A Multi-Dimensional Analysis of English tweets

Isobelle Clarke
Lancaster University, UK

Abstract
This paper applies Multi-Dimensional Analysis (MDA) to a corpus of English tweets to uncover the most common patterns of linguistic variation. MDA is a commonly applied method in corpus linguistics for the analysis of functional and/or stylistic variation in a particular language variety. Notably, MDA is an approach aimed at identifying and interpreting the frequent patterns of co-occurring linguistic features across a corpus, such as a corpus of spoken and written English registers (Biber, 1988). Traditionally, MDA is based on a factor analysis of the relative frequencies of numerous grammatical features measured across numerous texts drawn from that variety of language to identify a series of underlying dimensions of linguistic variation. Despite its popularity and utility, traditional MDA has an important limitation – it can only be used to analyse texts that are long enough to allow for the relative frequencies of many grammatical forms to be estimated accurately. If the texts under analysis are too short, then few forms can be expected to occur sufficiently frequently for their relative frequency to be accurately estimated. Tweets are characteristically short texts, meaning that traditional MDA cannot be used in the present research. To overcome this problem, this paper introduces a short-text version of MDA and applies it to a corpus of English tweets. Specifically, rather than measure the relative frequencies of forms in each tweet, the approach analyses their occurrence. This binary dataset is then aggregated using Multiple Correspondence Analysis (MCA), which is used much like factor analysis in traditional MDA – to return a series of dimensions that represent the most common patterns of linguistic variation in the dataset. After controlling for text length in the first dimension, four subsequent dimensions are interpreted. The results suggest that there is a great deal of linguistic variation on Twitter. Notably, the results show that Twitter is commonly used for self-commodification, as people manage their identities, engaging in practices of self-branding through stance-taking, self-reporting, promotion and persuasion, as well as broadcasting their message beyond their followership, distributing news and expressing opposition, and this often occurs in order to attract attention. Additionally, the results show that interaction is common, suggesting that Twitter is also used for social and interpersonal gain.

Corresponding author:
Isobelle Clarke, Linguistics and English Language, Lancaster University, Lancaster LA1 4YW, UK.
Email: i.clarke@lancs.ac.uk
Keywords
Corpus linguistics, functional linguistic variation, multiple correspondence analysis, multi-dimensional analysis, stylistic variation, Twitter

1. Introduction

Twitter is a social networking microblogging service, which allows users to post messages called ‘tweets’, to a network of associates, deemed ‘followers’. Twitter is based on ‘following’, which is like subscribing to someone’s updates. Following is non-reciprocal; a user does not have to follow someone back if they follow them. This leads to complex follower networks being formed of unidirectional and bidirectional connections with a variety of individuals, organisations (governmental and non-governmental) and media outlets (Weller et al., 2014: xxx).

Following and follower networks have been described as influencing the communicative practices on Twitter. Page (2012: 183) notes that the non-reciprocity of following has enabled some individuals to have millions of followers, especially celebrities, prominent figures and businesses. As a result, the size of a follower list is often taken as a sign of status and influence. Twitter’s CEO Jack Dorsey (2019) also attributed the importance assigned to follower size to the architecture of the platform because ‘following’ and ‘followers’ are presented in big and bold, which he suggests incentivises gaining followers. Regardless of why, because follower size is viewed as important, Page (2012: 183) and Dorsey (2019) note that there is a need to gain more followers. It has been suggested that this need has influenced practices of promoting, self-branding, micro-celebrity (Page, 2012: 183) and more oppositional, provocative and uncivil content (Anderson, 2019). Thus, the technological affordance of following and the meaning ascribed to follower size afforded by the design of the platform has influenced social and communicative practices.

Twitter is used by numerous people for a variety of communicative purposes in a variety of situations and therefore exhibits a whole host of linguistic variation. However, the patterns of linguistic variation across Twitter have not yet been examined. One common method for identifying patterns of functional and/or stylistic variation in a particular variety of language is Multi-Dimensional Analysis (MDA) (Biber, 1988). MDA is based on the notion of linguistic co-occurrence – frequent patterns of co-occurring linguistic features tend to reveal at least one shared underlying communicative function or style. Given this notion, MDA aims to identify the most common patterns of linguistic co-occurrence by subjecting the relative frequencies of numerous lexico-grammatical features across the texts of a corpus to a multivariate statistical technique, namely factor analysis. Factor analysis is a dimension reduction method, which is used in MDA to reduce a large number of measured variables (the relative frequencies of linguistic features in texts) into a smaller number of underlying or latent constructs (dimensions consisting of sets of linguistic features that co-occur often in the texts of a corpus). Based on the notion of linguistic co-occurrence, the dimensions of sets of co-occurring linguistic features are then interpreted for their underlying communicative functions or styles. When investigating functional linguistic variation from a register perspective, linguistic co-occurrence
patterns are interpreted for their underlying communicative function based on the assumption that linguistic features are functional and are often used to achieve particular communicative purposes in the situational contexts of the texts (Biber and Conrad, 2009: 2). When investigating stylistic variation, patterns of co-occurring linguistic features are often interpreted as reflecting aesthetic preferences associated with particular authors or historical periods (Biber and Conrad, 2009: 2). MDA has been applied to numerous languages, language varieties and discourse domains, such as spoken and written English registers (Biber, 1988) and the texts from a particular person’s social media account (Clarke and Grieve, 2019).

Notwithstanding the utility of MDA, it is limited to analysing texts that are long enough to allow for the relative frequencies of grammatical forms to be estimated accurately. If the texts under analysis are too short, then very few forms can be expected to occur sufficiently frequently for their relative frequency to be accurately estimated. The relative frequencies of features are measured in each text in MDA to place texts of different lengths on an equal footing. For example, imagine that Text A has 5000 words, Text B has 3000 words and Text C has 1000 words. All three texts have 50 nouns. Whilst they may all have the same raw frequency of nouns, nouns occur considerably more frequently in Text C (50 times per 1000 words), than Text B (5 times per 1000 words) and Text A (0.01 times per 1000 words). Thus, it is important to measure the relative frequencies of features so that text length does not confound the analysis. Tweets are characteristically short texts meaning that the relative frequencies of features in individual tweets cannot be accurately computed. For example, consider the average tweet length in the corpus used in the present study, which is 17 words. It is virtually impossible to measure the relative frequency of any word or grammatical form accurately in texts of this size – if a form occurs twice, we cannot assume that it is generally used twice every 17 words (118 times per 1000 words), because even the most frequent forms in a large corpus do not come anywhere near this rate. Moreover, if a form does not occur in a 17-word sample, we cannot assume that it generally never occurs at all.

The issue of relative frequencies in short texts is not new. It is an important methodological issue in stylometry (Stamatatos, 2009: 538), such as in authorship analysis (Grieve et al., 2019). Different solutions and methods have been offered, including looking at the presence or absence of linguistic features. This approach has been fruitful because the texts under analysis are so short that most features occur only once.

Despite this limitation, MDA has been used to analyse short texts, such as tweets (e.g. Berber-Sardinha, 2014, 2018; Titak and Roberson, 2013). These studies, however, have combined the short texts to form longer text chunks that are ultimately more suitable for frequency-based analyses. This approach is valid if texts are combined in a principled manner. However, the manner of concatenation in these studies is not always made clear and often seems that they are combined just to make up text samples of the required length for frequency-based analyses.

Moreover, concatenation tends to limit the kinds of research questions to comparing tweets generically or groups of tweets to other registers and varieties, as opposed to investigating variation between individual tweets. For example, previous research has explored how Twitter as a variety of language compares to the major patterns of linguistic variation of other online registers (Berber-Sardinha, 2018; Titak and Roberson, 2013) and
pre-internet registers (Berber-Sardinha, 2014). These studies have often produced differing results, potentially because of different datasets, but it may also be a consequence of concatenation.

For example, Titak and Roberson (2013) used MDA to identify the major patterns of linguistic variation across a corpus of online registers, including groups of aggregated Facebook and Twitter posts, emails, blogs and reader comments. They then compared the overall association of the different registers to the dimensions. In this study, Facebook and Twitter posts were concatenated and were revealed to be associated with a descriptive informational production in the first dimension and were most similar to online newspaper articles. Additionally, Facebook and Twitter posts also had a present tense orientation and a moderate level of involved and interactive discourse, but they were not associated with the complex statement of opinion function. Overall, they found that Facebook posts and tweets differ considerably from ordinary blogs, which are more narrative and more elaborated.

Berber-Sardinha (2018) also applied MDA to a corpus of online registers. This corpus similarly included Facebook and Twitter posts, emails and blogs; however, it varied in its inclusion of webpages as opposed to reader comments, and it distinguished concatenated Facebook posts from aggregated tweets. Three major dimensions of linguistic variation were identified and the overall association of the different online registers to the dimensions were compared. Unlike Titak and Roberson (2013), on the first dimension, tweets were found to be more associated with an involved and interactive function than an informational function. Other discrepancies appear with the next two dimensions. Berber-Sardinha (2018: 145) found that tweets were moderately associated with the second dimension, which was interpreted as reflecting the ‘Expression of stance: Interactional evidentiality’, where the texts most associated tended to be expressing their attitude towards knowledge. Additionally, Berber-Sardinha (2018: 146) found that Facebook posts were most strongly associated with the final dimension, which was interpreted as reflecting expressions of personal attitudes, emotions and feelings. This arguably counters Titak and Roberson (2013), who found that Facebook and Twitter posts were not associated with the complex statement of opinion function. Whilst the differences in the dimensions are a result of different datasets, the discrepancies between the associations of tweets and Facebook posts to the similar communicative functions represented in the dimensions across both studies are more likely to be a result of concatenation.

To address this, Friginal et al. (2018) distinguished Facebook posts from tweets. They also separated each group further according to topics, including politics, business, entertainment, personal, sports and weather. They then computed how associated each topic-specific platform-specific sub-corpus was to the dimensions of online registers revealed in Titak and Roberson (2013). They found that each group of topic-specific tweets varied in relation to the dimensions, thus demonstrating that Twitter is not a homogeneous entity and thus suggesting that even the concatenation of random tweets, such as in Berber-Sardinha (2018), may obfuscate variation across tweets.

This growing body of research has enabled rich descriptions of the ways in which online registers like tweets and even particular topic-focused groups of tweets compare to other online registers along the major dimensions of linguistic variation found across the online registers. Yet little is known about how individual tweets vary. Even though
Friginal et al. (2018) distinguished tweets according to topic, these topic-specific tweets were still aggregated together leaving patterns of variation between individual topic-specific tweets unknown. Moreover, these groups of tweets were compared to patterns of variation of internet registers. Whilst we know that there are a variety of different styles of tweets for a variety of different purposes, it is not yet known whether these differences are realised in actual linguistic distinctions.

As noted, this is largely because standard MDA, as used in the studies above, is not well-suited to analysing short texts. Consequently, in this study, I investigate the major patterns of linguistic variation across a corpus of tweets, by applying a short-text version of MDA to a corpus of tweets. In the rest of the paper, I describe the corpus of tweets (Section 2) and the short-text MDA approach and how it has been used to derive the major patterns of linguistic variation across them (Section 3). I then present the results of this analysis (Section 4) and discuss them with respect to Twitter’s architecture and social practices.

2. English Twitter Corpus

This paper aimed to collect a random corpus of English original tweets (i.e. not retweets). To collect the Twitter corpus, the ‘rtweet’ package (Kearney, 2018) in R was used. This package contains numerous functions for collecting and organising Twitter data. Specifically, the ‘stream_tweets()’ function was used on June 22nd, 2018, which is designed to return a small random sample of public tweets (around 1% at any particular time). The programme ran for approximately 4 hours, collecting 123,330 tweets. Each tweet is marked with various metadata, including whether it is a retweet, whether it is replying to another tweet, the number of times it has been liked and the language of the tweet (e.g. Italian, Arabic, English, etc.), among others. Non-English tweets and retweets were filtered out programmatically in R using this metadata. The resulting corpus contained 13,879 original English tweets, totalling 230,748-word tokens. The mean tweet length of this corpus of tweets is 17 words. The median tweet length is 11 words long with an interquartile range of 6-to-19-word tokens, which indicates that 50 percent of the tweets in the corpus are between this range. Overall, these values demonstrate that this is a corpus of short texts with most texts in the corpus not exceeding 20 words.

It is important to note that whilst this corpus represents a random corpus of English original tweets, it is restricted to representing English tweets from a single day in 2018. Other corpora containing tweets spanning a longer period, and even tweets from a single day in a different year or month may reveal different patterns of linguistic variation.

3. Short-Text Multi-Dimensional Analysis

Having collected a corpus of random English tweets, the next aim is to reveal the major communicative functions across the corpus. Given that this corpus comprises short texts, traditional MDA cannot be used because the relative frequencies of most grammatical forms are reliable only in texts around 1000 words long (Biber, 1993: 249). Consequently, a short-text version of MDA is applied (Clarke (2019); Clarke and Grieve, 2017). In this short text version of MDA, the relative frequencies of features are not measured. Instead,
the occurrence of features (presence/absence) are measured. The challenge with measuring the presence or absence of features is that factor analysis prefers continuous data (i.e. the relative frequencies of features), which means that factor analysis cannot be used because the presence or absence of features means that the dataset in the present study is categorical. Consequently, an alternative multivariate statistical technique is selected, called Multiple Correspondence Analysis (MCA), which is very similar to factor analysis, but is better suited for categorical data. Like factor analysis, MCA is used in MDA to return a smaller number of dimensions, comprising the major patterns of co-occurring grammatical features across the corpus of tweets.

3.1. Tagging the data and running the MCA

To conduct short-text MDA on the Twitter corpus, each tweet was tagged firstly using 0.3.2 version of Gimpel et al.’s (2011) Twitter tagger with the default tags in the Computational Natural Language Learning output format. This tagger tokenises and tags for different parts-of-speech and assigns a confidence measure to each tag between 0 and 1, where 1 indicates 100% confidence in the tag assigned to the token. Table 1 presents Example 1 and the output of this tagger on the example. Each token is placed on a new line and is assigned a tag, such as ‘@’ for mentioning – where other Twitter users are specifically mentioned in the tweet via their username.

The Gimpel et al. (2011) Twitter tagger can deal relatively well with the noisiness of Twitter spelling, as well as tag for features specific to Twitter, for example, URLs, hashtags and mentioning with a high accuracy (almost 90%). However, the overall feature set is nowhere near as detailed as the one used in standard MDA (see Biber, 1988: 211–245). Additionally, some features are conflated into one tag, such as prepositions and subordinators. Consequently, Clarke (2019) developed a Twitter-specific MDA tagger, called MDATT (Multi-Dimensional Analysis Twitter Tagger), which is based on the output of the Twitter tagger and tags each tweet for 124 linguistic features, according to the MDA feature set, including tense and aspect markers (past tense verbs, progressive aspect, perfect aspect) and specialised verb classes (e.g. public, private, stance and

| Tagger | Output |
|--------|--------|
| Raw Tweet: Example 1 | @username This is gold [thumbs-up emoji] |
| Gimpel et al. (2011) Twitter Tagger output | @username @ 0.9990 |
| | This O 0.9491 |
| | is V 0.9972 |
| | gold A 0.9007 |
| | [thumbs-up emoji] E 0.99 |
| Clarke (2019) MDATT output | @username_@_INITIAL-MENTION |
| | This_O_~PRO_~DEMONSTRATIVE-PRONOUN |
| | is_V_~VERB_~BE-MAIN-VERB |
| | gold_A_~ADJ_~PREDICATIVE-ADJECTIVE |
| | [thumbs-up emoji]_E_~EMOJI |
perception verbs), as well as other features related to Computer-Mediated Communication (CMC), such as different types of mentioning, like mentioning in the initial position of the tweet, which is used to direct the tweet to the mentioned user and interact with them, and non-initial mentioning, which can be used to refer to the Twitter user in the third person (e.g. Next up in Room 404 is @username presenting their research on puffins). Non-standard spellings, such as ‘dis’ for ‘this’ and ‘gonna’ for ‘going to’, were also included in the tagger because these are common on Twitter.

The MDATT is written in Perl. Firstly, it removes the confidence measures assigned by Gimpel et al.’s tagger and it appends the tag onto the token with an underscore. The MDATT then reassigns tags assigned by Gimpel et al.’s (2011) Twitter tagger to add more detail. For example, it separates subordinators and prepositions, which are assigned the same grammatical tag _P by the Gimpel tagger. Additionally, it assigns an additional tag with a tilde ‘∼’ and underscore ‘_’ in cases where adjectives and nouns have been tagged as a verb (e.g. a_D left_A leaning_V site_N to a_D left_A leaning_V ∼ADJ site_N).

Secondly, it assigns a basic part of speech tag to all features, which is often the same that has been assigned by the Gimpel et al. (2011) tagger, albeit the more detailed labels and fixed errors from stage 1 (e.g. a_D_∼DET left_A_∼ADJ leaning_V_∼ADJ site_N_∼NOUN).

Following this, the tagger incorporates the rule-based algorithms described in the appendix of Biber (1988: 211–245) and the particular lexicons specified in Quirk et al. (1985) and Biber et al. (1999), adapted to accommodate the Gimpel et al. (2011) Twitter tagger and common words, spellings and grammatical variations found across CMC (e.g. wanna for want to, more than 2 adverbs between HAVE and verb in perfect form). Specifically, rule-based algorithms are used to tag features by using specific words (e.g. a lexicon of different forms of the verb BE/HAVE), tags and sequences of words and tags, taking into account the various spelling variations inherent in CMC. For example, infinitives are tagged if the first word in a sequence is either ‘to’ or ‘2’, and if the second word in the sequence is tagged by the Twitter tagger as a verb (_V), and if that verb does not end in –ing (e.g. Look forward to seeing you). The tagger attaches the tag to the base tags assigned by the MDATT in the second step using another underscore and a tilde ‘∼’. Table 1 shows the output of the MDATT on Example 1. This tagger’s precision rate on this dataset is 95%.

Having tagged the data, a data matrix was computed which recorded the presence/absence of the features in each tweet, retaining only those that occurred in more than 5% of the tweets (63 features). This is because infrequent features can overly influence the results of the MCA (Le Roux and Rouanet, 2010: 39). The final feature set can be found in the Supplemental files. In addition, the number of word tokens per tweet was also added to the data matrix.

This data matrix was subjected to MCA using the package ‘FactoMineR’ (Husson et al., 2017) in R, specifying the 63 linguistic features as active variables, and the wordcount as a supplementary quantitative variable. Supplementary variables do not influence the results but can be used to assess if the patterns identified in the dimension are associated with the supplementary variable. The length of the tweet in words is defined as a supplementary variable to assess if text length has confounded the analysis. Because short-text MDA does not take the relative frequencies of features, there is the likelihood that text length could confound the analysis as it is not controlled for. By including
wordcount as a supplementary variable, it is possible to assess if text length is influencing any of the dimensions.

The MCA produced a series of dimensions, each with a positive and negative pole. Each category of a linguistic feature (e.g. presence of nouns vs. absence of nouns) and each text was assigned a coordinate (positive or negative) and a contribution for each dimension. The distance between the coordinates of tweets on each dimension indicates the dissimilarities in their linguistic composition with respect to the major pattern of variation that the dimension represents. Essentially, the shorter the distance between the tweets in the space indicates that these tweets are more similar in distribution; that is, the tweets tend to have the same categories of linguistic features. With respect to the categories of linguistic features, the coordinates reflect the nature of the association between the categories of linguistic features in terms of proximity, where linguistic features that are distributed in similar ways in the tweets will have coordinates closer to each other (Le Roux and Rouanet, 2010). Contributions show which categories of features and which tweets are the most important contributors to the dimensions. In this way, they are like factor loadings in factor analysis, as they reveal which variables are most strongly associated with the dimension. However, contributions are different to factor loadings because they do not have polarity and so the contributions need to be considered alongside the coordinates, which indicate which side of the dimension they are associated with, and with which features they co-occur. Le Roux and Rouanet (2010: 52) suggest that the categories of variables contributing above the average contribution should be interpreted, as these represent the most distinguishing patterns of variation.

Together, the contributions and coordinates of the tweets and categories of linguistic features returned by MCA reveal the range of the most important patterns of linguistic co-occurrence in the Twitter corpus, and the tweets most associated with these patterns. Like the results of the factor analysis in Biber’s MDA, the results of the MCA in short-text MDA are then interpreted. Benzécri (1992: 405) noted:

Interpreting an axis amounts to finding out what is similar, on the one hand, between all the elements figuring on the right of the origin and, on the other hand, between all that is written on the left; and expressing with conciseness and precision the contrast (or opposition) between the two extremes.

Thus, the features strongly contributing to the dimension with positive coordinates are interpreted in opposition to the features with negative coordinates. Starting with Dimension 1, this was repeated for each subsequent dimension until the dimensions were no longer readily interpretable. Other studies employing MCA have also similarly

| Dimension | 1   | 2   | 3   | 4   | 5   |
|-----------|-----|-----|-----|-----|-----|
| Eigenvalue| 0.108 | 0.048 | 0.027 | 0.027 | 0.023 |
| Modified Rates | 0.85 | 0.10 | 0.01 | 0.01 | 0.005 |
interpreted the coordinates and contributions of categories of variables in this way (e.g. Le Roux and Rouanet, 2010). Overall, five dimensions were interpreted, as these were interpretable, and based on the modified rates of the eigenvalues (Benzecri, 1992: 412), they explain 95% of the variance (see Table 2).

In the next sections, I present the interpretation of the dimensions, except for Dimension 1, which, like other studies employing short-text MDA (Clarke, 2019; Clarke and Grieve, 2017), primarily is associated with text length \( r = 0.84 \). Each section presents a table containing the features contributing the most to the respective dimension. This table is divided into features contributing strongly to the positive pole of the dimension and features contributing to the negative pole. Each section also describes these features and how they are used in the strongly associated tweets. The tweets most strongly associated with each dimension are also presented in a table. These tweets are presented verbatim and often include non-standard grammar and spelling. Notably, the patterns of linguistic variation that are identified in the dimensions are interpreted from a register and style perspective. Tweets are a language variety and Twitter is the situational context in which these patterns of linguistic co-occurrence are used and thus the linguistic features will serve particular functions within this context. For example, initial mentioning (i.e. mentioning in the initial position of the tweet) functions on Twitter to direct the tweet to the mentioned user to interact with them. However, because this study does not compare Twitter to other registers or situational contexts, but instead tweets are compared to each other, these patterns of variation can also be interpreted stylistically as reflecting styles of tweeting. In this way, I am using ‘style’ here in a broader sense of stylistics like Crystal and Davy (1969: 10–11), who understand style within the framework of general language variation, as opposed to the notion of idiolectal or evaluative style, which is more typical of literary scholars. Specifically, Crystal and Davy (1969: 3) talk about language as comprised of ‘many different “varieties” of language in use in all kinds of situations in many parts of the world’, which are ‘definably distinct from all the others’. They note that style can be analysed through language habits – those features which are restricted to particular contexts as they provide a particular function in the context, as opposed to other features (Crystal and Davy, 1969: 10). Thus, this study uses short-text MDA to find different varieties and/or styles of tweets by identifying frequent patterns of co-occurring linguistic features across a corpus of tweets and interpreting why they have been used over other features and what their functions are.

4. Results

4.1. Dimension 2: Informational versus Interactive production

Positive Dimension 2 is characterised by various noun types, such as nouns, nominalisations, proper nouns and numeral nouns, suggesting that the tweets associated with this side of the dimension have numerous referents (see Table 3). Non-initial mentioning is another noun type, which is often used by Twitter users to be reference another user in the third person via their username, as opposed to interacting with them directly (Honeycutt and Herring, 2009: 4). Additionally, several noun modifiers are associated with positive Dimension 2, enabling the concise and efficient integration of information (Biber, 1988: 104),
such as definite articles, numeral determiners and attributive adjectives, which are used for pre-nominal modification. Prepositions are also associated and are often used in complex noun phrases as post-nominal modification. A high degree of noun types and noun modifiers are associated with texts that have a high informational focus (Biber, 1988: 104).

Other features associated with positive Dimension 2 include text-based features, such as capitalisation and colons. Capitalisation can be used for emphasis, as well as to compact more information into the tweet, as they frequently signal an acronym, enabling longer strings to be expressed concisely, thereby freeing up space for additional information in a character-restricted context. Colons are often used to introduce content and information.

Positive Dimension 2 is also characterised by CMC-specific features, such as hashtags and URLs, both of which are associated with an informationally dense style. URLs allow unlimited space to extend the content of the tweet. Hashtags are often used to signal topic and broadcast the tweet to an audience beyond the author’s current followership, linking the content to feeds of tweets tagged in the same way without having to interact with people directly (Zappavigna, 2018: 132).

Finally, imperatives are also associated with positive dimension 2 and these are used to direct some form of action. Overall, these features are largely indicative of an informationally dense style, comprised of numerous specific referents and carefully

| Feature            | Coordinate | Contribution |
|--------------------|------------|--------------|
| Colon              | 1.5        | 4.927        |
| Hashtag            | 1.117      | 6.62         |
| Non-Initial_Mention| 1.043      | 3.695        |
| Numeral_Determiner | 0.988      | 3.001        |
| Numeral_Noun       | 0.856      | 2.422        |
| Nominalisation     | 0.703      | 1.977        |
| Capitalisation     | 0.663      | 3.469        |
| URL                | 0.637      | 5.813        |
| Proper_Noun        | 0.524      | 4.147        |
| Imperative         | 0.496      | 0.99         |
| Preposition        | 0.389      | 2.563        |
| Definite_Article   | 0.341      | 1.067        |
| Attributive_Adjective| 0.277    | 1.193        |
| Other_Noun         | 0.199      | 1.008        |
| First_Person_Pronoun| −0.302    | 1.177        |
| Subject_Pronoun    | −0.427     | 2.219        |
| Predicative_Adjective| −0.476  | 1.083        |
| Amplifier          | −0.513     | 0.907        |
| Analytic_Negation  | −0.531     | 1.435        |
| Contracted_Forms   | −0.585     | 1.776        |
| Interjection       | −0.638     | 2.293        |
| Auxiliary.DO       | −0.653     | 0.958        |
| Initial_Mention    | −0.688     | 5.638        |
integrated information, characteristic of texts constructed when there is time to plan and edit (Biber, 1988: 107). Additionally, there is a simultaneous broadcasting style, where content is shared to many and action is demanded from many, characteristic of procedural texts.

By contrast, negative Dimension 2 is characterised by features indicative of an interactive style (see Table 3). For example, initial mentioning – mentioning of other Twitter users through the ‘@’ symbol in the initial position of the tweet – is the most strongly contributing feature, which is used to direct the tweet to another Twitter user and interact with them. Subject pronouns, especially first-person pronouns, and predicative adjectives are associated with negative Dimension 2. First person pronouns and predicative adjectives often occur in interactive and involved texts, as they mark a personal focus and can be used to encode the author’s stance and beliefs (Biber, 1988: 105).

Negative Dimension 2 is also characterised by interjections and contracted forms, which are associated with informality and interactivity, as interjections can be used to acknowledge someone’s talk and encode a reaction, and contracted forms can be used to mirror the spoken realisation of the construction. Finally, Dimension 2 is characterised by auxiliary do and analytic negation, which often co-occur to negate a particular action or reject a previous statement, especially one that has been mentioned before, suggesting that interaction is taking place.

This opposition between an informationally dense broadcasting style with an interactive style is reflected in the tweets most strongly associated with positive and negative Dimension 2. Examples 2–4 in Table 4 are associated with positive Dimension 2. These tweets tend to be broadcasting information about current and future events and news stories, often providing an exceptional amount of detail in the limited space of a tweet through complex noun phrases and integrated structures.

Table 4. The tweets most strongly contributing to positive Dimension 2.

| Tweet                                                                 | Coord | Contrib |
|----------------------------------------------------------------------|-------|---------|
| 2 LISTEN: John Hodgson discusses “the emperor of all conjurers” RICHARD POTTER on @nhpr: https://t.co/ujPyaiWhYy (37 minutes in). Read more about this book which the @WSJ called a “provocative record of Potter’s odds-defying climb” here: https://t.co/zoAxS24sfK #magic #celebrity | 0.811 | 0.099   |
| 3 farmdoc Webinar: @jt_hubbs and @ScottIrwinUI will review @USDA’s June 29 Grain Stocks and Acreage reports and considers balance sheet and price implications for both old and new crop #corn and #soybeans. Register here: https://t.co/pqqvI299qH https://t.co/dLtJZRM5HY | 0.802 | 0.097   |
| 4 VIDEO – @odublast – “Building Leaders for Advancing Science and Technology.” High school students from across the Commonwealth spent 3 days learning about climate change, sea level rise and cybersecurity through STEM. WATCH NOW: https://t.co/YdaYT87fjw #odu @educationODU https://t.co/DfCATdItVd | 0.757 | 0.086   |
Example 2, for instance, is broadcasting information on a news story. This tweet manages to include a description of the story (John Hodgson discusses [...] where it was discussed (on @nhpr), a link to the radio show, and where to find the particular section (37 minutes in), an instruction to read more, a review of the book they are discussing from another source (which the @WSJ called [...] and another link to where interested parties can find out more, all in the space of 280 characters. Overall, Examples 2–4 are not interacting with individuals, but are focused on trying to reach the broadest audience possible, especially through hashtags.

Alternatively, Examples 6–8 in Table 5 are associated with negative Dimension 2. These tweets are considerably more interactive. The person tweeting is directly conversing with (an)other Twitter user(s) through initial mentioning. These interactive tweets tend to form part of a Twitter conversation and are replying to something that has been said in a previous tweet, as opposed to forming the initial turn/tweet to which others can reply. For instance, Example 7 is replying to a tweet sent by McFly member @DougiePoynter describing how he nearly killed all his band mates in a music video. Analytic negation and auxiliary do are used in Example 7 to acknowledge what was said in @DougiePoynter’s tweet.

Many of the tweets are written as if they were spoken by including reduced and informal forms like contractions and interjections. The interactive tweets also tend to mark a personal focus and encode personal stance through first person pronouns and predicative adjectives, such as Example 7 (I’m so glad).

Tweets on the negative side are also shorter than the informational tweets on the positive side. Positive Dimension 2 is the only dimension beyond Dimension 1 that is slightly correlated to tweet length ($r = 0.21$) indicating that the informational broadcasting tweets tend to be longer than interactive tweets. This may offer support to the interpretative labels assigned to this dimension because, on the one hand, whilst it is possible for others to respond to the tweets on the positive side of the dimension, like academic writing, there is no invitation for dialogue, which may mean that the tweets have been designed to integrate large amounts of content. For example, in academic writing, the author compacts as much content into the article

| Tweet | Coord | Contrib |
|-------|-------|---------|
| 6 @btykiwi ill spam u dont worry ! and its ok let’s just support our faves tho that’s why we’re here but yeah i feel u ‘sometimes’ 😅 | -0.623 | 0.058 |
| 7 @DougiePoynter @TomFletcher @harryjudd I’m so glad you didn’t killed them 😂 | -0.618 | 0.057 |
| 8 @TomArnold Lol haha your so stupid! Didn’t you know if anything exists Mule’R has it. And guess what that means when evidence is sealed? You’re going to fall. . . you believed him 😅 | -0.616 | 0.057 |
predicting and acknowledging questions and objections, being as concise and ef-
cient as possible. Although being a highly integrated style, it may nonetheless, in
the context of tweets, take up most of the space afforded to them by using more
words.

Interactive tweets, on the other hand, are not required to compact everything into one
tweet; rather, the message can be drawn out amongst several tweets in a conversational
thread. Despite being more fragmented and choosing longer structures like predicative
adjectives, interaction is governed by a mechanism of exchange, which generally refers to
individual participants alternating in offering relatively short bursts of information (Holler
et al., 2015: 1), suggesting that interactive turns are shorter than non-interactive turns.
Thus, the slight positive correlation may be used to support the interpretation that this
dimension opposes informationally dense tweets with interactive tweets.

Overall, Dimension 2 reflects the two main forms of communication on Twitter:
directed public conversation (one-to-one or one-to-few) and the general broadcast of
information to the entire network (one-to-many) (Yaqub et al., 2017: 614). Moreover, this
opposition mirrors the three bullet points displayed on the Twitter homepage before one
logs on, which are: (1) ‘Follow your interests’; (2) ‘Hear what people are talking about’;
and (3) ‘Join the conversation’. Specifically, (1) and (2) are related to the general
broadcast of information, whilst (3) is associated with interactive tweets.

The opposition of texts oriented towards the integration of information with texts
oriented towards interaction found here has been consistently identified as one of the most
important dimensions of linguistic variation in numerous MDA studies of various lan-
guages and language varieties, leading Biber to suggest that this could be a universal
dimension (Biber, 2014: xxxvi). Finding this pattern across individual tweets lends
support to this theory.

4.2. Dimension 3: Personal versus Other Description

The most contributing feature to positive Dimension 3 is object pronouns, which pre-
dominantly occur in the tweets associated with the dimension in the first person (e.g. me),
which is also associated with Dimension 3 (Table 6). Object pronouns are used to indicate
that some action is acting or has acted upon the author or other agents and often functions
to encode experience. For example, the object pronoun occurs in Example 10 in Table 7 to
describe the author’s personal experience of someone criticising them (stop bashing me).
First person pronouns in object and subject position are used to involve the author into the
discourse by explicitly referring to the self, enabling one’s personal stance and experience
to be encoded (see Examples 9–11 in Table 7).

Positive Dimension 3 is also characterised by numerous verb types, including stance
verbs, private verbs, public verbs, phrasal verbs, other more general verbs, infinitives,
past tense verbs and auxiliary DO. Stance verbs are often used in the tweets to express the
author’s desires. Private verbs are used to encode knowledge, thoughts and beliefs. Public
verbs are used to report on speech, and past tense verbs are used to narrate a particular
personal past event, such as ‘showed up’ in Example 9. Infinitives occur in the tweets to
complete the meaning of the verb phrase, such as ‘to explore’ in Example 11. Auxiliary
DO often co-occurs with analytic negation in the tweets most strongly associated with
positive Dimension 3 in order to negate an action. All these verb forms are often used in the tweets most associated with positive dimension 3 to self-report about various events, feelings, thoughts and ideas. For instance, Example 10 uses the public verb ‘bashing’ (meaning ‘criticising’) to report on another Twitter user who is criticising them for speaking their mind in order to request that other Twitter user to stop.

Finally, imperatives are also strongly associated with positive Dimension 3, and they tend to be used in the tweets to encode the author’s desires by making a personal request, such as Example 14 (Don’t allow me [...]). Overall, the features strongly associated with positive Dimension 3 co-occur in the tweets to self-report on a personal experience and share what the author is doing, thinking and feeling.

By contrast, the linguistic features contributing to negative Dimension 3 are typically used for describing and evaluating a subject that is external to the self (Table 6). The most contributing feature is stative forms, which are used to introduce and characterise a subject, often by encoding a judgement, such as Example 12 in Table 8, which evaluates Dalmazzi.

Negative Dimension 3 is also characterised by predicative adjectives, amplifiers and graduation (comparative and superlative forms), which are used for evaluating and encoding stance and descriptions of subjects. For instance, the predicative adjective ‘good’ occurs in Example 14 to evaluate and describe the football player Coutinho, and the comparative form ‘better’ is used in Example 13 to evaluate a subject. Amplifiers, such

| 3          | Feature                        | Coordinate | Contribution |
|------------|--------------------------------|------------|--------------|
| +          | Auxiliary_DO                   | 1.034      | 4.272        |
|            | Object_Pronoun                 | 0.951      | 6.098        |
|            | Imperative                     | 0.61       | 2.657        |
|            | Phrasal_Verb                   | 0.528      | 1.65         |
|            | Stance_Verb                    | 0.502      | 1.665        |
|            | Public_Verb                    | 0.492      | 1.781        |
|            | Analytic_Negation              | 0.459      | 1.908        |
|            | First_Person_Pronoun           | 0.348      | 2.781        |
|            | Private_Verb                   | 0.345      | 0.91         |
|            | Infinitive                     | 0.333      | 0.984        |
|            | Past_Tense_Verb                | 0.276      | 1.133        |
|            | Other_Verb                     | 0.243      | 2.056        |
| –          | IT                             | −0.422     | 1.379        |
|            | Third_Person_Singular_Verb     | −0.466     | 3.543        |
|            | Amplifier                      | −0.536     | 1.762        |
|            | Demonstrative_Pronoun          | −0.545     | 1.547        |
|            | Contracted_Forms               | −0.707     | 4.605        |
|            | Graduation                     | −0.764     | 2.161        |
|            | Stative_Form                   | −0.891     | 14.37        |
|            | Predicative_Adjective          | −1.3       | 14.328       |
as ‘so’ in Examples 14 mark the intensity of the description and link directly to the author’s personal scale. All three features mark the author’s stance and enable the description of a subject.

Pronoun *IT* and *demonstrative pronouns* are also contributing to negative Dimension 3, and are used to refer to the subject, especially one that is external to the self. These forms tend to indicate a shared communicative context, as the thing encapsulated in the pronoun is often only deducible from the context. Overall, the tweets on the negative side

---

**Table 7.** The tweets most strongly associated with positive Dimension 3.

| Tweet                                                                 | Coord | Contrib |
|----------------------------------------------------------------------|-------|---------|
| 9 TYLER JUST SHOWED UP AT MY WORK WITH COFFEE AND BREAKFAST FOR EVERYONE AND IF I DONT MARRY THIS MAN PLEASE SHOOT ME HOLY HELL WHAT DID I DO TO DESERVE HIM | 0.672 | 0.12    |
| 10 @seekingBushra How about you just hop off my ass and stop bashing me for voicing my thoughts? If you don’t like my tweets that much, then stop reading them. | 0.651 | 0.113   |
| 11 Don’t allow me to explore you if we are just going to end up like strangers | 0.606 | 0.098   |

**Table 8.** The tweets most strongly associated with negative Dimension 3.

| Tweet                                                                 | Coord | Contrib |
|----------------------------------------------------------------------|-------|---------|
| 12 @steve_vladeck Also, FWIW, it’s hilarious that Dalmazzi is no longer one of “THE CONSOLIDATED CASES.” | −0.657 | 0.115 |
| 13 @WonderBread941 That’s a flaw in the system, like saying good service is perfection. 4 out of 5 stars should mean really good but keep striving to be better. | −0.609 | 0.099 |
| 14 @MagesticSalah @The_Koxy @Thfc_Scops @JamesPearceEcho Coutinho is so fucking good. He’s a top 10 player in the world, dude is lights out. | −0.548 | 0.08    |
of the dimension often encode the author’s opinion, description and evaluation of a subject that is neither the author nor (if relevant) the addressed Twitter user(s) are the subject of the tweet.

Overall Dimension 3 reflects an opposition between tweets that are personal self-reports with tweets that are topic-focused and characterising other entities. Like Dimension 2, this pattern of linguistic variation is not necessarily new. Grieve et al. (2010: 320), for instance, used MDA to investigate the linguistic variation in blogs and found that there was a distinction between personal blogs and thematic blogs. The distinction between personal and topic-focused tweets in this dimension of Twitter may therefore support the common characterisation that Twitter is a micro-blogging platform because blogs similarly vary along this dimension.

In addition to reflecting blogging, the pattern of linguistic variation identified in Dimension 3 could also be opposing traditional micro-blogging with more contemporary uses of Twitter. Micro-blogging was designed for status updates – self-reports about what one is doing, thinking and feeling. Positive Dimension 3 therefore aligns with the platform’s original promoted uses. Additionally, many people nowadays perceive Twitter as a medium for instantaneous news distribution (Sveningsson, 2014: 221). Characterising external entities is important in the process of describing and disseminating newsworthy information. This dimension may therefore be opposing traditional micro-blogging with one of Twitter’s modern day uses – for disseminating news.

4.3. Dimension 4: Promotional versus Oppositional focus

Positive Dimension 4 is characterised by CMC and Twitter-specific features like hashtags, URLs and non-initial mentioning (see Table 9), which can be promotional resources of tweets (Page, 2012: 198). Hashtags are promotional as they are often used to promote the message to an audience beyond one’s followers (Page, 2012: 199).

URLs can be promotional, especially when they are used to promote and share content, such as Example 17, which is a link to a live stream of an online gamer. Additionally, URLs can support the tweet’s content, such as the link to an image of a girl with a cereal box in Example 16. Non-initial mentioning can also be a promotional resource because it is often used in the tweets to mention other Twitter users to promote their feeds to others, as in Example 17. Moreover, rather than interact with a user via initial mentioning, which directs the tweet to the user and reduces the tweet’s visibility to those who follow both the author and the mentioned user, non-initial mentioning can be used to increase visibility. Specifically, the tweet with non-initial mentioning is posted to the mentioned Twitter user and the timelines of all of one’s followers, increasing the chance of it being noticed. All these features are used to promote and increase visibility, not only of the content of the tweet, images and other multimedia content, but also to promote other Twitter users and events.

Positive Dimension 4 is also characterised by first person pronouns, stance verbs, perception verbs, predicative adjectives and emojis, all of which are associated with encoding personal stance. For example, perception verbs mark a sensory description, enabling descriptions of viewpoints and states, and emojis are often used to encode emotional states and support the sentiment expressed in the tweet. For instance, the emoji
with two palms together is used in Example 17 to plead to their followers to watch them play.

Positive Dimension 4 is also characterised by amplifiers, exclamation marks and capitalisation, which are associated with an emphatic production and suggest a confident and excitable conviction. Finally, time adverbs are associated with positive Dimension 4. Time adverbs provide temporal information and are often used in the tweets most associated with positive Dimension 4 to promote the time of an event, like ‘tomorrow’ in Example 15.

Overall, these features are connected by an underlying promotional function. Specifically, the tweets are often making endorsements, whereby the author encodes their stance and expresses approval, support and gratitude for an entity. Moreover, the tweets often promote entities (self or other) by incorporating recommendations.

Alternatively, the features most strongly associated with negative Dimension 4 (see Table 9) are used in the tweets to oppose a stance or entity. For example, negative Dimension 4 is characterised by interactive features, such as initial mentioning, WH-words and question marks, which are used in the tweets associated with negative

Table 9. The linguistic features strongly contributing to positive and negative Dimension 4 (coordinate; contribution).

| Feature                          | Coordinate | Contribution |
|---------------------------------|------------|--------------|
| Contracted Forms                | 0.787      | 5.811        |
| Non-Initial Mention             | 0.757      | 3.523        |
| First_Person_Pronoun            | 0.663      | 10.24        |
| PerceptionVerb                  | 0.587      | 1.407        |
| Stance Verb                     | 0.586      | 2.301        |
| Subject_Pronoun                 | 0.545      | 6.515        |
| Emoticon/Emoji                  | 0.504      | 3.028        |
| Amplifier                       | 0.501      | 1.565        |
| Hashtag                         | 0.476      | 2.172        |
| Object_Pronoun                  | 0.47       | 1.513        |
| Predicative_Adjective           | 0.464      | 1.858        |
| Exclamation_Mark                | 0.455      | 1.896        |
| Time_Adverb                     | 0.422      | 0.99         |
| URL                             | 0.284      | 2.09         |
| Capitalisation                  | 0.242      | 0.835        |
| Attributive_Adjective           | -0.185     | 0.959        |
| Third_Person_Singular_Verb      | -0.269     | 1.2          |
| Definite_Article                | -0.307     | 1.566        |
| Analytic_Negation               | -0.394     | 1.427        |
| Nominalisation                  | -0.462     | 1.544        |
| Initial_Mention                 | -0.478     | 4.91         |
| Passive                         | -0.594     | 1.082        |
| Auxiliary_DO                    | -0.62      | 1.563        |
| WH-Word                         | -0.688     | 3.015        |
| Question_Mark                   | -0.9       | 4.755        |
Dimension 4 to introduce a counterargument and challenge a statement from another Twitter user, such as Example 20 (see Table 11). Additionally, questions are used to express disbelief to ridicule their addressee’s previously mentioned position, such as ‘What are you talking about?’ in Example 19. A present tense orientation is marked through *third person singular verb forms*, which are often used when the topic is of immediate relevance (Biber, 1988: 105), characteristic when opposing, as the topic becomes relevant and needs to be countered.

*Analytic negation* and *auxiliary DO* are also associated and are used to negate an action so as to oppose it. Additionally, *auxiliary DO* often occurs in the tweets associated with negative dimension 4 to emphasise the verb and counter the position. For example, in Example 20, *auxiliary DO* is used to oppose a previous statement: ‘But all models do pose somewhere nude’.

### Table 10. The tweets most strongly associated with positive Dimension 4.

| Tweet | Coord | Contrib |
|-------|-------|---------|
| TGIF #itscalledfashionsweatylookitup and I’ll be in all day tomorrow too – | 0.673 | 0.123 |
| come see us if you want this di*khead lewk 🥰✨ @ Claw and Co. https://t.co/DJ82bMx2Ht | | |
| Breakfast/lunch time! I’m feeling really sick and this is all I could stomach so thank goodness for these crunchy flakes! Now send me to @Lovelsland please! kelloggs cer ealdater #hungry #crispyflakes | 0.667 | 0.121 |
| https://t.co/zW9L6kyDKX | | |
| IM LIVE! SO PEEPS COME SAY HELLUR! ❤️🧣 | 0.655 | 0.116 |
| Also check out these peeps! – @ArockMusic @Alyssa_Rxse @chasifrass 🌴@PROJ3CTGD 🌴@_OnlyVic_ 🌴@MrSynnn 🌴@Joshh9761 🌴@xLeaahh 🌴@OrbzTV 🌴@KingTylerish 🌴@lolgabiz 🌴@Elthirlwell | | |

### Table 11. The tweets most strongly associated with negative Dimension 4.

| Tweet | Coord | Contrib |
|-------|-------|---------|
| @zingwa @CryptoPseudonym @Crypto_Trogdor @cryptostardust @Elbrusco3 @mrdmrts @cripdoesimpson 🍄Can you explain why anyone would pay the astronomical price to fake the launch of hundreds of satellites, and hundreds of thousands of similar videos/images? What benefit is gained by such an extraordinary, and arbitrary, conspiracy? | -0.626 | 0.106 |
| @iamkhaleed No, this country wasn’t founded by immigration... What are you talking about? The British came here and conquered the land, then maintained it. | -0.619 | 0.104 |
| @TanaLara @Punkster1011 But all models do pose somewhere nude. What do you think PlayBoy Magazine is where almost most famous people pose nude? Isn’t that “Adult” magazine? | -0.57 | 0.088 |
Features indicative of a detached style and abstract informational focus, such as passive constructions and nominalisations are also associated with negative Dimension 4. Detached language involves subjects that are abstractions. For instance, ‘benefit’ in Example 21. Passive constructions enable the entity affected by the action to be emphasised to provide a counter argument. In Example 21, the ‘conspiracy maker’ is not mentioned in the sentence, but instead the beneficiary of the conspiracy is emphasised, as a way to suggest that there is no beneficiary. This functions to oppose the addressee and imply that they are wrong. The tweets on the negative side of Dimension 4 are considerably unsupportive and instead function to challenge and dispute.

Overall, Dimension 4 reveals a distinction between promotional tweets and oppositional tweets. Unlike the previous dimensions, which are patterns previously found in MDA research, this dimension identifies a specific yet important function of Twitter – as a platform for gaining attention and positioning oneself. This dimension may be influenced by following one’s interests on Twitter. Individuals may be more inclined to express support towards and promote one’s interests and may at the same time meet interests at variance to theirs and feel the need to oppose.

Another reason for the oppositional function may be due to the pervasiveness of trolling on Twitter. One trolling strategy has been to take the opposing view to provoke a response. At the same time, if trolls manage to provoke, the victim’s reaction can be oppositional too. Thus, the oppositional style may be a result of the interaction that takes place between trolls and their victims.

The promotional function reflected in positive Dimension 4 supports Page (2012: 198), who has suggested that practices of self-branding and micro-celebrity are frequent on Twitter. The high degree of promotional tweets may be because Twitter is used by individuals and businesses whose main purpose is to promote their brand, activity and products.

The promotional tweets, however, are not only self-promotional, but they are also endorsing and recommending entities, including other Twitter users. One of Twitter’s major uses is for following interests, including celebrities, businesses, organisations and topics. This practice of showing support for other entities may therefore be associated with demonstrating affiliation to such interests (Zappavigna, 2018: 122).

Another explanation for this dimension may be due to the architectural characteristics of Twitter. This dimension opposes promotional with oppositional tweets, both of which have been described as practices that have been influenced by the importance placed on follower size (Anderson, 2019; Dorsey, 2019; Page, 2012). Overall, the need for more followers is offered as a possible explanation for both styles of tweets opposed in this dimension, and this need is a result of the importance placed on the size of followers – architecturally and semiotically. It can be argued that because follower size is regarded as a sign of status, gaining followers has become an incentive amongst users. This incentive has therefore influenced the high frequency of promotional and oppositional tweets, which are important styles for gaining attention.
4.4. Dimension 5: Persuasive versus Non-Persuasive concern

Positive Dimension 5 is characterised by features associated with an interpersonal function and a shared communicative context, such as imperatives, WH-words, question marks, second person pronouns, interjections, demonstrative pronouns, private verbs and complementation (see Table 12). Additionally, capitalisation and exclamation marks, which can be used for emphasis or to mimic oral production, are also strongly associated with positive Dimension 5.

Positive Dimension 5 is also characterised by hashtags and non-initial mentioning, which are used to broadcast the tweet to feeds and mention other users in the third person. The occurrence of these features amongst interpersonal features suggests that these tweets are not necessarily interacting with people directly, but are perhaps purposely employing an interpersonal style and presupposing a shared communicative context to imply friendliness and intimacy, and address a particular, albeit unspecified, audience.

The tweets associated with positive Dimension 5 (see Table 13) show these features are connected by an underlying persuasive function. Specifically, the tweets aim to bring about something, often employing imperatives to demand some action, such as Examples 21 and 23. Although imperatives direct future action, they are not always persuasive. However, certain kinds of imperatives can contain presuppositions to persuade the reader, such as the imperative in Example 22, which presupposes that the reader is overpaying for

Table 12. The linguistic features most strongly contributing to positive and negative Dimension 5.

| Feature                        | Coordinate | Contribution |
|-------------------------------|------------|--------------|
| Imperative                    | 1.127      | 10.735       |
| Exclamation_Mark              | 0.972      | 10.085       |
| Question_Mark                 | 0.971      | 6.432        |
| Complementation               | 0.818      | 2.827        |
| WH-Word                       | 0.759      | 4.272        |
| Second_Person_Pronoun         | 0.735      | 8.679        |
| Hashtag                       | 0.606      | 4.098        |
| Auxiliary_DO                  | 0.528      | 1.317        |
| Capitalisation                | 0.459      | 3.497        |
| Demonstrative_Pronoun         | 0.454      | 1.271        |
| Interjection                  | 0.444      | 2.334        |
| Private_Verb                  | 0.389      | 1.366        |
| Non-Initial_Mention           | 0.357      | 0.91         |
| Other_Noun                    | −0.124     | 0.827        |
| First_Person_Pronoun          | −0.238     | 1.532        |
| Subject_Pronoun               | −0.27      | 1.86         |
| Indefinite_Article            | −0.306     | 1.414        |
| Third_Person_Pronoun          | −0.379     | 1.552        |
| Past_Tense_Verb               | −0.437     | 3.352        |
| Profanity                      | −0.55      | 1.314        |
| Passive                       | −0.802     | 2.29         |
Additionally, many of the interpersonal features associated with positive Dimension 5 that were described previously are used in the tweets most associated with positive Dimension 5 to persuade. For example, second person pronouns often occur presuppositionally (e.g. Example 22) in the sense that the addressee is assumed to read the tweet, thus enabling the author to target readers without mentioning them specifically. Such a technique is used commonly in advertising (Cook, 2001: 129), to imply friendliness and intimacy and persuade them to trust what is being said. Overall, the tweets associated with positive Dimension 5 are persuasive, often trying to bring about something in the future.

**Table 13.** The tweets most strongly associated with positive Dimension 5.

| Tweet                                                                 | Coord  | Contrib |
|----------------------------------------------------------------------|--------|---------|
| 21 ARE YOU HUMAN TOO? 💖 Where did you find that extra charm?        | 0.702  | 0.156   |
| You’re killing me softly! WHY SO HOT?! PLEASE EXPLAIN 😊😊          |        |         |
| @BTS_twt @bts_bighit #BTSxLotteFamilyConcert https://t.co/0UhE8dkYZy |        |         |
| 22 Find out how much you could save switching to BrandStencil for your | 0.644  | 0.131   |
| artwork creation – use our ROI calculator! https://t.co/vro5K44Q2n    |        |         |
| #MarketingAutomation                                                 |        |         |
| 23 Attn, #mixers! Stop what you’re doing and listen to @LittleMix and | 0.643  | 0.13    |
| @CheatCodesMusic’s #OnlyYou 🔥                                       |        |         |
| 🔥🔥🔥†#littlemixONLYYOU https://t.co/AjkcT0zWqV                      |        |         |

**Table 14.** The tweets most strongly associated with negative Dimension 5.

| Tweet                                                                 | Coord  | Contrib |
|----------------------------------------------------------------------|--------|---------|
| 24 b**IT**CH MY ACCOUNT JUST GOT TEMPORARILY LOCKED. I JUST HAD A MINI | −0.455 | 0.065   |
| HEART ATTACK                                                         |        |         |
| 25 Free watching of young male gay sex I had him undress all the way  | −0.43  | 0.058   |
| https://t.co/ynr3m3YHPC telugu hot gay porn zeb atlas fuck twink     |        |         |
| videogaysex afganistangayporn gay straight bait bus video full 3gp     |        |         |
| chloroform muscled gay studs voyeur gay medic                         |        |         |
| 26 We love our team development days at Lighting up Learning once    | −0.427 | 0.057   |
| every six weeks. Such unified and synchronous thinkers whilst deep   |        |         |
| respect enables them to challenge each other, including our Director.|        |         |
| Even Womble the Therapy Dog gets involved. https://t.co/5KoNLSxUes    |        |         |
By contrast, the linguistic features associated with negative Dimension 5 are indicative of personal narratives. Specifically, *past tense verbs* and *passive constructions* tend to occur in the tweets to refer to past events and experiences. For example, past tense verbs and passive construction occur in Example 24 in Table 14 to report that the author’s ‘ACCOUNT JUST GOT TEMPORARILY LOCKED’.

*Subject pronouns* are also associated with negative Dimension 5, especially *first person* and *third person pronouns*, and these are used to mark personal narratives, as well as narratives about external entities (e.g. Example 25). *Indefinite articles* and *general nouns* are also associated, which are indicative of new content being described, and are used in the tweets to provide detail on the experience.

The final feature strongly associated with Dimension 5 is *profanity*. Whilst profanity can have several functions, it is used in the tweets mainly to evaluate the event, often as an immediate response (e.g. Example 24), or incorporated into the description of the event (e.g. Example 25), or as an overall analysis.

Overall, Dimension 5 distinguishes between tweets on the positive side that are persuasive and bringing something about in the future and tweets on the negative side that are narrative and evaluating past events. This dimension suggests that Twitter is used for economic or non-economic gain, as a persuasive style is used to obtain something and bring about something. Persuasive tweets incorporate a range of different text types, including advertisements, political tweets, as well as tweets marking the author’s opinion. Twitter has become a platform for political expression, not only by politicians, but also ordinary citizens. Rhetoric is expected in political tweets, whereby the ideologies of one party are expressed to influence particular outcomes.

Moreover, Twitter is a commercial enterprise, which receives most of its revenue from advertisers by selling the ability to promote products, tweets, accounts and trends to consumers. Advertisers pay Twitter to broadcast content to the right audience. Promoted tweets are seen by all users; however, the promoted tweets users experience will vary, as they are designed and tailored to particular identities using algorithms. Because promoted tweets, regardless of the content, are seen by all users, their style may set a precedent for users. Thus, not only does the purpose of Twitter as a commercial enterprise influence the number of promoted tweets, which are often characterised by a persuasive style, but because promoted tweets are observed, their style may be used as a linguistic frame.

One explanation for a high frequency of narrative tweets could be that news reportage is shared on Twitter with great ease. For example, online news sites provide a small Twitter logo on the article’s page, which can be clicked to share the story with one’s followers and add commentary. Finally, many of the tweets associated with narrativity are characteristic of status updates, where personal recent events are described. Posting tweets is also referred to as updating one’s status. Therefore, one explanation for narrative tweets may be the result of the notion of ‘update’, which demands bringing one’s readers up to speed by mentioning not only what is happening, but also what has happened since the last update.
5. Conclusion

This paper sought to identify and describe the major patterns of linguistic variation across Twitter by applying short-text MDA to a corpus of tweets. After excluding the first dimension from interpretation because of its strong correlation with text length, 4 subsequent dimensions were identified and interpreted for their underlying communicative function and/or style.

The results suggest that there is a great deal of linguistic variation on Twitter. Like Honeycutt and Herring (2009: 9) and Yaqub et al. (2017: 614), the results show that interaction and the public broadcast of information are important communicative goals of tweets. Additionally, similar to Berber-Sardinha (2018: 149), the results in the present study suggest that stance-taking is important across Twitter, as we encode stance when we report on our thoughts and feelings, when we characterise external entities, when we promote particular things and express opposition, and when we persuade people to do something. Stance-taking is a means for identity construction, as through positioning ourselves, we construct an identity that aligns or disaligns with them.

Finally, the range of linguistic variation found across this Twitter corpus in this analysis is indicative that Twitter is used commercially for social and economic gain, including gaining money, attention, support, followers and social interaction, suggesting that Twitter users have common goals, which exploit the technology’s potential of reaching audiences.

On Twitter, and social media generally, individuals are given the opportunity to manage how they present themselves. Each person creates a profile, and they can construct their identity through their profiles and messages. Whilst this ‘is in many ways liberating’, it has also ‘brought pressure to create compelling identities that attract attention’ (Baym, 2014: 222), and in the case of Twitter, attract followers. This suggests that many people view their profiles and their messages as something to be consumed. The range of stylistic and functional variation found across this Twitter corpus supports this.

Twitter is commonly used for self-commodification, as people manage their identities, engaging in practices of self-branding through stance-taking, self-reporting, promotion and persuasion, as well as broadcasting their message beyond their followership, distributing news and expressing opposition. This often occurs to attract attention and gain something. Additionally, the results show that interaction is important, suggesting that Twitter is also used for social and interpersonal gain.

This study shows the utility of short-text MDA approach for identifying patterns of linguistic variation across a corpus of short texts. This study has thus provided new opportunities for future research to apply the approach to other short text corpora, such as poems, to uncover the major patterns of stylistic and/or functional variation.

Acknowledgments

I would like to thank Jack Grieve for his advice and support throughout my PhD, which led to this research. I also would like to thank the reviewers and Tony McEnery for their helpful comments on this paper.
Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs
Isobelle Clarke  https://orcid.org/0000-0001-5541-6327

Supplemental Material
Supplementary Material for this article is available online.

Notes
1. Eigenvalues indicate how much variance there is in the data in a particular direction. More variation equals higher eigenvalues. See Le Roux and Rouanet (2010) for further description of eigenvalues.
2. Crystal and Davy’s (1969) definition of style is dissimilar to Biber and Conrad’s (2009) conception of style (see section 1) and aligns more with Biber and Conrad’s (2009) register perspective. However, Crystal and Davy (1969) show that style can be conceived more broadly to account for different varieties within situations. Thus, I perceive style to be more appropriate in the present study than just register as the study focuses on linguistic variation within Twitter and does not compare to other registers.

References
Anderson C (2019) How Twitter Needs to Change TED Interview with Jack Dorsey, Vancouver: TED. Available at: https://www.ted.com/talks/jack_dorsey_how_twitter_needs_to_change/transcript?language=en (accessed 1 July 2019).
Baym NK (2014) The perils and pleasures of tweeting with fans. In: Weller K, Bruns A, Burgess J, Mahrt M and Pushmann C (eds) Twitter and Society. New York: Peter Lang, pp. 221–236.
Benzécri JP (1992) Correspondence Analysis Handbook. New York: Dekker.
Berber-Sardinha T (2014) Comparing internet and pre-internet registers. In: Berber-Sardinha T and Pinto MV (eds) Multi Dimensional Analysis, 25 years on: A Tribute to Douglas Biber. Amsterdam: John Benjamins Publishing Company, pp. 81–108.
Berber-Sardinha T (2018) Dimensions of variation across internet registers. International Journal of Corpus Linguistics 232: 125–157.
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): 243–257.
Biber D (2014) Multi-dimensional analysis: A personal history. In: Berber Sardinha T and Pinto MV (eds), Multi Dimensional Analysis, 25 years on: A Tribute to Douglas Biber. Amsterdam: John Benjamins Publishing Company, pp. xxvi–xxviii.
Biber D and Conrad S (2009). *Register, Genre, and Style (Cambridge Textbooks in Linguistics)*. Cambridge: Cambridge University Press.

Clarke I (2019) Functional linguistic variation in Twitter Trolling. *The International Journal of Speech Language and the Law* 26(1): 57–84.

Clarke I and Grieve J (2017) Dimensions of Abusive Language on Twitter. *Proceedings of the 1st Workshop on Abusive Language Online*: 1–10. Vancouver: Association for Computational Linguistics.

Clarke I and Grieve J (2019) Stylistic variation on the Donald Trump Twitter account. *Plos One* 14(9): 1–27.

Cook G (2001) *The Discourse of Advertising*. 2nd edition. London, UK: Routledge.

Crystal D and Davy D (1969) *Investigating English Style*. London, UK: Routledge.

Dorsey J (2019) *How Twitter Needs to Change TED Interview with Jack Dorsey*. Vancouver: TED. Available at: https://www.ted.com/talks/jack_dorsey_how_twitter_needs_to_change/transcript?language=en (accessed 1 July 2019).

Friginal E, Waugh O and Titak A (2018) Linguistic variation in Facebook and Twitter posts. In: Friginal E (ed), *Studies in Corpus-Based Sociolinguistics*. New York: Routledge, pp. 342–363.

Gimpel K, Schneider N, O’Connor B, et al. (2011) Part-of-speech tagging for Twitter: Annotation, features, and experiments. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Short Papers*, Portland, OR, 19–24 June 2011, Portland, Oregon: Association for Computational Linguistics (ACL).

Grieve J, Biber D, Friginal E, et al. (2010) Variation among blogs: A multi dimensional analysis. In: Mehler A, Sharoff S and Santini M (eds), *Genres on the web: Computational models and Empirical Studies*. New York: Springer-Verlag, pp. 303–322.

Grieve J, Clarke I, Chiang E, et al. (2019) Attributing the Bixby letter using n-gram tracing. *Digital Scholarship in the Humanities* 34(3): 493–512.

Holler J, Kendrick KH, Casillas M, et al. (2015) Editorial: Turn-taking in human communicative interaction. *Frontiers in Psychology* 6: 1919.

Honeycutt C and Herring SC (2009) Beyond microblogging: conversation and collaboration via Twitter. *Proceedings of the Forty-Second Hawai’i International Conference on System Sciences* (HICSS-42), Los Alamitos, CA: IEEE Press.

Husson F, Josse J, Le S, et al. (2017) Package ‘FactoMineR’. Available at: http://factominer.free.fr (accessed 30 June 2017).

Keeney MW (2018) Package rtweet. Available at: https://cran.r-project.org/web/packages/rtweet/rtweet.pdf (accessed 22 June 2018).

Le Roux B and Rouanet H (2010) *Multiple Correspondence Analysis*. Thousand Oaks, CA: SAGE Publications, Inc.

Page R (2012) The linguistics of self-branding and micro-celebrity in Twitter: The role of hashtags. *Discourse & Communication* 6(2): 181–201.

Quirk R, Greenbaum S, Leech G and Svartvik J (1985) *A Comprehensive Grammar of the English Language*. London, UK: Longman.

Stamatatos E (2009) A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* 60: 538–556.

Sveningsson M (2014) “Not the real thing”: Young people’s experiences of getting civic and political news through social media. In: *Proceedings of Culture, Technology and Communication*, Oslo, Norway, Norway: University of Oslo, pp. 221–241.

Titak A and Roberson A (2013) Dimensions of web registers: An exploratory multi-dimensional comparison. *Corpora* 8(2): 235–260.
Weller K, Bruns A, Burgess J, et al. (2014) Twitter and society: An introduction. In: Weller K, Bruns A, Burgess J, Mahrt M and Pushmann C. (eds), Twitter and Society. New York: Peter Lang. pp. xxix–xxxviii.

Yaqub U, Chun SA, Atluri V, et al. (2017) Analysis of political discourse on Twitter in the context of the 2016 US presidential elections. Government Information Quarterly, 34: 613–626.

Zappavigna M (2018) Searchable Talk: Hashtags and Social Media Metadiscourse. London: Bloomsbury.

Author biography

Isobelle Clarke is a Leverhulme Trust’s Early Career Research Fellow in Lancaster University’s Centre for Corpus Approaches to Social Science. Isobelle completed her PhD at the University of Birmingham, where she developed a version of Douglas Biber’s Multi-Dimensional Analysis (MDA) for the analysis of short texts and applied this technique to corpora of tweets, including general tweets and trolling tweets. More recently, Isobelle has been developing a method for approaching discourse through corpora, called Keyword Co-occurrence Analysis (KCA), which is essentially a short-text MDA of keywords produced in a keyword analysis, which she has applied to corpora of press articles and conspiracy and pseudo-science websites. Isobelle also works within the areas of forensic linguistics and language teaching and assessment.