays and romantic fiction, among many other such registers.

The major steps involved in a typical MD analysis are as follows:

1. Corpus design and compilation. The corpus should be designed as a representative sample of the registers of interest. The text is the central unit for corpus construction in MD analysis. The boundaries between the texts must be preserved (i.e. the individual texts should not be lumped into single files), and the counts of linguistic features must be taken for each individual text rather than for whole sections or the corpus as a whole. In the MD analysis framework, the total word count of the corpus is less important than the total number of texts (in each of the different sections of the corpus);
2. Selection of linguistic features. The analyst should select the linguistic features of relevance based on the previous literature;
3. Tagging. The corpus must be tagged for the relevant linguistic features, with a reliable automatic tagger;
4. Frequency counts. After tagging, the features are counted for each text, and the counts are normed to a fixed rate (e.g. 1,000 words) to enable frequency comparisons across texts of different lengths;
5. Factor analysis. The normed counts are submitted to factor analysis, a statistical technique that enables the identification of factors and sets of correlated linguistic features that correspond to latent (unobserved) variables in the corpus;
6. Scoring. For each factor, a score is computed for each individual text, based on the counts of the linguistic features loading on the factor. Mean scores are calculated for the relevant sections of the corpus;
7. Interpretation of the factors. The factors are interpreted qualitatively so as to determine the underlying dimensions of variation. Each dimension receives an interpretive label that captures its essential communicative properties.

These major steps were followed for the analysis reported in this chapter. However, because social media has particular typographical and linguistic characteristics that set them apart from other registers, extra steps were needed to prepare the corpus for tagging, which are detailed next.

A corpus was collected consisting of texts posted in English on *Twitter*, *Instagram* and *Facebook* between 2018 and 2019 (see Tables 46.1 and 46.2). Researchers should consider the role of images and audio features in social media and gauge the impact of retaining or removing such features from their corpus. Removing the non-textual materials from a corpus of social media can lead to the criticism that posts are often direct responses to images or audio, and therefore, the text is impoverished by separating them. Spoken conversational corpora have also been criticised for de-contextualising the language. To prevent such criticism, one solution is to code the extra-linguistic features in the corpus files; however, manual coding is time-consuming, so researchers first need to evaluate its cost-effectiveness. In the current study, metadata for the visual and audio content would have no added benefit because the focus is on the grammatical characteristics of the texts. Therefore, the corpus was restricted to the actual posts only – namely, the written messages posted by the account owners on the platforms (the text plus any in-text graphic content; i.e. emojis), while photographs, sound and video content were not included, nor were the comments to the posts. The texts were obtained

Table 46.1 Corpus used in the study: breakdown by platform

<i>Platform</i>	<i>Texts</i>	<i>Tokens</i>	<i>Mean tokens</i>	<i>Min.</i>	<i>Max.</i>	<i>SD</i>
Facebook	14,468	498,222	34.44	1	1089	38.03
Instagram	13,904	592,884	42.64	1	429	47.79
Twitter	14,288	322,382	22.56	1	83	13.68
Overall	42,660	1,413,488	33.13	1	1089	36.95

Table 46.2 Corpus used in the study: breakdown by user group

<i>Platform</i>	<i>User group</i>	<i>Texts</i>	<i>Tokens</i>	<i>Mean tokens</i>	<i>Min.</i>	<i>Max.</i>	<i>SD</i>
Facebook	Celebrities	4706	106,441	22.61	1	410	23.36
	Corporations	4974	149,407	30.03	1	367	25.84
	Politics	4788	242,374	50.62	1	1089	52.16
Instagram	Celebrities	4388	124,276	28.32	1	416	38.00
	Corporations	4759	250,243	52.58	1	373	49.87
	Politics	4757	218,365	45.90	1	429	50.57
Twitter	Celebrities	4430	72,890	16.45	1	78	12.12
	Corporations	4915	102,979	20.95	1	72	12.14
	Politics	4943	146,513	29.64	1	83	13.29

using a variety of methods, including manually copying the posts, scraping the posts in bulk and downloading ready-made datasets from the Web. The corpus was designed around three major components: platform, user group and user. “Platform” is one of the three social media outlets where the text was posted, “user group” refers to a class of users (e.g. celebrities, corporations or political figures/groups) and “user” identifies the account where the posts originated. The user group “celebrities” includes pop artists, movie actors/actresses, athletes and TV personalities; “politics” comprises world leaders, political parties and political organisations; and “corporations” consists of international brands. The users were selected based on lists of the most followed personalities, companies and organisations available on the Web. With the exception of corporations, most users were based in the United States or Europe.

The corpus was normalised using the following tools:

1. Emojificate, for converting emojis to text, <https://pypi.org/project/emojificate/>;
2. Truecase, for fixing the capitalisation, <https://pypi.org/project/truecase/>;
3. Ekphrasis, for tokenisation and hashtag segmentation, <https://github.com/cbaziotis/ekphrasis>;
4. MoNoise: for spell correction and acronym conversion, <https://bitbucket.org/robvanderg/monoise/src/master>.

The abbreviations were expanded using a shell script developed especially for this project. Other scripts were further used to clean up and format the posts prior to tagging.

After lexical normalisation, the corpus was tagged for part of speech using the Biber Tagger, which is widely used in MD research (Gray 2019). It tags texts for hundreds of linguistic characteristics, including word class, clause types, discourse features, semantic categories and stance markers. Once the corpus was fully tagged up, it was run through the Biber Tag Count, a programme that counts the features in the tagged texts, normalises the counts to a rate per thousand words and returns a data file with counts for more than 100 different features.

Frequency normalisation is needed because raw frequency counts are sensitive to text length. Longer texts will naturally have a higher count of particular features simply because they are longer. To enable comparisons of the same feature across texts of different lengths, in MD studies, counts are normed to a fixed rate:

$$\text{Normed count} = (\text{Count of feature in the text} / \text{Length of the text in words}) \times 1,000$$

The relative rate of incidence of a feature in a long text may be less than in a short text. For instance, a two-word-long post containing a single verb would have a relative frequency of verbs per 1,000 words of 500 $((1/2) * 1000)$, but a ten-word-long post having two verbs would have a relative frequency of verbs equal to 200 $((2/10) * 1000)$. Therefore, text 1 has relatively more verbs than texts 2. But does it make sense to say that in relative terms text 1 has 2.5 times as many verbs as text 2, or is it sufficient to say that text 1 has more verbs per 1,000 words than text 2?

According to Biber (1993: 252), 'with regard to the issue of text length, [...] text samples should be long enough to reliably represent the distributions of linguistic features'. Very short texts fail to reliably represent the distribution of the features because the features that are present will be overrepresented, while the majority of the features will be underrepresented as missing. To avoid these problems, researchers can impose a minimum word count for texts; for instance, Biber and Egbert (2018: 13) set a minimum length of 75 words for texts to be included in their corpus of web documents. However, in social media, short texts are the norm; therefore, texts cannot be excluded based on their short length (in our corpus, the average post is 33 tokens long).

As short text length is a register feature in social media, the resulting feature counts will be skewed. To avoid this problem, one option is to convert the actual counts from an interval to an ordinal scale, thereby reducing the drastic differences among the texts with respect to the frequency counts. Table 46.3 illustrates the use of an ordinal scale to rank the texts in terms of the incidence of a feature. As can be seen, the ordinal scale reduced the distance among the texts to one rank, whereas the interval scale increased the distance up to tenfold.

In their study of Donald Trump's tweets, Clarke and Grieve (2019) used a binary scale to represent the data, whereby the features were coded as either present or absent. If a

Table 46.3 Interval and ordinal scale example

<i>Text</i>	<i>Text length</i>	<i>Feature count</i>	<i>Relative frequency per 1,000 words (interval scale)</i>	<i>Rank (ordinal scale)</i>
1	10	3	300	3
2	20	3	150	2
3	100	3	30	1

binary scale were applied to the data in Table 46.3, all texts would be coded as “present” for this feature, which would erase the distinction among the texts. No consensus exists in the literature as to whether an interval or ordinal or binary scale is more appropriate for representing the feature counts in short text corpora. In this study, we used an ordinal scale, which enabled us to preserve the distinctions among the texts and reduce the effect of text size on the frequency counts. Since ordinal variables were used, polychoric correlations were computed for the factor analysis.

To date, social media has not received much attention in corpus linguistics in general or in MD studies in particular. One MD study is Berber Sardinha (2014), which reported an additive MD analysis that compared five online registers (email messages, webpages, blog posts and two social media varieties, namely Facebook posts and tweets) to the dimensions of register variation identified by Biber (1988). The results showed social media to be involved (dim. 1), non-narrative (dim. 2), situation-dependent (dim. 3), non-persuasive (dim. 4) and not marked for abstraction (dim. 5).

In a later study, Berber Sardinha (2018) carried out a full MD analysis of the same online registers, which identified three dimensions of variation: involved, interactive discourse versus informational focus (dim. 1); expression of stance: interactional evidentiality (dim. 2); and stance: interactional affect (dim. 3). The two social media registers scored similarly on dimension 1 as very involved and interactive, but were distinguished on the stance dimensions: *Twitter* was more marked for evidentiality, whereas Facebook was more marked for affect. Emails scored more similarly to social media than the web registers on all dimensions. In summary, the study showed that social media outlets are interactive registers in which users routinely position themselves.

Clarke and Grieve (2019) examined a corpus of tweets by former US president Donald Trump, posted from 2004 to 2009 and identified five major dimensions: tweet length (dim. 1), conversational style (dim. 2), campaigning style (dim. 3), engaged style (dim. 4) and advisory style (dim. 5). The tweets were marked for the dimensions at different rates over time. For example, the conversational style was more prominent from the end of 2012 to mid-2013, whereas the campaigning style was dominant between 2016 and 2017.

4 Major dimensions of variation for social media

In the case study reported here, two dimensions were identified based on the factor analysis:

1. Formal, informational, argumentative discourse;
2. Informal, interactive, speaker-oriented discourse.

The first dimension is packed with features that enable a formal type of posting that generally conveys planned, edited, highly informational and argumentative content. With a total of 31 different linguistic features (Table 46.4), it includes specialised nouns (cognition, process, abstract, group), adverbial features (likelihood, certainty, attitudinal), conjunctions (coordinating as clausal connector, causative, subordinating), discourse features (downtoners, hedges, amplifiers), adjectives (topical, relational), modals (necessity/obligation, prediction/volition, possibility, permission, ability), clause types (*to* complement clauses controlled by nouns, *wh*-relative clauses) and pronouns (third person, *it*, demonstrative, nominal/indefinite). In addition, posts marked on this dimension are longer and display more lexical variety type-token ratio (TTR). The three

Table 46.4 Factor pattern for dimension 1

<i>Feature</i>	<i>Loading</i>
Type-token ratio	.790
Word count	.787
Likelihood adverbs	.538
Adverb within auxiliary	.522
Linking adverbials	.505
Coordinating conjunction as clausal connector	.489
Hedges	.480
Causative subordinating conjunction	.470
Certainty adverbials	.455
All wh-relative clauses	.446
Other subordinating conjunction	.425
Downtoner	.418
Third-person pronoun (except <i>it</i>)	.417
Adverb (excluding other types)	.392
Cognition nouns	.388
Attitudinal adverbs	.386
Amplifiers	.375
Modals of necessity or obligation	.374
Process nouns	.367
Pronoun <i>it</i>	.366
Topical adjectives	.364
Demonstrative pronouns	.351
Nominal / indefinite pronoun	.347
All passives	.346
Emphatics	.345
Modals of prediction or volition	.342
Modals of possibility, permission, and ability	.327
<i>to</i> complement clause controlled by stance nouns	.326
Abstract nouns	.318
Group/institution nouns	.300
Relational adjectives	.300

platforms are statistically identical with respect to this dimension, as the platform explains a mere 0.2 per cent of the variation (Figure 46.1). At the same time, the user groups explain 13 per cent of the variation (Figure 46.2), ranked as follows: politics, corporations and celebrities. The actual users, however, capture twice as much variation (26 per cent); Figure 46.3 shows the top ten highest- and lowest-scoring users.

Politicians use dimension 1 features to promote their political agenda, as in the following example, which shows the dense use of features for this dimension in a long Facebook post (128 words, 23.3 TTR) by a US Senator (in the examples, parentheses were added to indicate the features loading on the factor):

Now (adverb), more (adverb) than (subordinating conjunction) ever (adverb), it (it) is abundantly (adverb) clear that high quality (abstract noun), affordable child care (abstract noun) is absolutely (amplifier) essential for families, as well (adverb) as for employers and the economy (abstract noun) as a whole (abstract noun). Fortunately

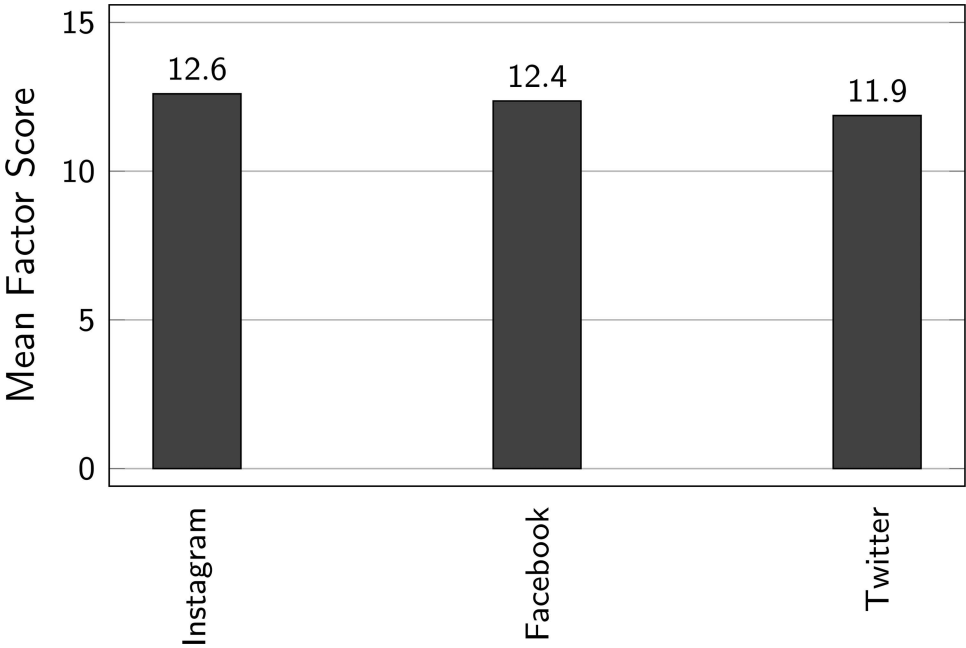


Figure 46.1 Means for platform dim. 1 ($R^2 = .2\%$; $F = 36.4$; $p < .0001$)

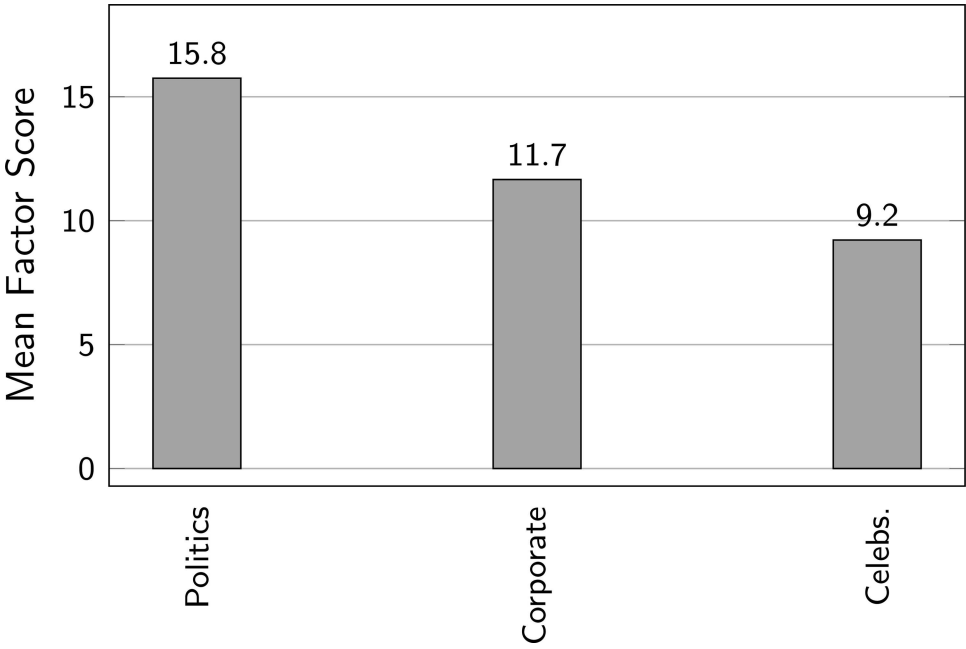


Figure 46.2 Means for user group dim. 1 ($R^2 = 13.04\%$; $F = 3198.53$; $p < .0001$)

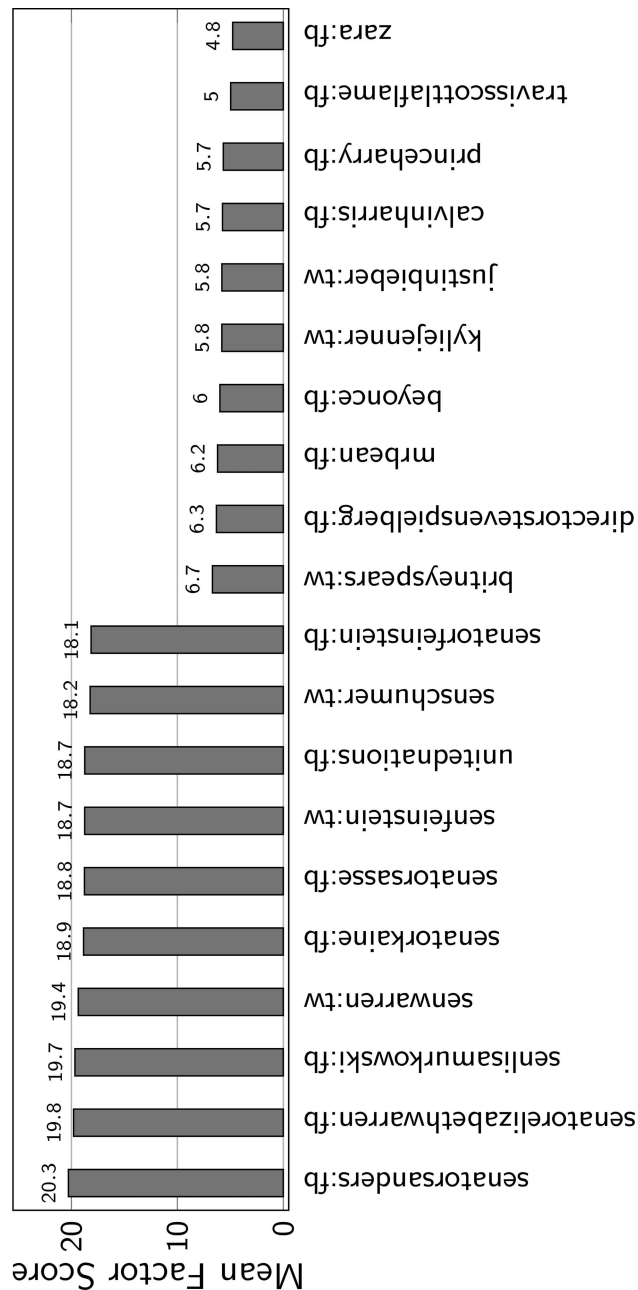


Figure 46.3 Means for user dim. 1 ($R^2=26.4\%$; $F=70.03$; $p<.0001$)

(linking adverbial), state and federal funding (abstract noun) is available to help families with the cost of child care (abstract noun) through Vermont's Child Care (abstract noun) Financial Assistance (abstract noun) Program. While (subordinating conjunction) we still (adverb) have a long way (abstract noun) to go to provide universal child care (abstract noun) to all and ensure our child care (abstract noun) providers are paid (passive voice) a fair wage for their (third-person pronoun) invaluable work (process noun), this program is an important tool for so many struggling families, including those (demonstrative pronoun) who (wh-relative clause) have experienced financial hardship (abstract noun) [...]

The following example, a United Nations post on Facebook, shows how abstract and process nouns can occur within hashtags, which the tagger was able to detect because the hashtags had been segmented:

Not one country has achieved gender (abstract noun) equality (abstract noun) yet (adverb). At the current rate (abstract noun) of progress, it will (prediction modal) take 99.5 years to close the global gender (abstract noun) gap. But (coordinating conjunction as clausal connector), it (*it*) doesn't have to (necessity modal) be this way (abstract noun). Join #GenerationEquality <generation>(process noun) <equality>(abstract noun) by supporting women and girls, defying gender (abstract noun) roles (abstract noun), fighting gender (abstract noun)-based violence, and (coordinating conjunction as clausal connector) standing up for equality (abstract noun). <url>

The next example (a tweet by a US Senator) shows how shorter posts can make intensive use of the dimension features. Note how linguistic features were detected by the tagger after abbreviations were expanded:

ICYMI<In><case><you><miss><it>(pronoun *it*) on @<handle> with @<user>: @<handle> is a monopoly and they (third-person pronoun) are abusing their (third-person pronoun) power (abstract noun) to (*to* complement clause controlled by noun) unfairly (adverb) target and censor conservatives. They (third-person pronoun) must (modal of necessity or obligation) be stopped (passive voice).

In contrast to politicians and political groups, corporations rely on this dimension for customer relations and brand management. The following example shows how the dimension is used to provide technical support:

@<handle> Hmm. An account can (possibility modal) be recovered (passive voice) for a limited time (abstract noun): <https://<url>>. You can (possibility modal) definitely (certainty adverbial; split auxiliary) create a new one here (adverb): <https://<url>>

Unlike the previous examples, posts with low scores on this dimension make scant use of these features, as shown in the following tweet by a singer:

I love u <you> guys. This (demonstrative pronoun) is funny as hell

Table 46.5 Factor pattern for dimension 2

<i>Feature</i>	<i>Loading</i>
<i>to</i> complement clauses controlled by adjectives	.723
<i>that</i> complement clauses controlled by adjective	.578
Adjectives in predicative position	.537
<i>that</i> deletion	.535
<i>that</i> complement clauses controlled by verb	.518
Verb (not including auxiliary verbs)	.515
First-person pronoun / possessive	.404
Contraction	.399
<i>to</i> complement clauses controlled by verbs	.360
Conditional subordinating conjunction	.338

The set of features loading on dimension 2 indicate informality, speaker orientation, engagement, and interaction (Table 46.5). It comprises many verb-based constructions, such as a high number of verbs, *that* complement clauses controlled by either adjectives or verbs (which are often used as stance devices), contractions, and *that* deletions. It also includes adjectives in predicative position, first-person pronouns, and subordinating conjunctions, which enable elaboration in clausal structures and a focus on the speaker.

As with the first dimension, the effect of the platform is negligible, as this accounts for less than 1 per cent of the variation (Figure 46.4). Similarly, the majority of the variation is accounted for by user groups and by the individual users, albeit to a lesser extent (Figures 46.5 and 46.6).

The next example (a tweet by a US congressman) shows how politicians make use of this dimension to rally support for a cause:

RT<retweet> if (conditional subordinating conjunction) you agree (verb) that (*that* complement clause controlled by verb) everyone risking their lives on the frontlines of this crisis deserves (verb) hazard pay. 🧑‍🚒 <man><raising><hand> https:<url>

The following is an example of a corporate tweet from a US car maker intended to provide customer support, which makes intensive use of the features loading on the dimension:

@<handle> Sorry to hear (verb) that, <name>. Please let (verb) your friend, Jo, know (verb) that (*that* complement clause controlled by verb) if (conditional subordinating conjunction) they ever have (verb) vehicle concerns, we (first-person pronoun) are (verb) only an email away. He would only need (verb) to (*to* complement clause controlled by verb) send (verb) an email to socialmedia@gm.com. We (first-person pronoun) are (verb) always happy (adjective in predicative position) to (*to* complement clause controlled by adjective) provide (verb) support.

The dimension features are often used by corporations to promote products and services, as in this tweet from an American fast-food chain:

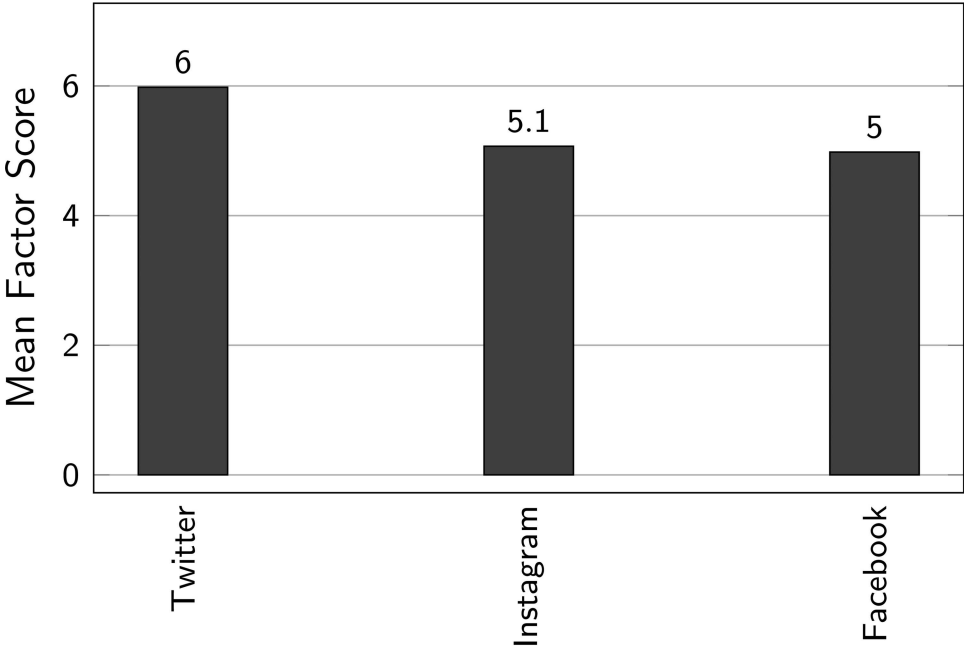


Figure 46.4 Means for platform dim. 2 ($R^2 = .98\%$; $F = 211.84$; $p < .0001$)

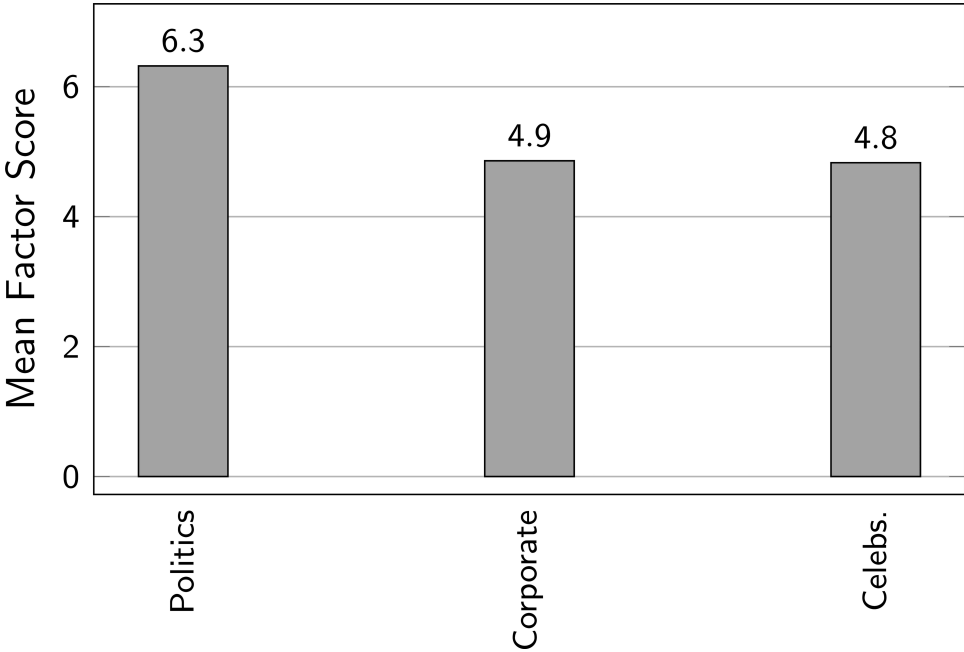


Figure 46.5 Means for user group dim. 2 ($R^2 = 2.35\%$; $F = 514.873$; $p < .0001$)

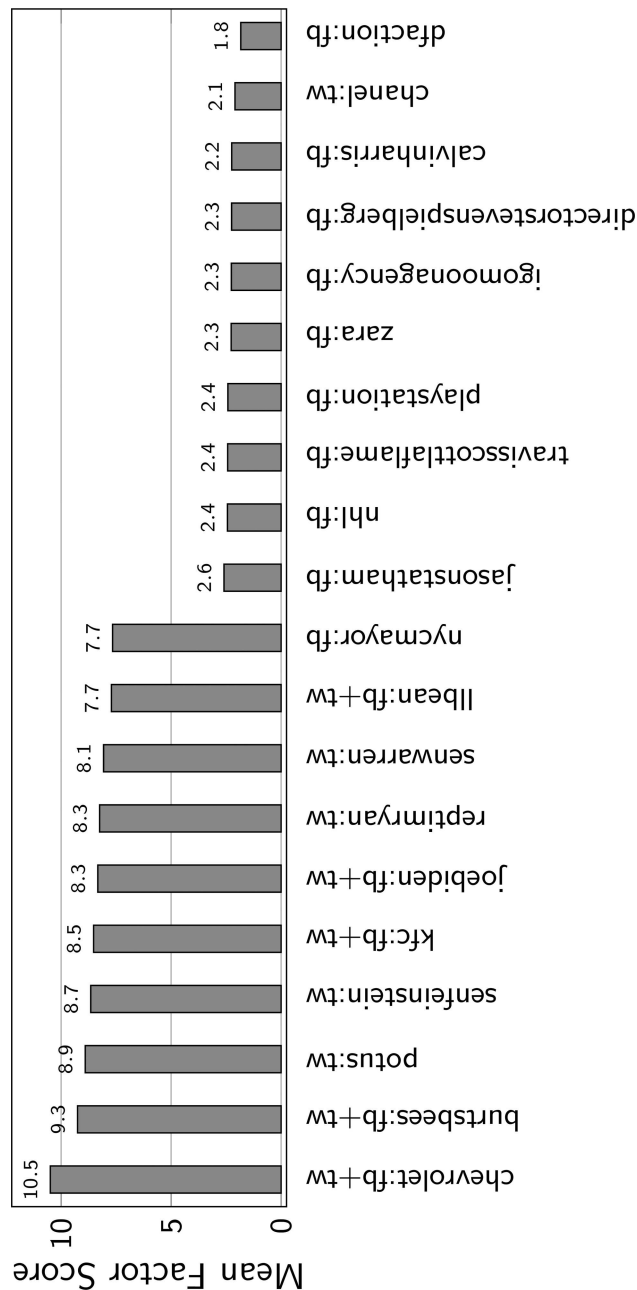


Figure 46.6 Means for user dim. 2 ($R^2 = 14.36\%$; $F = 32.8$; $p < .0001$)

They look (verb) similar (adjective in predicative position), but you'll (contraction) be surprised to learn (verb) that (*that* complement clause controlled by verb) one is (verb) full (adjective in predicative position) of helium, and one is full (adjective in predicative position) of 100 per cent white-meat chicken. #ColonelsRoadTrip <Coronel's><road><trip>#MacysParade <Macy's><parade>

Promotional posts can exploit the dimension features to create the illusion that companies are speaking with users, as in the following tweet from the same company:

Does anybody know (verb) how to (*to* complement clause controlled by verb) take (verb) care of a gold goldfish? I (first-person pronoun) have (verb) no clue what I (first-person pronoun) 'm (contraction) doing (verb). All I (first-person pronoun) 've (contraction) learned (verb) so far is (verb) that (*that* complement clause controlled by verb) it doesn't (contraction) eat (verb) fried chicken.

Celebrities can also take advantage of emulating a conversation, for example, to announce events that promote their careers, as in this tweet by an American TV personality:


I (first-person pronoun) 'm (contraction) so ready (adjective in predicative position) to (*to* complement clause controlled by adjective) see (verb) my fellow entrepreneurs at @<handle>#QBConnect <QB><connect> in San Jose next week! I (first-person pronoun) better C<see>(verb) U<you> there 😘 <face><throwing><a><kiss> http://<url>

As mentioned, the individual users play an important role in determining the variation across the posts: Some users choose a style more closely associated with dimension 1, while others choose a style more in line with dimension 2. However, some users choose a style consistent with both dimensions at the same time. To identify the individual preferences across the two dimensions, Pearson product correlations were computed for the dimension scores. The results showed that a minority of users (63 users, 21.5 per cent of the 293) scored a positive moderate correlation ($r \geq 0.5$) or higher between dimensions 1 and 2. Of these, only four had a high positive coefficient ($r \geq 0.7$). These users combine the formal, planned and information-based characteristics of dimension 1 with the informal, speaker-oriented, dialog-emulating characteristics of dimension 2. An example appears next, a post by an American performer which includes several dimension 1 and dimension 2 features:

=< <folded><hands><medium><skin><tone>(dim. 1: abstract noun) THANK (dim. 2: verb) YOU @<handle> for BRINGING LIFE (dim. 1: abstract noun) BACK TO OUR (dim. 2: first-person pronoun) HOME (dim. 1: group noun)!!! THIS (dim. 1: demonstrative pronoun) WAS (dim. 2: verb) AN HONOR (dim. 1: abstract noun) FOR ME (dim. 2: first-person pronoun) AND (dim. 1: coordinating conjunction as clausal connector) IM (dim. 2: first-person pronoun) (dim. 2: contraction) THANKFUL (dim. 2: adjective in predicative position) TO (dim. 2: *to* complement clause controlled by adjective) BE (verb) FROM VIRGINIA!!! ❤️; <heart> TIL (dim. 1: adverb) the wheels fall (verb) off

However, this style of post is rare, as the vast majority of users will write posts that are consistent with a single dimension at a time. In fact, the correlations show that, overall, the messages posted by an individual user will comprise a mixture of both dimension 1 marked posts and dimension 2 marked posts. For instance, a US apparel company (correlation $r = 0.27$) will often select a style that is consistent with dimension 1 for its posts, as in the first example shown, whereas at other times, it will prefer a style consistent with dimension 2, as in the second example:

The Sk8-Hi MTE infuses (dim. 2: verb) style (dim. 1: abstract noun) and function (dim. 1: process noun) to (dim. 1: *to* complement clause controlled by noun) get (dim. 2: verb) you through the winter ahead (dim. 1: adverb). Shop (dim. 2: verb) the All Weather MTE collection (dim. 1: abstract noun) at <url> or your nearest store (dim. 1: group noun).

We (dim. 2: first-person pronoun) had (dim. 2: verb) the best time (dim. 1: abstract noun) at the Stoke-O-Rama yesterday (dim. 1: adverb) down in Huntington Beach at the #VANSUSOPEN<Vans><US><Open>. Be (dim. 2: verb) sure (dim. 2: adjective in predicative position) to (dim. 2: *to* complement clause controlled by adjective) follow (dim. 2: verb) @<handle> to see (dim. 2: verb) more (dim. 1: adverb)!  <camera>: @<handle>

5 Summary and future prospects

Social media provides a vast space for human interaction, where millions of posts are exchanged every day by billions of individuals and private and government institutions. Society is increasingly dependent on social media for information, entertainment, governance and work. This is a quickly changing mode of communication, shaped by human and societal demands as well as technological developments.

In this chapter, we showed how corpus linguistics can be used to model the variation in the verbal language encountered in the social networks. The results showed that despite the enormous opportunity for variation in social media language that the abundance of texts and users afford, the social media posts in the corpus follow two basic styles: a formal, information-based, often argumentative style and an informal, interactive, engaging, speaker-oriented style. Each one of these dimensions represents a continuum of variation: Posts can be either more or less formal, information-based and argumentative while at the same time be more or less informal and person-oriented. The results showed that although few users will post messages highly marked for both dimensions at the same time, users do rely on both dimensions.

These findings paint a complex picture of social media texts in which the variation seems to be driven primarily by the individual users, rather than by the platform or the types of user. Two major conclusions can be drawn. First, it seems possible to generalise from one platform to the others (at least across those examined here). Second, it seems prudent to acknowledge the importance of the variation across individual users in the corpus. For corpus-based studies, an implication is that corpora of social media should be designed around samples of individual users and the texts should be identified by the account where the post originated. This will ensure that researchers are able to track down the source of the posts and determine the extent of the variation explained by the

users. Beyond users, future studies should be aware of confounding variables, like topic (Friginal *et al.* 2018) and time period (Clarke and Grieve 2019), which are often disregarded but can have a significant effect on the results.

The results reported here differ from those reported by Berber Sardinha (2014, 2018) to the extent that the current study found both a “literate” and a “persuasive” component (fused into dim. 1). The difference may be attributed to sampling, as the current study sampled heavily from “institutional” accounts, unlike Berber Sardinha (2014). The decision to sample from institutional accounts reflects the current social media environment, as social media has become an official communication channel for organised sectors of society. At the same time, our results support Clarke and Grieve’s (2019) dimension 2, which comprises literate and interactive poles.

Social media researchers should strive to preserve the multimodal content found in social media, such as pictures (including memes), video and sound, as these are part and parcel of the social media environment and are likely to become more pivotal as multimedia technology improves. However, few tools for automatically annotating auditory and visual features exist today (see for instance Google Cloud Vision), and the features that are annotated by such tools may be restricted; as a result, a great deal of manual work may be required for annotating a whole corpus of social media.

Social media represents an increasingly influential form of communication. Corpus linguistics can offer key insights into how social media language is patterned, develops, varies and changes. Yet the medium remains under-researched in our field due to the challenges involved in building, preparing and analysing social media corpora, especially from a multimodal perspective. More user-friendly methods and tools need to be developed so that researching social media becomes more appealing and accessible to more corpus linguists.

Acknowledgement

I want to thank CNPq (Brasília, DF, Brazil; Processo # 306994/2017-8 and Processo # 407788/2018-2) and PiPEq (PUCSP) for funding this research. In addition, I want to thank Joe Collentine and Rob van der Goot for their invaluable support with the lexical normalisation tools used in this project.

Further reading

- Baker, P. and McEnery, T. (2015) ‘Who Benefits When Discourse Gets Democratised? Analysing a Twitter Corpus around the British Benefits Street Debate’, in P. Baker and T. McEnery (eds) *Corpora and Discourse Studies: Integrating Discourse and Corpora*, Basingstoke: Palgrave Macmillan, pp. 244–65. (The authors analysed a corpus of tweets referring to the British TV show *Benefits Street* and to a televised debate about the programme. The show centred on people receiving government support [“benefits”] who lived in a poor area of Birmingham. The analysis detected some of the major discourses in the tweets, including “the idle poor”, which framed people in need as idle and undeserving. Overall, the study shows that social media corpora are valuable sources of data for corpus-based discourse studies.)
- Clarke, I. and Grieve, J. (2017) ‘Dimensions of Abusive Language on Twitter’, *Proceedings of the First Workshop on Abusive Language Online*, Vancouver, Canada, July 30 - August 4, 2017, pp. 1–10. (This paper presents an MD study of a corpus of 1,486 tweets coded for hate speech, such as racial, religious and sexist slurs. Through the MD analysis, the study shows that hate

speech in social media is patterned for grammar. In general, tweets displaying sexism are more interactive and attitudinal than tweets displaying racism.)

- Rüdiger, S. and Dayter, D. (eds) (2020) *Corpus Approaches to Social Media* (Studies in Corpus Linguistics, Vol. 98), Amsterdam: John Benjamins. (This edited collection comprises papers dealing with several important aspects of social media language, both verbal and visual. The volume includes case studies of different platforms such as Reddit, Twitter, WhatsApp and Facebook. The chapter by Christiansen, Dance and Wild tackles the challenges involved in analysing images, proposing the use of Google Artificial Intelligence tools to carry out visual constituent analysis.)

References

- Berber Sardinha, T. (2014) 'Comparing Internet and Pre-Internet Registers', in T. Berber Sardinha and M. Veirano Pinto (eds) *Multi-Dimensional Analysis, 25 years on: A Tribute to Douglas Biber*, Amsterdam/Philadelphia, PA: John Benjamins, pp. 81–107.
- Berber Sardinha, T. (2018) 'Dimensions of Variation across Internet Registers', *International Journal of Corpus Linguistics* 23(2): 125–57.
- Berber Sardinha, T. (2021) 'Discourse of Academia from a Multi-Dimensional Perspective', in E. Friginal and J. Hardy (eds) *The Routledge Handbook of Corpus Approaches to Discourse Analysis*, London: Routledge, pp. 298–318.
- Berber Sardinha, T. and Veirano Pinto, M. (eds) (2014) *Multi-Dimensional Analysis, 25 years on: A Tribute to Douglas Biber*, Amsterdam/Philadelphia, PA: John Benjamins.
- Berber Sardinha, T. and Veirano Pinto, M. (eds) (2019) *Multi-Dimensional Analysis: Research Methods and Current Issues*, London: Bloomsbury Academic.
- Biber, D. (1988) *Variation across Speech and Writing*, Cambridge: Cambridge University Press.
- Biber, D. (1993) 'Representativeness in Corpus Design', *Literary and Linguistic Computing* 8(4): 243–57.
- Biber, D. and Egbert, J. (2018) *Register Variation Online*, Cambridge: Cambridge University Press.
- Clarke, I. and Grieve, J. (2019) 'Stylistic Variation on the Donald Trump Twitter Account: A Linguistic Analysis of Tweets Posted between 2009 and 2018', *PLOS ONE* 14(9): e0222062. 10.1371/journal.pone.0222062.
- Friginal, E., Waugh, O. and Titak, A. (2018) 'Linguistic Variation in Facebook and Twitter Posts', in E. Friginal and J. A. Hardy (eds) *Studies in Corpus-Based Sociolinguistics*, London: Routledge, pp. 342–62.
- Gray, B. (2019) 'Tagging and Counting Linguistic Features for Multi-Dimensional Analysis', in T. Berber Sardinha and M. Veirano Pinto (eds) *Multi-Dimensional Analysis: Research Methods and Current Issues*, London / New York: Bloomsbury / Continuum, pp. 43–66.
- Miller, D., Costa, E., Haynes, N., McDonald, T., Nicolescu, R., Sinanan, J., Spyer, J., Venkatraman, S. and Wang, X. (2016) *How the World Changed Social Media*, London: UCL Press.
- Weller, K., Bruns, A., Burgess, J., Mahrt, M. and Puschmann, C. (eds) (2014) *Twitter and Society*, New York: Peter Lang.