Linguistic Aspects of Fictitious Product Reviews

Combining Automated Text Analysis and Experimental Design

Ann Kronrod [1], Jeffrey K. Lee [2], and Ivan Gordeliy [3]

[1] ***Manning School of Business, University of Massachusetts Lowell***

[2] ***NYU Shanghai***

[3] ***GNT, LNC2, DEC, École Normale Supérieure, INSERM, PSL Research University***

INTRODUCTION

UGC (User-Generated Content) is an important tool in digital. The main reasons for its success: **consumers value authenticity and they link authenticity with UGC**. However, there is also **a growing share of fraudulent UGC**. At the same time, it has been repeatedly shown that **users are not successful at evaluating the authenticity of written content and are naïve about its sincerity**. We study linguistic markers of deception in UGC with 2 goals in mind: understanding theoretically (starting with how Cognition works) how deceptive language will be different and developing computational methods to filter out fraudulent content.

Insights from Cognitive Theory

We focus on a particular type of deception: **recounting an event that one has not experienced**. Relying on studies of cognition and memory, we infer:

Fraudulent descriptions will be less concrete

To define concreteness of a text we rely on a lexical database WordNet *(Miller 1995, Fellbaum 1998).*

Concreteness of a text is defined as $D = \log \prod \binom{d+f}{f}$,

where d is the number of steps of a particular noun deeper from the word "entity"; f - how many times the noun recurs in a review; the number of factors is equal to the number of distinct nouns in a review. The expression under the logarithm roughly corresponds to the number of possible texts which are more general than the one being considered constructed by replacing each word with its hypernyms.

Fraudulent descriptions use fewer 'rare' words

We call a word-form rare if it appears in the whole corpus less often than a certain benchmark value.

TWO EXPERIMENTAL STUDIES

Study 1: Dataset Construction and Analysis of Concreteness and Rare-word Usage

Study Design:

"Hello,

We would like to ask you to write a review for a hotel in which you stayed in the past year (did not stay).

(We know you did not stay in this hotel, but we would still like you to write the review as if you did.)

Please make your review about 10 sentences long."

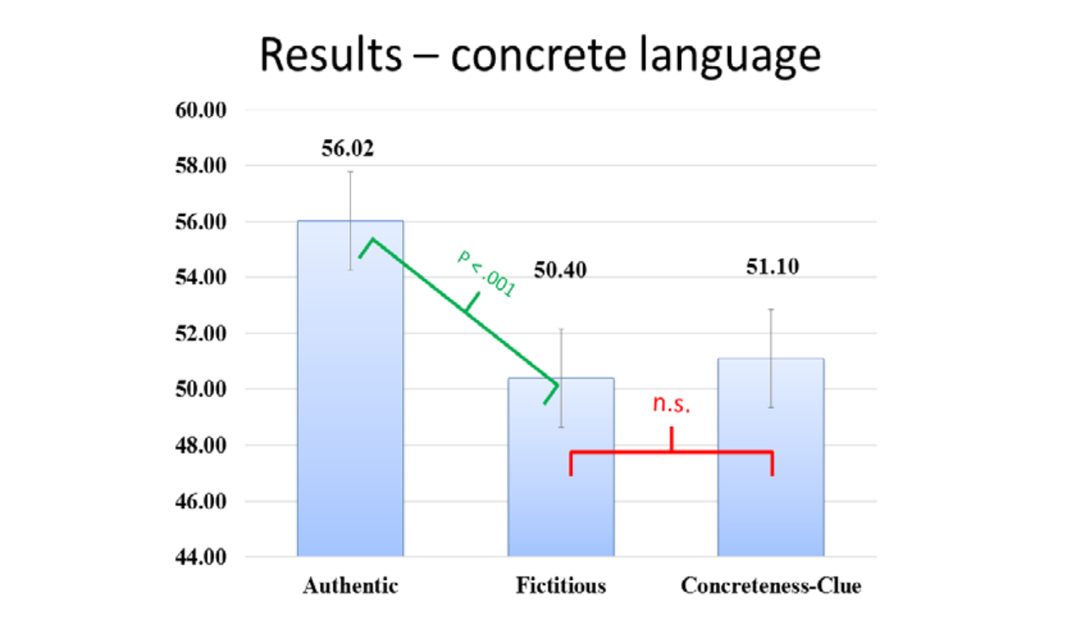
Some participants were asked to write a truthful review (174 participants), others – a fictitious one (202 participants), and there were several groups of review writers (~200 participants in each group) who received hints based on our theoretical expectations on how to imitate an authentic review better.

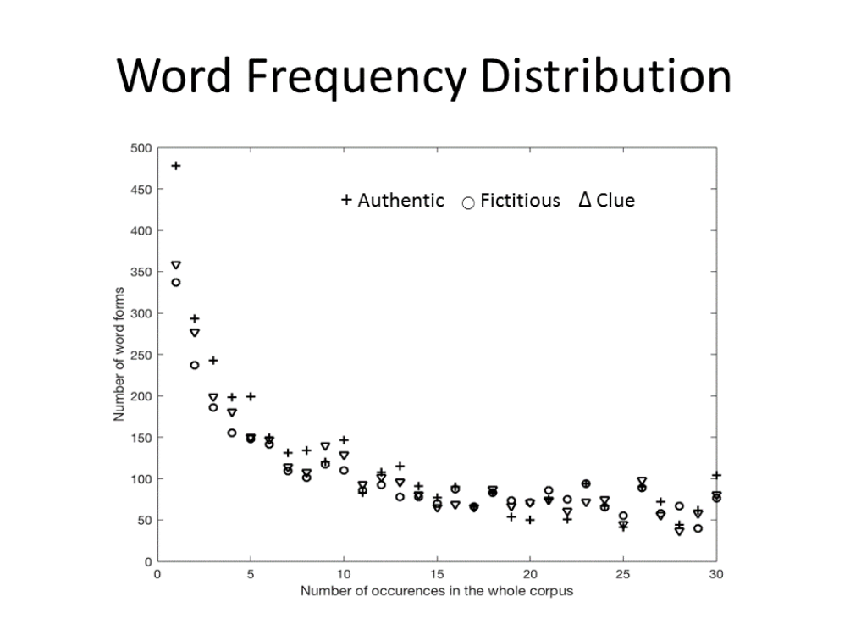
Example:

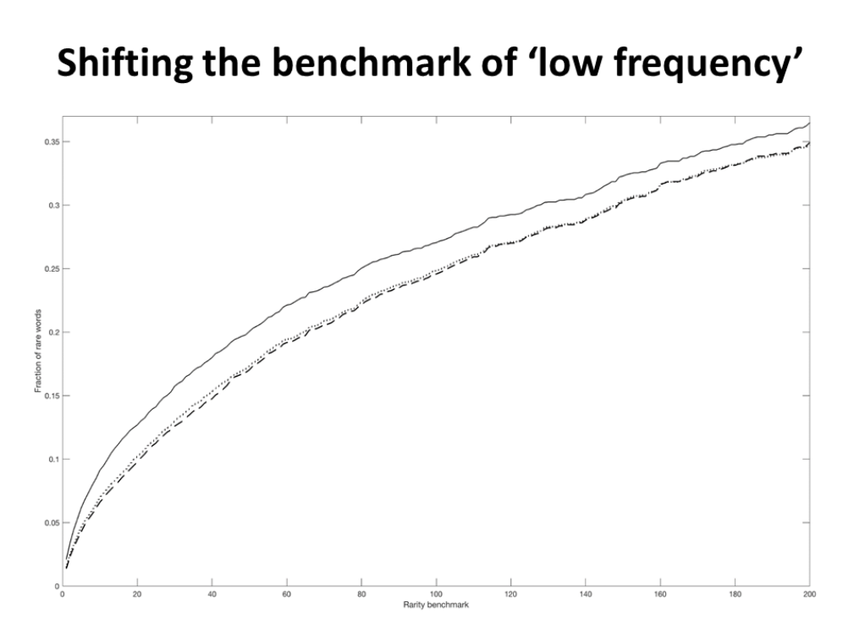
*"Please note: Scientists have discovered some characteristics of insincere (fake) reviews. One such characteristic is less use of low frequency words. A fake review is more likely to use frequent words that are common for that product (for example the word "keyboard" or "typing" for a computer). **When composing your fake review, we encourage you to use this clue to improve your text.**"*

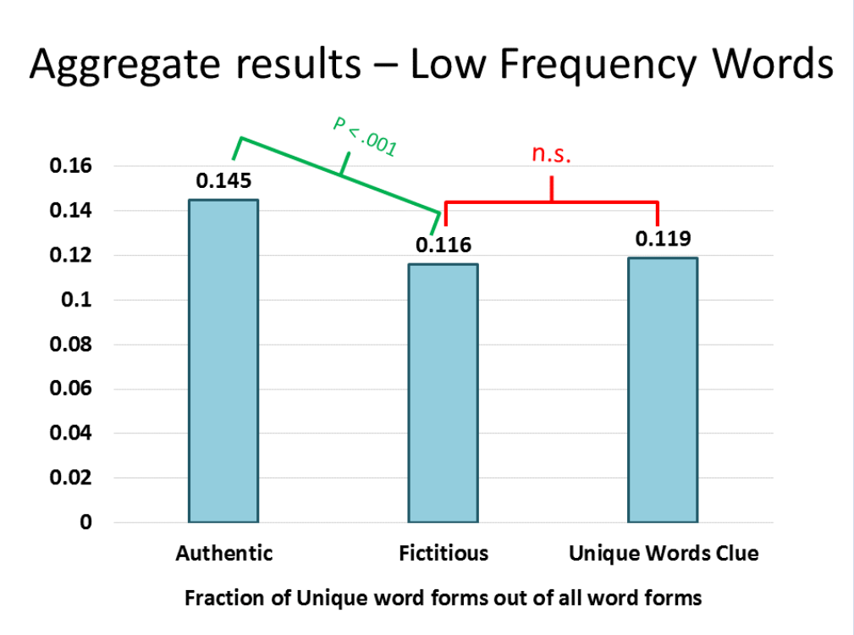
Results:

The results confirm the robustness of the two suggested linguistic features, i.e. 1) concreteness and low-frequency word usage do seem to be relevant features to distinguish authentic and fictitious reviews. 2) Review writers are unable to compensate for this even when provided with appropriate hints.


Results – concrete language


Word Frequency Distribution


Shifting the benchmark of 'low frequency'


Aggregate results – Low Frequency Words

Study 2: Human detection of insincerity with/without clues

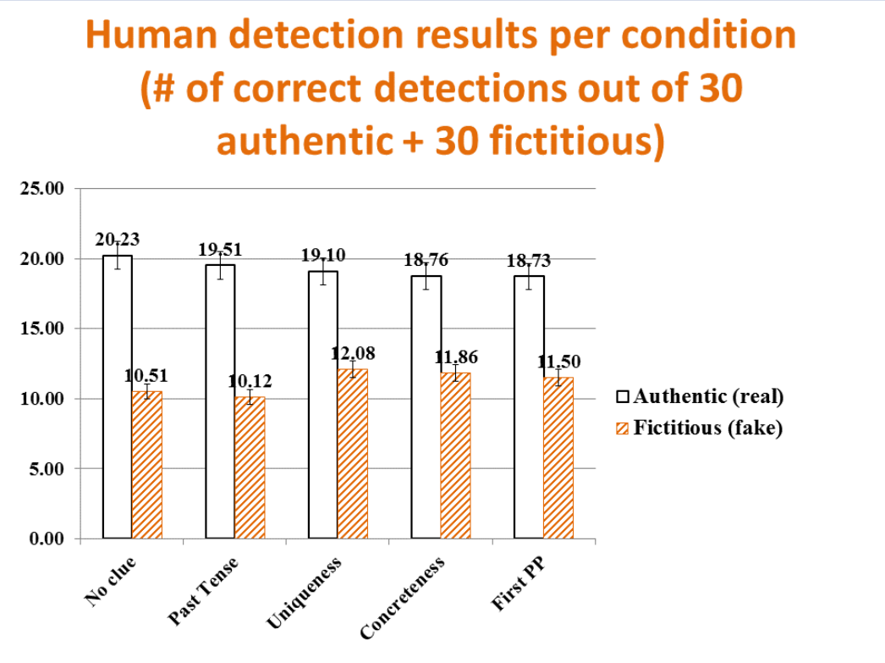
Study Design:

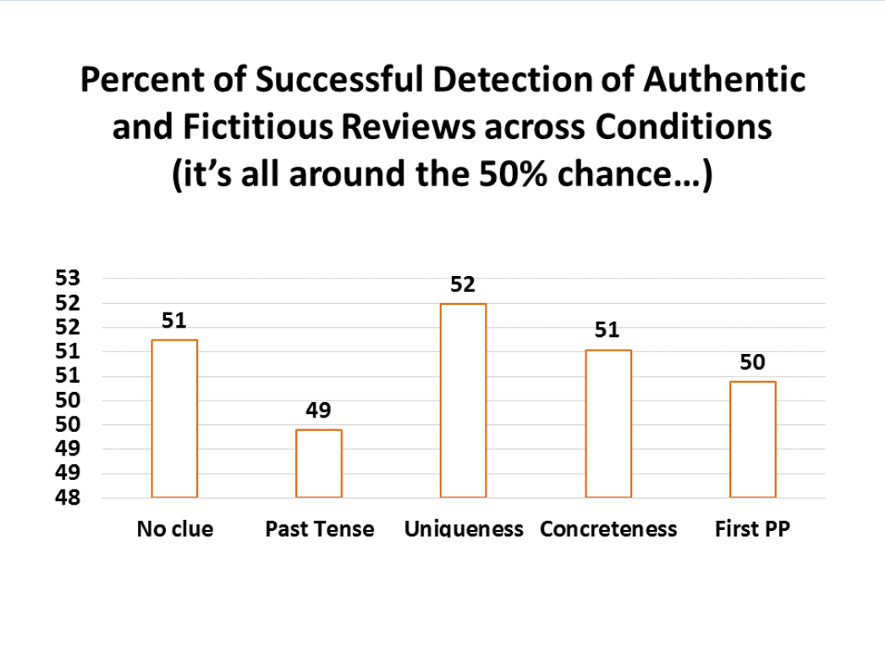
"In this assignment you will read 60 reviews for a hotel. For each review, you will make a decision whether it is a **true review** (the reviewer actually stayed at the hotel and wrote the review after that) or a **fake review** (the reviewer has not actually stayed at the hotel and made up the review)."

328 participants read a random subset of 60 reviews (half authentic and half fictitious) from our dataset, and for each review clicked a choice button for real/fake. Here again, we had five conditions: one where participants who did not receive any clues and several conditions in which participants received a hint based on our theoretical predictions.

Results:

Analysis confirms the well known features of human detection: 1) performance close to chance level, 2) Truth bias. Furthermore, **we find that when given hints on how to detect deception, participants become more suspicious, while still displaying a very strong truth-bias. Their overall accuracy does not improve.**


Human detection results per condition (# of correct detections out of 30 authentic + 30 fictitious)


Percent of Successful Detection of Authentic and Fictitious Reviews across Conditions (it's all around the 50% chance…)

Validation and Outlook

We have replicated our findings on the dataset of 800 authentic and 800 fictitious reviews by *Ott et al. 2011*. *Ott et al*. train a classification algorithm (using BIGRAMS+ and 80 LWIC parameters as features) and achieve high performance (up to 90% accuracy on their corpus). However, for some of the linguistic features (past tense, first person pronouns), which seem to matter a lot their findings, we find that these features are unreliable markers of deception (this explains as well a throve of conflicting results on this in literature).

Furthermore, we find that the algorithm developed by *Ott et al.* performs marginally better than human detectors on our dataset (58% accuracy) which may indicate a different approach may be required to construct algorithms which would not be as specific corpus dependent. We suggest, taking into account 'robust' (i.e. hard-to-imitate) and universal (independent from specific domain) features may be the way to go.

Acknowledgements

This work was supported by a grant from the Marketing Science Institute.

The authors would like to thank Seshadri Tirunillai for his help in developing the code and computational approach of an earlier version of this work. We thank Ravi Kiran and Wang Wan for their assistance in the preparation of this project. We would also like to acknowledge the valuable input from Alireza Alemi, and to thank Liudmyla Kushnir, Pantelis Leptourgos, Vasily Pestun, Vasil Khalidov and Alexey Arbuzov for fruitful discussions which promoted this work.