Detecting Fake Amazon Book Reviews using Rhetorical Structure Theory

Olu Popoola
University of Birmingham
England
United Kingdom
o.popoola@bham.ac.uk

ABSTRACT
This study explores the potential of a theory of discourse coherence relations to distinguish between truth and deception. It uses Rhetorical Structure Theory and logistic regression to build a deception model that achieves 78% accuracy on a sample of gold-standard Amazon book reviews drawn from the Deceptive Review corpus. It finds Contrast discourse relations to be a significant predictor of veracity and successfully tests a discourse mining method for their semi-automated extraction. These preliminary findings contribute to the development of a linguistic-based theory that can guide the design of computer-aided deception detection systems.

KEYWORDS
fake reviews, opinion spam, deception detection, Rhetorical Structure Theory, coherence relations, Amazon book reviews, big data veracity

1 INTRODUCTION
Two strands of computer-aided deception detection (CADD) research – linguistics and machine learning - use the textual content of reviews to assess veracity. Linguistic analysis distinguishes authentic and fake online reviews with moderate accuracy e.g. 75% in [27]. However, the surface features used - e.g. part-of-speech (POS), psychological lexicons such as LIWC and readability indices - are inconsistent in their predictive power across domains and communicative contexts. In addition, the findings often contradict the very deception theories (i.e. Reality Monitoring [18], Information Manipulation Theory [23], Interpersonal Deception Theory [4], Self-Presentational Theory [7]) used to rationalize feature selection [1] [2] [20]. In contrast, machine learning algorithms (typically using n-grams) have achieved notable accuracy rates for fake review detection (e.g. 90% in [27] [9]) but at the expense of diagnostic power, thus further obscuring our understanding of an increasingly prevalent form of illegal commercial activity1.

The focus in CADD on morphological, syntactic and lexicosemantic cues, initiated concurrently by the use of LIWC in [26] and stylometric features in [42], was originally only a matter of convenience since deep linguistic analysis at the discourse level was more difficult to automate [42]. Yet, over a decade later, there has still been little deception detection research using deeper linguistic features.

[31] was the first attempt to analyze deception at the level of discourse structure. They used Rhetorical Structure Theory (RST) [21] to distinguish true and fake narratives elicited under experimental conditions. RST annotation revealed systematic differences in coherence relations used. Evidence relations were significantly more frequent in true stories while Evaluation relations were significantly more frequent in deceptive stories; this difference might indicate different methods for presenting one’s communication as credible or authoritative. However, no underlying theory was proposed by the authors. [32] used RST less successfully to compare news stories categorized as true or fake in a ‘Bluff the Listener’ radio game show. There were no systematic differences in relations used; the shared communicative game show context, which meant all news stories were humorous and surprising regardless of veracity, may have made the supposedly true and fake news stories too similar. This

1 New York Attorney-General Eric Schneiderman described fake online reviews as “the 21st century’s false advertising” after the 2012 ‘Operation Clean Turf’ investigation.
research may constitute further evidence of the problem of using ‘pseudo-fake’ data – i.e. when research subjects are told to lie – for deception detection (cf. [25]). This study addresses some of the issues with previous research by analyzing a forensically compiled dataset of known fake and authentic reviews. This provides an opportunity for exploratory research designed to tackle the following questions:

1) What discourse coherence relations are used in fake vs authentic Amazon book reviews?
2) Can coherence relations effectively classify fake and authentic reviews?
3) What theories of deception are supported or suggested?

2 DATA AND METHOD

The reviews for this study were drawn from the Deceptive Review (DeRev) corpus [11]. DeRev is a ‘forensic corpus’ (as defined in [10]) compiled as a result of an investigation into fake review production conducted by renowned ‘sock puppet hunter’ Jeremy Duns [28] and journalist David Streitfield [35,36]. Fake reviews in DeRev were defined as reviews written for any books written by authors who had confessed to buying reviews or by any writers who had admitted to being paid to write reviews. DeRev also assigns each fake review an additional truth value based on the quantity of the following deception clues it contained: i) being part of a review cluster i.e. a group of at least two reviews posted within three days; ii) use of nickname by reviewer; iii) unverified purchase; iv) suspect book (i.e. reviews written by offending authors/writers). In the present study only reviews containing all four deception cues were labeled fake (n=628). DeRev’s authentic reviews are drawn from books written either by dead authors (e.g. Hemmingway) or established international best-selling writers (such as Ken Follett). For the present study, only the authentic reviews with 0 or 1 deception cues were used (n=942).

50 5-star reviews (25 authentic, 25 fake) were randomly sampled from the DeRev corpus and manually annotated with RST relations to create the DeRev-RST corpus (Popoola, 2017), using RSTTool and phpSyntaxTree [8]3. All reviews were between 50 and 150 words as a minimum length for analysis and convenient length for manual annotation. The protocol for RST annotation outlined in [37] was followed. The essential steps are i) divide text into elementary discourse units (EDUs), which are typically clauses; ii) mark adjacent pairs of EDUs with an RST relation, making sure all four constraints are satisfied (see Figure 1 for example); iii) look at larger adjacent text spans and apply relations recursively until all the text is accounted for as a tree structure (see Figure 2 for example).

The RST macro-relations outlined in [5] were used (see Figure 3). These group relations that fulfil a similar informational or pragmatic function and so minimize the impact of ambiguous relations on coding consistency. The reviews were randomized by a third party in order to conceal their veracity label prior to annotation by the author. DeRev-RST contains 4931 tokens and 490 RST relations in total. The average number of relations per review was consistent across authentic and fake reviews.

![Figure 2: Example RST hierarchical (tree) annotation using RSTTool](image)

![Figure 3: RST macro-relations used and their definitions.](image)

---

2 Since the best performing automated RST parsers [3] and [19] only obtain around 50% accuracy, manual annotation is currently the best approach for theory development.

3 DeRev-RST corpus is freely available on request from the author.
In addition to descriptive statistics, logistic regression (conducted with SPSS software) was used for data analysis with review veracity as the dependent variable (1 = authentic, 0 = fake). RST relations were used as predictors. The textual measures were frequency of RST relations per review (normalized by review length); these were continuous variables since each review contained 0 to as many as 17 of each relation. The overall performance of the model was checked using Nagelkerke's $R^2$ with $p=0.05$ used for significance and odds ratio to assess the impact of each relation on review veracity. The model performance was benchmarked against that of Review Skeptic (http://reviewskeptic.com) on the same data set. Review Skeptic was trained using the algorithms developed in [27] for 90% accuracy on a specific corpus of hotel reviews but weak cross-domain performance [25]; consequently, an inferior performance from the RST model would be proof of its ineffectiveness.

3 RESULTS

Almost every review has some Elaboration and Evaluation relations; two-thirds of reviews contain a Joint relation. Contrast and Explanation relations occur in just over half of the reviews while over a third include a Background relation (Figure 4a). These six relations make up 90% of all relations used. The frequency of relations in the corpus follows a Zipfian distribution (see Figure 4b), with reviews generally containing multiple Elaboration relations (e.g. describing book plot/content) in addition to some Evaluation (i.e. some form of recommendation) and Contrast (i.e. argumentation in support of stance towards the book) relations (see Figure 4c or examples).

Figure 4a: Distribution of RST relations in all reviews. N=50

Figure 4b: Frequency of RST relations in all reviews. N=490

Figure 5 shows that fake reviews use over 50% more Elaboration relations, whereas true reviews contain three times as many Contrast relations. Although overall use of Evaluation relations does not substantially differ between true and fake reviews, true reviews had an equal proportion of Elaboration and Evaluation relations, while fake reviews used over twice as many Elaboration as Evaluation relations.

Only the six most frequent relations were used in the logistic regression model. These accounted for half of the variability ($R^2=0.50$). Overall accuracy of 78% (Figure 6) is a substantial improvement on Review Skeptic’s performance (Figure 7) on the same set of reviews; similar levels of precision and recall indicate a balanced model.

<table>
<thead>
<tr>
<th>Observed</th>
<th>Predicted</th>
<th>% correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fake</td>
<td>True</td>
</tr>
<tr>
<td>Fake</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>True</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>Overall</td>
<td>5</td>
<td>39</td>
</tr>
</tbody>
</table>

Figure 6: Classification of DeRev-RST corpus using RST relation features
Contrast relations are significant predictors of authenticity, \( p=0.004; \ exp(B)=0.165 \), while repeated Elaboration relations are strong signs of deception \( p=0.066; \ exp(B)=2.419 \). Comparison relations were only found amongst authentic reviews so were excluded from the model.

## 4. DISCOURSE MINING FOR CONTRAST

These results, although promising, were achieved with a small sample of 50 reviews. In order to explore the hypothesis that Contrast was a significant predictor of review veracity, a further 1570 5-star reviews were analyzed. These consisted of 942 ‘true’ reviews (all remaining 5-star reviews containing either zero or one deception cues) and 628 ‘fake’ reviews (all remaining 5-star reviews containing the maximum four deception cues). Since full annotation of 1570 RST reviews is prohibitively expensive, a ‘discourse mining’ technique was used [22] – [34]. Potential Contrast relations were mined by extracting all reviews containing the word ‘but’. Previous research has shown that ‘but’ is the most common discourse marker of Contrast and that 30-40% of Contrast relations are signaled by ‘but’ [6] [38]. Thus, analysis of ‘but’ is the most efficient technique for finding Contrast discourse relations; although there will be more Contrast relations in the corpus, the frequency of such relations signaled by ‘but’ can be taken as indicative of their general frequency in reviews. Figure 8 shows that true reviews used ‘but’ substantially more frequently than fake reviews, making a prima facie case for Contrast being a key indicator of review veracity.

A random sample of text span pairs containing ‘but’ was then extracted and manually annotated by the author for Contrast relations using the following steps: 1) Each instance was coded for one of six interpretations of ‘but’ outlined in [17]: i) denial of expectation; ii) opposition; iii) correction iv) topic shift/cancellation; v) objection; vi) sequential. 2) Denial of expectation and opposition were aggregated as potential Contrast relations (corresponding to the Concession, Antithesis and Contrast RST relations cf. [15] [21] [23]. 3) In order to verify the Contrast relation, the co-text surrounding ‘but’ was coded using the formalisms detailed in [12] and [14].

In total, a subset of 125 authentic reviews and 134 fake reviews containing ‘but’ were annotated. Over 40% of the authentic reviews contained Contrast relations compared to less than 30% of fake reviews (Figure 9). The fact that authentic reviews use ‘but’ substantially more than fake reviews and that authentic reviews are more likely to use ‘but’ to signal Contrast relations indicates that Contrast relations are more strongly associated with authentic reviews than fake reviews.

In violation of this genre convention, the deceptive reviews in this study eschewed the nuance of argument-based evaluation and instead interleaved plot synopsis and PR materials with exclusively positive comments often connected simply by series of Elaboration relations (e.g. Figure 11 below). The fact that paid-for reviews use substantially more Elaboration relations reflects the deceptive context of communication. Fake review writers cannot engage in the type of evaluative contrast typical of the review genre because they haven’t read the book. Being paid £5 to £10 per review means that for the activity to be profitable, time must be spent on writing multiple reviews rather than reading

---

4 Concede and counter’ has also been noted as a common English language evaluation strategy in Appraisal Theory [24].
many books. This inevitably affects the quality of evaluation and appraisal of the books.

**Contrast** relations as a marker of veracity is also supported by the psycholinguistic deception detection literature, in which exclusive words and distinction markers have been found to be indicators of veracity [26] [13]. This suggests that **Contrast** is a pragmatic communicative act that is difficult to execute in a deceptive context because a liar cannot give a deceptive and balanced argument. The fact that none of the fake reviews contained **Comparison** relations further supports this (comparing and contrasting are similar pragmatic activities ⁵). Previous research demonstrating that deceptive reviews contain more extreme emotions (positive or negative) than authentic ones [20], and that positive authentic reviews contain more negative emotions than their deceptive equivalents [1], also suggests the presence of antonymous relations is indicative of authenticity.

_Titan Raines "Sol on Ice" is an intriguing story that delves beyond the race issue and what separates us. Here you will read about a personal journey, both physically and emotionally, about his travels and experiences with ayahuasca. This story is extremely interesting and thought provoking. It raises many questions and brings about many realizations. As you read you it becomes increasingly clear we really are not so different after all. Great read!_  

Figure 11: Fake review F0035

6. CONCLUSION

This exploratory study makes two methodological contributions to fake review detection. It demonstrates that coherence relations can assist the task of fake review detection and that RST provides a sensitive analysis framework, although using using fewer relations may be more effective. Furthermore, with automated RST relation annotation still a challenge, the discourse mining approach demonstrated here effectively estimates the impact of coherence on review veracity. Although the data set is small, this analysis suggests a genre-based linguistic theory can inform CADD system design. Future research should explore the effectiveness of using genre information as training data in the development of CADD algorithms.

REFERENCES


⁵ Contrast and Comparison are pragmatically close enough to be considered the same relation in alternative coherence relation frameworks [41].


