Appendix A: The Feature Sets and Decisions for Pooling

Le Roux and Rouanet (2010) advise that very infrequent features (e.g. those that occur in <5 per cent of the data) either need to be pooled with other related features or they might need to be discarded because infrequent features can overly influence the axes, as they contribute more to the overall variance. The list following presents the features occurring in fewer than 5 per cent of the turns in the TLC and the decisions that were made with respect to pooling or deleting features from the final dataset. The justifications for these decisions are also presented in the Table A1. Infrequent features that were specific types of a broader part-of-speech category were pooled into the broader part-of-speech or 'other' category. For example, the feature set distinguishes between different kinds of adverbs (e.g. place, time, downtoner, amplifier, quantifying adverbs), and then any other adverbs that are not one of these types are tagged as 'other adverb'. Quantifying adverbs do not occur in more than 5 per cent of the tweets. Therefore, this feature was pooled with the 'other adverb' category. Essentially, because it does not occur frequently, the feature is dropped from the feature set, meaning that if quantifying adverb wasn't included in the tagger, any occurrence of a quantifying adverb would be classed as an instance of 'other adverb'. Thus, this is the logical category in which to place it. When there was more than one option (deleting or pooling including the feature in multiple categories), many of these options were tested by running several MCAs on different feature sets depicting the different pooling options. For example, copular verbs that are not BE as a main verb did not occur in more than 5 per cent of the tweets, whereas BE as a main verb did. Both features are part of the broader category of 'stative forms', and so they could be pooled together into one broad category, or copular verbs could be deleted from the feature set. To test the effect of either decision, two data matrices were created and each was subjected to MCA: one with all other linguistic features but with copular verbs deleted, and the other involving copular verbs being pooled with BE as the main verb into the new category of 'stative forms'. Although the active variables in each MCA are

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 14 Feb 2025 at 05:47:30, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009208932.011

Table A1 *The full feature set.*

Features <5 per cent of the TLC	Decision	Justification
Adj+that complements clause	Deleted	All specific types of complement clauses occurred in fewer than 5 per cent of tweets. Even if the specific types were combined to form one broad category of complementation, they still did not occur frequently enough. As a result they were deleted.
Adj+to complement clause	Deleted	Even if the specific types were combined to form one broad category of complementation, they still did not occur frequently enough. As a result they were deleted.
Adverbs of frequency/ usuality	Pooled with general adverbs	Adverbs are divided into different types and all other adverbs not specified are grouped into a broader 'other adverbs' category. If the specific type of adverb occurs infrequently then each instance can be recombined with the 'other adverb' feature.
Agentless passives	Deleted	Passive constructions were divided into different types, yet even by recombining to form the broader category of passives, they did not occur frequently enough and so they were deleted.
By-passive	Deleted	Passive constructions were divided into different types, yet even by recombining to form the broader category of passives, they did not occur frequently enough and so they were deleted.
Comparative	Deleted	Comparatives could be pooled with superlatives to form a broader 'gradation' category. However, they do not occur enough times when combined and so they were deleted.
Concessive subordinator	Pooled with general subordinators	Subordinators are divided into different types and all other subordinators are grouped into an 'other subordinator' feature. Therefore, if a specific type does not occur frequently it can rejoin the 'other subordinator' feature category.
Conditional subordinator	Pooled with general subordinators	Subordinators are divided into different types and all other subordinators are grouped into an 'other subordinator' feature. Therefore, if a specific type does not occur frequently it can rejoin the 'other subordinator' feature category.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 14 Feb 2025 at 05:47:30, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009208932.011

Table A1 (*cont.*)

Features <5 per cent of the TLC	Decision	Justification
Copular verbs (not BE)	Pooled with BE as main verb and existential there	We could have deleted this feature from the feature set, or combined either with 'General/other verbs' or pooled with BE as a main verb and existential there into a category of stative forms. We tested all by running the analysis on all types and then correlated the coordinates of the individual turns from the results. We also compared the contribution and coordinate of 'Stative forms' and 'BE as a main verb', as well as 'General/other verbs' pre- and post-pooling in each analysis on each dimension to observe if the pooling led to any substantial difference. No substantial differences were found so they were combined to form the category stative forms.
Demonstrative determiner	Pooled with general determiners	As a type of determiners, they were pooled with the general determiner category. We could have pooled with demonstrative pronouns to form a demonstrative category. However, as pronouns occurred frequently enough to stand on their own, we decided to maintain this distinction.
Downtoner adverb	Pooled with general adverbs	Adverbs are divided into different types and all other adverbs not specified are grouped into a broader 'other adverbs' category. If the specific type of adverb occurs infrequently then each instance can be recombined with the 'other adverb' feature.
Existential there	Pooled with BE as a main verb and copular verbs	We could have deleted this feature from the feature set, or combined either with 'General/other verbs' or pooled with BE as a main verb and existential there into a category of stative forms. We tested all by running the analysis on all types and then correlated the coordinates of the individual turns from the results. We also compared the contribution and coordinate of 'Stative forms' and 'BE as a main verb', as well as 'General/other verbs' pre- and post-pooling in each analysis on each dimension to observe if the pooling led to any substantial difference. No substantial differences were found so they were combined to form the category stative forms.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 14 Feb 2025 at 05:47:30, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009208932.011

Table A1 (*cont.*)

Features <5 per cent of the TLC	Decision	Justification
Gerund	Deleted	No applicable broader category.
Indefinite/ quantifying pronoun	Deleted	Quantifying pronouns could have also been grouped with other quantifiers of different parts of speech (e.g. quantifying-determiners, quantifying-pre-determiners, quantifying-adverbs). They were not grouped this way because all instances did not meet the 5 per cent turn threshold. Additionally, there was no broader pronoun category without losing the distinction between other pronouns, such as first/second/third.
Initial verb	Deleted	Whatever the verb is, it would also be classified as either one of the verb types or in the 'other verb' category. Therefore it does not need to be pooled with broader verb category. We could have combined with other initial verbs. However, we tested this by running the analysis on the feature combined with other initial verbs as well as with this feature deleted. Overall, the new initial verb feature influenced the dimensions too substantially and made the dimensions far less interpretable.
Initial verb	Deleted	This feature is already counted as third-person singular verb form regardless of initial position. We could have combined it with other initial verbs. However, we tested this by running the MCA on one feature set where the feature combined all initial verbs, as well as another feature set where this feature was deleted. Overall, the new initial verb feature influenced the dimensions too substantially and made the dimensions far less interpretable.
Initial verb BE	Deleted	We could have combined this into a broader category of initial verbs with other initial verb instances. However, we tested this by running the MCA on one feature set where the feature combined all initial verbs, as well as another feature set where this feature was deleted. Overall, the new initial verb feature influenced the dimensions too substantially and made the dimensions far less interpretable.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 14 Feb 2025 at 05:47:30, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009208932.011

Table A1 (*cont.*)

Features <5 per cent of the TLC	Decision	Justification
Initial verb DO	Deleted	We could have combined into a broader category of initial verbs with other initial verb instances. However, we tested this by running the MCA on one feature set where the feature combined all initial verbs, as well as another feature set where this feature was deleted. Overall, the new initial verb feature influenced the dimensions too substantially and made the dimensions far less interpretable.
Initial verb HAVE	Deleted	We could have combined into a broader category of initial verbs with other initial verb instances. However, we tested this by running the MCA on one feature set where the feature combined all initial verbs, as well as another feature set where this feature was deleted. Overall, the new initial verb feature influenced the dimensions too substantially and made the dimensions far less interpretable.
Initial verb *-ing*	Deleted	It could be an auxiliary omission and thus in progressive form or it could be a gerund. Rather than check each instance to clarify, and rather than group these instances into either general verbs or general nouns and misclassify some, it was decided to just delete the feature altogether from the feature set. We could have combined with initial verbs (but it might not have been one). However, we tested this by running the MCA on one feature set where the feature combined all initial verbs, as well as another feature set where this feature was deleted. Overall, the new initial verb feature influenced the dimensions too substantially and made the dimensions far less interpretable.
Initial verb *modal*	Deleted	The type of modal is counted regardless of whether it is positioned initially. We could have combined it with other initial verbs. However, we tested this by running the MCA on one feature set where the feature combined all initial verbs, as well as another feature set where this feature was deleted. Overall, the new initial verb feature influenced the dimensions too substantially and made the dimensions far less interpretable.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 14 Feb 2025 at 05:47:30, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009208932.011

Table A1 (*cont.*)

Features <5 per cent of the TLC	Decision	Justification
Initial verb *past*	Deleted	This feature is already counted as past tense verb regardless of initial position. We could have combined it with other initial verbs. However, we tested this by running the MCA on one feature set where the feature combined all initial verbs, as well as another feature set where this feature was deleted. Overall, the new initial verb feature influenced the dimensions too substantially and made the dimensions far less interpretable.
Initial verb *question*	Deleted	Too few instances to combine with other question features. We could have combined with other initial verb types. However, we tested this by running the MCA on one feature set where the feature combined all initial verbs, as well as another feature set where this feature was deleted. Overall, the new initial verb feature influenced the dimensions too substantially and made the dimensions far less interpretable.
Laughter interjection	Pooled with general interjections	All specific types of interjections occurred in fewer than 5 per cent of tweets. These specific types were therefore combined into a broader category of interjections.
Modal of necessity	Pooled with general verbs	We could have combined all modals together but this would lose the distinction between possibility and prediction modals, which contribute differently on different dimensions. As a result we combined it with the general verbs category.
Negative interjection	Pooled with general interjections	All specific types of interjections occurred in fewer than 5 per cent of tweets. These specific types were therefore combined into a broader category of interjections.
Noun + that complements clause	Pooled with complementation	All specific types of complement clauses occurred in fewer than 5 per cent of tweets. Even if the specific types were combined to form one broad category of complementation, they still did not occur frequently enough. As a result they were deleted.
Numeral determiners	Pooled with general determiners	Not enough instances of either kind of numeral to combine to create a numeral feature. We combined with other determiners into a general determiner category (e.g. demonstratives, ordinal determiners, pre-determiners).

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 14 Feb 2025 at 05:47:30, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009208932.011

Table A1 (*cont.*)

Features <5 per cent of the TLC	Decision	Justification
Numeral noun	Pooled with general nouns	Not enough instances of either kind of numeral to combine to create a numeral feature. We combined with other nouns in a general noun category.
Ordinal determiner	Pooled with general determiners	Not enough instances of either kind of ordinal to combine to create an ordinal feature. We combined with other determiners (e.g. demonstratives, ordinal determiners, pre-determiners).
Ordinal noun	Pooled with general nouns	Not enough instances of either kind of ordinal to combine to create an ordinal feature. We combined with other nouns to form a general noun category.
Other conjunctions	Deleted	Too few instances to combine with other frequent and more specific kinds of conjunctions.
Perception verbs	Pooled with general verbs	Different types of verbs were distinguished from general verbs and therefore infrequent types can be recombined with broader verb category.
Phrasal verbs	Pooled with general verbs	Different types of verbs were distinguished from general verbs and therefore infrequent types can be recombined with broader verb category.
Pied-piping relative	Deleted	Not enough instances of either kind of relative clause to combine into a broader category of relatives
Place adverb	Pooled with general adverbs	Adverbs are divided into different types and all other adverbs not specified are grouped into a broader 'other adverbs' category. If the specific type of adverb occurs infrequently then each instance can be recombined with the 'other adverb' feature.
Place subordinator	Pooled with general subordinators	Subordinators are divided into different types and all other subordinators are grouped into an 'other subordinator' feature. Therefore, if a specific type does not occur frequently it can rejoin the 'other subordinator' feature category.
Possessive noun	Pooled with possession	The only feature denoting possession that occurred in more than 5 per cent of the tweets was possessive determiner. These features were combined to create one variable of possession.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 14 Feb 2025 at 05:47:30, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009208932.011

Table A1 (*cont.*)

Features <5 per cent of the TLC	Decision	Justification
Possessive pronoun	Pooled with possession	The only feature denoting possession that occurred in more than 5 per cent of the tweets was possessive determiner. These features were combined to create one variable of possession. The type of possessive pronoun is also counted (e.g. first, second, third or *it*).
Possessive proper noun	Pooled with possession	The only feature denoting possession that occurred in more than 5 per cent of the tweets was possessive determiner. These features were combined to create one variable of possession.
Pre-determiner	Deleted	Too few instances to combine with quantifying pre-determiners to create general pre-determiner category.
Progressive	Delete	No broader category with which to pool.
Pro-verb DO	Pooled with general verbs	Different types of verbs were distinguished from general verbs and therefore infrequent types can be recombined with a broader verb category.
Quantifying adverb	Pooled with general adverbs	Adverbs were divided into different types and all other adverbs that do not fall in these particular categories are grouped into a category called 'other adverbs'. Therefore, if a specific type does not occur frequently it can rejoin the 'other adverbs' category. Quantifying adverbs could have also been grouped with other quantifiers of different parts of speech (e.g. quantifying-determiners, quantifying-pre-determiners, quantifying-pronouns). They were not grouped this way because they did not occur enough times.
Quantifying determiner	Pooled with general determiners	Quantifying determiners are a kind of determiner, thus we pooled with a general determiner category. Quantifying adverbs could have also been grouped with other quantifiers of different parts of speech (e.g. quantifying-determiners, quantifying-pre-determiners, quantifying-pronouns). They were not grouped this way because they did not occur enough times.
Quantifying pre-determiner	Pooled with general determiners	Too few instances to combine with other pre-determiners to create general pre-determiner category. As a kind of determiner, they were pooled with general determiners.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 14 Feb 2025 at 05:47:30, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009208932.011

Table A1 (*cont.*)

Features <5 per cent of the TLC	Decision	Justification
Reflexive pronoun	Deleted	No applicable broader category, though 'pronouns'. Reflexive pronouns are counted according to first, second or third person, or *it*.
Relative clause object gap	Deleted	Not enough instances of either kind of relative clause to combine into a broader category of relatives
Relative clause subject gap	Deleted	Not enough instances of either kind of relative clause to combine into a broader category of relatives
Split infinitive	Pooled with infinitives	Split infinitives were separated from infinitives as a particular type and so therefore were recombined with the broader category.
Suasive verb	Pooled with general verbs	Different types of verbs were distinguished from general verbs and therefore infrequent types can be recombined with broader verb category.
Subordinator with ellipted subject	Deleted	No applicable broader category. If it is a specific type of subordinator it will be classified as such as well.
Superlative	Delete	Comparatives could be pooled with superlatives to form a broader 'gradation' category. However, they do not occur enough times when combined and so they were deleted.
Synthetic negation	Deleted	No applicable broader category, albeit 'negation', meaning that we could have combined analytic negation with synthetic negation. However, we did not want to conflate this distinction as previous research has found this to be an important feature (e.g. Biber, 1988; Clarke and Grieve, 2017; Clarke, 2018).
Time adverb	Pooled with general adverbs	Adverbs were divided into different types and all other adverbs that do not fall in these particular categories are grouped into a category called 'other adverbs'. Therefore, if a specific type does not occur frequently it can rejoin the 'other adverbs' category.
Time subordinator	Pooled with general subordinators	Subordinators are divided into different types and all other subordinators are grouped into an 'other subordinator' feature. Therefore, if a specific type does not occur frequently it can rejoin the 'other subordinator' feature category.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 14 Feb 2025 at 05:47:30, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009208932.011

Table A1 (*cont.*)

Features <5 per cent of the TLC	Decision	Justification
Title	Deleted	No applicable broader category.
Verb *-ing* (not in standard progressive form)	Deleted	It could be an auxiliary omission and thus in progressive form or it could be a gerund. Rather than check each instance to clarify, and rather than group these instances into either general verbs or general nouns and misclassify some, it was decided to just delete the feature altogether from the final feature set.
Verb + that complements clause	Pooled with complementation	All specific types of complement clauses occurred in fewer than 5 per cent of tweets. Even if the specific types were combined to form one broad category of complementation, they still did not occur frequently enough. As a result they were deleted.
WH-word + contracted verb	Pooled with pronoun + contracted verb in a 'contracted forms' variable	We could have either deleted this feature or combined it with 'pronoun with contracted verb' into a broader category of contracted forms. We tested both conditions by running two different MCAs: one where it was deleted and the other where it was combined with pronoun with contracted verb. We correlated the coordinates of the individual tweets for the first ten dimensions from both sets of results and this revealed that they were strongly correlated, suggesting that there was little effect by pooling. We also compared the contribution and coordinate of 'contracted forms' and 'pronoun with contracted verb' in each analysis on each dimension to observe if the pooling led to any substantial difference. There was none. One of the benefits from including this feature by pooling it was an increase in percentage of explained variance from the eigenvalues. It was therefore decided that the feature would be pooled.
WH-clause	Pooled with complementation	All specific types of complement clauses occurred in fewer than 5 per cent of tweets. Even if the specific types were combined to form one broad category of complementation, they still did not occur frequently enough. As a result they were deleted.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 14 Feb 2025 at 05:47:30, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009208932.011

Table A2 *The short-text MDA feature set.*

TLC feature set	Feature description (incl. pooled features)
Amplifier	Refers to adverbs used to intensify the verb/adjective.
Analytic negation	Refers to 'not' plus contracted forms
Attributive adjective	Adjectives that come before the noun and any other adjective not tagged as predicative.
Auxiliary DO	Refers to any form of DO that is followed by (up to three adverbs and) a verb.
Cause subordinator complementation	Verb + *that* complement clause, noun + *that* complement clause, adjective + *that* complement clause, adjective + *to* complement clause, WH-clause.
Contracted forms	Refers to when a pronoun has the verb contracted and when the WH-word has the verb contracted
Contrastive conjunction	Refers to conjunctions that signal a contrast is being made.
Coordinating conjunction	Refers to coordinating conjunctions.
Definite article	Refers to the use of the definite article.
Demonstrative pronoun	Refers to the use of this, that, these, those as a pronoun; that is *not* followed by noun.
First person	Refers to pronouns: subject/object/possessive/reflexive and possessive determiners that refer to the first person: singular and plural plus contracted forms.
General adverb	Refers to other adverbs that are not tagged as amplifiers, downtoners, time and place adverbials, quantifying adverbs, adverbs of usuality. However, downtoner, place adverb, time adverb, quantifying adverb, adverbs of frequency/usuality are pooled.
General determiner	Refers to *this*, *that*, *these*, *those* followed by a noun (which can be preceded by adjectives, adverbs).
General interjection	Refers to tagged as an interjection, as well as negative interjections and laughter.
General noun	Refers to numeral nouns, ordinal nouns and other nouns not specified as nominalisations or proper nouns.
General subordinator	Refers to all subordinators, including those indicating time, place, condition and concession, except for cause subordinators.
General verb	Refers to any verb not specified as public, private, stance, modal (prediction, possibility), HAVE/BE main verb and auxiliary DO. (Pro-verb DO, suasive verbs, perception verbs, phrasal verbs, and modals of necessity are pooled.)
HAVE as main verb	Refers to occasions when any form of HAVE is the main verb.
Indefinite article	Refers to use of indefinite article.
Infinitive	Refers to verbs in infinitive form that are not adjective + to complement clauses. Also refers to split infinitives: verbs in infinitives form separated by adverb(s).

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 14 Feb 2025 at 05:47:30, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009208932.011

Table A2 (*cont.*)

TLC feature set	Feature description (incl. pooled features)
Initial filler	Refers to any filler (erm, em, um) at the initial position of a turn.
it	Refers to any form of pronoun *it*: contracted, reflexive, possessive and possessive determiner.
Modal possibility	Refers to modals indicating probability/possibility/ability.
Modal prediction	Refers to modals indicating prediction and BE + going to construction.
Nominalisation	Refers to when verbs/adjectives are converted into nouns.
Non-initial filler	Refers to any filler (erm, em, um) that does not occur at the initial position of a turn.
Object pronoun	Refers to use of pronouns in their objective form.
Past tense verb	Refers to verbs in their past tense form that are not in perfect aspect.
Positive interjection	Refers to 'yes'/'yeah'/etc. in the initial position of a turn.
Possession	Refers to determiners, pronouns, proper nouns and nouns which indicate possession.
Predicative adjective	Refers to adjectives which come after a copular verb.
Preposition	Refers to the use of prepositions.
Private verb	Refers to private verbs: used to encode feelings, opinions, emotions, cognition.
Proper noun	Refers to anything tagged as a proper noun.
Public verb	Refers to public verbs: used to report on speech.
Second person	Refers to pronouns: subject/object/possessive/reflexive and possessive determiners that refer to the second person: singular and plural plus contracted forms.
Stance verb	Refers to verbs used to encode stance.
Stative form	Refers to when BE is the main verb and when BE is in its copular form; that is, when it is followed by a predicative adjective. Also it refers to the use of there in its existential form and thus not as a place adverb, and also includes other copular verbs in their copula form: followed by predicative adjective.
Subject pronoun	Refers to pronouns in their subject form.
Third person	Refers to pronouns: subject/object/possessive/reflexive and possessive determiners that refer to the third person: singular and plural plus contracted forms.
Third-person singular verb	Refers to verbs ending in –s.
WH-word	Refers to use of WH-words.

different, the individual turns are the same, meaning that they can be compared. Consequently, the coordinates and contributions of the individuals in each MCA were correlated to the other to observe if there was a substantial difference between the two feature sets. For the most part, the decision to

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 14 Feb 2025 at 05:47:30, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009208932.011

delete a feature or pool it with other categories or broader features made little difference to the position of the turns, where the dimensions (at least the first ten) from one MCA were strongly positively correlated to the corresponding dimensions in the other MCA with regard to the contributions and coordinates of the individual tweets.

The following are the features occurring in fewer than 5 per cent of the turns of the TLC. These are listed with the decisions and justifications for inclusion/exclusion in the final feature set.

After this pooling process was completed, each turn was analysed for the presence or absence of the following linguistic features.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 14 Feb 2025 at 05:47:30, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009208932.011