Predicting Review Helpfulness

A Machine Learning & Natural Language Processing based Approach

Ashwin Ittoo
Asst Prof
HEC, Management School,
University of Liège, Belgium
Background

- Online reviews
  - Information source assess product/quality
Background (cont)

• However...
• Huge number of human-authored online reviews
  – Hard for consumers to discern really helpful reviews
• Lack of editorial control
  – Great variations in review quality
• Emergence of online peer-reviewing
  – Consumers vote for review helpfulness

• Helpfulness votes
  – Indicator of review diagnosticity (Liu & Park, 2015, Mudambi & Schuff, 2010)
  – Facilitate efficient review filtering (Ghose & Ipeirotis, 2008)
Research Questions (RQ)

• « What makes a (customer) review helpful?
  – Fundamental issue yet to be addressed

• Main RQ addressed
  – Linguistic devices contributing to review helpfulness
  – Predicting reviewing helpfulness

• Approach based on
  – Linguistics, Natural Language Processing (NLP)
  – Machine Learning (ML)

• Development of novel algorithms
  – Automatic detection of linguistic structures
  – Predict review helpfulness
Motivations

• Automatically assessing review helpfulness

• Practical/Industrial motivations
  – Positively impact customer purchase decisions
  - Improve website social presence; attract customers, increase sales (Kumar & Benbasat 2006)

• Scientific motivations
  – Novel NLP/ML algorithms
  – Assessing deeply embedded textual features (helpfulness)
Presentation Structure

• Related studies
  – Automatically assessing review quality

• Proposed approach
  – Feature engineering
  – Machine learning experiments
  – Results
  – On-going/future work
Related Studies – Review Helpfulness

- 3 main areas of related research
  1. Spam, fake review detection
     - Detecting duplicate reviews (Jindal & Liu, 2007)
     - Pattern-based spam detection (Mukherjee et al., 2013)
  2. Influence of reviews on sales
     - Positive reviews and book sales positively correlated (Chevalier & Mayzlin, 2006)
     - Positive reviews, average review ratings and box-office sales positively correlated (Dellorcas et al., 2007)
Related Studies – Review Helpfulness

• Clarification of extant literature

• 3 main areas of related research

1. Spam, fake review detection
   – Detecting duplicate reviews (Jindal & Liu, 2007)
   – Pattern-based spam detection (Mukherjee et al., 2013)

2. Influence of reviews on sales
   – Positive reviews and book sales positively correlated (Chevalier & Mayzlin, 2006)
   – Positive reviews, average review ratings and box-office sales positively correlated (Dellorcas et al., 2007)
3. **Predicting review helpfulness/quality**

- Factors contributing to review helpfulness
- Identity disclosure (Sussman et al., 2003; Forman et al., 2008)
  - Improves review credibility; more helpful review
  - Presence of photos, email addresses, ...

- Reviewers’ characteristics (Gilly et al., 1998, Liu&Park, 2015)
  - Expertise, reputation
  - Review from expert likelier to influence purchase decision, more helpful
  - Number of followers/friends, number of reviews, replies

- Review star ratings (Danescu et al., 2007)
  - Reviews with extreme ratings (1,4) more helpful
  - Various statistics (mean, variance) from star ratings

- Perceived enjoyment in reading review (Liu&Park, 2015)
  - Number of clicks « Cool » button (only for Yelp! Data)
The work was done in a timely manner. A call back was necessary - the follow up was above average - both customer service people were very courtesy. The workman Tom call back within a reasonable time. They are guarantie their work. We can recommend them and will use them again.
Related Studies – Review Helpfulness (cont)

• Shortcomings of existing research

• Extrinsic factors (identity disclosure, reviewers’ characteristics)
  – Not accurately predictive of review helpfulness

• Factors hard to operationalize, formalize, measure
  – Coarse-grained approximates, oversimplify actual values
  – Identity disclosure as binary {0,1}
  – Expertise as number of friends

• Factors not always available
  – Number of fans, « cool » button
Linguistic factors for Review Helpfulness

• Reviews’ textual contents, styles
  – Intrinsic factor, inherent to reviews

• Theoretical underpinnings
  – Textual contents, styles determine message persuasiveness (Schindler, & Bickart, 2012)
  – “Source effect” of message (Janis et al. 1959)
Linguistic factors for Review Helpfulness

• Zhang & Varadarajan (2006) predicting Amazon review helpfulness
  – Cosine similarity schemes on TF IDF scores
  – Shallow syntactic features (surface forms of words, lemmas, parts of speech tags)

\[
similarity = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}
\]

\[
w_{i, j} = tf_{i, j} \times \log \left( \frac{N}{df_i} \right)
\]

\(tf_{ij} = \text{number of occurrences of } i \text{ in } j\)

\(df_i = \text{number of documents containing } i\)

\(N = \text{total number of documents}\)
Linguistic factors for Review Helpfulness (cont)

• Review elaborateness to predict helpfulness (Mudambi&Schuff, 2010)
  – More elaborate $\rightarrow$ more helpful
  – Elaborateness = number of words

• Stylometric features (Viswanathan et al., Lee&Choeh, 2014)
  – N-grams, syntactic dependencies, sentiments

• However
  – Lack scientific basis
  – Ad-hoc features
  – No consensus in feature definition
Proposed Approach

• Grounded in *argumentation theory*

• Hypothesis
  – Argumentative structures useful predictor of review helpfulness

• Rationale for Argument as predictor:
  – Argument as linguistic device for persuasion
  – Found in dialogue, where one party convinces another
  – Reader votes review as helpful only if persuaded of utility of review
What is an “Argument”?  

• Argument (Palau & Moens, 2009)  
  – Set of premises/evidence/facts  
  – Claim/conclusion  
  – Claim: proposition, either true or false  

• 2 steps in proposed approach  
  1. Identify arguments from review texts  
  2. Determine whether arguments effectively predict review helpfulness
Argumentation Theories

• Several argumentation theories

• Toulmi’s formalization (Toulmi, 2003)
  – 6 components of arguments: claim, data, warrant, backing, qualifier, rebuttal
  – Complex formalization, difficult to “learn” and to identify in text (ML, NLP)

• Rhetorical Structure Theory (RST)
  – Several rhetorical relationships between 2 text spans; nucleus-satellite
  – Nucleus: central to the discourse
  – Satellite: interpretable only w.r.t. nucleus
  – More suitable for computational purposes
Formalizing the “Argument”

• 2 components of argument
  1. Premise
  2. Conclusion

• Components essential in human cognition (Paulus & Moens, 2009)
  – Distinguish argumentative structures vs. normal statements

• Premise-conclusion relation
  – Various syntactic mechanisms
  – Subordination, coordination

• [I love this book]_{conclusion} as [it reflects the authors’ trials and tribulations....]_{premise}
Feature Engineering

• Formal definition for defining lexical features
  – Signal argumentative structures in text

• Prominent feature from discourse theory
  – Discourse connectives
  – [I love this book]_{conclusion}, \textcolor{red}{as} [it reflects the authors’ trials and tribulations....]_{premise}
Feature Engineering (cont)

• Subordinating conjunctions (because, since,...)
  – *Since* I am a fan of autobiography, I could not resist but to buy ...

• Coordinating conjunctions (but, and, or,...)
  – *This is not a best-seller, *but* I still love...*

• Adverbials (however, as a result, otherwise,...)
  – I had a bad experience with Motorola. *As a result*, I switched to Samsung...

• Conjoined connectives (when-if, if-when)
  – I will be satisfied *if and when* I receive a refund...
Experiments

• 3 datasets of consumer reviews
  1. Amazon.com (books, movies)
  2. Yelp! (restaurants)
  3. TripAdvisor (restaurant, hotels)

• Standard, used in previous research (Viswanathan et al., Jindal & Liu, 2008, Wang et al., 2011)

• 100,000 reviews randomly selected from each dataset
Experiments – Data Cleaning

• Data in JSON/XML format

```
review

{
    'type': 'review',
    'business_id': (encrypted business id),
    'user_id': (encrypted user id),
    'stars': (star rating, rounded to half-stars),
    'text': (review text),
    'date': (date, formatted like '2012-03-14'),
    'votes': {(vote type): (count)}
}
```

• Meta-data of interest
  – Review text
  – Number of votes (review helpfulness)
Experiments – Data Cleaning (cont)

• Following previous studies (Ghose, A., & Ipeirotis, 2011)
• If >=60% of users found review X helpful
  – Then, classify X as “true” (helpful)
• Else
  – Classify X as “false” (not helpful)

The other great new thing about this case is that the Voyage attaches to it magnetically. It attaches and detaches very easily, unlike the Paper attached to the standard case. This is great because if you decide you want to read the device and appreciate how light/thin it is without a case.

It's pricey but worth it.

PagePress

---

One of the new features is PagePress, which has sensors on the outer edge of the device that you can squeeze to turn pages. Read more

175 Comments 4,354 of 4,503 people found this helpful. Was this review helpful to you? Yes No Report abuse
Experiments – Data Cleaning (cont)

• Text data inherently noisy
• Requires further cleaning/pre-processing
  – Stopwords removal
  – Morphological analysis (lemmatization), e.g. “loved”, “love”, “loves” → “love”

• Reduce dimensionality via feature selection
• Best results with cube mutual information

\[
\log \left( \frac{f(x, y)}{N} \right)^3 \cdot \frac{f(x)}{N} \times \frac{f(y)}{N}
\]

• Degree of correlation between word x (e.g. “book”) in the review and class y (helpful/not helpful) of the review
• Result: find best words characterizing each class
Experiments – Machine Learning

• Adopted ML formulation
  – Given a review,
  – Predict its likeliest class (helpful or not)

• Various ML paradigms/classifiers investigated
  – ZeroR
  – SVM
  – Random Forest (RF)
Experiments – Machine Learning (cont)

• SVM classifier
  – Optimal hyperplane separating 2 classes
Experiments – Machine Learning (cont)

• RF classifier
  – Combines predictions of multiple decision tree
  – Makes best possible prediction
Experiments - Class Imbalance Issue

• Datasets highly imbalanced
  – Helpful class >> not helpful class

• Hard predicting minority class (not helpful)

• High accuracy with dumb classifier
  – If 80% of data belongs to helpful class (20% not helpful)
  – Accuracy(Predicting everything as helpful)=0.8
  – Results are misleading

• Need to generate a balanced distribution
Experiments - Class Imbalance Issue (cont)

- SMOTE (Synthetic Minority Oversampling Technique) (Chawla et al. 2002)
  - Resolve class imbalance issue
  - Generates synthetic (artificial) examples of minority samples
  - From K-Nearest Neighbors

- Operates in feature space
- Better performance vs. under/oversampling techniques in data space
  - E.g. replicate minority class examples
Experiments – Set up

• Baseline
  – For comparing contribution of argumentative patterns
  – Helpfulness prediction with argumentative patterns vs. with baseline features

• Each classifier (SVM, RF) evaluated over data twice
  – Setup1: using baseline features
  – Setup2: using argumentative patterns

• Measure accuracy of helpfulness prediction (per setup)

• Determine contribution of argumentative patterns
Experiments – Baseline Features

• Features employed in previous research

• Star rating (numerical, 1-5)

• Extreme rating
  – 1 iff rating ∈ {1,5}, 0 otherwise

• Avg. sentence length

• Number of 1\textsuperscript{st}, 2\textsuperscript{nd} pronouns

• Sentiment scores (numerical, computed from SentiWordNet)
Experiments – Baseline Features (cont)

• Baseline features: Readability metrics
  
  • SMOG (Mc Laughlin, 1969)
    – Polysyllables: words $\geq$ 3 syllables
      \[
      \text{SMOG} = 1.0430 \sqrt{\text{number of polysyllables} \times \frac{30}{\text{number of sentences}}} + 3.1291
      \]
  
  • Flesch-Kincaid (Kincaid et al., 1975)
    \[
    206.835 - 1.015 \left( \frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left( \frac{\text{total syllables}}{\text{total words}} \right)
    \]
  
  • Gunning fog (Gunning, 1952)
    – Complex words = polysyllables
    \[
    0.4 \left[ \left( \frac{\text{words}}{\text{sentences}} \right) + 100 \left( \frac{\text{complex words}}{\text{words}} \right) \right]
    \]
  
  • Coleman-Liau (Coleman&Liau, 1975)
    – $L =$ avg. number of letters per 100 words
    – $S =$ avg. number of sentences per 100 words
    \[
    0.0588L - 0.296S - 15.8
    \]
Measuring Prediction Accuracy

- Standard metrics
  - Precision (P)
  \[
  \frac{\text{true_positive}}{\text{true_positive} + \text{false_positive}}
  \]
  - Recall (R)
  \[
  \frac{\text{true_positive}}{\text{true_positive} + \text{false_negative}}
  \]
  - F1-score
  \[
  \frac{2 \times P \times R}{P + R}
  \]

- Estimated via 10-fold cross-validation
  - Divide data into 10 subsamples
  - Use 9/10 for training
  - Test over 1/10 (no SMOTE on test set)
Experiments - Results

• Best results with RF
  – Outperforms SVM in predicting review helpful/not
  – Over all 3 datasets
<table>
<thead>
<tr>
<th>Feature set</th>
<th>Precision (true)</th>
<th>Recall (true)</th>
<th>F1 (true)</th>
<th>Precision (false)</th>
<th>Recall (false)</th>
<th>F1 (false)</th>
<th>Precision (overall)</th>
<th>Recall (overall)</th>
<th>F1 (overall)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating + length + readability</td>
<td>0.734</td>
<td>0.77</td>
<td>0.752</td>
<td>0.758</td>
<td>0.721</td>
<td>0.739</td>
<td>0.746</td>
<td>0.746</td>
<td>0.745</td>
</tr>
<tr>
<td>Rating + length + extremity + readability</td>
<td>0.758</td>
<td>0.727</td>
<td>0.742</td>
<td>0.758</td>
<td>0.727</td>
<td>0.742</td>
<td>0.748</td>
<td>0.747</td>
<td>0.747</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.736</td>
<td>0.779</td>
<td>0.757</td>
<td>0.765</td>
<td>0.721</td>
<td>0.743</td>
<td>0.751</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>Argumentation</td>
<td>0.78</td>
<td>0.78</td>
<td>0.78</td>
<td>0.77</td>
<td>0.76</td>
<td>0.765</td>
<td>0.775</td>
<td>0.77</td>
<td>0.77</td>
</tr>
<tr>
<td>Baseline + argumentation</td>
<td>0.81</td>
<td>0.79</td>
<td>0.80</td>
<td>0.78</td>
<td>0.76</td>
<td>0.77</td>
<td>0.795</td>
<td>0.775</td>
<td>0.78</td>
</tr>
</tbody>
</table>
### Tripadvisor

<table>
<thead>
<tr>
<th>Feature set</th>
<th>Precision (true)</th>
<th>Recall (true)</th>
<th>F1 (true)</th>
<th>Precision (false)</th>
<th>Recall (false)</th>
<th>F1 (false)</th>
<th>Precision (overall)</th>
<th>Recall (overall)</th>
<th>F1 (overall)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating + length + readability</td>
<td>0.733</td>
<td>0.772</td>
<td>0.752</td>
<td>0.759</td>
<td>0.719</td>
<td>0.738</td>
<td>0.746</td>
<td>0.745</td>
<td>0.745</td>
</tr>
<tr>
<td>Rating + length + extremity + readability</td>
<td>0.73</td>
<td>0.778</td>
<td>0.753</td>
<td>0.763</td>
<td>0.713</td>
<td>0.737</td>
<td>0.746</td>
<td>0.745</td>
<td>0.745</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.735</td>
<td>0.788</td>
<td>0.761</td>
<td>0.771</td>
<td>0.717</td>
<td>0.743</td>
<td>0.753</td>
<td>0.752</td>
<td>0.752</td>
</tr>
<tr>
<td>Argumentation</td>
<td>0.765</td>
<td>0.78</td>
<td>0.77</td>
<td>0.78</td>
<td>0.73</td>
<td>0.75</td>
<td>0.77</td>
<td>0.76</td>
<td>0.76</td>
</tr>
<tr>
<td>Baseline + argumentation</td>
<td>0.77</td>
<td>0.80</td>
<td>0.78</td>
<td>0.76</td>
<td>0.75</td>
<td>0.75</td>
<td>0.765</td>
<td>0.775</td>
<td>0.77</td>
</tr>
<tr>
<td>Feature set</td>
<td>Precision (true)</td>
<td>Recall (true)</td>
<td>F1 (true)</td>
<td>Precision (false)</td>
<td>Recall (false)</td>
<td>F1 (false)</td>
<td>Precision (overall)</td>
<td>Recall (overall)</td>
<td>F1 (overall)</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>------------------</td>
<td>---------------</td>
<td>-----------</td>
<td>-------------------</td>
<td>----------------</td>
<td>------------</td>
<td>---------------------</td>
<td>------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Rating + length + readability</td>
<td>0.686</td>
<td>0.738</td>
<td>0.711</td>
<td>0.716</td>
<td>0.662</td>
<td>0.688</td>
<td>0.701</td>
<td>0.7</td>
<td>0.7</td>
</tr>
<tr>
<td>Rating + length + extremity + readability</td>
<td>0.692</td>
<td>0.743</td>
<td>0.716</td>
<td>0.722</td>
<td>0.669</td>
<td>0.694</td>
<td>0.707</td>
<td>0.706</td>
<td>0.705</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.711</td>
<td>0.754</td>
<td>0.732</td>
<td>0.738</td>
<td>0.693</td>
<td>0.715</td>
<td>0.725</td>
<td>0.724</td>
<td>0.724</td>
</tr>
<tr>
<td>Argumentation</td>
<td>0.745</td>
<td>0.76</td>
<td>0.75</td>
<td>0.75</td>
<td>0.736</td>
<td>0.75</td>
<td>0.75</td>
<td>0.748</td>
<td>0.75</td>
</tr>
<tr>
<td>Baseline + argumentation</td>
<td>0.75</td>
<td>0.77</td>
<td>0.76</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td>0.76</td>
<td>0.75</td>
</tr>
</tbody>
</table>
Results’ Discussion

- $F_1_{\text{argumentation\_features}} > F_1_{\text{baseline\_features}}$
- $F_1_{\text{argumentation\_features}} \sim F_1_{\text{baseline\_features+argumentation\_patterns}}$

- *Argumentation patterns useful predictor of review helpfulness*
Ongoing/Future Work

• Argumentation Mining algorithms
  – Automatically detecting argumentation structures (premises, conclusions) from review texts
  – Classifying premises (support or attack)

• Incorporate meta-data as features
  – Product type (hedonistic vs. utilitarian), age of reviews,…

• Deep Learning with Neural Networks
  – Move from word embedding to argument embedding
ありがとう
References


References (cont)


References (cont)

- Viswanathan, V., Mooney, R., & Ghosh, J. Detecting Useful Business Reviews using Stylometric Features.