A social-event based approach to sentiment analysis of identities and behaviors in text

Kenneth Joseph *[1], Wei Wei †[1], Matthew Benigni ‡[1] and Kathleen M. Carley§[1]

[1]Societal Computing Program, Carnegie Mellon University, 5000 Forbes Ave. , Pittsburgh, PA, USA

Abstract

We describe a new methodology to infer sentiments held towards identities and behaviors from social events that we extract from a large corpus of newspaper text. Our approach draws on Affect Control Theory, a mathematical model of how sentiment is encoded in social events and culturally-shared views towards identities and behaviors. While most sentiment analysis approaches evaluate concepts on a single, evaluative dimension, our work extracts a three-dimensional sentiment "profile" for each concept. We can also infer when multiple sentiment profiles for a concept are likely to exist. We provide a case study of a large newspaper corpus on the Arab Spring, which helps to validate our approach.

1 Introduction

Let us define a *social event* in the sense of Heise (2007) as a situation in which an *actor* enacts a *behavior* on an *object*. Further, let us assume that both the actor and the object are *identities*, which we will define as nouns that are commonly used to allude to a social category (Tajfel & Turner, 1979). Finally, we assume that each identity has a particular affective meaning, or sentiment.

In theory, an infinite number of social events could occur between two identities on an everyday basis. For instance, there are few concrete barriers that prevent all strangers that pass each other on the street from shaking hands. In practice, however, there are many constraints on the social events we are willing to engage in and those we will observe in our everyday lives. Some of these constraints are hard, or physical. Geospatial distance, for example, acts as a barrier that restricts the types of identities that come in contact. Others are "soft", existing within our perceptions of cultural norms. These soft constraints are particularly interesting, as they passively define the "right" way to interact without any actual physical restrictions.

*kjoseph@cs.cmu.edu
†weiwei@cs.cmu.edu
‡mbenigni@cs.cmu.edu
§kathleen.carley@cs.cmu.edu

For example, assume that you are a new elementary-school teacher. It is unlikely that your first action will be to "beat up" your students, even though you are perfectly capable of doing so. Rather, you would be considerably more likely to, for example, "advise" them. "Advising" is an act that fits your, and almost everyone's, intuitions for the identity of a teacher, the identity of a student and the relationship between them. Now consider the situation where you observe a policewoman roughly handling a suspect. Depending on your views on the police, your cultural upbringing and the context in which you observe this act (among other things), your perception of this event may range from an actor carrying out one's duty as an officer to purely inhumane behavior. Thus while some soft constraints, such as those imposing our views on teachers and children, are almost universal (at least within a particular culture), the variability in our emotional response to other events suggests the incredible complexities that can arise in understanding how an individual perceives and engages in social events.

The present work is interested in developing a methodology that allows for a better understanding of how one particular form of these soft constraints, *affective* constraints, mediate perceptions of a particular set of identities engaging in a particular set of behaviors across many social events. Specifically, we develop an approach that is able to infer affective meanings, or sentiments, of identities and behaviors from a large text corpus. These affective meanings, we will show, serve as strong constraints on perceptions of social events within the corpus. In pursuing such a method, three chief issues must be overcome.

Issue 1. While there are an increasing number of databases and tools for extracting world events (e.g. GDELT; Leetaru & Schrodt, 2013) and social behaviors of individuals (e.g. social media, mobile phone records), there is a surprisingly limited amount of data and computational methodologies supporting the extraction of social events engaged in by the generalizable social identities of interest (e.g. "teacher"). In the present work, we present a partial solution to this problem. We first use dependency parsing (Kübler, McDonald, & Nivre, 2009) of newspaper data to extract social events. We then manually clean the resulting output to pull out interesting identities and behaviors from the noisy result of the dependency parse. While we are far from the first to use dependency parsing to extract events from text (for a recent example, see O'Connor, Stewart, & Smith, 2013), few, if any, have considered the goal of extracting events with the aim of using them to infer affective characteristics of identities.

Issue 2. In order to make use of these extracted events, we must then address the issue of how to model the affective constraints that restrict actions in and perception of social events. In the present work, we use Affect Control Theory (ACT) (Heise, 1987, 2007; Robinson, Smith-Lovin, & Wisecup, 2006), which provides a formal social psychological model that describes the following (among many other things):

- The dimensions of sentiment along which we perceive identities and behaviors
- How social events change our perceptions of others
- How we engage in and think about social events in a way that confirms our sentiments

ACT is a "control" theory in that it assumes humans seek to *maintain* preexisting, culturally shared perceptions of identities and behaviors in transient impressions that are generated when social events are observed or carried out. While we may try to maintain these meanings through various methods, our efforts are all carried out in an attempt to reduce the *deflection*, or difference, between our impressions of individuals and culturally-shared sentiments of the identities they represent. ACT assumes that events we expect, or that we are more willing to carry out, are generally low in deflection, as these events are easy to incorporate into our current world-views. For example, the statement "the teacher advises the student" has been estimated to have a deflection of approximately 0.8, while the statement "the teacher beats up the student" has a deflection of 15.4[1].

In ACT, the deflection of a social event is estimated based on two sources of data. First, ACT scholars maintain a large database of survey results that serve as estimates for culture-wide sentiments of a host of identities and behaviors[2]. These sentiments are defined within a three-dimensional affective latent space that has been both theoretically and empirically validated (Osgood, 1975). Second, ACT scholars have developed a set of *change equations* which mathematically describe how the observation of a particular social event changes our perception of the actor, behavior and object involved in it. Given the position of these three entities in the affective space and a change equation, the (unnormalized) Euclidean distance between the affective meanings of the entities before versus after the event defines the level of deflection for the event. We can thus use deflection to understand the relative likelihood of different social events, implicitly giving us an understanding of the affective constraints imposed within the social system of interest.

Unfortunately, while ACT's dictionaries already encompass thousands of identities and behaviors, the data within them are difficult to apply directly to a specifically themed corpus. While collecting new data on new identities and behaviors of interest is possible, it currently requires lengthy survey procedures. Additionally, ACT makes the tenuous assumption that point estimates are sufficient to describe the affective meanings of identities and behaviors (Hoey, Schröder, & Alhothali, 2013a). Finally, while the theory has been tested using survey methodology with individuals in several large cultural groups (e.g. nations), how best to identify any possible differences within these cultures without additional surveys remains an open question. ACT thus holds the potential to be used to provide insight into the affective constraints that shape social events and their perceptions. However, both methodological and data issues prevent a direct application of the theory in many settings of interest to scholars.

Issue 3. The third issue at hand is thus bow best to adapt the concepts involved in ACT into a model that can overcome, at least in part, it's current limitations. The primary contribution of the present work is a probabilistic graphical model (Koller & Friedman, 2009) that provides an initial and substantial step forward in this direction. The model we introduce has four desirable features in that it:

[1] these values were computing using the INTERACT Java program (Heise, 2010a) with the Indiana 2002-2004 dictionary (Francis & Heise, 2006)

[2] `http://www.indiana.edu/~socpsy/ACT/data.html`

- infers affective meanings for identities and behaviors not currently in the ACT dictionaries
- incorporates prior knowledge from existing ACT dictionaries
- infers where multiple "senses" of a particular identity or behavior exist within our dataset
- provides a variance for the sentiment of each sense of each identity and behavior

Our approach is the first effort we are aware of to apply ACT concepts in an automated way to full text data. The statistical model we develop can be applied in a semi-automated fashion to any text corpus from which social events can be extracted, making it a potentially useful tool across a host of sociological domains. Further, from a natural language processing (NLP) perspective, while many other approaches exist to extract sentiment from text (see, e.g. Pang & Lee, 2008), such approaches typically exist on a single "good/bad" dimension. In comparison, our use of ACT allows for a multi-dimensional approach to understanding sentiment in text. This approach is critical in fully interpreting perceptions of identities, behaviors and social events (Osgood, 1975).

After describing the inner workings of our model, we provide a case study on a set of social events extracted from a corpus of approximately 600,000 newspaper articles relevant to the Arab Spring. After rigorous cleaning, the dataset contains 102 identities and 87 behaviors of interest that engage in 10,485 social events over the span of 30 months. Of the 189 identities and behaviors, only 84 (44%) exist in the original Affect Control dictionaries. Thus, we obtain new EPA profiles for many of the important identities and behaviors of interest in our dataset, and can use these to better understand how the English speaking news media perceived these identities and behaviors as the social movement evolved.

Naturally, our understanding is limited by the quality of model output. To this end, we provide a rigorous quantitative analysis of the effectiveness of the model on the task of predicting the behavior one identity enacts on another. Our model's performance improves over several baseline approaches on the prediction task, though struggles with issues of data sparsity in comparison to the strongest of our baselines. Still, the final model we present gives meaningful affective meanings for the identities and behaviors within the dataset, which none of the baseline models are able to provide.

As the quantitative analysis suggests that the model fits the data reasonably well, we also (cautiously) consider the implications of model output on our understanding of news media coverage of the Arab Spring. Most interestingly, we observe a discrepancy in the way major English-speaking news outlets portrayed the generic Muslim identity as opposed to the more specific Sunni and Islamist identities.

2 Related Work

In this section, we first provide background on Affect Control Theory. As our methodology also draws comparisons to a variety of other tasks in the NLP literature, we also touch on efforts in this domain, in particular existing approaches to sentiment mining.

2.1 Affect Control Theory

2.1.1 Overview

Affect Control Theory, originally introduced by Heise (1979, 1987, 2007), presents a compelling model of how perceptions of the affective meaning of identities and behaviors develop. ACT also details how these perceptions can simultaneously exist for social categories and generic behaviors as well as for specific individuals and individual behavioral acts. In detailing these processes, ACT uses a host of ideas beyond the aforementioned concepts of identities, behaviors, the change equation and deflection. We discuss here only the portions of the theory relevant to our model, which center mainly around these four basic components. For those interested in a more complete discussion, we refer the reader to the chapter by Robinson et al. (2006) and the book by Heise (2007).

For matters of convenience, we will use the term *entities* where we are discussing something that applies to both identities and behaviors. All entities have an affective, or sentimental, meaning in a three dimensional latent space, the dimensions of which draw from early work by Osgood (1975) on measuring numeric profiles of affective meanings. The first dimension is evaluative, which describes the "goodness" or "badness" of an identity or behavior. The second dimension is potency, which describes the "powerfulness" (weakness) of an entity. The final dimension is activeness, which defines the level of energy or excitedness of a given entity. Each dimension is defined on the continuous interval $[-4.3, 4.3]$.

Combined, these three dimensions form what can be referred to as the EPA space. Within EPA space, all entities hold a particular position that defines their *EPA profile*. For example, the EPA profile of the identity "teacher" is (0.72, 1.87, 1.41), indicating that teachers are relatively good, powerful and active. In contrast, the EPA position of a student is (1.49, 0.31, 0.75), which shows students are "more good" than teachers, but much less powerful and active[3]. EPA positions of these entities, and many others, have been estimated by Affect Control researchers through a vast collection of survey experiments run across individuals from a host of cultures[4]. These values define the *fundamental*, culturally-shared meanings of identities and behaviors.

ACT assumes that the EPA profile for an individual instantiation of an entity may differ from the generic entity's EPA position. Thus, one may perceive a particular teacher as being "less good" than teachers in general. While in general fundamental meanings are assumed consistent throughout a culture, the theory also allows for the possibility that two different people may have different perceptions of the EPA profile of a generic identity or behavior. Variations of this sort are generally assumed to occur at the boundaries of social groups and social institutions. For example, Smith-Lovin and Douglas (1992) show that individuals in a gay, religious institution had uniquely positive views of the gay, cleric and congregation identities. As the authors state, these individuals "transformed both religious and gay identities so that the homosexual person [could] participate in religious rituals while not abandoning his or her gay identity".

[3]Values from the Indiana 2002-2004 sentiment dictionary Francis and Heise (2006)

[4]The methodology involved in these surveys has evolved over time, and a thorough discussion can be found in (Heise, 2010b)

Thomas and Heise (1995) provide a broader exploration of these multiple senses, showing that systematic differences do exist across social groups (e.g. gender) and via the extent to which individuals are embedded in multiple networks. The model in the present work only partially deals with the fact that different sentiments of the same entity may exist across social groups. On the one hand, the model we use allows for the possibility of multiple perceptions of the same generic identity. However, we also assume that one of the various possible perceptions for each entity in our event data is representative of an American cultural standpoint as provided by a particular ACT dictionary.

In addition to defining culturally shared sentimental meanings of entities, ACT also defines how one's perception of, for example, an individual teacher may develop as the teacher is observed carrying out different social events. The perception of a particular actor, behavior and object that we have before an event is known as the *pre-event transient*. The *post-event transient* describes our perception of the entities in the event after the event has been completed. In general, social events can be chained together such that transient impressions of a previous event become pre-event impressions for the next. In the present work, however, we will assume that each event occurs in isolation, and therefore that the pre-event impression are equal to their fundamental meaning.

The changes in impressions due to a particular social event are calculated in ACT using a change equation, which gives the post-event impression from a function of the pre-event one. The change equation mathematically defines the intuitive way in which pre-event impressions are altered by the social event that is observed. For example, a teacher should be seen as "less good" after beating up a child, and beating up should also be seen as less bad of an action. ACT postulates that the greater the difference, or *deflection*, between culturally-shared, fundamental impressions and post-event impressions is, the less likely an event is to occur. Thus, ACT postulates that we "prefer" to perceive and engage in social events in a way that aligns with our fundamental beliefs.

Affect Control researchers have used survey data to estimate the form of and parameters for the change equation (Heise, 2007; Smith-Lovin, 1987). They have found that the form of the change equation may differ depending on national culture. In the present work, we assume for computational purposes that there exists only a single, universal change equation. This assumption, while still likely flawed, is supported by recent work that suggests differences in change equations across cultures may be due at least in part to weaknesses in earlier estimation techniques rather than to differences in the data (Heise, 2014).

2.1.2 Mathematical Model

Having given an overview of ACT, we now turn to the mathematical model given by the theory. To do so, we first introduce the form of the pre-event transient vector for a social event. Equation (1) gives this vector, which contains the EPA profiles associated with the three entities (*a*ctor, *b*ehavior, *o*bject) in a social event.

$$f = \begin{bmatrix} a_e & a_p & a_a & b_e & b_p & b_a & o_e & o_p & o_a \end{bmatrix} \tag{1}$$

Given the form of this vector, we can now describe how a social event changes these pre-event impressions to produce a post-event impression. This change occurs via the application of the change equation to a particular vector f. Though a variety of estimation methodologies have been used to estimate the change equation (Heise, 2007, 2014), the form of the equation is expected to define a polynomial, multiplicative combination of the pre-event transients. Thus, we can represent the change equation with two parts. First, the function $G(f)$ gives the subset of terms in the power set of f that have been estimated to impact the formation of the transient. As of the writing of this article, $G(f)$ is the following:

$$\begin{aligned} G(f) = [1 \quad & a_e \quad a_p \quad a_a \quad b_e \quad b_p \quad b_a \quad o_e \quad o_p \quad o_a \quad a_e b_e \quad a_e o_p \quad a_p b_p \quad a_a b_a \\ & b_e o_e \quad b_e o_p \quad b_p o_e \quad b_p o_p \quad a_e b_e o_e \quad a_e b_e o_p] \end{aligned} \tag{2}$$

Second, for each element of $G(f)$, we can define a set of coefficients, M, that describes the extent to which the element modifies the value of each element in f. The matrix M is thus a two-dimensional matrix with $|f|$ rows and $|G(f)|$ columns. The $M_{i,j}$ element of M describes the extent to which the jth coefficient of $G(f)$ impacts the ith element of the transient.

The deflection of a particular event is a measure of the squared Euclidean difference between the post-event transients and fundamental impressions. Because the pre-event impression is set equal to the fundamental, we can equivalently define deflection as the squared Euclidean difference between the pre and post-event transient impressions, as shown in Equation (3). In the equation M_{i*} is the ith row of M.

$$Deflection = \sum_{i}^{9} (f_i - M_{i*}^T G(f))^2 \tag{3}$$

It is important to note that because of the way the deflection equation is constructed, one can reassemble it as a quadratic function of the form $c_0 f_i^2 + c_1 f_i + c_2$ for any single element of f, f_i, if all other elements of f are considered to be constant. That is, if we were to actually replace M_{i*} and $G(f)$ with the regression model provided by ACT scholars, perform the multiplication of the squared term, all addition and all simplifications possible, we would end up with a long expression consisting of linear and quadratic combinations of all elements in f, plus a constant (e.g., $1.3 + .5 f_1 f_2^2 + .3 f_5^2 + ...$).

If we treat all $f_j, j \neq i$ as known (as constants), then this massive quadratic equation will reduce to the three term quadratic equation above with constants c_0, c_1 and c_2. The values of c_0, c_1 and c_2 can be computed using the equations above and will consist of nonlinear combinations of constants, including those elements of f, $f_{j,j\neq i}$, that we assume constant. This observation is vital in developing the Gibbs sampling equations for our model[5].

[5]Note that while one could, quite simply, provide a regression model in which either the constant c_0 or c_1 is zero, in practice such an occurrence is unlikely and in any case it can be shown via simple algebra that the equations used here do not fit this case.

2.1.3 Bayesian Affect Control Theory

Recently, Hoey and colleagues (Hoey et al., 2013a; Hoey, Schröder, & Alhothali, 2013b) converted aspects of ACT's mathematical model into one piece of a Partially Observable Markov decision process (POMDP). Their POMDP is used to train an intelligent tutoring tool, and thus their efforts are in a distinctly different vein. However, insights from the their efforts are directly relevant to our model. Most important, perhaps, is Hoey et. al's observation that one can exponentiate and negate the deflection equation to produce a true probability distribution. In doing so, a rearranging of the terms produces a multivariate normal distribution that makes Bayesian analysis feasible. While our model uses substantially different techniques, the relationship between the exponentiated form of the deflection equation and the normal distribution also plays an important role in the development of the model.

2.2 Other related approaches

The extraction of the sentimental meaning of different terms in a text is far from novel in the NLP community. Such efforts typically fall under the domain of sentiment analysis, defined as the extraction of emotional content from text, often in combination with other forms of data suitable for machine learning approaches. For a slightly dated but still very much relevant review of sentiment analysis techniques, we refer the reader to (Pang & Lee, 2008). In general, our approach differs in two important ways from previous sentiment mining approaches. First, there exists only a single other previous work that uses data made available by ACT researchers. Ahothali and Hoey (2015) apply an ACT-based model to social events extracted from news headlines. Their work differs in that they use only news headlines, use manual coding (via Mechanical Turk) to extract social event structure from text rather than the semi-automated approach defined here, and only extract a single EPA profile for each entity. Second, and perhaps most importantly, our approach moves beyond sentiment analysis tools that extract sentiment along single, evaluative dimension. Instead, we place identities and behaviors into a more empirically consistent three-dimensional latent, affective space. Beyond the work of Ahothali and Hoey (2015), few efforts have been made in this direction in the NLP community.

While unrelated to sentiment analysis, our use of the affective latent space draws comparisons to techniques like Latent Semantic Analysis (LSA) (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990) and more recent approaches involving neural networks and "deep learning" (Lee, Grosse, Ranganath, & Ng, 2009) that place words into latent spaces that are representative of their meaning. Such approaches have been shown to be useful in both understanding meaning and in prediction problems. For example, recent convolutional neural network models have been developed that are able to solve analogies via simple algebra and distance models (Mikolov, Yih, & Zweig, 2013), not unlike the methods for finding optimal behaviors for social events in ACT that we will describe below.

Finally, existing NLP tools, perhaps most notably Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003), have formalized many of the difficulties involved with Bayesian analysis of text

data and have shown the effectiveness of considering terms as belonging to many latent "topics" (or in our case, entities having multiple latent "senses"). The most relevant of these models is the work of O'Connor et al. (2013), who infer classes of behaviors that countries enact on each other over time. The authors use dependency parsing to extract events in which one country enacts a behavior on another. They then develop a model that jointly infers types of behaviors between countries and the extent to which the relationship between different countries is described by these classes of behaviors.

3 Extracting Social Events from Text

Table 1: Countries of interest to the present work and number of newspaper articles relevant to them

Country	Num. Articles	Country	Num. Articles
Algeria	11,059	Bahrain	21,314
Egypt	111,779	Iran	138,343
Iraq	101,147	Jordan	23,060
Kuwait	14,559	Lebanon	35,071
Libya	92,101	Morocco	27,153
Oman	8,581	Saudi Arabia	59,406
Syria	96,893	Tunisia	28,485
United Arab Emirates	73,029	Yemen	21,146

In order to use Affect Control Theory, we require a set of social events engaged in by entities of interest. In the present work, we are interested in understanding news media perceptions of the Arab Spring. In order to extract the requisite social events, we rely on a corpus of approximately 600K newspaper articles that we have collected. This large news corpus provides valuable information about numerous social events throughout the Arab Spring, and analyses of these texts can provide insight into behavior Joseph, Carley, Filonuk, Morgan, and Pfeffer (2014). The newspaper articles were extracted from LexisNexis Academic's corpus of "Major World Publications"[6] and were written between July 2010 and December 2012. We only extract articles written in English, and only consider articles that LexisNexis has indexed using its proprietary algorithms as being relevant to one or more of sixteen countries involved, either directly or tangentially, in the Arab Spring. These countries, and the number of articles relevant to them, are listed in Table 1[7].

Extraction of social events from text requires extracting information about "who did what to whom". While we expect such social events to be rampant in the text, the extraction of this type of information is an area of on-going research in the NLP community and has been studied within the subdomains of both dependency parsing (Kübler et al., 2009) and semantic role labeling

[6] `http://www.lexisnexis.com/hottopics/lnacademic/?verb=sf&sfi=AC00NBGenSrch&csi=237924`

[7] Note that a news article may be indexed by LexisNexis as being relevant to more than one country; hence the values in Table 1 sum to a value greater than 600,000

(Carreras & Màrquez, 2005). Here, we use dependency parsing, as methodologies for dependency parsing are more readily available for the type of data we use. Specifically, we use the Stanford CoreNLP pipeline (Manning et al., 2014) to perform dependency parsing on our full set of data with the recently implemented, state-of the art model-based parser (Zhu, Zhang, Chen, Zhang, & Zhu, 2013). For more details on the general techniques and ideas behind dependency parsing, we refer the reader to (De Marneffe & Manning, 2008; Kübler et al., 2009). For an online example of dependency parsing, visit `http://nlp.stanford.edu:8080/parser/index.jsp`.

Quite simply, dependency parsing uses a variety of statistical techniques to extract from each sentence in our corpus the ways in which different terms are linguistically dependent on others. We run the dependency parser on all sentences from our corpus and extract all relations where we find both the subject and direct object of a verb. The subject, verb and object of the dependency parse are lemmatized[8] to their normalized form and then output for further processing.

This procedure allows us to extract social events from the text. For example, from the sentence "The teacher advised the student" the dependency parser (and post-parsing lemmatization) would extract the relationship "teacher advise student". Naturally, this process also extracts a host of noun-verb-noun relationships that are *not* social events, i.e. cases where either of the nouns are clearly not identities ("sanction help talk"), cases where the behavior is ambiguous ("husband say wife") and cases where the dependency parser appears to simply get confused ("issue hold talk"). To filter these events out from the data as best we can, we use a two-pass approach to cleaning. The first pass engages a variety of heuristics to remove highly irrelevant results. We then use a second, manual pass to further increase the relevancy of our data.

We use five heuristics in our first pass cleaning over the data. First, we ignore any events which do not contain at least one ACT identity or behavior. Though this is not required for our model, we find this serves to remove a host of uninteresting dependency parsing outputs. Second, we remove from the data any events appearing in highly similar sentences on the same day. This acts as a crude form of shingling (Rajaraman & Ullman, 2011), which helps in ensuring that we do not double count events from articles that contain nearly the same exact content reiterated by different outlets (O'Connor et al., 2013). Two sentences are similar if the noun-verb-noun relation extracted, along with any terms dependent on these three words (e.g. adjectives) are the exact same. Third, we ignore any relations where the subject or object is a pronoun. Such relations may be useable in the future if co-reference resolution is performed (Soon, Ng, & Lim, 2001), but due to the computational complexity of doing so and the relatively high level of noise this process tends to induce, we do not use it at this point. Fourth, we ignore any results in which we observe the term "not" before the verb of the dependency parse, as we were unsure how to use ACT equations under this negation. Finally, we ignore any behaviors and identities that appear less than 25 times in our dataset, recursively removing events until all terms satisfy this requirement.

[8]Lemmatization is a process by which words are normalized in a deterministic fashion to facilitate analysis. Lemmatization includes stemming (e.g. changing "walking" to "walk" but also other steps, like synonym replacement, (e.g. replacing "better" with "good"). It is standard practice in the NLP literature to perform lemmatization before analysis

After this first pass of cleaning, we were left with approximately 5300 possible identities, 1300 possible behaviors and approximately 1.3M (of 1.7M total) social events. At this point, it was feasible to manually validate that all remaining entities extracted from the dependency parsing were indeed things we considered to be identities or behaviors, even if we could not consider the feasibility of each event individually. While future work may allow for more noise in the data, the present work was chiefly focused on model development, and thus we err on the side of caution in deciding whether to include or exclude entities. For identities, we included only terms in the current ACT dictionaries, terms representing national identities and/or governments (including national leaders), well-known social groups and general identities that were deemed to be interesting by the majority of co-authors (e.g. "protestor"). We only included behaviors that were unambiguous in the action being taken and that could feasibly be expected to relate two identities. For example, we chose not to use the term "have" as a behavior, as it is unintuitive to consider one identity "having" another.

After finishing this processing, we again ensured that all identities and behaviors in our data occur at least 25 times in the cleaned dataset to ensure there was enough data to provide a reasonable estimate of their EPA profile. The final dataset we use has 102 identities, 87 behaviors and 10,485 social events.

4 Model Description

Figure 1 depicts the probabilistic graphical model used in the present work using standard plate notation[9]. In this section, we introduce the model in accordance with its generative structure, working roughly from the top of Figure 1 to the bottom. Although the model is visually complex, we will show here that it is comprised of two rather straightforward pieces. First, the variables θ, ϕ, Q and their predecessors define a simple *language model* (Charniak, 1996), or a model which assigns probabilities to a sequence of words based on their distribution within a corpora of text. This language model governs the probabilities of drawing a particular actor/behavior/object combination for a social event. Second, the variables $\mu_0, \sigma_0^2, \pi, \mu_a, \mu_b, \mu_o, d, z$ and their predecessors define a sort of *Gaussian mixture model* (GMM) that uses ACT, which we will refer to as ACT-GMM. All variables we use, along with a brief description, are listed in Table 2. In reviewing the model, the reader may find Table 2 helpful in that it provides summaries of the mathematical constructs described here.

The model takes three forms of data as input. First, it accepts the set of social events N extracted from the dependency parser. Each social event in N consists of an actor a_n, a behavior b_n and an object o_n. For ease of notation, our discussion below assumes the n subscript on a, b and o is implicit. Second, model hyperparameters m_0 can be set to incorporate EPA profiles of entities appearing in N that also appear in the ACT dictionaries. Finally, the model accepts a change equation, used to calculate deflection. This equation is considered to be static and thus is

[9]For an introduction to plate notation, and to Bayesian modeling more generally, we refer the reader to the general texts from Gelman et al. (2013). We will here, out of necessity, assume some familiarity with such models

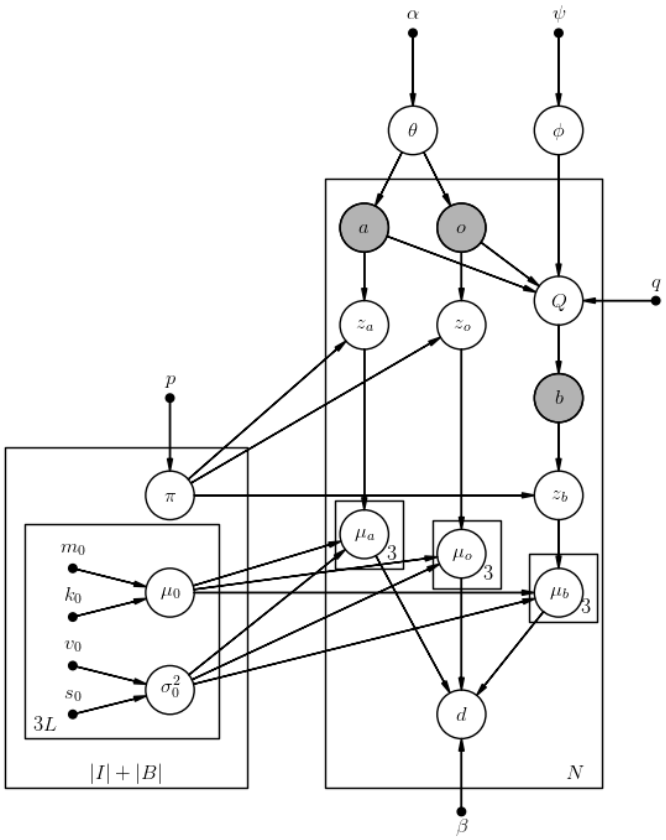


Figure 1: A depiction of the probabilistic graphical model used in the present work using standard plate notation

not updated in any fashion during model inference, nor is it explicitly referenced in Figure 1. The change equation we use is an average of the most recent female and male change equations as given by Interact as of December 30th, 2014.

4.1 Language model component

All entities are assumed to be drawn from a simple language model. The simplest feasible model would be to allow a, b and o to each be drawn from their own Categorical distributions, each of which has a Dirichlet prior. The use of a Dirichlet prior to "smooth" a multinomial or Categorical distribution is known as Laplace smoothing (Zhai & Lafferty, 2001). However, such a model would make poor use of both the data and the theory. With respect to the data, modeling the distributions of actors and objects separately ignores the fact that both draw from a common distribution over identities. Empirically, drawing from a single distribution over identities provides a significant improvement in the model's predictive capabilities. Additionally, ACT is often concerned with

Table 2: Variables used in the description of the model

Variable	**Description**
n	Each social event, n, consists of the triple $< a_n, b_n, o_n >$
a_n, o_n, b_n	The actor, object and behavior for the nth social event, respectively
$z_{a_n}, z_{b_n}, z_{o_n}$	The latent sense in which a_n, b_n or o_n is used
$\mu_{z_{a_n},epa}$	The current expected E,P, or A value epa for the latent sense z_{a_n} for actor a_n. Similar entries exist for z_{b_n}, z_{o_n}
d_n	The difference between the actual deflection of the nth social event and the deflection expected from the fundamental
$\mu_{0,z_e,epa}, \sigma^2_{0,z_e,epa}$	The prior expected mean and variance for the EPA value epa for latent sense z_e of entity e.
$m_{0,z_e,epa}, k_{0,z_e,epa}$	Hyperparameters for $\mu_{0,z_e,epa}$
$v_{0,z_e,epa}, s_{0,z_e,epa}$	Hyperparameters for $\sigma^2_{0,z_e,epa}$
π_e	Distribution governing the likelihood of the different latent senses for entity e
p	Hyperparameter for π
α, ψ	Hyperparameters governing the prior likelihood of any identity or behavior, respectively
θ, ϕ	The estimated likelihood of any identity or behavior, respectively
I, B	The set of all identities or behaviors existent in all events in N, respectively
L	The assumed number of latent senses per identity and behavior
β	The expected scale of the distribution around d
Q	The parameter used to determine the likelihood of a behavior for a particular event
q	Dirichlet smoothing parameter for the conditional distributions $p(b\|o)$ and $p(b\|a)$

the behavior that connects two identities, and thus it makes more sense in the generative model to include an assumption that the behavior for an event is reliant on the actor and object. This assumption should exist in the language model, we believe, above and beyond similar assumptions wired into the ACT-GMM portion of the model.

In the language model, actors and objects are thus both assumed to be drawn from the same Categorical distribution θ, which defines a likelihood of the identity occurring in any given social event. Given an actor and an object, we then draw a behavior to connect them. We assume that the most likely behavior for this event will be influenced by the a and o selected, as well as the overall distribution of behaviors. This overall distribution of behaviors is encoded in the Categorical variable ϕ. The auxiliary variable Q, which is also Categorical, combines information in ϕ with Laplace smoothed estimates on the likelihood of b given a and o. We describe Q in more detail in the following section. Mathematically, these relationships can be expressed by the following:

$$\phi \sim \text{Dirichlet}(\psi)$$
$$\theta \sim \text{Dirichlet}(\alpha)$$
$$a \sim \text{Categorical}(\theta)$$
$$o \sim \text{Categorical}(\theta)$$
$$b \sim \text{Categorical}(Q)$$

4.2 ACT-GMM

Each identity and behavior in the dataset is assumed to have L possible EPA profiles in which it might be used within N, where L is set by the researcher and can be tuned empirically. Allowing for multiple EPA profiles for the same term is an important piece of our model, as the newspaper data we use is extracted from a variety of English-speaking cultures. Each culture may associate a unique EPA profile to a particular entity. We will refer to the different EPA profiles for a particular entity as its different *latent senses* in the sections below. The Categorical variable π governs the frequency with which each latent sense is expected to be used for each entity; p is a hyperparameter for π.

Each latent sense for each entity is associated with three values in μ_0 and σ_0^2; one for each dimension of the EPA profile for that latent sense for that entity. Here, and throughout the article, the 0 subscripts (e.g. on μ_0) are used to represent a variable that is a prior or a hyperparameter to the Bayesian network model. A particular entry in the vector μ_0, which we will refer to as $\mu_{0,z_{ib},epa}$ (ib stands for "identity or behavior") exists at the $3*ib*z+epa$ location in μ_0. Here, z_{ib} is the index of the z_{ib}th latent sense for entity ib and epa is the index of the sentiment dimension. A similar indexing scheme is used for σ_0^2. Combined, these six mean and variance parameters determine the mean and variance of the three dimensions of the EPA profile for this particular latent sense z_{ib} of the entity ib.

All values in μ_0 are assumed to be drawn from a normal distribution governed by m_0, k_0 and σ_0^2, while σ_0^2 is assumed to be drawn from an Inverse Chi-squared distribution with parameters v_0, s_0. A key insight that is leveraged in our approach is that the values of m_0, as priors on the EPA profiles, can be used to set, or to "control", the EPA profiles for entities in the ACT dictionaries. For example, we might set the (static) value of $m_{0,1_{teacher},e} = 0.72$ to help ensure that the evaluative dimension of the first latent sense for the identity teacher is biased towards the "correct" value implied by the ACT dictionaries. More formally, we assume that the joint prior density for μ_0 and σ_0^2 follows a Normal Inverse Chi-squared distribution, which allows us to infer both values using Bayesian inference. This formulation is a common representation for Bayesian models where one wishes to infer both the standard deviation and the mean for a Normal distribution; as such we defer the reader to (Gelman et al., 2013, pp.76-68) for further details. Mathematically, our assumptions can be expressed as follows:

$$\pi \sim \text{Dirichlet}(p)$$
$$\mu_0 \sim \text{N}(m_0, \frac{\sigma_0^2}{k_0})$$
$$\sigma_0^2 \sim \text{Inv}{-}\chi^2(v_0, s_0)$$

For each social event, each actor, behavior and object is associated with a particular latent sense z of its corresponding entity. Once z_a, z_b and z_o are drawn, we can obtain the entities' EPA profiles μ_a, μ_b, and μ_o (respectively) by sampling an EPA profile from the Normal distributions governed by the relevant entries in μ_0 and σ_0^2. Once these values have been drawn, we can obtain a deflection score for that event.

One could define the deflection for an event as a deterministic function. To do so, the values of μ_a, μ_b and μ_o would be combined to form the pre-event impression f. We could then provide a deterministic deflection score for the event by substituting these values into Equation (3). Instead, however, we treat deflection as a stochastic process whose mean is this expected deflection but that has some variance, β. We feel this assumption is more reasonable than the deterministic one in our particular case, as it accounts for context of this particular social event beyond what we can account for with our mixture model. For example, the lack of incorporation of information about settings implies an inherent randomness in the deflection measured by our model, thus justifying the assumption of stochasticity. The distribution of deflection is assumed to be Laplacian, which makes model inference easier while still retaining the desired sociotheoretic meaning of deflection as a distance metric. Mathematically, our assumptions can be stated as follows:

$$z_a \sim \text{Categorical}(\pi) \qquad z_b \sim \text{Categorical}(\pi) \qquad z_o \sim \text{Categorical}(\pi)$$
$$\mu_a \sim \text{N}(\mu_{0,z_a}, \sigma_{0,z_a}^2) \qquad \mu_b \sim \text{N}(\mu_{0,z_b}, \sigma_{0,z_b}^2) \qquad \mu_o \sim \text{N}(\mu_{0,z_o}, \sigma_{0,z_o}^2)$$
$$d \sim \text{Laplace}(\sum_{i}^{9}(f_i - M_{i*}^T G(f))^2, \beta) \quad \text{where} \quad f = [\mu_{a_e}, \mu_{a_p}, \mu_{a_a}, \mu_{b_e}, \mu_{b_p}, \mu_{b_a}, \mu_{o_e}, \mu_{o_p}, \mu_{o_a}]$$

Algorithm 1: Model inference for ACT-GMM portion using Stochastic EM

Initialize all μ_0, σ_0^2, π
for $i = 0$ **to** $N\_ITERATIONS$ **do**
    E step:
        Sample all z using Gibbs sampling and Equation (7)
        Sample all μ using Gibbs sampling and Equation (8)
    M step:
        Update all σ_0 using MAP estimate of Equation (10)
        Update all μ_0 using Equation (12)
        Update all π using Equation (13)

4.3 Summary

Having fully defined our model, a useful summarization can now be provided by giving the formal generative process required by the model. To generate a new social event, the following process is carried out:

1. Draw an actor and an object; $a \sim \text{Cat}(\theta)$ $o \sim \text{Cat}(\theta)$
2. Draw a behavior; $b \sim \text{Cat}(Q)$
3. Draw a latent sense for a, b and o; $z_a \sim \text{Cat}(\pi_a)$ $z_b \sim \text{Cat}(\pi_b)$ $z_o \sim \text{Cat}(\pi_o)$
4. Draw EPA profiles for a, b, and o
    - $\mu_{a,e} \sim \text{N}(\mu_{0,z_a,e}; \sigma^2_{0,z_a,e})$ $\mu_{a,p} \sim \text{N}(\mu_{0,z_a,p}; \sigma^2_{0,z_a,p})$ $\mu_{a,a} \sim \text{N}(\mu_{0,z_a,a}; \sigma^2_{0,z_a,a})$
    - $\mu_{b,e} \sim \text{N}(\mu_{0,z_b,e}; \sigma^2_{0,z_b,e})$ $\mu_{b,p} \sim \text{N}(\mu_{0,z_b,p}; \sigma^2_{0,z_b,p})$ $\mu_{b,a} \sim \text{N}(\mu_{0,z_b,a}; \sigma^2_{0,z_b,a})$
    - $\mu_{o,e} \sim \text{N}(\mu_{0,z_o,e}; \sigma^2_{0,z_o,e})$ $\mu_{o,p} \sim \text{N}(\mu_{0,z_o,p}; \sigma^2_{0,z_o,p})$ $\mu_{o,a} \sim \text{N}(\mu_{0,z_o,a}, \sigma^2_{0,z_o,a})$
5. Draw a deflection score for the event
    - $d \sim \text{Laplace}(\sum_i^9 (f_i - M_{i*}^T G(f))^2, \beta)$ where $f = [\mu_{a_e}, \mu_{a_p}, \mu_{a_a}, \mu_{b_e}, \mu_{b_p}, \mu_{b_a}, \mu_{o_e}, \mu_{o_p}, \mu_{o_a}]$

The described process helps to explain how a new social event might be "generated" by the Bayesian network model described here, but also provides insight into how the model determines the likelihood of an event it is given. The likelihood of a particular event is a function of likelihood of the actor and object's identities overall (1.), the "semantic likelihood" of the behavior given these identities (2.), and the "affective likelihood" of the social event as a whole (3.-5.).

5 Model inference

Model inference is completed in two steps. First, we determine Maximum *a posteriori* (MAP) for the parameters of the language model, as they are straightforward enough to determine in closed form.

Second, we use the Stochastic EM (Tregouet, Escolano, Tiret, Mallet, & Golmard, 2004) (Bishop & others, 2006, p. 439) algorithm displayed in Algorithm 1 to draw inferences for parameters in the GMM portion of the model. Note that Algorithm 1 references several equations that will be introduced later in this section. In the Expectation ("E") step, we use Gibbs sampling to draw expected values for z_a, z_b and z_o and for μ_a, μ_b and μ_0 for all social events. In the Maximization ("M") step, we then update σ_0, μ_0 and π with their MAP estimates. Note that we do not explicitly sample d, as the value of this stochastic process is not of particular interest to us in the present work.

Below, we first derive the MAP estimate for the language model. We then derive the Gibbs sampling equations for all z and all μ and finally the MAP estimates for σ_0, μ_0 and π. In doing so, we introduce three additional pieces of notation. First, let μ_* represent all nine fundamental values drawn for an event. Second, let μ_*/x represent μ_* where all values of all elements are known except for x. Finally, we define Ω as the set of all parameters, and Ω_{-x} as the set of all parameters besides x.

5.1 MAP estimates for language model

MAP estimation for the language model portion of the model is relatively straightforward. The variables of interest are θ and Q_n. The distribution for θ is given in Equation (4), where $n(a_i)$ is a function that represents the number of times the identity i appeared as an actor in N and $n(o_i)$ the number of times i appeared as an object. The Dirichlet distribution is a well-known conjugate of the Categorical distribution, and we thus do not re-derive the posterior distribution here. Note, however, that we follow the notational convenience of absorbing the minus one in the second line of Equation 4 into the Dirichlet hyperparameter in all following statements about the posterior distribution and MAP estimates of the Categorical distribution.

$$\begin{aligned} p(\theta) &= p(\theta|\alpha) * \prod_{i=1}^{|I|} p(a_i|\theta)p(b_i|\theta) \\ &\propto \prod_{i=1}^{|I|} \theta_i^{n(a_i)+n(o_i)+\alpha_i-1} \\ &\sim \text{Dirichlet}(n(a_i) + n(o_i) + \alpha_i) \end{aligned} \tag{4}$$

Given $p(\theta)$ is distributed as in Equation (4), the MAP estimate for the posterior distribution of θ is given by Equation (5). The estimate is simply a normalized function of the number of times an identity appears plus the "pseudo-counts" from the Dirichlet prior α.

$$\hat{\theta} = \arg\max_\theta p(\theta) = \frac{n(a_i) + n(o_i) + \alpha}{\sum_{i \in I} n(a_i) + n(o_i) + \alpha} \tag{5}$$

The MAP estimator for Q is given in Equation (6). Note that because Q depends on the actor and objects for each event, there are actually $|N|$ values of Q. We will discuss the derivation for a particular entry of Q, Q_n here, as the derivation is the same for all events. The distributions $p(b_n|a_n)$

and $p(b_n|o_n)$ are Categorical distributions that give the conditional likelihood of the behavior b_n given a_n and o_n, respectively. To ensure that these values are never zero, we introduce a smoothing parameter q resulting in the distributions $p(b_n|a_n, q)$ and $p(b_n|o_n, q)$, respectively. Introducing the smoothing parameter q is equivalent to inserting an auxiliary variable for both $p(b_n|a_n)$ and $p(b_n|o_n)$ and putting a Dirichlet prior over each with the hyperparameter q. As the introduction of this variable would unnecessarily complicate notation, we do not use it here. The likelihood of any particular behavior as derived from the MAP estimate is thus simply the product of three Laplace smoothed Categorical variables, ϕ (smoothed by ψ), $p(b_n|a_n)$ and $p(b_n|o_n)$, both smoothed by the constant q. The distribution of b_n is thus Categorical with Q as the parameter.

$$\begin{aligned}\hat{Q_n} = \arg\max_{Q_n} p(Q_n) &= \arg\max_{Q_n} p(b_n|a_n, q) * p(b_n|o_n, q) * p(\phi|\psi) \prod_i^N p(b_{n,i}|\phi) \\ &= \frac{n(b_i|a) + q}{\sum_{b_i \in B} n(b_i|a) + q} * \frac{n(b_i|o) + q}{\sum_{b_i \in B} n(b_i|o) + q} * \frac{n(b_i) + \psi}{\sum_{b_i \in B} n(b_i) + \psi} \end{aligned} \tag{6}$$

5.2 "E" Step for ACT-GMM

For each document, we must draw z_a, z_b and z_o and all nine values for the fundamental, three each for μ_a,μ_b and μ_o μ. Because the sampling proceedure is analogous for all entities in a particular event and are the same for each event, we will focus here only on the agent for one specific event n.

5.2.1 Sampling z

The conditional probability that the variable z_{a_n} is equal to the latent sense t is specified in Equation (7):

$$p(z = t|\Omega_{-z}) = p(z = t|\pi) \prod_i^{[e,p,a]} p(\mu_{a_i}|\mu_{0,t,i}, \sigma^2_{0,t,i}) \tag{7}$$

Sampling from this conditional distribution is straightforward, as both the first and second terms of the probability function are easy to compute. The first term is simply the likelihood of latent sense t as given by the current value of π. The second term can be obtained by evaluating the likelihood of $\mu_{a,e}$,$\mu_{a,p}$ and $\mu_{a,a}$ relative to their expected distribution given the current state of μ_0 and σ_0. These three values are multiplied together to generate a likelihood for μ_a as a whole. Multiplying the result of this process by the first piece of the probability function, we can then normalize over all possible values of z and draw a new latent sense for the actor in this event from this Categorical distribution.

5.2.2 Sampling μ

The conditional distribution of $\mu_{z_{a_n},e}$, the value for the evaluative dimension of the actor's EPA profile for event n and latent sense z_{a_n}, is given below in Equation (8). Representation of the potency and activity dimensions are analogous, so we focus only on the evaluative dimension here. Also, we shorten z_{a_n} to z ease notation.

$$p(\mu_{z,e}|\Omega_{-\mu_{z,e}}) = p(\mu_{z,e}|\mu_{0,z,e}, \sigma^2_{0,z,e})p(d|\mu_*/\mu_{z,e}; \beta) \tag{8}$$

To infer the conditional distribution of $\mu_{z,e}$, Equation (8) shows we simply need to understand the prior distribution of $\mu_{z,e}$ and the distribution of d given all values except that of $\mu_{0,z,e}$. From the section above, we know that $p(\mu_{z,e}|\mu_{0,z,e}, \sigma^2_{0,z,e}) \sim \mathrm{N}(\mu_{0,z,e}, \sigma^2_{0,z,e})$. Thus, we are left with interpreting the distribution of $p(d|\mu_*/\mu_{z,e}; \beta)$. It can be shown, rather unexpectedly that evaluating the distribution of d given all values except $\mu_{z,e}$ results in a distribution which is normally distributed on $\mu_{\mathbf{z,e}}$ with a known mean and variance.

The proof is shown below; the derivation follows from the fact stated in Section 2.1.2 that the deflection score with one unknown variable is a quadratic in that variable. By completing the square and dropping constant terms that do not inform the conditional distribution for $\mu_{z,e}$, we are left with a function that defines a Normal distribution on $\mu_{z,e}$ with the given parameters.

$$\begin{aligned}
p(d|\mu_*/\mu_{z,e}; \beta) &\propto \exp(-\frac{|d - \sum_i^9 (f_i - MG(f_i))^2|}{\beta}) \\
&= \exp(-\frac{|d - (c_0\mu_{z,e}^2 + c_1\mu_{z,e} + c2)|}{\beta}) \\
&\propto \exp(-\frac{|(c_0\mu_{z,e}^2 + c_1\mu_{z,e})|}{\beta}) \\
&= \exp(-\frac{|c_0|(\mu_{z,e} + \frac{c_1}{2c_0})^2}{\beta}) \\
&= \exp(-\frac{(\mu_{z,e} + \frac{c_1}{2c_0})^2}{\frac{\beta}{|c_0|}}) \\
&\propto \mathbf{N}_{\mu_{\mathbf{z,e}}}(-\frac{c_1}{2c_0}, \frac{\beta}{2|c_0|})
\end{aligned} \tag{9}$$

There are two important points to note in the derivation shown in Equation (9). First, the result relies on the fact that there are no social events in which the same identity appears more than once. If this were to be the case, the equation would no longer be quadratic in $\mu_{z,e}$. Second, and perhaps more importantly, is that the resulting distribution is centered at the value of $\mu_{z,e}$ which minimizes the deflection of the social event given all other fundamental meanings as estimated by Maximum Likelihood Estimation (Heise, 2007, ch. 8). Though this result fits our intuition, we do not believe that this was an obvious outcome given the initial distribution.

Thus, when updating $\mu_{z,e}$ we are drawing from a product of two normal distributions. One of these distributions is centered at the current expected value of $\mu_{z,e}$ as given by $\mu_{0,z,e}$ and $\sigma^2_{0,z,e}$.

The second distribution, usefully, is centered at the value which will minimize deflection for the current event given all other values in the pre-event fundamental vector. It is well-known that the product of two normals is proportional to a new normal distribution[10], and thus we can sample a new value for $\mu_{z,e}$ from this new distribution, which has a mean of $\frac{\mu_{0,z,e}*\frac{\beta}{2|c_0|}-\frac{c_1}{2c_0}*\sigma^2_{0,z,e}}{\sigma^2_{0,z,e}*\frac{\beta}{2|c_0|}}$ and a variance of $\frac{\sigma^2_{0,z,e}*\frac{\beta}{2|c_0|}}{\sigma^2_{0,z,e}+\frac{\beta}{2|c_0|}}$. From this sampling distribution, it is clear that the new value is informed by both the prior information from μ_0 and σ^2_0 and information from the current event.

5.3 "M" Step for ACT-GMM

5.3.1 MAP estimates for μ_0, σ^2_0

Because all updates in μ_0 and σ^2_0 are analogous, we will consider the conditional distribution of the evaluative dimension of a particular latent sense z_i of a single identity i. To ease notation, we will refer to the relevant entry in μ_0, which is $\mu_{0,z_i,e}$, as simply μ_0, and the relevant entry in σ^2_0, which is $\sigma^2_{0,z_i,e}$, as simply σ^2_0. A similar shortening of notation will be applied to the four relevant hyperparameters $m_{0,z_i,e}, k_{0,z_i,e}, s_{0,z_i,e}$ and $v_{0,z_i,e}$. Let us also define the set S, which consists of all events in which the latent sense z_i of the identity i is used in the current iteration of the inference algorithm. Formally, $S = \{n \in N : (a_n = i \,\&\, z_{a_n} = z_i) \,|\, (o_n = i \,\&\, z_{o_n} = z_i)\}$. The variable S is introduced as we need not worry about events outside of it; they will be irrelevant in evaluating the distribution of $\mu_{0,z_i,e}$.

The derivation for the MAP estimation of σ^2_0 can be easily obtained from its well-known posterior distribution, shown in Equation (10). In Equation (10), $\bar{s^2}$ is the sample variance deviation (that is, $\sum_n^S(\mu_n - \mu_0)^2$) and $\bar{\mu} = \frac{\sum_n^S \mu_n}{|S|}$.

$$\begin{aligned} p(\sigma^2_0|\Omega_{-\sigma^2_0}) &= p(\sigma^2_0|v_0, s_0)\prod_n^S p(\mu_n|\mu_0, \sigma^2_0) \\ &= \text{Inv}{-}\chi^2(v_0 + |S|, v_0 s_0 + (|S|-1) * \bar{s^2} + \frac{k_0 * |S|}{k_0 + |S|}(\bar{\mu} - \mu_0)^2) \end{aligned} \tag{10}$$

The expected value of this posterior distribution is then the MAP estimate of Equation (10). The MAP estimate is $\frac{xy}{x-2}$, where x is the location (first parameter) of the Inv$-\chi^2$ in Equation (10) and y is the scale (second parameter) of this distribution.

The distribution of μ_0 also consists of two parts, a straightforward prior and a posterior component that is the product across the events in S. This is shown in Equation (11):

$$p(\mu_0|\Omega_{old}) = p(\mu_0|m_0, \sigma^2_0, k_0)\prod_n^S p(\mu_n|\mu_0, \sigma^2_0) \tag{11}$$

The MAP estimate of a normal distribution in this form is well known, so we simply provide the resulting estimate in Equation (12). Note that σ^2_0, as used in Equation (12), represents the "new" version of σ^2_0 from Equation (10).

[10] for a formal proof, see (Bromiley, 2013)

Table 3: Model initialization details

Variable	**Initialization value/method**
β, α, ψ, q	1
v_0	2
s_0	.1
p	3
m_0	value from ACT dictionary, random otherwise
k_0	50 if from ACT dictionary, 10 or 1 otherwise. See text
L	varied for parameter tuning
π, μ_0, σ_0^2	drawn from Prior

$$\hat{\mu}_0 = \frac{\frac{k_0}{\sigma_0^2} m_0 + \frac{|S|}{\sigma_0^2} \bar{\mu}_n}{\frac{k_0}{\sigma_0^2} + \frac{|S|}{\sigma_0^2}} \tag{12}$$

5.3.2 MAP estimate of π

The MAP estimate of π reduces to a new Categorical distribution where the likelihood of each latent sense is the number of times this latent sense is "used" in the E step plus the "pseudo-counts" from the Dirichlet prior p. The derivation of this MAP estimate is, as we have mentioned, straightforward given previous, well-known results in the literature. Equation (13) gives the distribution of the new value of π.

$$\hat{\pi} = \frac{n(z_t) + p_t}{\sum_s (n(z_s) + p_s)} \tag{13}$$

5.4 Initializing the Model

All that remains to be introduced with respect to model inference is how parameters are initialized. Table 3 details the initialization of all parameters aside from Q, d, z and μ. An initialization of Q is unnecessary, as we simply compute it's MAP estimate once. The value of d is not of interest and does not affect the estimation of other parameters, thus initialization is unnecessary. The parameters z and μ are sampled before they are used, so also need not be initialized. In this section, beginning with the hyperparameters, we provide more details about the meaning of the statements in Table 3.

Hyperparameter tuning can have important implications on model performance, perhaps especially so in cases where we are dealing with language data (Wallach, Mimno, & McCallum, 2009). However, hyperparameters also reflect one's prior expectations, and thus we attempt here to balance a search for optimal parameters between our prior expectations and model performance. Further, given the number of hyperparameters for the model, we chose to only use heuristic searches to explore the parameter space. Thus, we set $\beta, q, p, \alpha, \psi, s_0$ and v_0 to specific values based on the

parameter settings which maximized performance on a single fold of the data and set m_0 and k_0 via a combination of informal model testing and heuristic methods.

For β, we assume that a reasonable prior for the variance of the deflection score around its ACT-implied value is 1. The variable q is set to 1, as testing on development data using a variety of language models suggested that this minimal pseudo-count led to the strongest model. We also set α and ψ to 1 based solely on results from testing, as these values performed better than other tested values of 3, 50 and 1000. We similarly set s_0 to .1 as opposed to other tested values of .5 and 1, and p to 3 as opposed to other tested values of 50, 400 and 1000. Finally, we set v_0, which can be thought of as our confidence in the prior s_0, to 2, which reflects a low level of confidence in the value of s_0. Setting the value to 1 caused high instability in the MAP estimates for σ_0^2; the value of 2 was the smallest at which they showed stability.

We use the hyperparameter m_0 to encode our prior belief of the EPA profile for each latent sense of each entity. We make the assumption that one of the latent senses is drawn from an American cultural perspective, as many of the media outlets within our dataset are based in America. Hence, we use data from the ACT website that is the best representation avaiable of this perspective. This data comes from a dictionary of 500 identities and 500 behaviors coded with EPA profiles from surveys in 2002-3 of undergraduate students at a large, public, American institution (Francis & Heise, 2006). These data have been used in a variety of ACT studies since it was introduced, and are used as the default values for the computer program Interact (Heise, 2010a), from which much ACT research is derived.

For entities ib that are in both the survey data and our set of social events, we initialize values of m_0 for its zeroth sense (i.e. $m_{0,0_{ib},e}, m_{0,0_{ib},p}, m_{0,0_{ib},a}$) to the values from the survey data. We then heuristically set the rest of the values of m_0 using an iterative algorithm for the zeroth sense of all entities that are not in the ACT dictionary. The algorithm takes as input the set of already known values in m_0 and uses the fact that the standard mathematical model of ACT can be used to solve for the EPA profile of the third entity in an event if the EPA profiles of the other two entities are already known (Heise, 2007, ch. 8). On the first round of the iterative algorithm, we extract all events where values for m_0 for two of the three entities are known from the ACT dictionary. We then compute the optimal EPA profile for the third entity in each of these events. We then take the average EPA profile for each entity we can obtain at least one EPA profile from in this process and use these values to initialize $m_{0,0_{ib},e}, m_{0,0_{ib},p}$ and $m_{0,0_{ib},a}$ for each of these entities.

Once we have set these values, we can iterate through N again, treating both the original m_0 values from the ACT dictionary and the new values from the first iteration of the algorithm as known. This iterative process continues until we can initialize m_0 for all entities. If we reach a point where no new information can be gleaned from the process above, we select one random event and set the EPA profile of one of the entities using uniform random values in the range $[-4.3, 4.3]$. This allows the algorithm to continue learning. In practice, the algorithm finishes in around two iterations, only having to set a random score zero or one times for entities that appear in the training data. For terms appearing in held-out test data but not in the training data, the appropriate entries for the zeroth sense in m_0 are initialized to uniform random values.

Table 4: Predictive distributions for the four baseline models and the full model

UNIGRAM	$p(b\|a,o) = p(b, s=1)$
BIGRAM	$p(b\|a,o) = p(b, s=1)p(b\|o, s=1)p(b\|a, s=1)$
ACTONLY	$p(b\|a,o) = p(b\|\phi, a, o, q)p(b\|a, o, d=0, m_0)$
NOACT	$p(b\|a,o) = p(b\|\phi, a, o, q)\mathbf{E}_{z,\mu}[p(z\|\pi)p(d=0\|\mu)p(\mu\|\mu_0, \sigma_0^2)]$
FULLMODEL	$p(b\|a,o) = p(b\|\phi, a, o, q)\mathbf{E}_{z,\mu}[p(z\|\pi)p(d=0\|\mu)p(\mu\|\mu_0, \sigma_0^2)]$

The zeroth sense for each entity thus represents some instantiation, real or heuristically defined, of the EPA value of that entity from an American cultural standpoint. Values in m_0 for all other latent senses of all entities (including those found in the ACT dictionary) are set using uniform random values on the interval $[-4.3, 4.3]$. While future work may attempt to be smarter in how these parameters are set, we currently use random values to insinuate no prior knowledge of the EPA profiles of other cultural groups whose perceptions may exist in the data. This lack of prior knowledge is reiterated with the initialization of k_0, which can be thought of as the number of observations that we associate with m_0 as a prior for μ_0. For the zeroth latent sense for entities in the ACT dictionary, we set k_0 to 50 (the number of respondents in the survey data). For the zeroth latent sense of entities initialized in the iterative algorithm, we set k_0 to 10. For all other latent senses of all entities, $k_0 = 1$.

Once we have initialized all hyperparameters, all that remains is to initialize are π, μ_0 and σ_0^2. We do so by drawing π, μ_0 and σ_0^2 from their respective prior distributions. This completes model initialization.

6 Approach to Model Evaluation

The model we present is a combination of three well-established methodological approaches- language modeling, Gaussian mixture models and ACT. While we have confidence that each component can extract useful information from the data, our extensions of current ACT methodology and the novel way in which we combine techniques requires a careful study of the extent to which our efforts produce parameter estimates that are truly representative of information within the data. To evaluate the quality of the estimates generated by our model, we use 10-fold cross-validation. In k-fold cross validation, the data is split into k "folds". We use $k-1$ folds as "training data" to train the model and carry out a prediction task on the "left out" test data. This process is repeated k times, leaving out a different chunk of the data, and then results on the prediction task across all folds are averaged. Here, we use perhaps the most common prediction task in the ACT literature, that of predicting the behavior between an actor and an object. That is, for a given left out event n, we give the trained model the actor and object in n and then attempt to predict the behavior.

In establishing the quality of our model's predictions, we can have more confidence that parameter estimates accurately represent processes within the social event data. Importantly, such an understanding requires some baseline for comparison. The four baseline models we compare our full

model's predictions to range from simple language models to more complex structural ablations of our full model. While only some of these models can actually help to infer EPA profiles of entities, they are important in giving us a sense of how "easy" the prediction task we are addressing is. When a simple baseline can predict the data perfectly, a complex model like the one we propose is more likely to learn patterns in the data that are largely noise, and thus post-hoc interpretations of parameter estimates may suffer.

To evaluate the success of each model on the prediction task, we compute the *log perplexity* of the model in determining the correct behavior across all test events. Log perplexity, or simply perplexity as it is often written, a measure of accuracy typically used in explorations of the predictive abilities of NLP models. This value is defined in Equation (14) below. In the equation, let TD be the set of held out data used for testing for a single fold. The log perplexity, averaged across all test events in all folds, gives us a sense as to how much weight in the model's predictive distribution that the model places on the correct behavior. The value $2^{log(perpl(TD))}$ can be thought of as the number of behaviors the model feels are equally as likely as the true answer. So, if $log(perpl(TD)) = 1$, the model would, on average, be "flipping a coin" between the correct answer and one other answer. If $log(perpl(TD)) = 4$, the model would be rolling a 16-sided die. Note that the metric is a measure of the extent to which a particular model is "confused" by the data, and thus a lower score represents a better model.

$$log(perpl(TD)) = \frac{-\sum_{n \in TD} log(p(b_n|a_n, o_n))}{|TD|} \tag{14}$$

All models we test and the predictive distributions they use to determine the likelihood of each behavior for a given test event are shown in Table 4. The first two models are simple, Laplace smoothed language models with a smoothing parameter $s = 1$. The first predicts the likelihood of a behavior by simply determining the likelihood of the single-word behavior label, or the behavior *unigram*, $p(b, s = 1)$. We call this model the UNIGRAM model. The second model uses the conditional likelihood of the behavior given the actor and the object independently, as well as the likelihood of the behavior itself. This is exactly the language model used in our full Bayesian model. Drawing from the language modeling literature, this model is termed the BIGRAM model, meaning that we draw information for the behavior from its distribution, its distribution conditioned on the actor and its distribution conditioned on the object.

Combined, these two baselines show how well events can be predicted by considering only the semantic relationships between identities and behaviors. Note that this semantic information, particularly in the BIGRAM model, will implicitly capture a significant amount of affective meaning as well - just because we are not explicitly modeling affecting meaning does not mean it isn't capture in semantic relationships within the text. We should therefore expect that these semantic models, which derive likelihoods from only connections between words, are strong predictive models. Adding in the ACT component of the model may or may not help in a predictive sense, but will serve the vital purpose of helping us to understand *why* these semantic relationships are occurring.

The third baseline we use removes the GMM portion of the full model, replacing it with what is

essentially the pure prediction model of current ACT methodologies. In this ACTONLY model, we use the values of m_0 initialized by the iterative algorithm described above to determine the sentiment for each entity. These values are therefore only roughly informed by the data, capturing only the heuristic information from the initialization algorithm. The ACTONLY model then uses these heuristically set EPA profiles in combination with the language model and the base mathematics of ACT to make predictions, no further statistical optimization is performed. Under the assumption that the actual deflection of each social event is zero, i.e. $d = 0 \,\forall\, n$, and that $\beta = 1$, the prediction reduces to a form of the probability model for deflection related to that proposed by Hoey et al. (2013a), as shown in Equation (15).

$$\begin{aligned} p(b|a,o) &= p(b|\phi,a,o,q) * p(b|a,o;m_0,d=0,\beta=1) \\ &= p(b|\phi,a,o,q) * \exp(\sum_{i}^{9}(f_i - MG(f_i))^2) \\ \text{Where} \quad f &= [m_{0,a_e} \;\; m_{0,a_p} \;\; m_{0,a_a} \;\; m_{0,b_e} \;\; m_{0,b_p} \;\; m_{0,b_a} \;\; m_{0,o_e} \;\; m_{0,o_p} \;\; m_{0,o_a}] \end{aligned} \tag{15}$$

The final baseline we use is one that effectively removes ACT from the full model by randomizing the change equation matrix and removing any information from the ACT dictionaries in the priors. This model is labeled the NOACT model in our results. While we expect performance of this model to be comparable to the full model, it loses a significant amount of value in qualitative analysis of results. Finally, we of course train our FULL model. Note that we run both of these models assuming a variety of values for L in order to understand how they perform with different numbers of assumed latent senses.

For both the NOACT and FULL models, we run Algorithm 1 for 200 iterations. Parameter estimates used in the prediction task are extracted from the final iteration of the algorithm. Once parameter estimates have been obtained, we also must account for the fact that in both models, the likelihood of a particular behavior for a test event is determined by averaging over all possible values of z_a, z_b and z_o and all values of μ_a, μ_b and μ_o. That is, we must compute $\mathbf{E}_{z,\mu}[p(d = 0|\mu)p(\mu|\mu_{0,z},\sigma^2_{0,z})p(z|\pi)]$, which when expanded becomes $\sum_z \int_\mu p(d = 0|\mu)p(\mu|\mu_{0,z},\sigma^2_{0,z})p(z|\pi)$. Note that we here condense all z and μ values for the actor, behavior and object into a single term to simplify notation.

We choose to estimate this expectation using Gibbs sampling. To do so, we can simply draw z_a, z_b and z_o and then μ_a, μ_b and μ_o. After making $|S|$ such draws and computing the value of $p_s(d = 0|\mu_s)$ for each draw s, we then can get an estimate of the likelihood of any actor/behavior/object combination. Formally, we use the fact that $\mathbf{E}_{z,\mu}[p(d = 0|\mu)p(\mu|\mu_{0,z},\sigma^2_{0,z})p(z|\pi)] \approx \frac{\sum_s^{|S|} p_s(d=0|\mu_s)}{|S|}$. We use 50 Gibbs samples each time we compute the expectation.

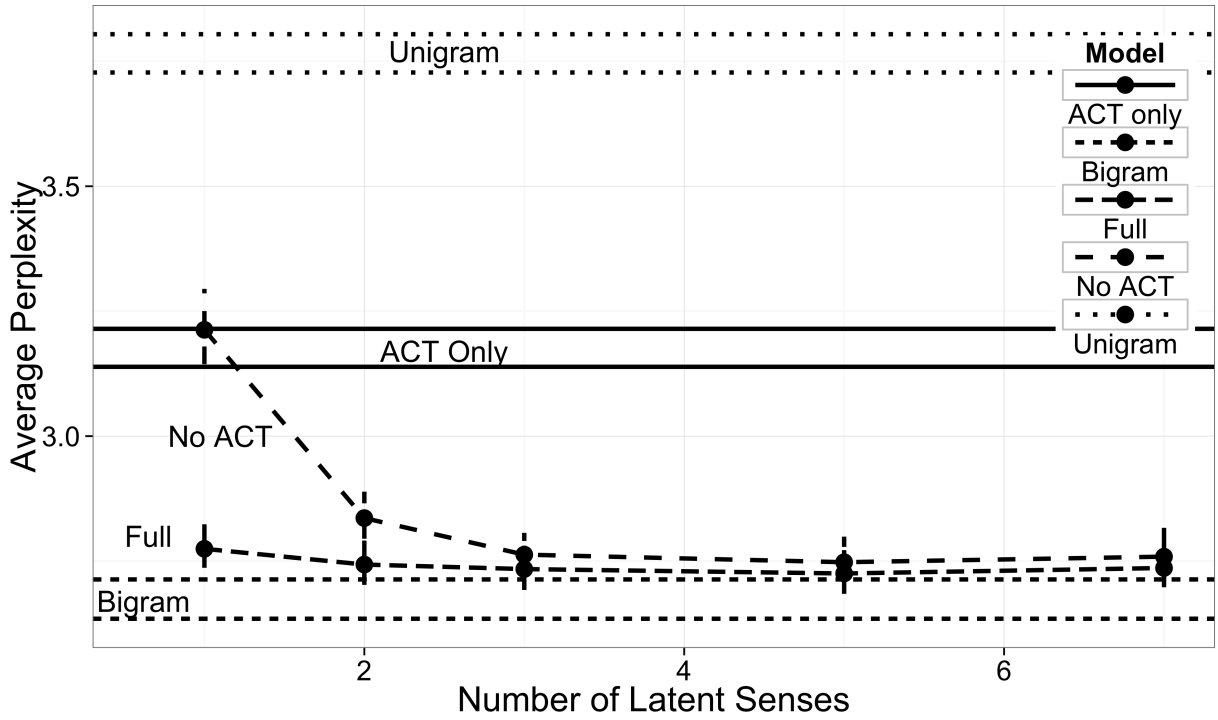


Figure 2: Average Perplexity for the four baseline models and the full model for varying numbers of topics. Where models do not use topics, we present two bands, the top and bottom of the 95% bootstrap CIs. For the FULL and NOACT models, 95% bootstrap CIs are also shown for the different values of L at which the models were trained.

7 Results

7.1 Prediction Task

Figure 2 displays results for log perplexity (recall, the equation for log perplexity was given in Equation (14)) for the four baseline models we used, as well as the full model. On the y-axis, the average perplexity across folds is given. Note that this value is only computed across nine of the ten folds; we ignore results from the fold we use to determine hyperparameter values. The x-axis represents the numbers of latent senses used in the model. For models that do not use multiple senses (the UNIGRAM, ACTONLY and BIGRAM models), Figure 2 shows two vertical bands which represent the upper and lower limits of the 95% bootstrapped confidence intervals. For both the NOACT and FULL models, we run iterations with 1, 2, 3, 5 and 7 latent senses and present confidence intervals for models evaluated at each.

Figure 2 shows that the worst performing model was the UNIGRAM model. As this is the simplest possible approach to behavior prediction, this is not surprising. However, the poor performance of the unigram model relative to the others is important in that gives us confidence that the prediction problem is non-trivial. The ACTONLY model, which implements the basic ACT model, improves our ability to predict behavior by almost an order of magnitude over the baseline UNIGRAM

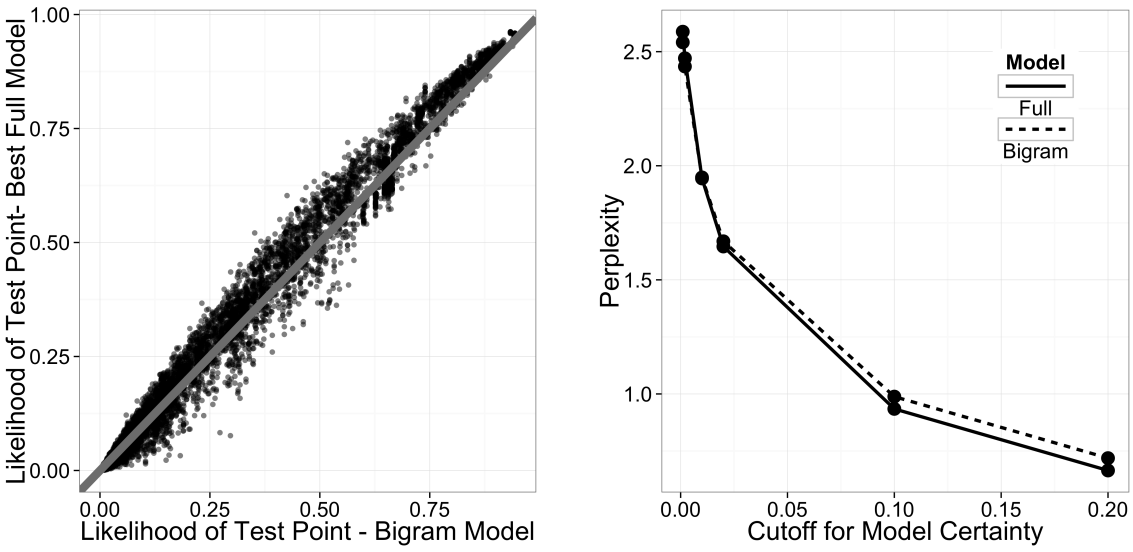
approach. Similarly, the the FULL model performs significantly better than the NOACT model when the number of latent senses is low.

More specifically, we see that affective information encoded in the ACT model improves perplexity by 16% with only one latent sense and by 3% when the number of latent senses per entity is assumed to be two. These gains, while modest are statistically significant - 95% CIs do not overlap at all in either case. However, Figure 2 also shows that as the number of latent senses increases, the difference in predictive power between these two models begins to decrease. Specifically, as the number of assumed senses extends beyond three, model performance on the prediction task becomes virtually indistinguishable. This shows that as the number of latent senses increases, the GMM portion of both the NOACT and FULL models is able to find parameter values that can reliably inform us of future events.

We therefore see that when the number of parameters in the model is low, the theory of ACT provides important guidance for how any assumed affective meaning is structured. As the number of free parameters in the model grows, however, the need for a theoretically driven model of affect decreases in order to accurately predict the data - the model is able to fit the data well in either case because it has enough parameters to "make up" for the lack of theoretically driven priors. Predictive accuracy in this case, however, is still sacrificed for the use of the resulting parameters. As noted in the sections above, only results for the FULL model are useful in interpreting EPA profiles of entities, as only the full model allows us to begin from a baseline of intuitive EPA values for at least a subset of the data. In other words, while the NOACT model assumes the existence of affective constraints, it is essentially free to "make up" its own cultural norms about the form of those constraints. Only in the FULL model, where affective constraints have been painstakingly estimated via decades of survey data, are the affective constraints estimated by the model likely to match our culturally-normed intuitions.

Finally, Figure 2 shows that neither the FULL nor the NOACT models perform as well as the BIGRAM model. This indicates that the ACT-GMM portion of the model actually *decreases* the predictive performance of the bigram language model on the training data. This stems from two factors. First, as mentioned above, semantic information from the BIGRAM model retains a large amount of the affective meanings which may drive these semantic connections. Again, though, it is only with the FULL model that we are able to better understand these affective relationships. Second, because the BIGRAM model is a probability model over only information in the training data, it does not need to "consider" information from the ACT dictionary, information which does not always confirm that provided in the dataset of social events.

Regardless of this difference in the function that these two models maximize, predictive ability on the testing data is the best tool available to understand how well parameter estimates represent the social event data. To better understand how the ACT-GMM portion of the model affects predictions from the language model, Figure 3a plots the likelihood given to the correct behavior for each test point across the nine folds. Each point on the graph represents a single test point. The y-value of a point represents the probability of the actual behavior in the event as evaluated by the best FULL model (where $L = 5$). The x-value provides the same probability, except from



(a) (b)

Figure 3: a) Comparison of the predictions for the best full model and the bigram model for all test points - a dashed line is drawn where x=y for clarity; b) Comparison of average perplexity of the full and bigram models for different likelihood cutoffs

the BIGRAM model. Points that fall above the grey diagonal line are those in which the full model assigned a higher probability to the actual behavior for the test point, while points below the line represent those where the bigram model outperformed the full model.

Figure 3a is constructed with alpha blending, so darker areas of the plot represent areas where more test points fell. Most test points fall near the two tails of the likelihood- that is, either both models believed the test point to be highly unlikely or highly likely. As we see, the bigram model performed significantly better at the lower tail. Thus, the bigram model put more weight in the posterior predictive distribution on the correct behavior in cases where both models believed the true behavior to be highly unlikely. In contrast, the full model performs much better on test points that both models believed were more likely than chance (any value of less than approximately .011 represents a point for which both models would do worse than a random guess). Consequently, we can have some confidence

Figure 3b emphasizes this point. On the y-axis, average perplexity is given for the best full model (solid line) and the bigram model (dashed line). On the x-axis, we provide a cutoff value for test points. As the cutoff increases, we remove test events where both models put less probability on the correct behavior than the cutoff value. Thus, for a cutoff of .02, we remove all test events where both models put less than 2% of the weight in their respective posterior predictive distributions on the correct behavior. Figure 3b shows that full model performance is slightly better than the simple bigram model on data points for which both models believe the actual behavior to be relatively likely.

Combined, Figure 3a and Figure 3b suggest that information from the ACT-GMM portion of

the full model aids the language model portion in predicting already likely behaviors, and detracts in cases where the behavior is already unlikely. This suggests that the model that we have developed struggles when given noisier data. It also suggests, however, that the affective meaning the model uses for prediction may provide important information in a predictive sense beyond what one can derive from pure semantics. This observation shows that a better pipeline of event extraction and future iterations of the model may be provide an important new avenue of predictive modeling of text as well.

As stated, however, the goal of the present work is not to predict behavior, but rather to infer affective meaning. Results in this section were therefore intended to show that the full model learns parameter estimates indicative of the data as proven by the model's ability to predict events it has not seen. To this end, we observe that the full model performs significantly better than all baselines except the bigram language model. We noted one reason for the model's inability to eclipse performance of the bigram model alone, and followed with results suggesting that the full model does better at appropriating higher likelihoods to test points that the language model component indicates are already somewhat likely.

Further, with respect to absolute metrics, the full model (and by extension, the bigram model) are highly accurate in their predictions. Across all test points, the median probability ranking of the correct behavior in the posterior predictive distribution for the best full model was third, and the correct behavior was in the top ten (out of 87) behaviors in 76.9% of the test events. Combined, all of these indicators give us confidence that our model is able to provide parameter estimates for EPA profiles that represent actual processes inherent in the data.

7.2 Perceptions during the Arab Spring

7.2.1 Behaviors

We now turn to a cautious interpretation of model results. Our focus is on parameters that give insights into how the English speaking news media portrayed the entities in our dataset during the Arab Spring. All results in this section are from parameter estimates of the full model run with three latent senses on the entire dataset. While the model performed slightly better with five latent senses we chose to use the model with three latent senses for parsimony. Qualitative conclusions are similar for both models.

Figure 4 displays 95% confidence intervals, as determined using μ_0 and σ_0^2, for the EPA profiles for all behaviors not already in the existing ACT dictionaries. Behaviors are ordered from left to right by their mean evaluative score, as this dimension tends to be the easiest to conceptualize. Importantly, results are shown only for latent senses having more than 10 samples, as including data from latent senses with fewer than this number made the plot difficult to read and also displayed data that was heavily influenced by random initial values of m_0 and s_0.

Figure 4 shows that the model inferred a single, dominant latent sense for all behaviors - only a single latent sense had more than 10 samples for each of the behaviors listed. Thus, the model believed that across all cultural domains incorporated in the news data, the behaviors of interest

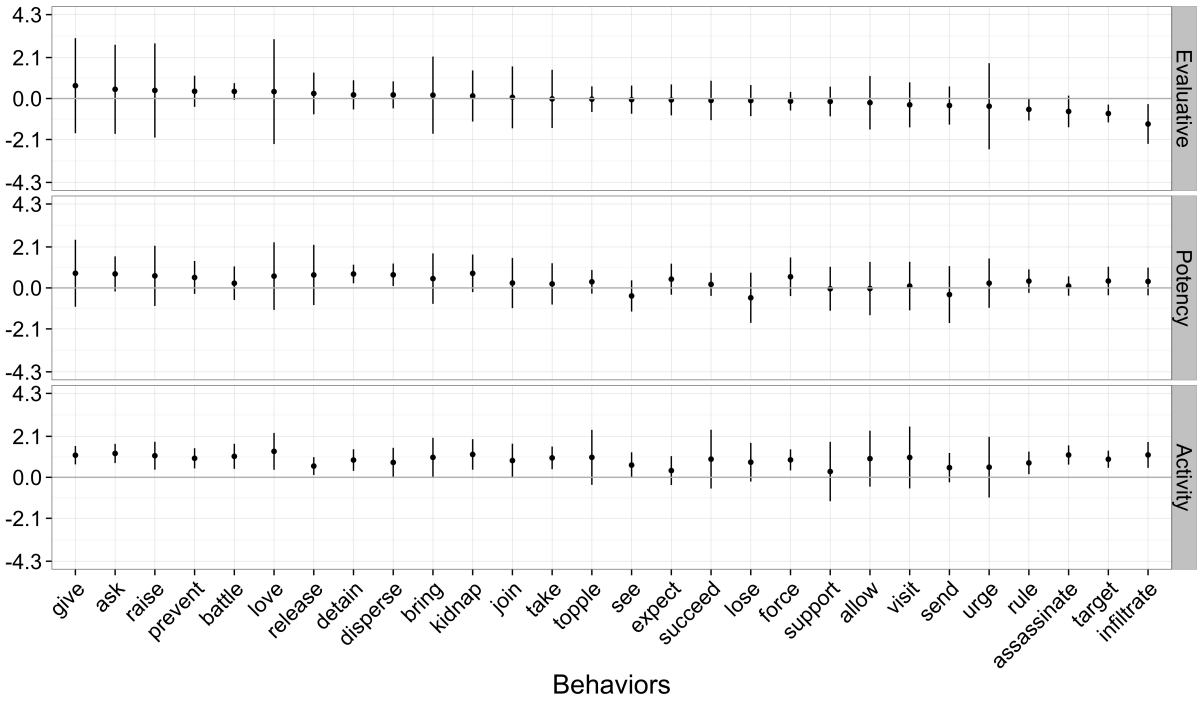


Figure 4: EPA Profiles for all behaviors used by the model *that were not already in the ACT dictionaries*. Confidence intervals are 95% intervals based on σ_0^2. A horizontal grey line is drawn at y=0 to delineate positive from negative values

had a relatively stable meaning. There are multiple reasons why this could be the case. From a mathematical perspective, GMMs operate to a certain extent on a "rich-get-richer" phenomena in that large clusters tend to attract more points. This may have an impact on the extent to which all behaviors clustered along one latent sense. However, as there are many identities that the model estimated to have multiple latent senses, it is plausible that other reasons exist for this observation. One possible reason is that behaviors simply have relatively stable and universal meanings across cultures. This finding may help to ground ACT analyses across cultures in the future.

Given the stability of the affective meanings of behaviors, we would expect that the EPA profiles of these terms fit our intuitive sentiments. We observe this to generally be the case, particularly for extreme values. The most positive behaviors in terms of mean evaluative score, "give", "ask" and "raise", intuitively seem to be things that a good identity would engage in towards another good identity. In contrast, "infiltrating", "targeting" and "assassinating" are indicative of behaviors that have a bad connotation. Similarly, behaviors at the higher end of the potency spectrum, "give", "ask" and "kidnap", are behaviors that more powerful identities could engage in towards lesser individuals, and the least powerful behaviors, "lose" and "see", are relatively powerless. Finally, while all behaviors reported by the media and captured by the model are, unsurprisingly, reasonably active, "love", "ask" and "kidnap" can be considered to be three of the more active ones.

Though any analysis of such results is almost by definition subjective, the model's views on behaviors at the ends of the EPA spectrums fit with at least our own intuitions. When considering

the "middle" of these distributions, however, two findings are also of interest. First, most behaviors are given more neutral values than what we expected. In comparison to data in the ACT dictionary used in the present work, the values are indeed slightly more neutral. For example, the 95% bootstrapped confidence interval around the Evaluative dimensions of the behaviors in Figure 4 is [-0.21,0.08] (mean -.05), a distribution which puts slightly more of its probability weight near zero than a similar band placed around all behaviors in the ACT dictionary [-0.01, 0.25], (mean .12). The cause of this is not immediately clear, and we return to this in the following section.

Second, we observe that there do exist behaviors for which results do not fit our intuitions. For example, kidnapping is actually a slightly *positive* behavior. While we cannot rule out other factors, the explanation for this seems to reside in what was considered newsworthy behavior during the Arab Spring. Although there are several instances of bad identities kidnapping good identities (e.g. "gunman kidnap woman"), the majority of the social events that involve kidnapping in our dataset are ones in which a good (or ambiguous) identity kidnaps another good identity ("father kidnap mother", "police kidnap child"). These events are newsworthy precisely because they are *un*expected (we would not generally expect fathers to kidnap mothers). Given information that good identities kidnap other good identities, however, the model is led to believe that the dominant sense of kidnapping is one of slightly positive evaluative sense.

This observation does not detract from the utility of the proposed approach - although this meaning for kidnap is unexpected, it is supported by this dataset. Future work will need to consider how to remedy, theoretically or methodologically, these differences between what is newsworthy and what is not. For example, one methodological remedy would be to modify assumptions about deflection. As newspaper articles likely include both culturally consistent information and more surprising, high-deflection events events, a bimodal distribution for deflection may be a more appropriate.

The fact that the model can estimate deviations around mean values for EPA profiles helps us to understand the extent to which the model is certain of its estimates. In the case of kidnapping, and many other behaviors in Figure 4, we see that the model is relatively uncertain of, for example, the "goodness" or "badness" of the behavior. Thus, for ambiguous cases (like kidnapping), the model responds with large deviations. The increase in this deviation may unfortunately lead to the masking of the existence multiple latent senses in the data, as these two sources of variation are in direct conflict during model inference. Future work will consider how best to account for this. However, current inference of both at least allows us to better understand how certain we can be of the different parameter estimates while still retaining the necessary theoretical components of ACT.

7.2.2 Identities

On some level, our analysis of the behaviors used by the news media was another exercise in model validation, as we observed that parameter estimates simply matched, for the most part, our intuitions. We now turn to an analysis of a small portion of the identities of interest to us and how

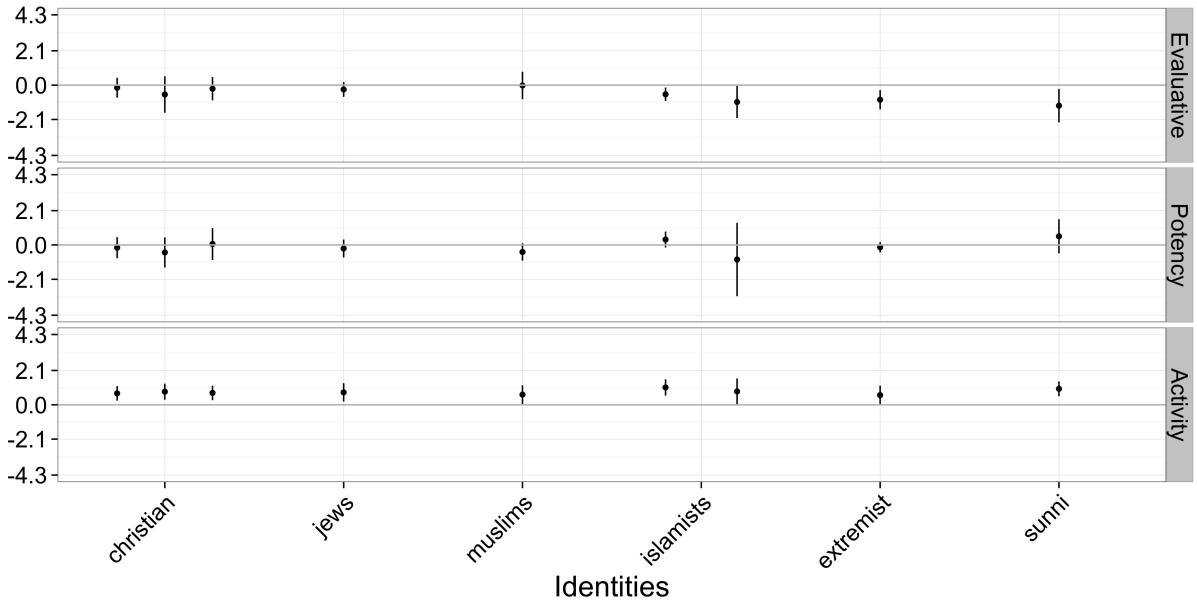


Figure 5: EPA Profiles for six identities of interest. Confidence intervals are 95% intervals based on σ_0^2. A horizontal grey line is drawn at y=0 to delineate positive from negative values

they were perceived by major English speaking news media outlets. As a detailed analysis of all 102 identities in our dataset is not feasible, we choose to focus on one of particular interest, specifically those relating to religious groups[11]. Figure 5 displays results for these six identities in the same fashion as Figure 4. Note that several of the identities were used in multiple latent senses more than ten times. Thus, certain identities have more than one EPA profile associated with them.

The identities portrayed in Figure 5 represent the three major religious groups in the Middle East (Judaism, Christianity and Islam) as well as the identities "Islamist", "Sunni" and "Extremist". Unfortunately, we did not have enough data to compare perceptions of Sunni Muslims to Shiite Muslims. We hope to determine some way of doing so in the future. The figure shows that the media outlets found in our dataset collectively had a relatively neutral perception of the evaluative nature of the generic Muslim identity. Muslims were considered to be almost exactly neutral, with a mean evaluative score of -.02. Being neutral and reliably powerless [-0.96, 0.10], the Muslim identity in general was thus portrayed as more the victim than the perpetrator by the English speaking news media. This identity was even more neutral than Jewish and Christian identities. In fact, Figure 5 shows that Jews and Christians were frequently viewed as being slightly "bad".

This result alone suggests that the news media did not focus on the religious aspects of the Arab Spring at the level of global religious identities. However, Figure 5 suggests that, far from identifying all Muslims as a neutral identity, the news media instead used more specific Muslim identities to connote a strongly negative view of the Muslim identity during the Arab Spring. The identities of Islamists, Muslims who believe politics should be driven by Islamic principles

[11] Results for all identities can be found online, at `https://www.dropbox.com/s/oas8rvlzgw4o6dj/all_identities.png?dl=0`

(Cammett & Luong, 2014), and Sunnis, the majority sect of Islam in the Arab world, are almost entirely negative and active. In all but one sense in which the Islamist identity is portrayed, these identities are viewed as being powerful as well. Even compared to the generic "extremist" identity, the Sunni and Islamist identities were strongly vilified.

It is relatively well-known that Islamist actors took advantage of the Arab Spring revolutions to gain power (Bradley, 2012), and that their ideological stance on government conflict with Western ideals of the separation of church and state. However, the vilification of the Sunni identity as opposed to the more generic Muslim identity is somewhat vexing, particularly in its extent.

8 Conclusion

In the present work, we introduced a new methodology that can extract social events from text and then use these social events to extract sentiment profiles for identities and behaviors. The chief contribution is a statistical model that, by using the concepts of Affect Control Theory (Heise, 2007), provides insight into the soft, affective constraints that influence how we perceive and enact social events. This represents one of the first attempts to ground the analysis of sentiment in text towards generalized identities and behaviors in a way that uses the rigorous theoretical ideals put forth by ACT. From an NLP standpoint, our approach is one of the first efforts to extract a multi-dimensional sentiment profile for concepts; moving beyond the traditional approach of evaluation along a single, evaluative dimension. Further, our work allows for the extraction of multiple sentiment profiles for the same concept within a single corpora.

After describing our model, we provided a case study of its use on how the news media perceived and portrayed identities and behaviors of interest during the Arab Spring. Two findings were of interest. First, while the model found several cross-cultural differences in sentiments of identities, sentimental meanings of behaviors were universal across data from a large number of English-speaking news outlets across the world. While more work is need to better understand this finding, the possibility of stable meanings of behaviors across cultures would be of significant use in anchoring studies of cross-cultural and inter-group identity meanings.

The second finding of interest from our case study was that more specific, connotative Muslim identities of Sunni and Islamists were vilified by major English speaking news outlets, whereas the generic Muslim identity was considered to be rather neutral, even in comparison to the Jewish and Christian identities. A complete understanding of these perceptions requires a detailed consideration of both the events that actually occurred on the ground as well as an understanding of how particular events were perceived by the news media. In addition to those mentioned above, other well-known factors can be expected to have played a role in this finding. These including but are not limited the perpetual Sunni-Shiite conflict, the majority position of Sunnis in the Arab world and their resulting role in the revolutions. However, a comprehensive analysis of the relative influence of each of these factors, and how the Western media was influenced by them, is beyond the scope of the present work.

In taking a step beyond present methodological boundaries in a variety of fields, we made a

host of decisions that had implications on our results. This was particularly true of our approach to event extraction. Three limitations should be noted in our current approach to extracting social events from text. First, we do not currently use the full subject or object, deciding to only use the single dependent term of a possibly multi-term entity (e.g. we only use "America", where the full subject might be "United States of America"). Similarly, we also ignore both settings and modifiers, and thus may lose significant semantic meaning. We assume these errors to be random at this point, and therefore that they "wash away" during model estimation. Second, in ignoring social events which do not have any existing ACT terms, we are removing a large set of potentially useful social events from our data. Future work is needed to better extract social events. Finally, we ignore the order of events in a document and over time. Accounting for temporal sequences might allow for a more accurate predictive model that does not rely on the assumption that all social events extracted begin with the same transient meanings.

Aside from social event extraction, limitations also exist in our statistical model, including the heuristic way in which model parameters are initialized and its relatively poor performance on unlikely events. As we observed in our results, the model also seems to be slightly biased towards neutral sentiments, something we are currently working to understand. Combined, these limitations suggest that, as used here, the current iteration of the model is useful as a descriptive tool, providing insight from large amounts of data that can then be used for more focused, specialized studies. This use of the model is therefore similar to how current NLP tools, such as LDA (Blei et al., 2003), are used in the sociological literature, and we hope that our model provides another methodology for interpreting information from widely available textual data sources.

In the future, we intend to improve model performance in several ways. First, we can use slice sampling on hyperparameters (Neal, 2003) and assume a Dirichlet Process (Teh, 2010) on the number of latent senses per entity to better formalize the notion that the number of latent senses for each entity is unknown and thus should be estimated from the data. Similarly, we could extend the mixture model to use other covariates in the data (such as the particular newspaper from which a social event was extracted), similar to the recent efforts of Roberts, Stewart, and Airoldi (2013). This work will aid in understanding the origin of different perceptions. As our model is agnostic to the source of social event extraction, we also hope to extend our efforts to consider different media beyond newspaper data.

The model could also be improved by a stronger relationship between the language model, which extracts semantic relationships, and the mixture model, which extracts affective relationships. We are currently exploring how the relationships between semantic "constraints" (Heise & MacKinnon, 2010) and affective "signals" can be formalized via cognitive modeling, an approach which may provide novelty in both the theoretical and methodological domains. Finally, we note that the ability to extract affect and relate it to behaviors is also an important extension to network text analysis (Carley, 1994). Future work could extend our model to this domain, perhaps utilizing ideas from relational event modeling (Butts, 2008), to extract more meaningful, valanced links between actors and between actors and objects and thereby expand the use of news data for network analytics.

The present work tackles a methodological question at the intersection of a variety of domains. Perhaps most importantly, we extend methodology surrounding Affect Control Theory to allow for the automated extraction of EPA profiles of entities from text. Given the vast expense of surveys in obtaining this information, the efforts described here, as well as those that build off them in the future, will lead to a stronger and more cost-efficient means of understanding how individuals perceive the actions and identities of others and how such affective constraints affect the way we think about and act towards others. Our work also provides researchers with the opportunity to perform historical analyses where, of course, surveys are not available. To this end, we intend to continue the public development of both code and documentation in a way that allows others to extend our work and use it without a strong programming background. The version of the code used for the present work is available at `https://github.com/kennyjoseph/act_paper_public`.

References

Ahothali, A., & Hoey, J. (2015). Good News or Bad News: Using Affect Control Theory to Analyze Readers Reaction Towards News Articles.

Bishop, C. M., & others. (2006). *Pattern recognition and machine learning* (Vol. 1). springer New York.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, *3*, 993–1022.

Bradley, J. R. (2012). *After the arab spring: How islamists hijacked the middle east revolts.* Macmillan.

Bromiley, P. A. (2013). *Products and convolutions of gaussian probability density functions.* Tina-Vision Memo.

Butts, C. T. (2008). A relational event framework for social action. *Sociological Methodology*, *38*(1), 155–200.

Cammett, M., & Luong, P. J. (2014). Is there an islamist political advantage? *Annual Review of Political Science*, *17*, 187–206.

Carley, K. (1994). Extracting culture through textual analysis. *Poetics*, *22*(4), 291–312.

Carreras, X., & Màrquez, L. (2005). Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the ninth conference on computational natural language learning* (pp. 152–164). Association for Computational Linguistics.

Charniak, E. (1996). *Statistical language learning.* MIT press.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS*, *41*(6), 391–407.

De Marneffe, M.-C., & Manning, C. D. (2008). Stanford typed dependencies manual. *URL http://nlp. stanford. edu/software/dependencies manual. pdf*.

Francis, C., & Heise, D. R. (2006). Mean affective ratings of 1,500 concepts by indiana university undergraduates in 2002–3. *Data in Computer Program Interact*.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (Vol. Third Edition). CRC press.

Heise, D. R. (1979). *Understanding events: Affect and the construction of social action.* CUP Archive.

Heise, D. R. (1987). Affect control theory: Concepts and model. *The Journal of Mathematical Sociology*, *13*(1-2), 1–33. doi: 10.1080/0022250X.1987.9990025

Heise, D. R. (2007). *Expressive order.* Springer.

Heise, D. R. (2010a). *INTERACT: Introduction and software.*

Heise, D. R. (2010b). *Surveying cultures: Discovering shared conceptions and sentiments.* John Wiley & Sons.

Heise, D. R. (2014). Determinants of normative processes: comparison of two empirical methods of specification. *Quality & Quantity*, 1–18.

Heise, D. R., & MacKinnon, N. J. (2010). *Self, identity, and social institutions.* Palgrave Macmillan.

Hoey, J., Schröder, T., & Alhothali, A. (2013a). Affect control processes: Intelligent affective interaction using a partially observable markov decision process. *arXiv:1306.5279 [cs]*. (arXiv: 1306.5279)

Hoey, J., Schröder, T., & Alhothali, A. (2013b). Bayesian affect control theory. In *2013 humaine association conference on affective computing and intelligent interaction (ACII)* (pp. 166–172).

Joseph, K., Carley, K. M., Filonuk, D., Morgan, G. P., & Pfeffer, J. (2014). Arab Spring: from newspaper data to forecasting. *Social Network Analysis and Mining*, *4*(1), 1–17.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques.* MIT press.

Kübler, S., McDonald, R., & Nivre, J. (2009). Dependency parsing. *Synthesis Lectures on Human Language Technologies*, *1*(1), 1–127.

Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th annual international conference on machine learning* (pp. 609–616). ACM.

Leetaru, K., & Schrodt, P. A. (2013). Gdelt: Global data on events, location, and tone, 1979–2012. In *of: Paper presented at the isa annual convention* (Vol. 2, p. 4).

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: System demonstrations* (pp. 55–60).

Mikolov, T., Yih, W.-t., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *HLT-NAACL* (pp. 746–751). Citeseer.

Neal, R. M. (2003). Slice sampling. *Annals of statistics*, 705–741.

O'Connor, B., Stewart, B. M., & Smith, N. A. (2013). Learning to extract international relations from political context. In *ACL (1)* (pp. 1094–1104).

Osgood, C. E. (1975). *Cross-cultural universals of affective meaning.* University of Illinois Press.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in*

information retrieval, *2*(1-2), 1–135.

Rajaraman, A., & Ullman, J. D. (2011). *Mining of massive datasets*. Cambridge University Press.

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2013). *Structural topic models* (Tech. Rep.). Working paper.

Robinson, D. T., Smith-Lovin, L., & Wisecup, A. K. (2006). *Affect control theory*. Springer.

Smith-Lovin, L. (1987). Impressions from events. *Journal of Mathematical Sociology*, *13*(1-2), 35–70.

Smith-Lovin, L., & Douglas, W. (1992). An affect control analysis of two religious subcultures. *Social perspectives on emotion*, *1*, 217–47.

Soon, W. M., Ng, H. T., & Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, *27*(4), 521–544.

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In *The social psychology of intergroup relations* (W Austin & S. Worche ed., pp. 33–47). Monterey, CA: Brooks/Cole.

Teh, Y. W. (2010). Dirichlet process. In *Encyclopedia of machine learning* (pp. 280–287). Springer.

Thomas, L., & Heise, D. R. (1995). Mining Error Variance and hitting pay-dirt: Discovering systematic variation in social sentiments. *The Sociological Quarterly*, *36*(2), 425–439.

Tregouet, D. A., Escolano, S., Tiret, L., Mallet, A., & Golmard, J. L. (2004). A new algorithm for haplotype-based association analysis: the stochastic-EM algorithm. *Annals of human genetics*, *68*(2), 165–177.

Wallach, H., Mimno, D., & McCallum, A. (2009). Rethinking LDA: Why priors matter. *Advances in Neural Information Processing Systems*, *22*, 1973–1981.

Zhai, C., & Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 334–342). ACM.

Zhu, M., Zhang, Y., Chen, W., Zhang, M., & Zhu, J. (2013). Fast and accurate shift-reduce constituent parsing. In *ACL (1)* (pp. 434–443).