The Language of Dialogue Is Complex

Alexander Robertson,[1] **Luca Maria Aiello,**[2] **Daniele Quercia**[2]
[1]University of Edinburgh, [2]Nokia Bell Labs
alexander.robertson@ed.ac.uk, luca.aiello@nokia-bell-labs.com, daniele.quercia@nokia-bell-labs.com

Abstract

Integrative Complexity (IC) is a psychometric that measures the ability of a person to recognize multiple perspectives and connect them, thus identifying paths for conflict resolution. IC has been linked to a wide variety of political, social and personal outcomes but evaluating it is a time-consuming process requiring skilled professionals to manually score texts, a fact which accounts for the limited exploration of IC at scale on social media. We combine natural language processing and machine learning to train an IC classification model that achieves state-of-the-art performance on unseen data and more closely adheres to the established structure of the IC coding process than previous automated approaches. When applied to the content of 400k+ comments from online fora about depression and knowledge exchange, our model was capable of replicating key findings of prior work, thus providing the first example of using IC tools for large-scale social media analytics.

1 Introduction

Integrative complexity (IC) is a psychometric that measures the degree to which a person has engaged in two cognitive processes, given a particular topic or issue (Suedfeld, Tetlock, and Streufert 1992). The first is differentiation: the recognition of multiple perspectives for the issue at hand. The second is integration: having identified such perspectives, the person demonstrates how these perspectives are connected. The lowest end of the IC spectrum is associated with inflexible, fixed perspective thinking and the highest end with integrating groups of perspectives in an elaborate, hierarchical fashion. Table 1 outlines the seven IC bands as described by Baker-Brown et al. (1990).

IC has been applied to a wide range of source materials, including diplomatic communications, political speeches, personal correspondence and legal judgments (Suedfeld, Tetlock, and Streufert 1992). As a result, it has been presented as a powerful predictor for a variety of outcomes, such as whether an international crisis will end in conflict (Suedfeld and Tetlock 1977; Suedfeld, Tetlock, and Ramirez 1977) or how far along in a term a president is (Thoemmes and Conway 2007). In addition, IC levels have been linked with many other factors, such as aggression (Winter 1993) and political preferences (Conway et al. 2016). Despite varied and interesting findings, IC is argued to be under-utilized in research (Conway et al. 2014). This is attributed to the time-consuming nature of manual scoring of texts. Furthermore, becoming qualified to determine IC requires several weeks of intensive training. Remedying these issues should therefore see IC used more often and at much larger scales. An attractive proposition, it has motivated the development of automated approaches to IC scoring. This is in spite of the perceived difficulty of the task by experts: while they state that an automated system able to perform IC scoring would be a major advance, they simultaneously warn that IC "does not rely on simple content-counting rules" and "cannot be reduced to a simple [...] content analysis system" (Baker-Brown et al. 1990).

IC	Differentiation	Integration	Details
1	None	None	No evidence of IC.
2	Emergent	None	Some acknowledgment of differing views
3	Explicit	None	At least two perspectives stated.
4	Explicit	Emergent	Connections suggested, not stated.
5	Explicit	Explicit	All perspectives connected in a new perspective.
6	Explicit	Explicit	High level of integration
7	Explicit	Explicit	Overarching perspective, detailing relationship between alternatives.

Table 1: Description of the seven levels of integrative complexity and the degree to which they exhibit evidence of cognitive differentiation and integration.

This has not discouraged attempts. Following a brief description of IC, we describe these prior attempts (§2) and we then present our contributions:

- We build and make publicly available a machine learning model for automated scoring of IC which uses syntactic features that are theoretically well-motivated by the IC framework (§3);
- We test our model on the official IC scoring test and achieve state-of-the-art results, with a F1 score of almost 25% higher than previous approaches (§4);
- We conduct for the first time an analysis of IC at scale by applying our tool to over 400k textual snippets from Reddit. Results obtained on a support-based forum focused on

mental health match theoretical expectations, thus providing initial evidence of external validity to our tool (§5) and setting the stage for its usage in the context of large-scale social media analytics (§6).

2 Prior work

2.1 Linguistic style analysis

The study of linguistic style in text and conversations has been related to a number of outcomes.

On Twitter, researchers have investigated the use of specific markers that are predictive of the initiation of a conversation (Boyd, Golder, and Lotan 2010) and found that linguistic affinity between participants fosters continued engagement (Budak and Agrawal 2013). From text, one can also predict more intangible properties associated to verbal expressions. Multimodal features of online threads can predict the perceived interestingness of the themes discussed (De Choudhury et al. 2009). The combined use of topic detection and sentiment analysis on Twitter has been used to extract higher-level emotional properties such sympathy, apology, and complaint (Kim, Bak, and Oh 2012). Statistical stylometry has been used to evaluate the quality of literary writing and to identify successful pieces of literature (Ashok, Feng, and Choi 2013).

Conversation style has also implications on the social processes involving participants. The evolution of discussion topics over time unearths patterns of social identity and cohesion (Purohit et al. 2014). The language complexity, and emotions expressed in a conversation (Danescu-Niculescu-Mizil et al. 2012; Tchokni, Séaghdha, and Quercia 2014) echoes the power differential of participants. A considerable amount of work has been done to understand the connection between linguistic style and conflict. Linguistic cues such as markers of agreement and confidence distinguish between productive and unproductive discussions (Niculae and Danescu-Niculescu-Mizil 2016). Rhetorical prompts deployed in the very first conversation exchanges are predictive of emergence of conflict (Zhang et al. 2018). Antisocial behavior is also impacted by the mood of the context surrounding the discussion (Cheng et al. 2017) and exacerbated by individuals who attempt to steer the discourse towards irrelevant topics (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2015).

In this work, instead of studying linguistic markers that create conflict, we focus on how language can bring reconciliation and peace by recognizing different points of view and integrating them.

2.2 Automated IC methods

The success of prior work runs counter not only to the technological predictions of experts, but also to the theory that underpins IC. Specifically, that it is a measure of *structure* rather than *content*. Integrative Complexity, as scored by skilled humans, is concerned not with *what* we say, but *how* we say it. The two extant automatic coding methods, detailed below, are both focused on content.

Ambili and Rasheed (2014) trained models to predict low, medium or high IC using as features the text length, the vocabulary used, and a metric based on the semantic coherence of the text (Li, Bandar, and McLean 2003). Using the words in the first sentence of a text, pairwise comparisons with all subsequent words was performed on the basis of their connection in the WordNet (Miller 1995) knowledge-base. Specifically, the normalized product of the minimum path length between two words and the depth in the knowledge-base hierarchy of their lowest common hypernym. This value, along with text length and one-hot encoding of the words in the text, were used as features in a variety of classifier models. The system achieved an F1 score of around 0.8.

Conway et al. (2014) created a rule-based system based on the presence of specific vocabulary items thought to be indicative of differentiation or integration. The count of each word is then weighted by custom values. If the differentiation keywords do not reach a threshold of 3, then no integration keywords are considered in calculating the final score. Because weights are real-valued, the system does not output classes but real numbers between 1 and 7. The system was evaluated using correlation with human scorers, as well as by attempting to replicate prior outcomes on small human-coded datasets.

2.3 Challenges

Each of these approaches adjusts the structure and assumptions of the IC system. Ambili and Rasheed reduce the number of classes from seven to three bins: low (1,2), medium (3,4,5) and high (6,7). Although this reduction in task difficulty was motivated by lack of training data, the bins are not well aligned with the structure of the IC bands they represent. A band of 1 is not simply "low" but a complete absence of IC, while bands 3 to 5 represent very different levels of differentiation and integration.

Similarly, the real-valued output of Conway et al. does not strictly align with the bands of the IC scoring system. Each band is a label for the absence or presence of particular properties and there is no provision in the literature for non-integer scores (Baker-Brown et al. 1990). It is not appropriate to use linear correlation methods to compare these pseudo-continuous variables to categorical ones, since this may result in higher correlations than when the real-valued predictions are rounded to become integer values.

The seven IC bands in Table 1 are most properly treated as ordinal variables. Differences between bands cannot be numerically quantified. It is not possible to claim that the difference between an IC band of 2 and 3 is equivalent to that between 3 and 4. This enables us to cast the scoring task as a classification problem, with seven distinct class labels.

Last, and most importantly, both previous approaches focus on semantics and ignore syntax (*how* people express concepts). In the following, we will show that syntactic information is crucial to generalize automated scoring to text whose nature is very different from the text upon which the model training is performed.

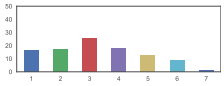
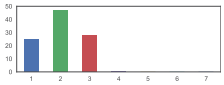
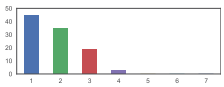
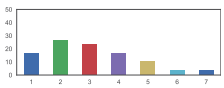
Material	Source	Texts	Tokens/text	Usage	Score distribution
Official IC Practice Sets	Suedfeld, 1992	156	μ 57.5 (σ 41.7)	Training	
Heritability	Conway, 2014	310	μ 92.7 (σ 52.4)	Training	
Early Christian Writings	Conway, 2014	173	μ 117.6 (σ 68.1)	Training	
Official IC Coding Test	Suedfeld, 1992	30	μ 72.7 (σ 30.7)	Evaluation	

Table 2: Datasets used in experiments, and their properties.

3 Methodology

3.1 Training data

For training and evaluation, we use the datasets outlined in Table 2. We follow Conway et al. in using some sources as training material, while leaving others aside for evaluation. The Official Practice Sets and Official Coding Test are taken from the Suedfeld's Electronic Complexity Workshop. The rest were kindly supplied by Conway: the Heritability texts cover student responses to prompts on a range of topics (religion, death, assertiveness); Early Christian Writings are randomly sampled from the New Testament. All IC scoring was performed by trained humans, with texts being scored by multiple scorers. The "unofficial" materials are somewhat longer and more numerous, but also show a distinct lack of variety in the range of IC bands represented. Where there was slight disagreement between scorers (the IC handbook permits a difference of up to 2) for a particular text, the average was taken—we have therefore rounded these scores to the nearest whole integer. Examples of texts belonging to the 7 IC bands are reported in Table 10, in the Appendix.

3.2 Features sets

The automated techniques described in §2.2 take a semantic approach. Semantic features include actual words or phrases, information about senses of words or phrases, or information about classes of words in terms of their meaning (e.g., whether they are positive or related to a particular topic). An alternative, closer to *how* than *what*, is a syntactic approach. We therefore distinguish two types of features along these lines. Syntactic features include more abstract properties, related to the way language is meaningfully structured. For example, the syntactic role played by a word within a text (e.g., noun, adjective) or the syntactic relations between words (e.g., direct object of a verb). Below, we detail both the semantic features (vocabulary-based) and the two families of syntactic features (POS tags and dependency subtrees) we considered. We extract syntactic features using the CNN-based tagger provided by the spaCy python package, which has an accuracy of 97% for POS tagging and 90% for dependency labeling on English texts.

Vocabulary. Following the approach of Conway et al., we use the IC handbook (Baker-Brown et al. 1990) to identify key phrases which are said to be associated with each particular IC band. The original set of words includes some "content-insensitive" words (adverbs like *however* or *yet*) and some words that directly refer to differentiation and integration processes (e.g., *compromise*, *compensate*, *reconciliation*). We expand this list with synonyms and related terms by searching for each key phrase in the ConceptNet knowledge-base (Speer, Chin, and Havasi 2017). We clean the resulting expanded list by manually filtering out items not likely to be linked to differentiation/integration. This is a binary feature, representing the presence or absence of a given vocabulary item. Rather than creating a feature for each form a word or phrase can take, we lemmatize all keywords and search for these in a lemmatized version of the input text. Text length has been shown to be vacuously predictive of IC (Baker-Brown et al. 1990), so we use binary features rather than counts or weighted counts as in Conway et al. (2014)—to lower the risk of implicitly encoding the length of the the text as a feature. Finally, we compute two binary features, indicating the presence of vocabulary related to differentiation or integration. We extract 312 semantic features.

POS tags. Using the Penn Treebank tag set, tokens in a text are labelled according to their syntactic part of speech (e.g., as particular types of nouns, adjectives or verbs). Counts of these tags are then normalized by the total number of words in the text. We extract 45 syntactic features.

Dependency subtrees. A dependency parser labels the relationship between words. This relationship shows which words modify others: if word A is modified by word B, then A is a child of B. These relations form a graph. By starting at each node in the graph and determining its descendants, subtrees can be extracted. In these subtrees, edges between words are labeled with the type of relationship between the two words. For example, the sentence "the cat sleeps" is converted to $sleeps \xrightarrow{\text{nsubj}} cat \xrightarrow{\text{det}} the$. The article *the* is the determiner (det) of the noun *cat* and *cat* is the nominal subject (nsubj) for the verb *sleeps*. The resulting features encode

only the labels on the edges, not the labels in the nodes. In the example, the feature extracted are *nsubj*, *det* (subtrees of length 1) and *nsubj_det* (subtree of length 2). Similar to vocabulary features, subtree features are binary, set to 1 if the subtree is detected at least once in the text. Due to the small size of our training datasets and having the goal of extracting complex syntactic structures, we extract subtrees with a number of edges up to 5. Prior work by Vosoughi and Roy (2016) used a maximum length of 2, but were working with tweets—not only a much larger dataset, but also much smaller individual texts. We keep as features only those subtrees which appear in the training datasets with a frequency of at least 5. Counts of subtrees are normalized by the total number of subtrees extracted. We extract 280 syntactic features.

LIWC. LIWC is a standard dictionary of 2,300 English words grouped in 72 categories (Tausczik and Pennebaker 2010). These categories are generally abstract and aim at capturing markers of emotional and psychological expressions that do not hinge on the particular topic of the text. Examples of categories include expressions focused on the future or verbs referred to the human perceptual sphere. Each word may belong to multiple categories. Similar to the vocabulary features, we use 72 binary scores rather than counts: each feature is set to 1, if at least a word in the respective LIWC category is present in the text.

4 Classification results

4.1 Experimental setup

Classifier. We evaluate each feature set by training an ensemble of decision trees with gradient boosting (Chen and Guestrin 2016), implemented by the xgboost package. This model is well-suited to the small datasets we are working with, makes it easy to interpret the contribution of individual features, and is able to ignore any vacuous features that may be present. This last property is useful to avoid overfitting since some feature sets are large.

Baselines. We compare our classifier against four baselines. First, we include a simple baseline that always predicts the most common class. Second, we consider a method that uses text length, measured as the number of words, as the only feature. Third, we use text sentiment. Among the many sentiment analysis tools available, we chose Vader (Gilbert 2014), a state-of-the-art sentiment analysis technique widely used on noisy text. Vader outputs a score from -1 (most negative) to 1 (most positive). Last, we compare our results against the AutoIC tool (Conway et al. 2014). AutoIC has been previously trained by Conway on the same data we use, and that makes it an appropriate choice for a fair comparison. AutoIC outputs real numbers rather than integers. However, IC bands as described in the IC handbook (Baker-Brown et al. 1990) represent whether particular properties are present/absent, rather than having a continuous value. Failing to reach a threshold value is therefore best considered as evidence of failing to detect sufficient evidence for that threshold. We therefore we apply the floor function to AutoIC's output.

Evaluation metrics. We compare different approaches and feature sets using the F1 score. For each of the 7 classes, F1 score is calculated. The harmonic mean is then weighted by the total number of examples of that class, with the overall average reported. Casting IC scoring as a discrete classification problem is justified by the official IC handbook, which makes no provision for interpreting IC as a continuous variable. Additionally, we provide confusion matrices for each feature set to broadly illustrate the strengths and weaknesses of each with regard to the seven IC classes. Finally, in order to measure how close a model's predictions are to the true class labels we calculate the Mean Squared Error (MSE) between the two. This is possible because the class labels are ordinal, even if not continuous, and have a sensible ordering. However, we acknowledge that the magnitude of the difference between adjacent class labels is not likely to be linear. Misclassifying an IC text as 3 instead of 2, where the difference is based purely on degree of differentiation, should not be penalized to the same degree as misclassifying as 4 instead of 3, where the difference is now due to integration.

4.2 Cross-validation on training data

In the first experiment, we use 5-fold cross-validation on the training set and report the mean F1 score (with standard deviation) achieved across all folds, for each feature set individually and some of them in combination, along with the baselines. Figure 1 shows the mean F1 score over all five cross-validation folds. The sentiment and word-count approaches are the worst-performing ones after the naive baseline. When combined with other feature sets, they never yield any performance improvement. Syntactic features perform better (LIWC especially) and work also well in combination, being able to achieve a F1 of .445 when put together. In this case, semantic approaches perform roughly as well as syntactic features combined. Merging syntactic and semantic feautures yields no improvement on the results. The imbalanced dataset is problematic: few models can correctly predict IC above 4. The only notable exception is Vocabulary, which manages to identify 4 (out of 37) such documents. AutoIC fares better: 5/37 for band 4, 1/20 for band 5, 1/2 for band 7 (Figure 2). This may be attributed to the method used by AutoIC to weight vocabulary items in terms of their contribute to differentiation/integration, compared to our binary approach. AutoIC may also have a broader list of keywords which are coincidentally present in the higher IC texts. The mean MSE values across all cross-validation folds for selected feature subsets are shown in the first row of Table 3. AutoIC's lower error is due to the fact that their model predicts a wider range of bands, whilst still failing to classify them correctly. In general, model predictions are within a reasonably tight range of the true labels.

4.3 Prediction on heldout data

In the second experiment, we retain all the data used in the first experiment for training and tuning. We fine-tune model parameters through cross-validated grid search. A large ensemble of relatively deep trees (500, with a maximum depth of 6), with 80% subsampling, gave the best performance. Subsampling sets the proportion of training data, selected

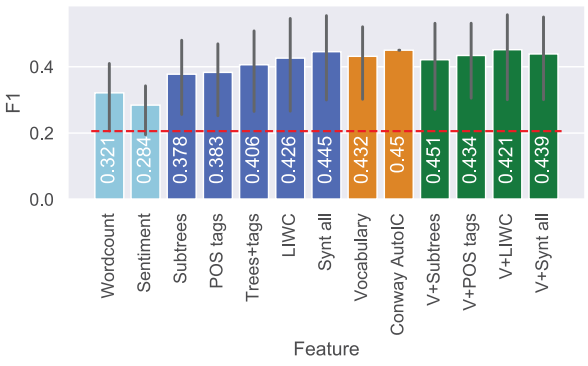


Figure 1: Classifier performance on the training dataset on 5 cross-validation folds, in terms of F1 score. Error bars show variance across folds. Wordcount and sentiment baselines are in light blue, syntactic features in blue, semantic in orange, joint in green. The red line shows the baseline result for the most frequent class predictor.

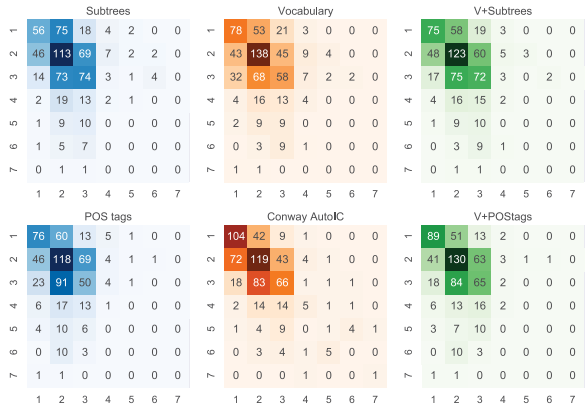


Figure 2: Confusion matrices for each classifier trained on syntactic (blue), semantic (orange) or joint (green) features, evaluated on the training datasets summed across cross-validation folds. Rows are the true classes, columns are classifiers' predicted classes.

	AutoIC	Subtrees	POS tags	Vocab.	V+Subtrees	V+POStags
Cross-val.	1.08	1.63	1.70	1.59	1.43	1.48
Heldout	1.33	2.87	2.13	3.40	2.60	1.80

Table 3: Mean squared error for predictions made in the cross-validation and heldout experiments. Result for Conway's AutoIC is shown for reference.

at random from all available data, that each decision tree receives during training and helps to reduce the effect of overfitting a model to the data. We then train a new model, one per feature set, with the best parameters. These models are evaluated using the official IC coding test (last row in Table 2). Crucially, the creators of AutoIC did not use the official coding test to manually identify any differentiation/integration words or phrases, making this an especially fair comparison of systems.

Figure 3 shows the F1 score on the official IC scoring test. Unlike in the first prediction experiment, some syntax-based models—in particular, POS Tags and all syntactic features combined—outperform vocabulary features and are comparable to AutoIC. LIWC's F1 drops considerably compared to the cross-validation setting. The best-performing model uses both semantic and syntactic features to achieve an F1 score of 0.462. This is a non-negligible improvement over AutoIC, which scores an F1 of 0.400.

The confusion matrices in Figure 4 show that while most models correctly classify IC band 1, no model correctly predicts an IC band higher than 4. Also, dependency Subtrees are the only features which result in a model producing many predictions above IC band 4, though not with any accuracy. MSE is shown in Table 3. Again, Conway's AutoIC makes predictions which are numerically closer to the real class labels (again, due to the wider range of incorrect class labels predicted) but this does not translate into better F1 scores. This outcome highlights the importance of reporting multiple relevant metrics in order to show a more nuanced picture of model performance.

For the best-performing model, a classification report is shown in Table 4. Performance is generally high for classes 1 to 4, but no predictions above these bands are made. This is not especially surprising, since these higher bands are generally very rare and appear mainly in the official training materials for IC scoring. When compared to Conway's AutoIC, a similar situation for high IC bands emerges. In general, however, the combination of semantic and syntactic features significantly outperforms the purely semantic AutoIC system.

4.4 Feature analysis

We now examine the role played by semantic and syntactic features. For the Vocabulary and V+POStags feature sets, we look at three aspects. First, which features are *never* used by any decision trees in the ensemble. Second, the importance attached to features which *are* used as measured by averaging the information gain of the feature across all decision trees in the ensemble. Last, in instances of correct classifications by a model, examples of which features are most discriminative for a particular IC band.

Vocabulary. Of the 313 features, only 89 are used. Unused features are either extremely common words or not present at all. The majority of top features are differentiation keywords. The Vocabulary model correctly classified test items in three IC bands only (Table 5). For band 1, the most important feature is that denoting the absence of any differentiation-related words. For band 2, this same feature also plays a role though does not appear in the top 10. Here, the presence of the differentiation terms themselves are of more importance. And for band 3, the binary feature for integration-related vocabulary, set to 0, is most important alongside the differentiation keywords. The failure of the Vocabulary model to correctly classify the band 4 text is

		V+POStags			Conway AutoIC			Difference		
IC band	Support	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
1	5	0.71	1.00	0.83	0.57	0.80	0.67	+0.14	+0.20	+0.16
2	8	0.45	0.62	0.53	0.45	0.62	0.53	+0.00	+0.00	+0.00
3	7	0.33	0.43	0.38	0.40	0.29	0.33	-0.07	+0.14	+0.05
4	5	1.00	0.6	0.75	0.5	0.40	0.44	+0.50	+0.20	+0.31
5	3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Average	0.48	0.53	0.49	0.39	0.43	0.40	+0.09	+0.13	+0.09

Table 4: Classification report for the best-performing model, V+POStags, on the official 30 item IC coding test. Results of Conway's AutoIC tool are provided for reference.

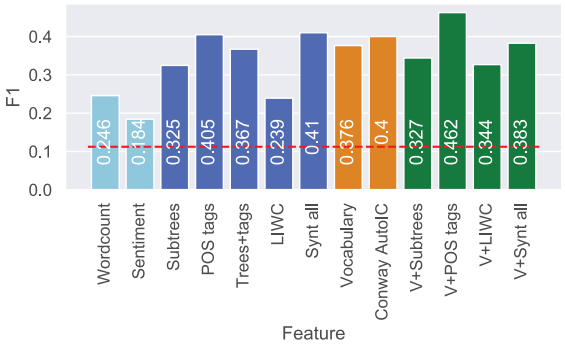


Figure 3: Classifier performance on the heldout dataset, measured by F1 score. Word count and sentiment baselines are in light blue, syntactic features in blue, semantic in orange, joint in green. The red line shows the populous baseline.

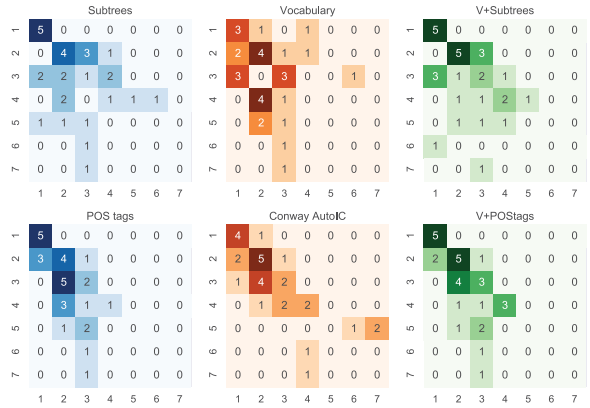


Figure 4: Confusion matrices for each classifier trained on syntactic (blue), semantic (orange) or joint (green) features, evaluated on the heldout dataset, summed across cross-validation folds. Rows are the true classes, columns are classifiers' predicted classes.

easily explained: it contains very few of the expected vocabulary items.

V+POStags. A total of 98 features are used. Of the 45 POS features, 10 are never used because too frequent or too rare. Only 63 vocabulary items are used, compared to the 89 used in the vocabulary-only model. Overall, the most important syntactic features are adjectives and adverbs. This is not surprising, given that the majority of differentiation/integration keywords are in this category. Predeterminers, which are found in phrases such as "all this", "what a" and "many times the", are also important. Even though semantic features result in the greatest information gain during training, syntactic features turn out to be the most useful in the evaluation. Table 6 shows the features which most contributed to a correct classification. The band 1 text lacks many syntactic features (shown as having a value of 0.000) and even though there is some evidence of differentiation, these features are negatively weighted. This may be attributed to the fact that the presence of keywords is not necessarily evidence of IC. For IC bands 2 and 3, the distinct lack of integration features is important, along with the presence of some syntactic features. The band 4 text, by comparison, has fewer zero-value features and may be considered a more complex text in terms of syntax, which here also belies its higher level of IC.

4.5 Sensitivity analysis

To adhere to the original theoretical framework as much as possible, our operationalization of IC follows a 7-class scale. In addition to that, we check the robustness of our results to coarser IC aggregations, also motivated by previous work in which fewer and wider bands of IC were used (Ambili and Rasheed 2014). We first tried a 4-way classification: no IC (class 1 only), low IC (classes 2 and 3 joint), medium IC (4 and 5), high IC (6 and 7). We then collapsed the low and medium classes to try a ternary classification. We always kept class 1 separate (and did not merge it with class 2, for example) because it has a very distinctive meaning: it is the only class that exhibits no sign of differentiation or integration.

Results are presented in Figure 5. For brevity, we report the results for the main baselines and for the best performing

IC = 1			IC = 2			IC = 3		
Feature	Value	Contribution	Feature	Value	Contribution	Feature	Value	Contribution
has_diff	0.0	+0.953	dif_too	1.0	+1.155	Bias term	1.0	+0.872
Bias term	1.0	+0.419	Bias term	1.0	+0.594	has_int	0.0	+0.450
dif_but	0.0	+0.114	dif_consider	1.0	+0.491	dif_may	1.0	+0.393
dif_because	0.0	+0.058	dif_however	1.0	+0.186	dif_but	1.0	+0.306
dif_how	0.0	+0.033	dif_how	1.0	+0.092	dif_hope	1.0	+0.215
dif_yet	0.0	+0.033	dif_hope	0.0	+0.027	dif_while	0.0	+0.060
int_unity	0.0	+0.029	dif_perhaps	0.0	+0.019	dif_rather	0.0	+0.058
dif_depend	0.0	+0.024	dif_almost	0.0	+0.018	dif_too	0.0	+0.033
dif_hope	0.0	+0.022	dif_sometimes	0.0	+0.012	dif_seem	0.0	+0.030
dif_rather	0.0	+0.022	dif_although	0.0	+0.011	dif_differ	0.0	+0.026
has_int	0.0	-0.009	dif_while	0.0	-0.032	int_remain	0.0	-0.024
dif_close_to	0.0	-0.009	dif_rather	0.0	-0.033	dif_separate	0.0	-0.026
dif_seem	0.0	-0.010	dif_different	0.0	-0.050	int_weigh	0.0	-0.028
dif_consider	0.0	-0.011	dif_often	0.0	-0.050	dif_possible	0.0	-0.029
int_account	0.0	-0.012	dif_each	0.0	-0.050	int_unity	0.0	-0.031
dif_secret	0.0	-0.013	dif_either	0.0	-0.069	dif_however	0.0	-0.036
dif_differ	0.0	-0.018	dif_about	0.0	-0.078	dif_often	0.0	-0.037
dif_usually	0.0	-0.019	dif_both	0.0	-0.079	dif_about	0.0	-0.068
int_remain	0.0	-0.024	dif_because	0.0	-0.284	dif_though	0.0	-0.073
dif_may	0.0	-0.036	dif_but	1.0	-0.575	dif_because	0.0	-0.102

Table 5: Top ten and bottom ten features used in successful IC classifications using vocabulary features. Differentiation and integration terms are prefixed with *dif* and *int*, while has_dif and has_int are the binary features for whether any differentiation/integration terms are present at all. The bias term is the averaged sum of the value associated with each root node in the ensemble.

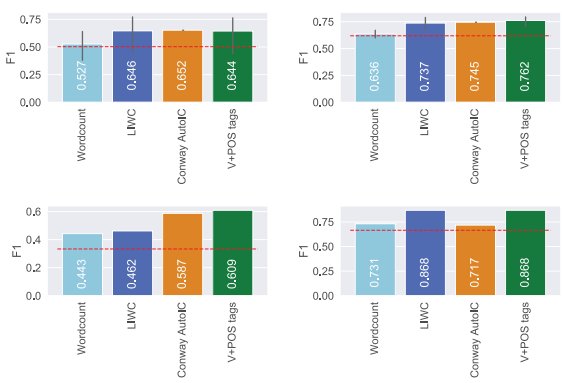


Figure 5: Classification results by aggregating the 7 IC bands into fewer classes. Left: 4-class classification (1, 2+3, 4+5, 6+7). Right: 3-class classification (1, 2+3+4+5, 6+7). Top: results of cross-validation. Bottom: results of classification on the heldout dataset.

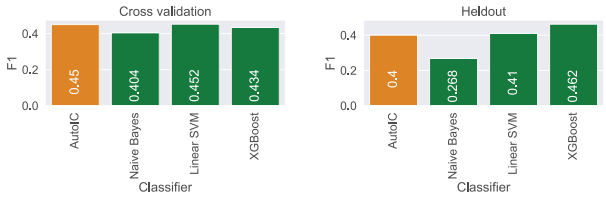


Figure 6: Comparison between classifiers in cross validation (left) and heldout classification (right) using the V+POS Tags features. AutoIC is reported for comparison.

model. The pattern is consistent with that of the 7-band classification. Word count is the weakest approach; LIWC, Conway's IC, and V+POS Tags perform comparably in cross validation; V+POS Tags is always the best approach in the classification on the heldout dataset.

In addition, we also experiment with other classification models to check the robustness of our approach. We compare xgboost with Naive Bayes as a simple baseline and with linear SVM, which achieves better generalization than decision trees. Results are reported in Figure 6. As expected, Naive Bayes is the weakest approach. Linear SVM is second to xgboost and keeps a rather decent performance, being able to slightly beat AutoIC both in the cross validation and heldout setups.

4.6 Classification summary

A purely syntax-based approach to IC scoring, which is well-motivated given the theory behind IC, performs reasonably well on held-out evaluation data. However, this is not sufficient to surpass the performance of the current state of the art as represented by the semantic-based approach of Conway's AutoIC tool. This is only possible by combining syntax and semantics together in one model. This allows the model to look beyond the surface forms of language, the

IC = 1			IC = 2			IC = 3			IC = 4		
Feature	Value	Contribution	Feature	Value	Contribution	Feature	Value	Contribution	Feature	Value	Contribution
Bias term	1.000	+0.894	Bias term	1.000	+0.786	Bias term	1.000	+1.150	Noun, singular	0.194	+0.506
Verb, present	0.000	+0.672	dif_too	1.000	+0.743	Verb, base form	0.130	+0.308	Adj.	0.081	+0.441
Coord. conj.	0.022	+0.642	Verb, past	0.000	+0.208	Noun, plural	0.056	+0.235	Subord. conj.	0.145	+0.394
Particle “to”	0.000	+0.513	Adj.	0.051	+0.207	Noun, singular	0.056	+0.230	Verb, base form	0.016	+0.340
Preposition	0.000	+0.497	Noun, singular	0.120	+0.143	Subord. conj.	0.074	+0.132	Verb, past part.	0.016	+0.336
Verb, past	0.067	+0.338	Verb, base form	0.043	+0.129	Verb, past part.	0.000	+0.129	Comp. adj.	0.016	+0.186
Proper noun	0.222	+0.317	Preposition	0.060	+0.121	Coord. conj.	0.037	+0.109	Verb, n3ps pres.	0.016	+0.163
Subord. conj.	0.111	+0.259	Possessive	0.000	+0.052	dif_but	1.000	+0.081	Determiner	0.081	+0.149
Determiner	0.111	+0.161	Modal verb	0.034	+0.030	has_int	0.000	+0.078	dif_but	1.000	+0.136
Comp. adj.	0.000	+0.136	has_int	0.000	+0.027	dif_too	0.000	+0.066	Verb, present	0.000	+0.133
Modal verb	0.000	-0.020	Noun, plural	0.026	-0.113	Verb, n3ps pres.	0.056	-0.098	dif_yet	0.000	-0.034
Adv., superlative	0.000	-0.021	Determiner	0.094	-0.129	Proper noun	0.000	-0.120	Particle “to”	0.016	-0.035
Particle	0.000	-0.060	Adverb	0.060	-0.133	Verb, past	0.000	-0.122	Existential “there”	0.000	-0.047
has_diff	1.000	-0.251	Verb, past part.	0.017	-0.133	Adverb	0.093	-0.142	dif_how	0.000	-0.063
Adverb	0.022	-0.255	Verb, present	0.026	-0.137	Preposition$	0.019	-0.152	Verb, gerund	0.048	-0.077
Verb, gerund	0.022	-0.281	Verb, n3ps pres.	0.017	-0.205	Modal verb	0.074	-0.178	Coord. conj.	0.065	-0.134
Verb, past part.	0.000	-0.342	Proper noun	0.017	-0.296	Particle “to”	0.056	-0.262	Wh-adverb	0.000	-0.196
Verb, n3ps pres.	0.000	-0.501	Possessive pronoun	0.043	-0.341	Adj.	0.111	-0.268	Modal verb	0.000	-0.258
Adj.	0.111	-0.716	Subord. conj.	0.162	-0.341	Determiner	0.056	-0.282	Noun, plural	0.113	-0.524
Wh-determiner	0.022	-0.861	Adj., superlative	0.017	-1.046	Verb, present	0.000	-0.315	Bias term	1.000	-0.818

Table 6: Top ten and bottom ten features used in successful IC classifications using V+POStags features. The bias term is the averaged sum of the value associated with each root node in the ensemble.

what of what is being said, and leverage the syntactic properties of the text to access the *how* of what is being said. Limited training examples and stylistic differences within the training data that is available accounts for why syntax alone is insufficient and must be bootstrapped, to some degree, by a lexicon of differentiation and integration. The best-performing model represents an almost 25% improvement over the current state of the art. As a contribution to the community, we make our model public and open-source (social-dynamics.net/ic).

5 Measuring IC in social media

After testing our model on annotated data, we apply it on a larger set of posts and comments from Reddit with the goal of building initial evidence about its external validity. Based on previous literature, we set an hypothesis about the level of IC that we expect to find under certain conditions. This is the first time an analysis of Integrative Complexity is conducted at this scale. Although Conway's AutoIC tool has been used in research (McCullough and Conway III 2018a; 2018b), it has not been tested in this manner and on large-scale data.

5.1 Data

Reddit is a social media site focused on news aggregation and user discussion. It is organized into themed *subreddits* where users submit posts for others to both comment and vote on. We focus on the three subreddits in Table 7, which allow only text-based posts and are focused on particular forms of discussion. Using the Reddit API to collect all posts and comments made between January 2018 and August 2018, we gathered data from /r/depression, a support-based subreddit focused on mental health, as well as two subreddits where the focus is on high quality responses to specific questions. The lower number of posts/comment in /r/AskHistory and /r/AskScience is due to both stricter controls on quality of submissions and a much narrower focus.

Subreddit	Subscribers	Posts	Comments
/r/depression	380k	8k	212k
/r/AskHistorians	790k	1.6k	42k
/r/AskScience	15.9m	1.2k	143k

Table 7: Data collected from Reddit.

5.2 Hypothesis

Findings of Suedfeld and Bluck 1993 show that IC for personal writings made during periods of severe personal distress (e.g., following the death of a loved one or a betrayal) was higher than those written under normal conditions, before the negative event. Positive events (e.g., marriage, career success) had no effect. Based on these findings, we hypotesize that texts from the /r/depression subreddit, where users write about their experience of depression, often triggered by difficult personal circumstances, grief, and other traumas (De Choudhury and De 2014), exhibit higher IC than what is measured in other discussions about non-dysphoric experiences. We therefore compare texts from /r/depression with other two communities, /r/AskScience and /r/AskHistorians, which are focused on knowledge exchange rather than sharing negative experiences and providing social support. As the latter two subreddits include threads rich of content and debates about complex and possibly controversial themes (e.g., how the universe was originated and will end), we expect them to contain a non-negligible number of posts with some level of integrative complexity. However, because the theory predicts that being in a state of psychological distress adds an additional layer of complexity to a person's reasoning, we predict that texts from /r/depression will exhibit markedly higher IC scores.

5.3 Results

We apply the best-performing classifier (V+POStags features), trained on all available training data, to the post and comment text of the subreddits. The resulting probability

Figure 7: Probability distribution of IC values in the posts (left) and comments (right) of the three subreddits considered.

Subreddit	Posts	Comments
/r/depression	μ 2.06 (σ 0.74)	μ1.61 (σ 0.71)
/r/AskHistorians	μ 1.34 (σ 0.67)	μ1.47 (σ 0.74)
/r/AskScience	μ 1.44 (σ 0.73)	μ1.52 (σ 0.71)

Table 8: Mean IC scores for posts and comments, per subreddit.

distributions over IC scores is shown in Figure 7. Even if posts with very high scores are globally rare, we found a variety of scores in all the subreddits. Not surprisingly, the majority of Reddit posts analyzed belong to class 1—no sign of IC at all. The distribution in /r/depression is more right-skewed than those from the other two subreddits, which provides a first indication of higher IC levels in depression-related discussions. The average IC scores in posts and comments, per subreddit, are summarized in Table 8. Examples of high and low IC texts for the different online communities are shown in Table 9.

Our classifier does not take into account text length to estimate IC classes and, as a result, short texts can be scored as highly complex. In Table 9, for example, we report two comments from /r/depression where the shorter one has much higher complexity than the longer one. However, it is known that text length might correlate with IC (Baker-Brown et al. 1990). It is harder to compress a lot of information about differentiation and integration into a short piece of text and, as a result, longer texts will be associated with higher IC, on average. To make sure that the difference in average IC across subreddits is not due only to differences in text length, we contrast the level of IC of posts and comments across subreddits for texts of comparable length (Figure 8). Consistent with what theory would suggest, we observe that depression-related posts have higher IC than history- or science-related posts. A similar trend emerges for comments. The gap between them is not prominent for very short posts (those roughly under 10 words, represented in the first bin in Figure 8). As the text length increases, the overall levels of IC rise but sensibly more so for depression-related posts. Although text length and IC correlate in this specific Reddit sample, large differences in IC levels for texts with similar length do exist.

Last, in addition to looking at the differences in IC across subreddits, we also look at how users reactions are associated to these differences as a first example of exploratory

Source	IC	Text
Depression	H	I am so lucky to have friends who understand me. But I regret telling my recent ex about my depression. He used to understand, but I mean I get it. He was mentally unstable too.
Depression	L	A lot of people are lurkers. Doesn't mean they don't care, they just don't know what to say and sometimes that's better than saying something bad. Just write whatever you want, like let it all out, maybe someone out there feels the same way and they just didn't want to write it too.
AskHistory	H	The Lewis and Clark parallel is spot on. Not everyone in 1830s America was wilderness-capable, but the percentage would be much higher than it is today. Hunting and foraging was common in the frontier, which in this era started around Wisconsin, Illinois, and west. It's also not too hard for a single person with a rifle and a fair amount of survival skills to scrabble through in the wild. Lansford Hastings was a politician in the Republic of California - he knew just enough trailblazing to sell a plausible trail, but not to ensure its safety.
AskHistory	L	So I'm looking to teach myself an ancient language, and would like to translate text that has yet to be translated yet. Are there any projects out there that need help?
AskScience	H	Nonsense. Both your esophagus and trachea are valve protected (and the pressure would be preferentially released through your nose and lips but even if it wasn't) there are sphincters around your stomach and there is your anal sphincter and all of these are well able to retain an atmosphere of pressure. If anything your difficult in this scenario would be exhaling without atmospheric pressure to help your diaphragm and supporting muscles cause your rib cage to contract.
AskScience	L	What happens exactly with the stability of therapeutic proteins when kept at room temperature?

Table 9: Examples of Low ($<$3) and High ($\geq$3) IC texts from specific subreddits.

analysis that could be enabled by the application of our tool at scale.

In Reddit, the community votes on content, with popular content having a higher score than bad. Figure 9 shows the distribution of Reddit scores per IC band. Because of the very long tail of these scores, we use a value plot (Hofmann, Wickham, and Kafadar 2017). This non-parametric visualisation displays more percentiles than a standard boxplot, with each percentile represented as a box whose height denotes the range of values within that percentile. To emphasize the trend we display a linear regression line. In /r/AskHistorians and /r/AskScience, better answers tend to be more integratively complex. These are competitive subreddits where users spend time writing informative comments which meet the rules of the community and answer the question at hand and complex answers are likely to be rewarded. On the contrary, /r/depression is a support subreddit where there is no explicit motivation to reward complex comments: the main purpose of comments in this subreddit is that of social support rather than of discussion enrichment with new perspectives (De Choudhury and De 2014).

6 Discussion and conclusion

The language of dialogue is complex: people who are able to recognise the differences between conflicting points of view and to identify paths for their reconciliation use not only the right words but they combine them wisely. We capture that by designing a system for automatic scoring of Integrative Complexity that blends semantic and syntactic features. Its accuracy is sufficient to replicate findings from prior work, showing that people experiencing personal crisis exhibit higher levels of IC than those who are not. This

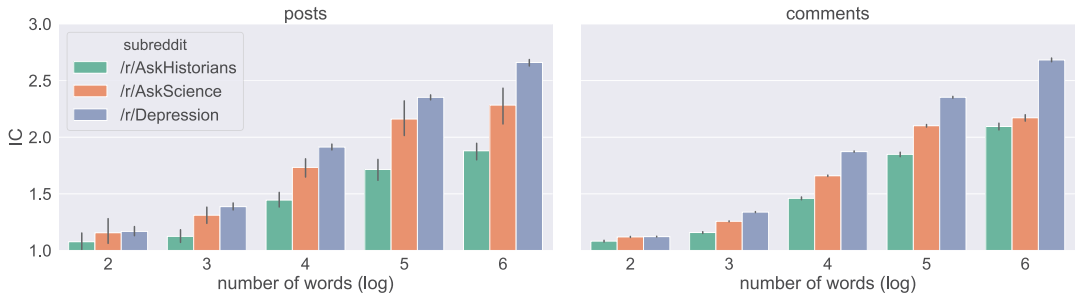


Figure 8: Average IC of posts (left) and comments (right) in the three subreddits considered, binned by text length (log of the number of words, rounded to the next integer). Depression-related posts and comments have higher IC compared to texts of comparable length from the other two subreddits. Bars show the 95% confidence intervals of the mean.

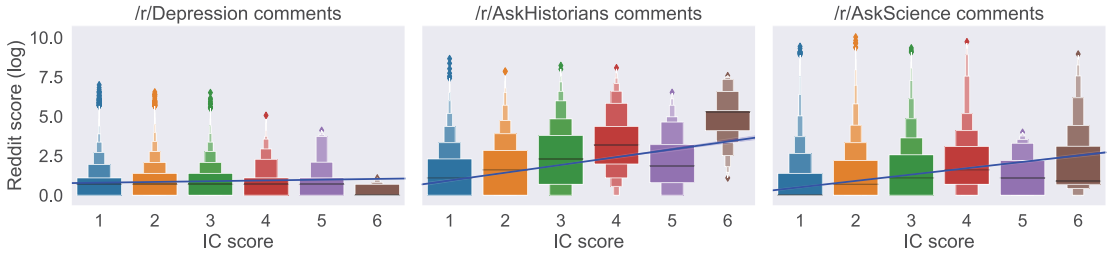


Figure 9: Value plots showing the distribution of Reddit comment scores for each IC band, for each subreddit. The black lines shows the mean, with each box representing a percentile. The blue regression line shows the linear relationship between IC and Reddit score.

validated advance will hopefully encourage the application of this measure of cognitive processes to a wider range of situations. Next, we discuss some of the limitations of our work and some of the opportunities it opens up.

6.1 Implications

This work results in theoretical and practical implications. From the theoretical standpoint, this work reinforces the evidence that IC can be operationalized and that it can be done most effectively when language syntax is brought into the equation. From a methodological perspective, by casting the IC scoring task as a classification problem and generating labels rather than real-valued output, as in Conway et al. (2014), our approach more closely follows the methodology laid out in the IC coding manual (Baker-Brown et al. 1990).

By opening our method to the research community, we start paving the way towards important practical applications in social media analytics. In the last decades, thanks to the diffusion of tools to calculate psychological attributes easily and fast, the application of psychometrics have expanded its scope from small laboratory experiments to large-scale studies with online data. The shortened Big-5 personality test (Gosling, Rentfrow, and Swann Jr 2003) and the LIWC dictionary (Pennebaker, Francis, and Booth 2001) are just some examples. Our method aims at complementing these type of approaches by providing the first open tool to calculate IC fast, at scale, and on any type of text. Our tool does not require subjects to take a task-specific test and does not rely completely on wordlists and can therefore detect cognitive processes quite flexibly. In our Reddit study, we were able to code around 400k texts in hours—a task that would take human coders an incredible amount of time.

6.2 Limitations

Our model is built using a training dataset of limited size, narrow range of genres and styles, and with scarce coverage of higher IC bands. The NLP methods used for feature extraction are general purpose and not optimized for the particular domain under study. This limitation does not prevent our method from obtaining good classification results but partially constrains the contribution that different types of syntactic features combined could add to the overall performance. To increase the number and variety of datapoints, we plan to collect more labels from trained human coders. In that respect, to ease this labelling process, our tool can be used to select candidate texts that are likely to belong to a variety of IC classes.

The external validation of our model is limited to one case-study of depression. Further replications of prior work should be undertaken to produce more evidence in support for or against the validity of any automatic system. These could use the original datasets, where possible, and also extend application to new datasets drawn from social media. For a better understanding of the potential and limits of our method, a deeper analysis on how automated IC scoring behaves when applied to different types of textual input—

IC	Text
1	I have strong negative feelings towards this topic in a society where we are supposed to be somewhat evolved socially and economically. Killing a man for his crimes does not seem to be the most evolved or consistent thing. To do an eye for an eye has existed for ages and you think the United States would be over that by now.
2	My opinion is negative because the death penalty attempts to right a wrong with a similar wrong action. A person should be punished by other means which are more humanistic and thoughtful.
3	I think that being assertive is important but at the same time being too assertive can be a bad thing. Being assertive can get things done in a much more effective manner. I am personally an assertive person and think that it can lead to problems with others. However, just because a person is not an assertive person doesn't mean that they should possess that quality. Every person is different and has different characteristics.
4	I am a loving, caring, humorous individual who can be vicious, vindictive, and thoughtless. at times i'm impatient, insightful, and demanding. I recognize what can be and what is and usually am able to suggest how to bridge the gap. I feel that I have learned these qualities from my family and experience with the world.
5	My family existed during the great depression of the 1930's and like most people we lived on very little money. No welfare or pensions then but lots of love and sharing with others. That, with my nursing experiences later and friendships contributed to my present firm belief that money and possession of things do not necessarily result in happiness.
6	I grew up in Canada in a medium sized white middle class family. As such part of what makes me who I am are the values and beliefs I was taught, and have subsequently questioned, of a western patriarchical society. I very much understand myself to be part of a larger scheme of things partly a product of my culture and of my upbringing. However, I also see myself as an individual a product of my own unique experiences and views. So I guess what makes me "me" is the culture I come from, the family I was raised in the time period I'm living in and the other individuals and experiences I've encountered during my life.
7	Fortunately, the goals of deterrence of defense and of arms control are not always in conflict. For example, when we improve our command and control systems we improve our deterrent to aggression and at the same time we decrease the chance of a completely uncontrolled war. Should deterrence fail we have installed a number of both administrative and physical safeguards for our nuclear weapons which reduce as far as possible the chances of unauthorized use. The great emphasis we have placed on forces which can survive a nuclear attack from the Soviet not only serves to deter Soviet aggression but also greatly reduces the pressure on us to act precipitately in a crisis thus decreasing the danger of inadvertent or accidental war.

Table 10: Examples of paragraphs exhibiting different levels of integrative complexity. The texts are taken from the training set and have been scored by certified professionals.

topic of discussion, length of text, cultural background of the authors—is also in order. The use of our tool to analyze Reddit should be considered just as a preliminary attempt to illustrate its potential. Further validation of the methodology is needed before it can be applied at scale and "in the wild". We therefore recommend to the researchers that may use our tool in the future to do it with caution. Most of all, the tool in its current stage should not be used for ethically-sensitive tasks such as decision- or policy-making.

Given these limitations, this work only scratches the surface of the studies that the method we propose could eventually enable. Since previous research has shown that Integrative Complexity is a good predictor of the richness of dialogue (Conway et al. 2016), we believe that IC has a role in tackling the resolution of conflicts in an increasingly polarized social media space.

Appendix

Some examples of texts from the training set belonging to the 7 IC bands are reported in Table 10.

References

Ambili, A. K., and Rasheed, K. M. 2014. Automated scoring of the level of integrative complexity from text using machine learning. In *Proceedings of the 13th International Conference on Machine Learning and Applications (ICMLA)*, 300–305. IEEE.

Ashok, V. G.; Feng, S.; and Choi, Y. 2013. Success with style: Using writing style to predict the success of novels. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1753–1764.

Baker-Brown, G.; Ballard, E. J.; Bluck, S.; De Vries, B.; Suedfeld, P.; and Tetlock, P. E. 1990. *Coding manual for conceptual/integrative complexity*. University of British Columbia and University of California.

Boyd, D.; Golder, S.; and Lotan, G. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Proceedings of the 43rd Hawaii International Conference on System Sciences (HICSS)*, 1–10. IEEE.

Budak, C., and Agrawal, R. 2013. On participation in group chats on twitter. In *Proceedings of the 22nd International Conference on World Wide Web (WWW)*, 165–176. ACM.

Chen, T., and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining (KDD)*, 785–794. ACM.

Cheng, J.; Bernstein, M.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of*

the Conference on Computer Supported Cooperative Work and Social Computing (CSCW), 1217–1230. ACM.

Cheng, J.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2015. Antisocial behavior in online discussion communities. In *Proceedings of the 9th International Conference on Weblogs and Social Media (ICWSM)*. AAAI.

Conway, L. G.; Conway, K. R.; Gornick, L. J.; and Houck, S. C. 2014. Automated integrative complexity. *Political Psychology* 35(5):603–624.

Conway, L. G.; Gornick, L. J.; Houck, S. C.; Anderson, C.; Stockert, J.; Sessoms, D.; and McCue, K. 2016. Are conservatives really more simple-minded than liberals? the domain specificity of complex thinking. *Political Psychology* 37(6):777–798.

Danescu-Niculescu-Mizil, C.; Lee, L.; Pang, B.; and Kleinberg, J. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, 699–708. ACM.

De Choudhury, M., and De, S. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM)*. AAAI.

De Choudhury, M.; Sundaram, H.; John, A.; and Seligmann, D. D. 2009. What makes conversations interesting?: themes, participants and consequences of conversations in online social media. In *Proceedings of the 18th International Conference on World Wide Web (WWW)*, 331–340. ACM.

Gilbert, C. H. E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM)*. AAAI.

Gosling, S. D.; Rentfrow, P. J.; and Swann Jr, W. B. 2003. A very brief measure of the big-five personality domains. *Journal of Research in personality* 37(6):504–528.

Hofmann, H.; Wickham, H.; and Kafadar, K. 2017. Value plots: Boxplots for large data. *Journal of Computational and Graphical Statistics* 26(3):469–477.

Kim, S.; Bak, J.; and Oh, A. H. 2012. Do you feel what i feel? social aspects of emotions in twitter conversations. In *Proceedings of the 6th International Conference on Weblogs and Social Media (ICWSM)*. AAAI.

Li, Y.; Bandar, Z. A.; and McLean, D. 2003. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on knowledge and data engineering* 15(4):871–882.

McCullough, H., and Conway III, L. G. 2018a. "and the oscar goes to...": Integrative complexity's predictive power in the film industry. *Psychology of Aesthetics, Creativity, and the Arts* 12(4):392.

McCullough, H., and Conway III, L. G. 2018b. The cognitive complexity of miss piggy and osama bin laden: Examining linguistic differences between fiction and reality. *Psychology of Popular Media Culture* 7(4):518.

Miller, G. A. 1995. Wordnet: A lexical database for english. *Communications of the ACM* 38(11):39–41.

Niculae, V., and Danescu-Niculescu-Mizil, C. 2016. Conversational markers of constructive discussions. In *Proceedings of the Conference of the North American Association for Computational Linguistics (NAACL-HLT)*.

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. Linguistic inquiry and word count. *Mahway: Lawrence Erlbaum Associates* 71.

Purohit, H.; Ruan, Y.; Fuhry, D.; Parthasarathy, S.; and Sheth, A. P. 2014. On understanding the divergence of online social group discussion. In *Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM)*. AAAI.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*, 4444–4451.

Suedfeld, P., and Bluck, S. 1993. Changes in integrative complexity accompanying significant life events: Historical evidence. *Journal of personality and Social Psychology* 64(1):124.

Suedfeld, P., and Tetlock, P. 1977. Integrative complexity of communications in international crises. *Journal of Conflict Resolution* 21(1):169–184.

Suedfeld, P.; Tetlock, P. E.; and Ramirez, C. 1977. War, peace, and integrative complexity: Un speeches on the middle east problem, 1947–1976. *Journal of Conflict Resolution* 21(3):427–442.

Suedfeld, P.; Tetlock, P. E.; and Streufert, S. 1992. Conceptual/integrative complexity. In Smith, C. P., ed., *Motivation and personality: Handbook of thematic content analysis*. Cambridge University Press. 393–400.

Tausczik, Y. R., and Pennebaker, J. W. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology* 29(1):24–54.

Tchokni, S. E.; Séaghdha, D. O.; and Quercia, D. 2014. Emoticons and phrases: Status symbols in social media. In *Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM)*. AAAI.

Thoemmes, F. J., and Conway, L. G. 2007. Integrative complexity of 41 us presidents. *Political Psychology* 28(2):193–226.

Vosoughi, S., and Roy, D. 2016. Tweet acts: A speech act classifier for twitter. In *Proceedings of the 10th International Conference on Weblogs and Social Media (ICWSM)*. AAAI.

Winter, D. A. 1993. Slot rattling from law enforcement to lawbreaking: A personal construct theory exploration of police stress. *International Journal of Personal Construct Psychology* 6(3):253–267.

Zhang, J.; Chang, J.; Danescu-Niculescu-Mizil, C.; Dixon, L.; Hua, Y.; Taraborelli, D.; and Thain, N. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Meeting of the Association for Computational Linguistics*, 1350–1361. ACL.