The L-I-A-R Project: Linguistic Indicators of Authentic Reviews

INTEGRATING TEXT ANALYSIS, MACHINE LEARNING AND EXPERIMENTAL DESIGN TO DISTINGUISH BETWEEN AUTHENTIC AND FAKE REVIEWS BY THE LANGUAGE THEY EMPLOY

Ann Kronrod
Jeff Lee
Ivan Gordeliy

In this work, we combine empirical and theoretical approaches to understand linguistic features that differentiate word-of-mouth based on authentic experiences versus word-of-mouth based on experiences that did not happen. We focus on the context of consumer reviews, but our theoretical foundations apply to other contexts, such as news, user-generated content, or police reports and evidence.

We first develop a theory linking cognition to language and suggest that people who recount an authentic experience use language that expresses episodic memory of past events. Conversely, when telling about something they have not experienced, people rely on semantic memory, which comprises words and concepts associated with the topic. We next propose that retrieving information from different types of memory can lead to differences in the language people use in these two situations. We hypothesize based on literature linking cognition and language that fake product reviews will employ less concrete language and will use fewer low-frequency words.

Subsequently, we operationalize the language features that reflect semantic versus episodic memory usage by proposing precise mathematical definitions and developing algorithms that calculate the corresponding values for any piece of text. Specifically, we introduce the following measure of text concreteness. We consider a simplified model of a text as a collection of words in it and rely on Wordnet ontology to define the concreteness of individual words. Then to define the concreteness of a set of words, we count the number of texts one can generate, which will be more 'general' than the text under consideration. We generate those texts by replacing any word in the original text by any of the hypernyms. We compare word-frequencies to the average frequency of words in the given corpus.

We next employ an experimental approach to collect a dataset through crowdsourcing (through MTurk) and test our predictions on this dataset. We collect 12,000 reviews total. We also conduct multiple tests to explore whether: (1) humans can imitate an authentic review

style when writing fake reviews and receiving information about our markers of deceptive writing; (2) human judges can discern authentic from fake reviews after receiving the same information. These tests allow us to confirm that our theoretically predicted language features (concreteness and low-frequency word usage) are hard to imitate and hard for human judges to detect, and thus valuable for being included as features in automated classification algorithms. These experiments also illustrate the importance of developing automated classification tools versus relying solely on detection methods involving human judges.

We replicate our findings on the dataset of 800 authentic and 800 fictitious reviews by Ott et al. (2011). We discuss why, although the datasets are very similar, the algorithm proposed in Ott et al. (2011) performs extremely poorly on our dataset (58% accuracy), even though it achieves best-of-class accuracy on their dataset, and suggest how to address this problem. We develop a set of features based on our original predictions to construct such a classification algorithm. To introduce additional features related to the concreteness of language, we can consider different parts-of-speech separately. Furthermore, we can use different parts-of-speech to generate a range of low-frequency wordforms-based features. We discuss a variety of methods of how to generate additional features of relevance.

To test our model of memory tapping when performing a task (of writing a review), we conduct an additional study. We design a product (an app) that participants could not have experienced previously. Then we have participants of the study go (or not go) through the app and write real (or fake) reviews of their experience. Along the way, we measure their reliance on different types of memory. To achieve this, we develop an entirely new way of measuring the ability of human subjects to tap into their episodic memory in real-time (at the time of performing the task of interest).

Our work offers several notable advances in detecting deceptive consumer word-of-mouth. First, we rely on a theory, which allows avoiding interpretability issues inherent to purely data-driven approaches, and may improve generalization to other databases and contexts. Second, we develop new text analysis techniques that do not rely on dictionary lists. Instead of looking at the word-level, we define and operationalize summary variables at the text/sentence level, which allows us to go beyond the bag-of-words model of a text. Third, we complement theory and text analysis with experiments designed to test predictions of our model, probe its behavioral foundations, as well as to demonstrate the value of (and the

mechanism behind) our proposed language features in discerning authentic versus fake consumer word-of-mouth.